I Was Blind but Now I See: Implementing Vision-Enabled Dialogue in Social Robots

Giulio Antonio Abbo
IDLab-AIRO
Ghent University – imec
Ghent, Belgium
0000-0001-6301-0028

Tony Belpaeme
IDLab-AIRO
Ghent University – imec
Ghent, Belgium
0000-0001-5207-7745

***Abstract*—In the rapidly evolving landscape of human-robot interaction, the integration of vision capabilities into conversational agents stands as a crucial advancement. This paper presents a ready-to-use implementation of a dialogue manager that leverages the latest progress in Large Language Models (e.g., GPT-4o mini) to enhance the traditional text-based prompts with real-time visual input. LLMs are used to interpret both textual prompts and visual stimuli, creating a more contextually aware conversational agent. The system's prompt engineering, incorporating dialogue with summarisation of the images, ensures a balance between context preservation and computational efficiency. Six interactions with a Furhat robot powered by this system are reported, illustrating and discussing the results obtained. The system can be customised and is available as a stand-alone application, a Furhat robot implementation, and a ROS2 package.**

***Index Terms*—Large Language Model, Vision Language Model, Dialogue, HRI, Conversation, Prompt Engineering, ROS**

Fig. 1. Four interactions with a Furhat robot powered by our system.

I. Introduction

In the ever-evolving landscape of Human-Robot Interaction, the quest for more intuitive and immersive experiences has driven advances in natural language processing and artificial intelligence. Conversational agents play a pivotal role in this progression and as they become increasingly integrated into daily life – spanning home assistants, help desks, elderly care, and teaching – the demand for richer, context-aware conversations has grown more pronounced.

Large Language Models (LLMs) have demonstrated remarkable abilities in generating human-like text. However, their traditional reliance on purely textual inputs creates a gap in achieving a holistic understanding of the user's context. Human communication naturally incorporates visual cues, non-verbal expressions, and environmental context to enhance interaction and collaboration [1]. For instance, when faced with an unfamiliar tool, one might naturally ask: "How do you use this?" A robot responding to this query must discern whether the user is pointing at an object, identify the object in question, understand its use, and communicate the information effectively. Prior work on visual grounding in conversational systems has shown promise but often lacks real-time adaptability or requires extensive customization for HRI applications.

Funded by Horizon Europe VALAWAI (grant agreement 101070930).

The challenge lies in integrating textual and visual information within the LLM's prompt structure. Currently, such capabilities are often re-implemented on a per-project basis, leading to redundancy and inefficiency. Although HRI-specific toolkits like ROS4HRI [2] provide foundational support for multimodal interaction, they do not incorporate the latest advancements in LLMs – particularly their capacity for visual input – into conversational components.

This paper addresses this gap by presenting a ready-to-use module for spoken interactions that incorporates both text and image processing. The proposed tool leverages the vision capabilities of LLMs, enabling a conversational agent to process textual inputs while also assimilating and responding to visual stimuli in real time. By capturing images from a live video feed, the system enhances contextual awareness, fostering more natural and immersive conversations.

Our implementation is fully customisable, allowing users to select strategies for integrating visual elements into the LLM's prompt, choose suitable models, and manage prompt growth over time. A real-time summarisation pipeline processes image frames, balancing scalability and contextual relevance. We

provide an implementation[1] that can function as a stand-alone application with a webcam, as a ROS2 [3] node, or with a Furhat robot [4].

We used the system in six different sessions in a lab, a bedroom, a bathroom, and an entrance to a home, as illustrated in Figure 1. Four subjects interacted with the Furhat robot powered by our implementation, seeking assistance, asking for suggestions, and engaging in small talk. The robot's performance demonstrates that the system is effectively grounded in reality, delivering context-aware responses without requiring additional information in the prompt.

II. Background

A. Large Language Models for Dialogue Generation

Traditionally, LLM-based systems relied solely on textual inputs to infer context [5], [6]. For instance, Janssens et al. [7] proposed a system where captions generated by an image model served as input for a text-based LLM. However, such approaches do not leverage the full amount of information available in the original image.

Vision Language Models (VLMs) are a major step forward in artificial intelligence [8], combining the ability to recognise images with the language understanding of LLMs. This integration enables tasks like image description, visual question answering, and ongoing dialogues about visual content. VLMs operate by linking visual details from images with corresponding text-based information. For example, in a model like LLaVA [8], a pre-trained visual encoder like CLIP [9] processes the image. The extracted features are then transformed into a format compatible with a language model like LLaMA [10] through a trained projection matrix, allowing image inputs to be included in prompts.

