2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Grounding Conversational Robots on Vision Through Dense Captioning and Large Language Models

Lucrezia Grassi, Zhouyang Hong, Carmine Tommaso Recchiuto, Antonio Sgorbissa

Abstract— **This work explores a novel approach to empowering robots with visual perception capabilities using textual descriptions. Our approach involves the integration of GPT-4 with dense captioning, enabling robots to perceive and interpret the visual world through detailed text-based descriptions. To assess both user experience and the technical feasibility of this approach, experiments were conducted with human participants interacting with a Pepper robot equipped with visual capabilities. The results affirm the viability of the proposed approach, allowing to perform vision-based conversations effectively, despite processing time limitations.**

I. Introduction

Conversational social robots have become increasingly prevalent in various applications, from companionship for children and the elderly to customer service in shops [1]–[3]. These robots encompass a range of forms, including physical embodiments, avatars on mobile devices, and even virtual assistants. However, despite their versatility, existing social robots encounter limitations in context awareness, language understanding, memory, and emotional recognition, impeding purposeful and multi-modal conversations [4].

In November 2022, OpenAI introduced ChatGPT, built upon Generative Pre-trained Transformer 3.5 (GPT-3.5) with exceptional natural language comprehension and context-aware responses. A few months later, in March 2023, OpenAI launched GPT-4, demonstrating improved reasoning, intricate text comprehension, and logical response generation.

This research, informed by the latest advancements in the field of Artificial Intelligence (AI), has two main objectives:

1) Pioneering a novel approach for enabling robots to perceive and interpret the visual surroundings using textual descriptions.
2) Assessing the performance of the proposed approach in terms of user experience and processing times to determine its real-world applicability.

To achieve our first objective, initial investigations were conducted using the ChatGPT website[1]. Here, we provided ChatGPT with custom-tailored information to simulate visual perception, resulting in promising responses. As we delved deeper, we recognized the potential of Large Language Models (LLMs), especially the latest GPT-4. To leverage this potential, we integrated its capabilities with dense captioning — a challenging task involving detecting visual elements in images and generating coherent natural language descriptions. This fusion stands at the core of our methodology, which was later deployed on a Pepper robot.

To address our second objective, we conducted experiments where human participants interacted with the Pepper robot, equipped with visual capabilities. Our goal was to evaluate the effectiveness of the approach both in terms of user experience and technical feasibility. The analysis of participants' questionnaire responses provided valuable insights into their experiences, allowing us to compare them with state-of-the-art interaction systems. Additionally, we examined the processing times of the main operations performed by the system, including image processing, speech recognition, and response generation.

The article is structured as follows. Section II covers state-of-the-art tools and applications. Section III outlines the system architecture and delves into the prompt design. Section IV describes the experiments and discusses the results. Finally, Section V summarizes the findings and offers insights from the study.

II. State of the Art

A. Large Language Models

AI has evolved significantly, transitioning from symbolic AI to Deep Neural Networks (DNNs). Transformer-based deep neural networks have become dominant, surpassing older models like Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTMs) networks, especially in speech and natural language processing tasks [5]. Transformers also excel in image processing, especially when combined with Convolutional Neural Networks (CNNs) [6].

Examples of transformer-based models that have achieved state-of-the-art performance are BERT, T5, and the standout GPT. The Huggingface platform[2] has made these advanced models accessible to a broader audience, eliminating the need for in-depth knowledge of how they work. GPT, or Generative Pre-trained Transformer is a LLM designed for diverse Natural Language Processing (NLP) tasks. LLMs have evolved significantly, from ELMo with 94 million parameters in 2019 to PaLM with 540 billion parameters in 2023. This evolution introduced influential models like GPT-2 (1.5 billion parameters) in 2019 and GPT-3 (175 billion parameters) in 2020[3]. However, the official scale of GPT-4 is still undisclosed.

All authors are with the Department of Informatics, Bioengineering, Robotics, and Systems Engineering, University of Genoa, Via all'Opera Pia 13, 16145 Genoa, Italy.
Corresponding author's email: lucrezia.grassi@edu.unige.it

[1]https://chat.openai.com/

[2]https://discuss.huggingface.co/

[3]https://cmte.ieee.org/futuredirections/2023/04/24/how-much-bigger-can-should-llms-become/

In recent years, there has been growing interest in using LLMs to enhance robot interaction capabilities. Microsoft engineers employed ChatGPT for robotics applications, emphasizing prompt engineering and dialogue strategies [7]. LINE Corporation[4] researchers addressed LLMs limitations by creating a task-breakdown dialogue system for robots, outperforming rule-based systems in tasks like tourist-spot recommendations [8].

