ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

Embodying Language Models in Robot Action

Connor Gäde, Ozan Özdemir, Cornelius Weber and Stefan Wermter *

University of Hamburg - Dept of Informatics
Vogt-Kölln-Straße 30, 22527 Hamburg - Germany

Abstract. Large language models (LLMs) have achieved significant recent success in deep learning. The remaining challenges in robotics and human-robot interaction (HRI) still need to be tackled but off-the-shelf pre-trained LLMs with advanced language and reasoning capabilities can provide solutions to problems in the field. In this work, we realise an open-ended HRI scenario involving a humanoid robot communicating with a human while performing robotic object manipulation tasks at a table. To this end, we combine pre-trained general models of speech recognition, vision-language, text-to-speech and open-world object detection with robot-specific models of visuospatial coordinate transfer and inverse kinematics, as well as a task-specific motion model. Our experiments reveal robust performance by the language model in accurately selecting the task mode and by the whole model in correctly executing actions during open-ended dialogue. Our innovative architecture enables a seamless integration of open-ended dialogue, scene description, open-world object detection and action execution. It is promising as a modular solution for diverse robotic platforms and HRI scenarios.

1 Introduction

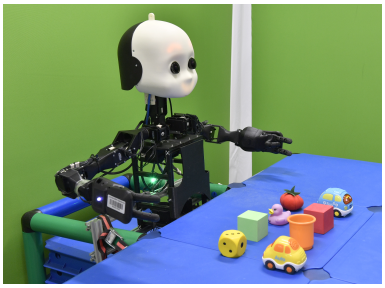

Fig. 1: NICO in our scenario.

The recent advances made in deep learning have led to successful applications of artificial intelligence such as chatbots, visual object detection, image captioning and speech recognition. The advent of the Transformer architecture [14] has brought about the possibility of training ever larger models with enormous amounts of data. Omnipresent large language models (LLMs), for example, have usually billions of parameters and are trained on trillions of text tokens. Although these large models perform well on tasks they are trained on, training them on such large datasets requires a considerable amount of computational power, which is not available to small labs or end users. Fortunately, the availability of pre-trained models renders training from scratch unnecessary.

In this work, we propose a modular approach, ELMiRA[1] (**E**mbodying **L**anguage **M**odels **i**n **R**obot **A**ction) that integrates highly capable foundation mod-

*This work was supported by the German Research Foundation (DFG) under Project TRR 169 Crossmodal Learning (CML) and LeCAREbot. Philipp Allgeuer contributed software.

[1]Our project website with an exemplary video can be found at `https://knowledgetechnologyuhh.github.io/ELMiRA`

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

els trained on big datasets for a conversational human-robot interaction (HRI) scenario. ELMiRA uses pre-trained models of automatic speech recognition (ASR), text-to-speech (TTS), vision-language (VLM) and object detection, which do not require further training. To adapt ELMiRA to our robotic setup, we employ a visuospatial coordinate transfer network and an inverse kinematics (IK) solver. Moreover, we devise a specific motion planner to perform several actions within our object manipulation scenario. By utilising multiple off-the-shelf models, our humanoid robot NICO [7] can manipulate objects on a table while conversing with a human in an open-ended fashion (Fig. 1). This brings capabilities of pre-trained models such as dialogue, zero-shot open-vocabulary object detection and scene description into a robotic application.

2 Related Work

Recently, many approaches have been proposed utilising LLMs or VLMs in the context of robotic manipulation [6, 13]. SayCan [1] uses LLMs to split high-level instructions into executable actions, which are evaluated by a value function in terms of affordance. It chooses actions with a combined score of high common-sense relevance and affordance. Likewise, ViLa [5] employs GPT4-V [11] to decompose high-level instructions into low-level executable actions for object manipulation. Jointly processing vision and language by a VLM results in a task-focused understanding of the current scene based on the given instruction. ViLa outperforms SayCan in common-sense tasks and works with multimodal input instructions, i.e. providing a goal image or a combined image-language goal. TidyBot [16] exploits common-sense knowledge inherent in LLMs for generalising to object types according to user preferences in a room-tidying task. It can generalise to objects regarding their attributes like colour or purpose. PIVOT [10] leverages VLMs for robotic control by casting robotic tasks as VQA problems. Scene images are annotated with possible movements for spatial control and fed to a VLM with an accompanying query for the best actions. The actions are iteratively optimised based on the VLM's answers without learning. However, ambiguities in depth information degrade PIVOT's performance, indicating a need for an adaptive mechanism. Generally, the state-of-the-art approaches focus on action planning, while we capitalise on LLMs to facilitate open-ended multi-turn dialogue. Moreover, the zero-shot generalisation capabilities of ELMiRA allow us to detect and localise all objects as well as describe any given scene independently of our specific scenario.