Despite these advancements, challenges remain. VLMs often overemphasise visual inputs, describing image content in detail rather than incorporating it naturally into the dialogue. Mixing dialogue and images in prompts can disrupt the conversational thread, and processing the additional data can introduce latency, affecting real-time applications. Addressing these issues is critical for deploying VLMs effectively in dialogue systems. Finally, LLMs' and VLMs' safety is tied to their training data, making assessing their alignment an important part of the development process [11], [12].

B. ROS Packages for Dialogue Generation

ROS (Robot Operating System) [3] is a versatile framework for robotic development, offering modularity, standardized communication, and extensive libraries for tasks like control, sensing, and motion planning. Its open-source nature and active community make it a useful tool for robotics research, industry, and education.

Several ROS packages provide conversational capabilities, though most rely on traditional natural language processing approaches. For example, DialogFlow integrations, such as GB-dialog[2] and Dialogflow ROS2[3], enable spoken interactions but lack the contextual depth provided by LLMs and are challenging to extend to multimodal inputs.

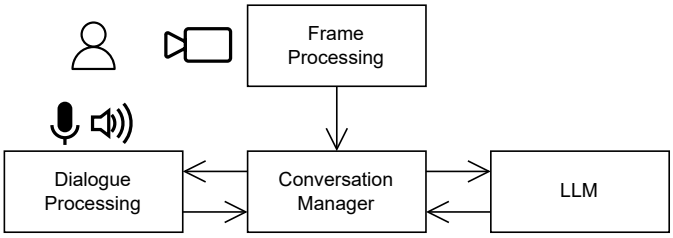

Fig. 2. Components: the conversation manager receives data from the frame and dialogue processing components, using an LLM to produce a response.

Other packages, like ros2_nanollm[4], support running LLMs and VLMs and interacting through ROS messages. These however do not handle dialogue management. Similarly, tools like ROScribe[5] [13]–[17] focus on robot programming or movement control through language rather than conversational dialogue. ROOTED[6] aims to develop adaptable dialogue systems for social robots by combining rule-based generation, web search, and LLMs. However, it does not currently support VLMs and targets specifically the Plantroid Robot.

Among HRI-specific toolkits, HRItk [18] uses traditional NLP approaches, with their limitations mentioned above, and is not actively maintained, while ROS4HRI [2] offers robust packages for multimodal interaction but does not yet include tools for LLM or VLM dialogue generation.

III. Implementing Vision-Enabled Dialogue

The system proposed empowers a conversational agent with vision capabilities. When a user interacts with the system, the responses will be grounded in reality and aware of the context, thanks to additional visual input. This visual input consists of frames from a video captured as the conversation takes place, which are weaved into the conversation. The LLM that produces the output is instructed to interpret these images as its own sight sense. For this implementation, we chose GPT-4o mini as the underlying LLM, as it offers a good balance between costs, speed and accuracy. While the system can work on its own, using a webcam and the terminal's text interface, to make the demonstration more realistic we have chosen to use a Furhat robot. We also provide a ROS2 compatible implementation.

A. Components and Implementation

The system is composed of four components as shown in Figure 2: the frame and dialogue processing components, the conversation manager and an external LLM.

[1]https://github.com/giubots/vision-enabled-dialogue – A release of the code at the time of writing is published at https://doi.org/10.5281/zenodo.14627887.

[2]https://github.com/IntelligentRoboticsLabs/gb_dialog
[3]https://github.com/Juancams/dialogflow_ros2
[4]https://github.com/NVIDIA-AI-IOT/ros2_nanollm
[5]https://github.com/RoboCoachTechnologies/ROScribe
[6]https://gitlab.com/AntonioGCGonzalez/rooted

The *frame processing* component is in charge of retrieving the frames from a video feed and sending them to the conversation manager. The frequency with which the frames are sent can be configured. We have found that a good compromise between speed and conversation quality is one frame every five seconds. The component can be configured to source frames using four approaches: the video feed from the built-in camera of the Furhat robot, a webcam, a video file, or a ROS source. Additional details on the ROS implementation are omitted for brevity and are documented in the linked code repository.