Furthermore, GPT-3 integration with Aldebaran Pepper and NAO robots facilitated open verbal dialogues [9]. GPT-3 is also explored for interactive learning with educational social robots as tutors [10]. LLMs also simplify task planning for robots, enabling action generation from natural language instructions [11]. In household cleanup tasks, TidyBot, a mobile robot, adapts to user preferences through language-based planning and LLMs [12].

Another noteworthy approach is the SayCan method, proposed by roboticists at Google. This method exploits the powerful 540B PaLM model along with robotic skills to enable robots to execute natural language instructions [13]. However, it lacks real-world grounding as no information about the environment is provided to the language model.

B. Prompt engineering

Generative models have powerful capabilities, but their full potential can be expressed only when they are guided by carefully crafted prompts. Prompts act as directives for LLMs, enabling them to follow guidelines, automate tasks, and control generated content. However, designing prompts for LLM-based chatbots often lacks a systematic approach among non-AI experts, leading to undesirable outputs [14].

The influence of the prompt on the performance of LLMs is so profound that it has given rise to a specialized field known as "prompt engineering" [15], [16]. Prompt engineering encompasses various techniques, including Zero-Shot, One-Shot, and Few-Shot Learning [17]. These techniques involve providing the model with different amounts of examples, ranging from none (zero-shot) to one (one-shot) or a few (few-shot), influencing its task comprehension. Another method, Chain of Thought (CoT) [18], enables complex reasoning through intermediate steps.

Combining CoT with few-shot prompting improves performance, particularly in tasks requiring thoughtful reasoning. Notably, adding a directive like "Let's think step by step" enhances zero-shot reasoning abilities. Experimental results demonstrate that zero-shot-CoT outperforms simple zero-shot LLM performance in various reasoning tasks [19].

C. Visually grounded dialogue

Integrating visual information into human-robot dialogues is essential for effective and natural interactions. A recent architecture for visually grounded human-robot dialogue combines neural networks for vision and language processing with symbolic reasoning mechanisms [20], enabling generalization to unseen objects and adaptability for diverse

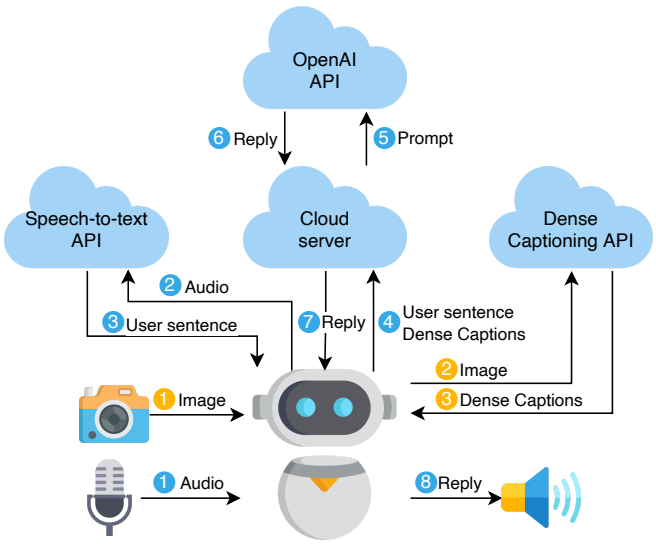


Fig. 1. System architecture showing the sequence of operations performed during the conversation. The blue and orange numbers indicate two processes that are executed in parallel.

applications. Another study introduced a context-aware approach to human-robot interaction [21], using audio-visual perception to detect contextual anomalies. These anomalies trigger various events, such as robot expressions, gestures, or utterances. Experimental results validate the robot's proficiency in generating suitable responses.

A visual scene-aware dialogue system for human-machine interaction that extracts semantic information from complex data was presented by [22]. The system employs a transformer-based neural network framework guided by user questions and outperforms baselines on the Audio-Visual Scene Aware Dialog (AVSD) dataset [23]. The ARI humanoid robot, recently deployed in a hospital reception area [24], integrates visual context through Facebook's Detectron2 framework[5]. Its dialogue capabilities are provided by the Alana chatbot [25], combining rule-based and machine-learning systems. Notably, most of these approaches do not exploit the power of recent LLMs for response generation. Moreover, they primarily focus on object detection, lacking dense captioning integration for effectively grounding conversations in the environment.