3 ELMiRA: A Modular HRI Approach

To leverage state-of-the-art models in different domains, we devise the modular ELMiRA architecture, shown in Fig. 2. It is composed of an ASR, a VLM, an object detector, an visuospatial coordinate transfer unit, a motion planner, an IK solver and a TTS model. We evaluate it in a tabletop object manipulation task-oriented HRI scenario where a user communicates verbally with the robot. The

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

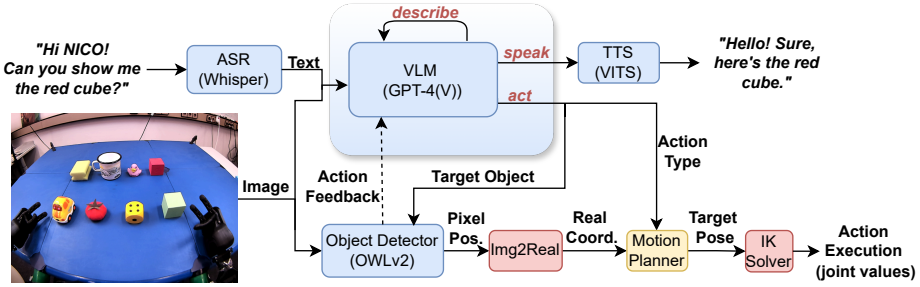

Fig. 2: ELMiRA architecture. It accepts as input human user speech and images from the robotic eye camera; it outputs speech and executes actions. Blue modules denote pre-trained off-the-shelf models, red modules denote robotic platform-specific models and the yellow module is task-specific. Akin to a conductor of an orchestra, the VLM decides whether to output speech (*speak* mode), describe the current scene (*describe* mode) or trigger a robot action (*act* mode).

conversation can range from chitchat to task-related visual processing requiring action commands. The robot needs to know when to speak and when to execute an action according to the user instruction.

Vision Language Model The VLM assumes the role of a dialogue manager. It receives user input in textual format via the ASR model (Whisper [12] adapted from Allgeuer et al. [2]) and sends textual output to the TTS model (VITS [8]) to produce it as speech. We use GPT-4V as VLM, while also employing the GPT-4 LLM via the OpenAI API. We utilise both the text-only GPT-4 and the image-allowing GPT-4V since the GPT-4V is not yet available as a conversational assistant that can manage a dialogue. We therefore deploy GPT-4 as the chatbot that has conversation memory, while triggering a one-time response only GPT-4V instance when visual processing is required[2]. Based on the user input and visual observation, GPT-4 chooses one of the three task modes:

- *speak*: When the given user input is not related to the tabletop scenario, ELMiRA generates a textual output via GPT-4, which is passed to the TTS model to produce speech.
- *describe*: When the user asks the robot to describe what it sees on the table, the current scene image is fed to the GPT-4V with the user input. The GPT-4V then passes the description of the tabletop scene to GPT-4 which in turn triggers the TTS to produce speech.
- *act*: When the user asks the robot to manipulate an object, the VLM triggers the action mode and infers the action type and target object. The target object name is passed to the object detector alongside the current

[2]The recent multimodal GPT-4o can replace GPT-4 and GPT-4V together.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

image to localise the target object in pixel space. The pixel positions are then passed to the *Visuospatial Coordinate Transfer* model that locates the real-world coordinates of the object. These coordinates and the action type are fed to the motion planner to find the target arm pose including orientation, which are finally given to the IK solver to produce the joint angle values of the suitable arm for action execution.

Object Detection In the *act* mode, after GPT-4 has extracted the target object name from the user action command, OWLv2 [9] detects the target object on the table. It receives the current scene image through the robot eye camera as well as the target object name as input. OWLv2 can detect multiple instances of the same object but we choose the one with the highest score and extract the pixel coordinates of the bottom centre of the bounding box. In case the target object cannot be found on the table or is placed in an unreachable position, the object detection module informs the VLM that the action cannot be executed.

Visuospatial Coordinate Transfer The Img2Real module transforms the pixel coordinates to 3D real-world coordinates. It uses a multi-layer perceptron (MLP) trained as an implicit energy-based model (EBM) [3] with an InfoNCE loss function and sampling-based derivative-free optimisation for inference. The implicit MLP is trained in advance by distinguishing a set of uniformly distributed real-world coordinates from random counter-examples based on the corresponding points in image pixel space. During inference, it finds the real-world coordinates for given pixel positions of the target object by iteratively resampling random candidates based on their predicted probabilities and adding Gaussian noise.

Motion Planner Our motion planner decides whether to use the left or right arm of the robot based on the real-world coordinates of the target object and outputs the corresponding end-effector trajectory of the chosen arm for a given action type (i.e. *show*, *touch*, *push-forwards*, *push-leftwards* and *push-rightwards*). Each target pose in the trajectory consists of the Cartesian (x, y, z) position of the end-effector and its target orientation as a unit quaternion.

Inverse Kinematics Solver The joint configurations needed to execute the robot arm's trajectory are computed using EvoIK [4], an evolutionary IK solver which aligns the forward kinematics of its population with the target pose by minimising the weighted sum of the Euclidean distance between the position vectors and the geodesic distance between the orientation quaternions. The trajectory is computed iteratively, using the previous joint angles as the initial centre of the population for the following step to accelerate the computations and guide the algorithm to find solutions which are close to each other.