The *dialogue processing* component provides the user input to the conversation manager and shows the output back to the user. This component runs in parallel with the previous one, meaning that the conversation manager can receive dialogue and frame inputs in any order. There are four implementations available for this component: a text-based input using the terminal, a file-based input for testing purposes, a Furhat and a ROS implementation. The Furhat implementation instructs the robot to look at the user and uses the built-in speech-to-text capabilities of the robot to obtain input from the user. While the input is being elaborated, the robot looks away, to signal that the robot is not listening. When the result is ready, the robot looks again at the user and tells the answer leveraging its text-to-speech module, which also controls the mouth movements.

The *conversation manager* is the most important component, as it is in charge of managing the prompt that generates the responses of the system. When a frame or a message from the user is received, they are added to the prompt, following a customisable strategy: either using the most recent frame, or a sequence of frames interleaved in the prompt.

The conversation manager can be configured to summarise some frames to reduce the prompt length. In this case, when necessary, a VLM will be used to summarise the frames' contents into a textual description which will replace the frames in the prompt. Similarly, this module can be configured to summarise the conversation contents, to shorten the prompt.

To generate a response, the conversation manager can use a default model or choose one from a set. In the second case, it will use another LLM to choose between using an LLM or a VLM, based on the user message. The default models are GPT-4o mini as a VLM and GPT-3.5 Turbo as the model-chooser and text-only LLM. The response is then returned to the dialogue processing module.

B. The Prompt

The prompt initially consists of a list of frames and dialogue lines, preceded by the following instructions. *You are impersonating a friendly kid. In this conversation, what you see is represented by the images. For example, the images will show you the environment you are in and possibly the person you are talking to. Try to start the conversation by saying something about the person you are talking to if there is one, based on accessories, clothes, etc. If there is no person, try to say something about the environment, but do not describe the environment! Have a nice conversation and try to be curious!*

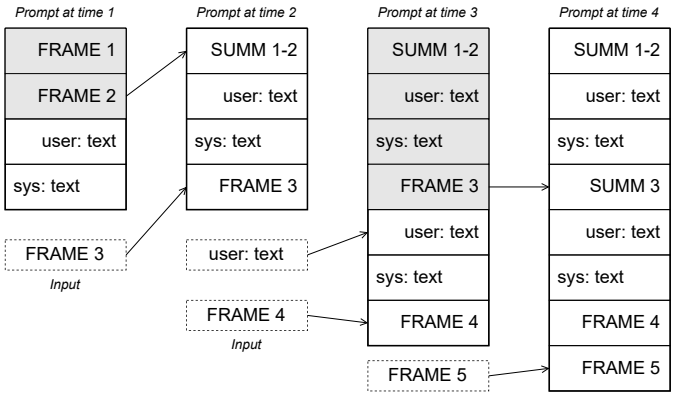

Fig. 3. Example of the summarisation process. Considering $n = 3$ and $m = 2$. In the first step, a frame is added. The number of frames is now n, so the algorithm summarises the first m (two). Then a dialogue line is added, with the system response, and then another frame is added. In the final step, `FRAME 5` triggers another summarisation. This time, only `FRAME 3` is summarised, as including the following frame would disrupt the ordering of the elements. In grey the elements used to obtain the summary at each step: notice that the previous part of the conversation is included.

It is important that you keep your answers short and to the point.

Impersonating a kid has been found to improve the quality of the answers, reducing unwanted messages about the capabilities of the model and other disclaimers. Then, the prompt tells the model how to interpret the images in the prompt, we found this wording to be the most effective so far, compared to more technical explanations. The rest of the sentences are necessary to reduce the loquacity of the system, and to keep the output relevant and salient.

C. Frames Summarisation

Continuously adding frames to the prompt leads to an increase in its size, with longer computation times and costs. To shorten the prompt we propose to summarise its frames.

A naive solution would be to send the first part of the dialogue and frames to a LLM and ask for a summary. However, this would impact negatively on the conversation quality: the majority of the summary would be devoted to a description of the frames, leaving less room for a summary of the conversation. and this problem would be made even worse with high frame rates. Furthermore, since the focus of the interaction is the dialogue, it does not make sense to summarise this with the frames, which serve only as context.

Our solution summarises the frames separately and keeps the ordering of the frames and dialogue lines. To achieve this, when a frame is received, the conversation manager checks how many frames are in the prompt, if a configurable limit n is reached it performs a summarisation routine. This routine will scan the prompt, and summarise the first m consecutive frames, as shown in Figure 3. Setting $m < n$ ensures that at least one frame remains in the prompt, to maintain context-awareness. We have found that keeping at most $n = 4$ frames in the prompt and summarising in chunks of $m = 3$ frames yields satisfying results.