III. Methodology

A. System Architecture

The system is based on a client-server architecture depicted in Figure 1. The cloud server hosts the Dialogue service responsible for conversation management. This service takes the user sentence as input, enriched with contextual details, such as image captioning results. Then, it formulates a prompt, discussed further in Section III-B, and sends it to the GPT-4 model through the OpenAI API. Upon receiving the response, it returns it to the client. These steps are denoted by the blue numbers in Figure 1 ranging from 4

[4] https://linecorp.com/en/

[5] https://github.com/facebookresearch/detectron2

to 7. Although the Dialogue service currently handles a straightforward task, we have deployed it on a cloud server to facilitate future integration with the CAIR cloud server — a knowledge-based autonomous interaction system [26], developed by some of the authors. This integration aims to enhance the dialogue capabilities, enabling the system to manage diversity-aware [27] and multi-party [28] interactions.

The client device collects audio and visual data through its camera and microphone. During the interaction, two parallel processes, denoted as P_1 and P_2, run on the robot.

P_1, whose sequence of operations is illustrated by the orange numbers in Figure 1, handles visual data acquisition. It takes an image as input from the camera (1) and sends a request to the Microsoft Azure AI Vision API (2). The API provides the dense captions, offering detailed descriptions of complex visual scenes. These descriptions are associated with precise locations in the image through bounding boxes. While the open-source model supports both local and cloud-based utilization [29], preliminary tests have indicated its tendency to produce less precise descriptions. Once the outcome of the dense captioning task is received by the client (3), it is stored locally. This process runs throughout the interaction to ensure that the captioning result is consistently updated. The update frequency of visual information is determined by the time it takes for the client to capture an image, transmit it to the API, and receive the response.

Simultaneously, P_2, denoted by the blue numbers in Figure 1, acquires user input and provides the reply. It captures the user's spoken words (1), sends the audio to the Microsoft Azure Speech to text API (2), and retrieves the transcribed text (3). Then, the user sentence, along with the dense captions that are constantly updated by P_1, are sent to the cloud server (4). At this point, as already explained, this information is used by the Dialogue service on the cloud to generate the prompt that is sent to the language model (5) to obtain the reply (6), which is then returned to the client (7). Eventually, the client employs speech synthesis to convert text into sound and play it through its speaker (8). The text is reproduced by the robot's speech synthesizer, which, although of lower quality compared to the latest text-to-speech models, offers the advantage of being immediate and cost-free.

B. Prompt design

The response obtained from the OpenAI API in step (5) of P_2 varies based on the prompt design. To enable the robot to engage in natural conversations with users while considering visual information, we provided textual descriptions to GPT-4, instructing it to embody a visually-aware entity.

The OpenAI API interface utilizes three prompt fields: "system," "user," and "assistant," respectively defining the system's philosophy, user input, and assistant content. The system field content plays a pivotal role in anchoring the conversation to visual cues, as illustrated in Figure 2. To maintain conciseness, only selected portions of the system field content are presented.

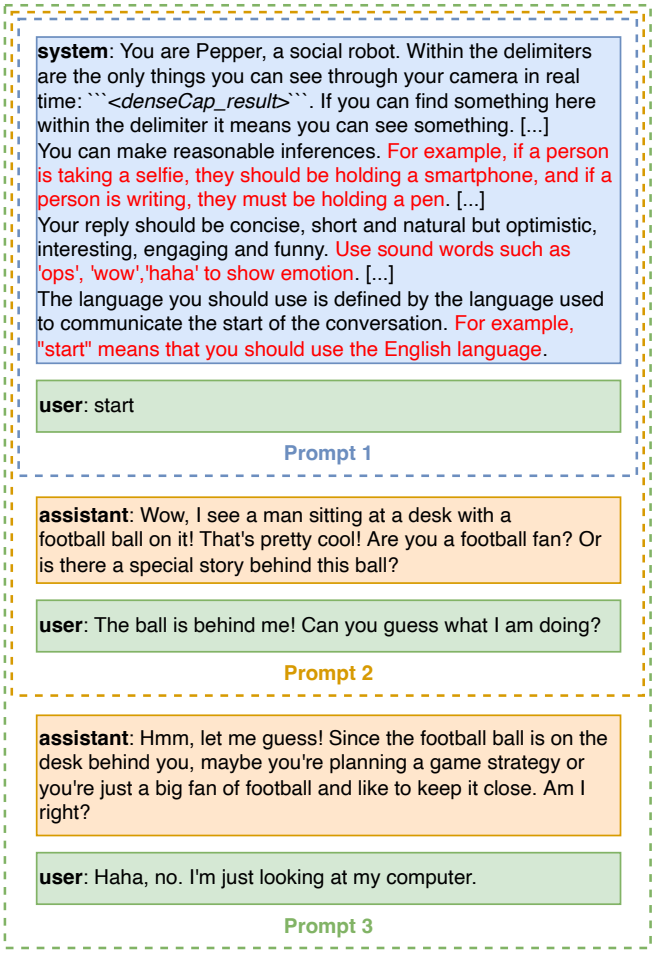
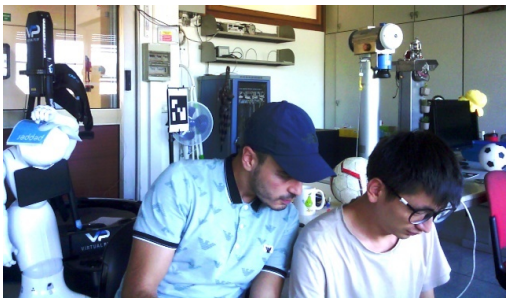