The model components, which are connected by the robot operating system (ROS), have the following response times, averaged over 60 trials: VLM 8.98s including GPT-4 together with GPT-4V, object detection 0.57s, Img2Real 0.05s, motion planner <1ms, IK solver 0.26s.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

Table 1: Mode Detection Success

Act	Describe	Speak
86.67±4.71	46.67±28.67	100.0±0.0

Table 2: Action Execution Success

Forwards	Leftwards	Rightwards
79.17	86.96	72.0

4 HRI Experiments with NICO

We conduct two sets of experiments with the NICO robot on our tabletop setup, aimed at evaluating the robustness of ELMiRA in two aspects: mode selection and action execution.

Mode Selection Experiments We test the mode selection performance of our method by having three instances of a scripted human-robot conversation, involving 60 turns, where we check whether the VLM detects the correct mode intended by the user. The detection success rate for each mode is given in Table 1. In most cases (77% on average), ELMiRA understands the user's intention and switches to the correct mode. However, frequently the VLM refuses to describe a scene in *describe* mode and goes into *speak* mode instead. In the few cases, when it refuses to trigger the *act* mode, it also chooses the *speak* mode.

Action Execution Experiments The action execution experiments involve three directions of push (*forwards*, *leftwards* and *rightwards*) on 8 different objects (a sponge, a die, a rubber duck, a toy tomato, a green cube, a toy car, a cup and a red cube), with a total of 72 executions. We use a minimum threshold of 2 cm displacement in the direction of the intended push action as an objective success metric. Table 2 shows the action execution results per push direction. Overall NICO executes the given action with an average accuracy of 79%. The *push-leftwards* action is the most successful, while *push-rightwards* is the least successful action. Fig. 3 displays the distribution of target object displacements in the intended direction; samples over 2 cm (dashed line) are considered correct. NICO clearly pushes the target object into the right direction in most cases, with few exceptions in which the target is moved in the wrong direction due to mistakes in the IK solution. Apart from the push actions, we tested ELMiRA with show and touch actions, both of which were distinguishably and correctly executed.

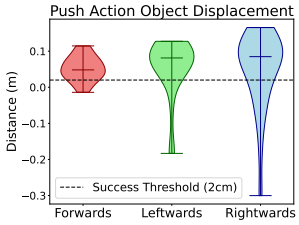

Fig. 3: Object displacements per push direction.

5 Conclusion

We have proposed a novel modular architecture leveraging recent progress in LLMs, open-vocabulary object detection and speech recognition. Specifically,

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

we have used pre-trained general ASR, VLM, object detection and TTS models alongside small robotic-platform- and task-specific modules for a dialogue-based tabletop object manipulation scenario. ELMiRA is a starting point for facilitating full-fledged dialogue and robotic action through a seamless integration of open-world object detection, scene description and general conversational skills. By drawing inspiration from neuroscience concepts [15] and leveraging increasingly capable foundation models, our modular approach can widely benefit HRI.

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, et al. Do as I can, not as I say: Grounding language in robotic affordances. In *Conference on Robot Learning*, pages 287–318, 2022.

[2] Philipp Allgeuer, Hassan Ali, and Stefan Wermter. When robots get chatty: Grounding multimodal human-robot conversation and collaboration. In *International Conference on Artificial Neural Networks (ICANN)*, 2024.

[3] Pete Florence, Corey Lynch, Andy Zeng, Oscar A Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. In *Conference on Robot Learning*, pages 158–168. PMLR, 2022.

[4] Connor Gäde, Jan-Gerrit Habekost, and Stefan Wermter. Domain adaption as auxiliary task for sim-to-real transfer in vision-based neuro-robotic control. In *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024.

[5] Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, and Yang Gao. Look before you leap: Unveiling the power of GPT-4V in robotic vision-language planning. *arXiv preprint arXiv:2311.17842*, 2023.

[6] Wenlong Huang, Fei Xia, Ted Xiao, et al. Inner monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning*, pages 1769–1782, 2023.

[7] Matthias Kerzel, Erik Strahl, Sven Magg, Nicolás Navarro-Guerrero, Stefan Heinrich, and Stefan Wermter. NICO—Neuro-Inspired COmpanion: A developmental humanoid robot platform for multimodal interaction. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 113–120. IEEE, 2017.

[8] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR, 2021.

[9] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36, 2024.

[10] Soroush Nasiriany, Fei Xia, Wenhao Yu, Ted Xiao, et al. PIVOT: Iterative visual prompting elicits actionable knowledge for VLMs. *arXiv preprint arXiv:2402.07872*, 2024.

[11] OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023.

[12] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[13] Chao Tang, Dehao Huang, Wenqi Ge, Weiyu Liu, and Hong Zhang. GraspGPT: Leveraging semantic knowledge from a large language model for task-oriented grasping. *IEEE Robotics and Automation Letters*, 2023.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[15] Stefan Wermter, Günther Palm, and Mark I. Elshaw. *Biomimetic Neural Learning for Intelligent Robots. Intelligent Systems, Cognitive Robotics and Neurosci.* Springer, 2005.

[16] Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. TidyBot: Personalized robot assistance with large language models. *Autonomous Robots*, 47(8):1087–1102, 2023.