To obtain the summary, a VLM is prompted with the full conversation and previous summaries up to the frames to summarise and is asked to provide a brief description. The frames are then removed from the conversation and substituted by their summary.

IV. Interactions

To demonstrate the capabilities and the results obtained, we report and discuss the settings and main highlights of six interactive sessions. For these sessions, we used the system previously described to empower a Furhat robot with vision and dialogue capabilities. Since the dialogues generated are highly context-dependent, we ran the sessions in five different environments (see Figure 1): a lab, a kitchen, the home entrance, a bathroom, and a bedroom.

Session 1: in a lab with desks and a window; it is dark outside. The system recognises the environment correctly, identifying that the user is in a lab or a workplace. It then asks the user if he is working on something interesting, and recommends not to work late hours even if the project is exciting. Unexpectedly, the system deduced that it was late, probably from the dark windows.

Session 2: in a kitchen, in front of a counter. The system recognises that the person in front of it is cooking, and when asked is able to come up with suggestions on what to prepare. This setting is probably one of the most realistic use cases for a vision-enabled assistant in the home and showcases the intrinsic knowledge contained in LLM. Provided a higher frame rate, we can imagine the system being able to follow the actions of the user and guiding her step-by-step through countless recipes.

Session 3: in a kitchen, in front of a coffee machine. In this session, the user opens the conversation with a direct question: "Hi, can you help me with this?" The system recognises that the appliance in question is the coffee machine, and provides detailed instructions on how to use it. This example shows that the LLMs are able to disambiguate the user's request, without the need of providing additional information.

Session 4: in the home entrance; wearing a rain jacket. In this case, a bright-coloured rain jacket immediately attracts the robot's attention. The robot asks whether it is raining outside and proceeds to have a conversation about the weather. More and more frequently conversational agents and social robots are used for entertaining and keeping company, improving the well-being of isolated people. This session is an example of how much more engaging conversation with these systems can be when powered by images together with text.

Session 5: in a bathroom; a person is lying on the ground. The person in the shot starts the interaction asking for help. As expected, the extremely rational response of the system and the calm voice of the speech synthesiser are in contrast with the criticality of the moment. However, what is relevant is that the system is able to understand that the situation is problematic, and offers advice on how to solve the problem, fully knowing the limitations of its capabilities.

Session 6: in a bedroom; holding a jacket and a t-shirt. The user tells the robot that it is raining and she has to choose what to wear. The robot is able to recognise that the person is undecided between the two pieces of clothes held, and sees that the jacket has a hood. It then proceeds to suggest to wear the jacket and keep the other to stay inside.

The six interactive sessions highlight the system's potential to enhance user interaction by grounding conversations in visual context. For example, the system's ability to recognise environments and infer contextual details, such as identifying a coffee machine or a rain jacket, leads to more natural and engaging dialogues. However, several limitations were observed during the trials, offering insights for future improvements.

A primary limitation was response speed, which varied between 1 and a maximum observed of 25 seconds. This could be solved using faster, less powerful, self-hosted models. Another gain in speed could be obtained by adopting more performant [19] transformer-based speech-to-text techniques, which would easily add support for multiple languages.

Another limitation was the temporal resolution. The system currently struggles to capture fine-grained gestures, limiting its ability to interpret dynamic cues, such as body language or subtle movements, and group interactions. Future work could explore higher frame rates or the implementation of a memory mechanism that allows the system to "look back" at previous frames for improved context awareness.

Reducing image resolution to save on prompt size did not significantly affect the quality of responses. However, the loss of detailed visual information in certain contexts suggests the need for more sophisticated image compression or summarisation techniques.

V. Conclusion

The implementation of a vision-enabled dialogue system represents a significant step forward in conversational agents by integrating real-time visual information into interactions. This fusion of language and vision enhances the system's contextual awareness, providing a more immersive and responsive conversational experience.

We propose a ready-to-use customisable implementation of such a system, tailored for use in HRI applications. We support important features such as image and text input, model selection, and prompt summarisation. The system is available as a stand-alone application, powering a Furhat robot, or as a ROS package.

While we have observed intriguing results from interactive sessions with a Furhat robot powered by this system, in the future we want to run a comprehensive evaluation of the implementation's speeds and response quality, including more dynamic (outdoor) environments and multiple users.