Fig. 2. Prompt engineering using state-of-the-art approaches such as zero-shot one-shot and few-shot learning, in combination with the chain of thought reasoning technique.

The prompt engineering incorporates the methods outlined in Section II, including zero-shot, one-shot, few-shot learning, and chain of thought approach, emphasized in red in Figure 2. The CoT prompting method serves as the foundation of the design, with the entire system field content serving as a sequence of instructions, with intermediate steps for the model to consider in generating its response.

The model is instructed to assume the persona of the Pepper social robot, as used in our experiments, equipped with vision capabilities thanks to its camera. The *denseCap_result* contains a list of the dense captions linked to the latest robot-captured image (Figure 3). The model is guided to make inferences based on perceived objects or individuals during interactions, with illustrative examples to enhance response accuracy. Furthermore, instructions are provided regarding the tone, expression, and language the robot should employ throughout the interaction. Notably, the language is determined by the content of the initial user field, indicated in green in Figure 2.

The first prompt, referred to as `Prompt 1` in Figure 2, encloses the system and the user fields. The latter serves as the starting point for GPT to initiate the conversation in the



denseCap_result: ['a man sitting at a desk with a ball on it',
'a man wearing a hat',
'a person playing a videogame',
'a football ball on a table',
'a white robot with a black tablet']

Fig. 3. Dense captioning result of an image.

Fig. 4. Experimental setup showing a participant interacting with the Pepper robot, the external PC housing the microphone, and several objects (i.e., a football ball, a cup, a mouse, a marker, some LEGO bricks, and a pair of pliers) on the table, which can be used during the conversations.

appropriate language. GPT's first reply is meant to capture the user's interest by selecting intriguing elements from image descriptions. The model's response is encapsulated within the "assistant" field, depicted in orange in Figure 2, and is the one received by the robot in step 7 of P_2.

A crucial concept involves integrating context-containing information with the current sentence before presenting it as input to the language model. This contextual information serves as a form of memory. In this work, context awareness is achieved using a basic retention approach, which entails keeping a limited number of recent dialogue items while discarding older ones. During each conversation turn, the latest assistant reply and user utterances are added to the previous dialogue items, as shown in Figure 2.

For instance, in response to the initial request performed with `Prompt 1`, GPT-4 replies with a message that leverages visual information and conveys curiosity about the environment. This response is then appended to the prompt, alongside the subsequent user reply, creating `Prompt 2`. This iterative process continues with each conversational turn. The progressive modifications in the prompt are visually depicted in Figure 2.

IV. Experiments and Results

To address the second objective of this study, which involves evaluating the impact of integrating visual information into our conversational system, we conducted experiments where participants interacted with a social robot using vision information. Then, we analyzed their replies to a validated questionnaire and assessed the system's performance, with an emphasis on the time spent at different operational stages.

The experiments, as depicted in the setup shown in Figure 4, were conducted within a robotics laboratory setting, using the Pepper robot developed by Aldebaran[6]. Note that we used an external computer, connected to the robot via a socket, as a microphone due to the limitations of Pepper's internal microphone. Throughout the experiments, participants were positioned in front of the robot, while the computer with the microphone was placed nearby. On the left side of each participant, a table was arranged, featuring various objects.

We recruited a total of 14 university students aged between 20 and 30 years who expressed their interest in participating. There were no restrictions on their native language as GPT-4 could understand and generate replies in most languages. Importantly, participants were initially unaware of the robot's vision abilities. They were instructed to engage in open conversations with the robot and were encouraged to use the objects on the table. The language used during the conversations was determined by the participant's native language, encompassing English, Italian, and Chinese. A video showing interaction examples in these languages is available on YouTube [7].

In each experimental session, lasting approximately 15 minutes, participants engaged in a 10-minute conversation with the robot and spent around 5 minutes completing a questionnaire assessing their interaction experience. Throughout these interactions, we also collected data on computational time and content size for tasks such as image processing, speech recognition, and response generation.

It is important to note that no correlation is established between any individuals identified by the dense captioning and the dialog interlocutors. Currently, the system does not incorporate face detection, thus it is not feasible to correlate the location of recognized people with the individuals who are actually interacting with the robot.

A. System performance

This section analyzes system data gathered during the experiments, focusing on time measurements for image capturing, dense captioning, speech recognition, and model response generation.

1) Image capturing: The first examined indicator is the time interval between consecutive image captures. Long intervals may result in image data being collected after the response generation process has already begun, while overly frequent collections may impact the robot's performance and increase image processing costs. During this time frame, the Pepper robot engages in a series of tasks, encompassing image capture, storage, compression, transmission, and waiting

[6]`https://www.aldebaran.com`

[7]`https://youtu.be/B9iWC146Qc0`

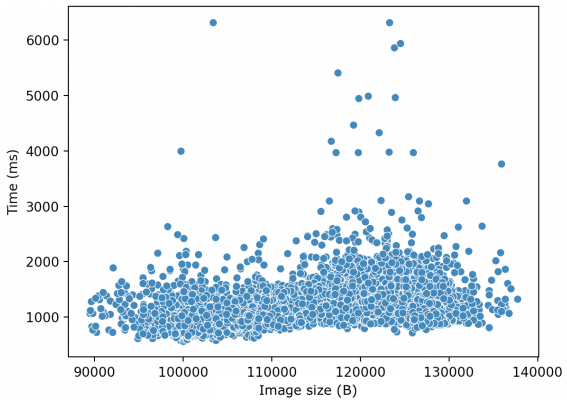


Fig. 5. Dense captioning task duration as the image size varies.

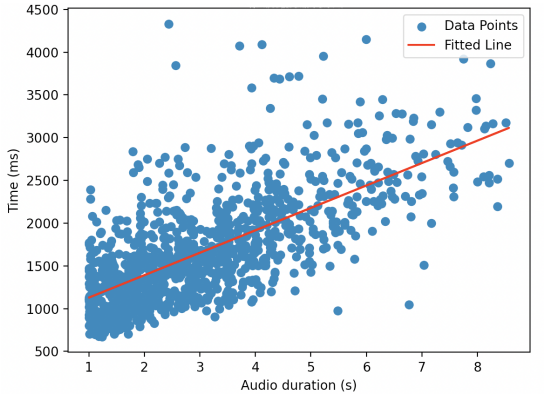


Fig. 6. Correlation between audio length and speech-to-text time.

for the dense captioning result before proceeding to capture the subsequent image. The analysis was performed over 5354 samples resulting in an average interval time between captured images of 2.6 s. Such a value is mainly due to the hardware limitations of the Pepper robot.

This result holds significant importance as it indicates the frequency with which visual information is updated during the dialogue. It is crucial to note that this interval is significantly affected by the resolution of the captured images, which was set at 640x480 pixels to ensure higher accuracy in the dense captioning task. However, opting for lower resolutions may reduce this time.

2) Dense captioning: In this section, we analyze the relationship between image size and the time required to obtain the captioning result, as illustrated in Figure 5. Specifically, to perform the dense captioning task we employed Microsoft Azure AI Vision API. It is worth noting that the image size variation is in the order of bytes due to the consistent resolution maintained throughout the experiments.

On average, the dense captioning task for an image, computed from a dataset of 5159 samples, takes approximately 1.2 s. An analysis using a Pearson correlation coefficient of 0.316 suggests a modest positive correlation between image size and captioning time. While this relationship may not strictly follow a linear pattern, it suggests that larger images may require slightly more time for captioning.

3) Speech recognition: This section explores the relationship between audio recording duration and the time required for voice recognition, along with our methodology for optimizing recognition times. Our approach involves transmitting data for recognition when the Root Mean Square (RMS) of the noise falls below a specific threshold for more than one second. If the silence period extends beyond two seconds, it is interpreted as the conclusion of the speaker's sentence. This approach aims to minimize user waiting time by processing content incrementally during speech, avoiding the need to recognize a substantial amount of speech all at once. The actual transcription task is performed by Microsoft Azure AI Speech to text API.

Figure 6 illustrates the correlation between audio length and speech recognition time, offering a clear insight into the trend. The transcription task, computed over 1092 samples, took on average 1.7 s, with the average audio segment duration being approximately 3 s. A Pearson correlation coefficient of 0.7 indicates a strong positive relationship between audio length and recognition time.