In conclusion, this vision-enabled dialogue system aims to simplify the development of context-aware conversational agents in HRI. By integrating visual cues, it opens new possibilities for applications where both textual and visual information are essential for a more engaging, interactive experience.

References

[1] D. Gergle, R. E. Kraut, and S. R. Fussell, "Using visual information for grounding and awareness in collaborative tasks," *Human–Computer Interaction*, vol. 28, no. 1, pp. 1–39, 2013.

[2] Y. Mohamed and S. Lemaignan, "Ros for human-robot interaction," in *2021 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2021, pp. 3020–3027.

[3] S. Macenski, T. Foote, B. Gerkey, C. Lalancette, and W. Woodall, "Robot operating system 2: Design, architecture, and uses in the wild," *Science robotics*, vol. 7, no. 66, p. eabm6074, 2022.

[4] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström, "Furhat: A Back-Projected Human-Like Robot Head for Multiparty Human-Machine Interaction," in *Cognitive Behavioural Systems*, ser. Lecture Notes in Computer Science, A. Esposito, A. M. Esposito, A. Vinciarelli, R. Hoffmann, and V. C. Müller, Eds. Springer, 2012, pp. 114–130.

[5] A. Zeng, M. Attarian, B. Ichter, K. Choromanski, A. Wong, S. Welker, F. Tombari, A. Purohit, M. Ryoo, V. Sindhwani, J. Lee, V. Vanhoucke, and P. Florence, "Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language," arXiv preprint, 2022.

[6] Y. Wang, D. Sun, R. Chen, Y. Yang, and M. Ren, "Egocentric Video Comprehension via Large Language Model Inner Speech," in *3rd International Ego4D Workshop*, 2023.

[7] R. Janssens, P. Wolfert, T. Demeester, and T. Belpaeme, ""Cool glasses, where did you get them?": Generating Visually Grounded Conversation Starters for Human-Robot Dialogue," in *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '22. IEEE Press, 2022, pp. 821–825.

[8] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," in *Thirty-Seventh Conference on Neural Information Processing Systems*, Nov. 2023.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, Jul. 2021, pp. 8748–8763.

[10] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," arXiv preprint, 2023.

[11] G. A. Abbo, S. Marchesi, A. Wykowska, and T. Belpaeme, "Social value alignment in large language models," in *Value Engineering in Artificial Intelligence*, N. Osman and L. Steels, Eds. Cham: Springer Nature Switzerland, 2024, pp. 83–97.

[12] G. A. Abbo and T. Belpaeme, "Vision language models as values detectors," To appear in Value Engineering in Artificial Intelligence, 2025.

[13] C. E. Mower, Y. Wan, H. Yu, A. Grosnit, J. Gonzalez-Billandon, M. Zimmer, J. Wang, X. Zhang, Y. Zhao, A. Zhai *et al.*, "Ros-llm: A ros framework for embodied ai with task feedback and structured reasoning," *arXiv preprint arXiv:2406.19741*, 2024.

[14] A. Raja and A. Bhethanabotla, "Operatellm: Integrating robot operating system (ros) tools in large language models," in *2024 IEEE 1st International Conference on Communication Engineering and Emerging Technologies (ICoCET)*. IEEE, 2024, pp. 1–4.

[15] R. Royce, M. Kaufmann, J. Becktor, S. Moon, K. Carpenter, K. Pak, A. Towler, R. Thakker, and S. Khattak, "Enabling novel mission operations and interactions with rosa: The robot operating system agent," *arXiv preprint arXiv:2410.06472*, 2024.

[16] A. Koubaa, A. Ammar, and W. Boulila, "Next-generation human-robot interaction with chatgpt and robot operating system," *Software: Practice and Experience*, 2024.

[17] B. Benjdira, A. Koubaa, and A. M. Ali, "Rosgpt_vision: Commanding robots using only language models' prompts," *arXiv preprint arXiv:2308.11236*, 2023.

[18] I. Lane, V. Prasad, G. Sinha, A. Umuhoza, S. Luo, A. Chandrashekaran, and A. Raux, "Hritk: the human-robot interaction toolkit rapid development of speech-centric interactive systems in ros," in *NAACL-HLT Workshop on Future directions and needs in the Spoken Dialog Community: Tools and Data (SDCTD 2012)*, 2012, pp. 41–44.

[19] R. Janssens, E. Verhelst, G. A. Abbo, Q. Ren, M. J. Pinto Bernal, and T. Belpaeme, "Child speech recognition in human-robot interaction: Problem solved?" in *Social Robotics*, 2024.