Note that the total time for speech recognition only surpasses the two-second silence threshold when transcribing the last audio segment requires more than one second. This happens because the transcription of the final audio segment begins after a shorter one-second pause following the user's speech. However, considering the common practice to include pauses of at least one second while speaking and the average transcription time of 1.7 s, it can be deduced that the entire speech recognition process takes, on average, 2.7 s. Consequently, the transcription result becomes available, on average, 0.7 s after the system has acknowledged the user's completion of speech.

4) Model response generation: Let us now analyse the response time, i.e., the duration between sending a request to OpenAI and receiving the result. It is important to note that the content length is measured by the number of characters in the response from OpenAI.

Figure 7 depicts the correlation between the length of the generated text, in terms of the number of characters, and the time users must wait for the result. On average, based on an analysis of 885 samples, users waited approximately 6.1 s to receive a reply. The average length of the generated text was around 150 characters, corresponding to roughly 83 tokens. A Pearson correlation coefficient of 0.63 indicates a reasonably strong connection between the length of the generated text and the processing time. Note that the length of the generated text can be influenced by providing appropriate instructions within the prompt. However, conducting a detailed analysis of this phenomenon falls outside the scope of the present work.

5) System response time: The system's overall response time spans from the user's completion of speech to when the robot begins delivering the model's reply. Mainly, it depends on two factors: the time required for speech recognition (averaging 2.7 s) and, mostly, the time taken by GPT-4 to generate a response (averaging 6.1 s).

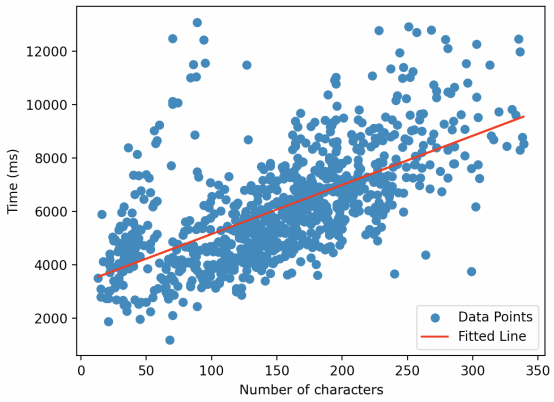


Fig. 7. Correlation between the number of characters in the generated text and the time required to receive the response from GPT-4.

To mitigate most of the delay we integrated brief phrases like "Let me think..." or "Give me a moment..." into the robot's responses. These expressions are triggered as soon as the robot receives the user's sentence resulting from the speech recognition (depicted as the blue step 3 in Figure 1). They serve a dual purpose: they promptly inform the user that the system has comprehended their input and is awaiting the model's response, while effectively filling any potential silence that may occur during the reply generation process. This approach reduces the perceived response time and enhances the user experience.

Note that the time taken for image capturing and dense captioning does not affect the response time. Rather, its significance lies in enhancing visual reactivity.

B. Questionnaire results

The questionnaire used in this study, adapted from the validated Godspeed questionnaire [30], assessed three key aspects: Likeability, Perceived Intelligence, and Perceived Safety. The Likeability scale comprised 5 items evaluating participants' feelings towards the robot, including their liking of the robot, its friendliness, kindness, pleasantness, and overall niceness. The Perceived Intelligence dimension consisted of 5 items aimed at assessing whether participants perceived the robot as competent, knowledgeable, responsible, intelligent, and sensible. Finally, the Perceived Safety section included 3 items measuring participants' feelings of relaxation, calmness, and surprise in the robot's presence. Responses were provided using a 1-5 Likert scale.

The analysis of the questionnaire yielded the following results: an average Likeability score of 4.3 (± 0.7), an average Perceived Intelligence of 4.0 (± 0.8), and an average Perceived Safety of 4.2 (± 0.9).

An exploratory factor analysis [31] revealed low eigenvalues and low reliabilities for certain components within the five Godspeed scales. However, this was primarily observed for the Animacy and Perceived Safety scales, whereas Anthropomorphism, Likeability, and Perceived Intelligence showed higher reliability.

To contextualize the obtained results for the reliable scales, we analyzed the state-of-the-art research that used the Godspeed questionnaire to evaluate human-robot interaction. Notably, a meta-analysis of 49 studies with 3,556 participants analyzed the Anthropomorphism and Likeability scores of the Godspeed questionnaire [32]. The Likeability scores ranged from 2.63 to 4.98, with a median of 3.92, indicating the positivity of our average Likeability score which is largely above the median. However, it is important to note that we could not find reference values for the Perceived Intelligence scale, but we could only compare our results with those of other researchers who employed this questionnaire [33]–[35]. The obtained values were between 4 and 4.4 leading us to hypothesize that, in general, average scores of 4 or above can be deemed satisfactory.

V. Conclusion

This study introduced an approach using LLMs for context-aware conversations enriched with visual information. We implemented a conversational system, combining dense captioning and the GPT-4 API. The system was deployed on the Pepper robot to assess its feasibility and time consumption at various stages of data processing.

Our assessment involved 14 participants engaging with the robot, equipped with visual perception capabilities. Participants responded to 13 questions extracted from the Godspeed questionnaire, evaluating Likeability, Perceived Intelligence, and Perceived Safety. Simultaneously, we recorded data related to time intervals between consecutive image captures, image captioning tasks, speech recognition processes, and model-generated reply times.

The experiments highlighted that GPT-4 content generation is the primary factor contributing to the overall system response time. This delay has the potential to influence the quality of the entire interaction. To tackle this issue, we implemented a practical solution. Our solution serves the dual purpose of signaling to users that the system has acquired their input while filling in pauses during the reply generation process for smoother interaction.

Despite the current responsiveness constraints of GPT-4 and the delay of $2.6\,\mathrm{s}$ in updating visual data, mainly due to Pepper robot's limitations, the questionnaire results affirmed a positive user experience. This highlights the potential of our approach and lays the foundation for further optimization.

As a limitation, it is worth mentioning that we have not established any correlation between individuals identified by the dense captioning and the dialogue interlocutors. This is due to the absence of face detection in our system, making it challenging to correlate the recognized individuals' locations with those who are actively engaging with the robot.

For future work, integrating face detection capabilities into our system could enable us to accurately associate recognized individuals with the participants interacting with the robot. This enhancement would facilitate the generation of context-aware responses during user interactions. Additionally, our goal is to integrate this technology with the CAIR cloud server [26], developed by some of the authors, to enhance its dialogue capabilities further.

REFERENCES

[1] J. Broekens, M. Heerink, and H. Rosendal, "Assistive social robots in elderly care: a review," *Gerontechnology*, vol. 8, no. 2, pp. 94–103, 2009.

[2] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, "Social robots for education: a review," *Sci. Rob.*, vol. 3, no. 21, 2018.

[3] M. Niemelä, P. Heikkilä, H. Lammi, and V. Oksman, *A Social Robot in a Shopping Mall: Studies on Acceptance and Stakeholder Expectations*. Cham: Springer International Publishing, 2019, pp. 119–144.

[4] N. Mavridis, "A review of verbal and non-verbal human–robot interactive communication," *Rob. Auton. Syst.*, vol. 63, pp. 22–35, 2015.

[5] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI Open*, vol. 3, pp. 111–132, 2022.

[6] F. Yuan, Z. Zhang, and Z. Fang, "An effective CNN and transformer complementary network for medical image segmentation," *Pattern Recognit.*, vol. 136, p. 109228, 2023.

[7] S. Vemprala, R. Bonatti, A. Bucker, and A. Kapoor, "ChatGPT for robotics: design principles and model abilities," *Microsoft Auton. Syst. Robot. Res*, vol. 2, p. 20, 2023.

[8] T. Yamazaki, K. Yoshikawa, T. Kawamoto, T. Mizumoto, M. Ohagi, and T. Sato, "Building a hospitable and reliable dialogue system for android robots: a scenario-based approach with large language models," *Adv. Rob.*, vol. 0, no. 0, pp. 1–18, 2023.

[9] E. Billing, J. Rosén, and M. Lamb, "Language models for human-robot interaction," in *Proc. of the ACM HRI, Stockholm, Sweden*. ACM Digital Library, 2023, pp. 905–906.

[10] S. Sonderegger, "How generative language models can enhance interactive learning with social robots." *International Association for Development of the Information Society*, 2022.

[11] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, "ProgPrompt: generating situated robot task plans using large language models," in *IEEE ICRA*, 2023, pp. 11 523–11 530.

[12] J. Wu, R. Antonova, A. Kan, M. Lepert, A. Zeng, S. Song, J. Bohg, S. Rusinkiewicz, and T. Funkhouser, "TidyBot: personalized robot assistance with large language models," 2023.

[13] M. Ahn, A. Brohan, N. Brown *et al.*, "Do as I can, not as i say: grounding language in robotic affordances," 2022.

[14] J. D. Zamfirescu-Pereira, R. Wong, B. Hartmann, and Q. Yang, "Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts," in *Proc. of the CHI Conference*, 2023, pp. 1–21.

[15] A. Gao, "Prompt engineering for large language models," *SSRN*, 2023.

[16] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, jan 2023.

[17] H. Dang, L. Mecke, F. Lehmann, S. Goller, and D. Buschek, "How to prompt? opportunities and challenges of zero- and few-shot learning for human-AI interaction in creative applications of generative models," 2022.

[18] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023.

[19] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," 2023.

[20] X. Sun, C. Weber, M. Kerzel, T. Weber, M. Li, and S. Wermter, "Learning visually vrounded human-robot dialog in a hybrid neural architecture," *Lect. Notes Comput. Sci.*, vol. 13530 LNCS, p. 258 – 269, 2022, cited by: 0.

[21] S. Paplu, H. Ahmed, A. Ashok, S. Akkus, and K. Berns, "Multimodal perceptual cues for context-aware human-robot interaction," *Mech. Mach. Sci.*, vol. 127, p. 283 – 294, 2023.

[22] F. Liu, B. Guo, H. Wang, and Y. Liu, "Visual scene-aware dialogue system for cross-modal intelligent human-machine interaction," *Commun. Comput. Inf. Sci.*, vol. 1682 CCIS, p. 337 – 351, 2023, cited by: 0.

[23] H. Alamri, V. Cartillier, A. Das, J. Wang, A. Cherian, I. Essa, D. Batra, T. K. Marks, C. Hori, P. Anderson, S. Lee, and D. Parikh, "Audio visual scene-aware dialog," in *Proc. of the IEEE/CVF Conference*, 2019, pp. 7558–7567.

[24] N. Gunson, D. Hernandez Garcia, W. Sieińska, A. Addlesee, C. Dondrup, O. Lemon, J. L. Part, and Y. Yu, "A visually-aware conversational robot receptionist," in *Proc.of the 23rd SIGdial*. Edinburgh, UK: Association for Computational Linguistics, Sep. 2022, pp. 645–648.

[25] A. C. Curry, I. Papaioannou, A. Suglia, S. Agarwal, I. Shalyminov, X. Xinnuo, O. Dusek, A. Eshghi, I. Konstas, V. Rieser, and O. Lemon, "Alana v2: entertaining and informative open-domain social dialogue using ontologies and entity linking," in *1st Proc. of Alexa Prize*, 2018.

[26] L. Grassi, C. T. Recchiuto, and A. Sgorbissa, "Sustainable cloud services for verbal interaction with embodied agents," *Intell. Serv. Robo.*, 2023, accepted for publication.

[27] L. Grassi, D. Canepa, A. Bellitto, M. Casadio, A. Massone, C. T. Recchiuto, and A. Sgorbissa, "Diversity-aware verbal interaction between a robot and people with spinal cord injury," in *Proc. of the IEEE RO-MAN Conference*, Busan, South Korea, 2023.

[28] L. Grassi, C. T. Recchiuto, and A. Sgorbissa, "Robot-induced group conversation dynamics: a model to balance participation and unify communities," in *Proc. of the IEEE IROS Conference*, Detroit, USA, 2023.

[29] V.-Q. Nguyen, M. Suganuma, and T. Okatani, "Grit: Faster and better image captioning transformer using dual visual features," in *Proc. of the ECCV Conference*. Springer, 2022, pp. 167–184.

[30] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *Int. J. Social Rob.*, vol. 1, pp. 71–81, 2009.

[31] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, "The robotic social attributes scale (RoSAS): development and validation," in *Proc. of the ACM/IEEE Conference*, 2017, pp. 254–262.

[32] M. Mara, M. Appel, and T. Gnambs, "Human-like robots and the uncanny valley," *Zeitschrift für Psychologie*, 2022.

[33] D. S. Syrdal, K. Dautenhahn, M. L. Koay, Kheng Leeand Walters, and W. C. Ho, "Sharing spaces, sharing lives–the impact of robot mobility on user perception of a home companion robot," in *Proc. of the ICSR Conference*. Springer, 2013, pp. 321–330.

[34] S. Tobis, J. Piasek-Skupna, and A. Suwalska, "The Godspeed questionnaire series in the assessment of the social robot TIAGo by older individuals," *Sensors*, vol. 23, no. 16, 2023.

[35] M. Maroto-Gómez, Á. Castro-González, M. Malfaz, and M. Á. Salichs, "A biologically inspired decision-making system for the autonomous adaptive behavior of social robots," *Complex Intell. Syst.*, pp. 1–19, 2023.