EMOROBCARE: A Low-Cost Social Robot for Supporting Children with Autism in Therapeutic Settings

Sara Cooper[1], Bartomeu Pou[1,2], Arnau Mayoral-Macau[1], Alberto Redondo[3], David Rios[3], and Raquel Ros[1]

[1] Artificial Intelligence Research Institute (IIIA-CSIC), Bellaterra, 08193, Spain
{name.surname}@iiia.csic.es
[2] Universitat Politècnica de Catalunya (UPC), Carrer de Jordi Girona, 31, Les Corts, 08034 Barcelona
[3] Mathematical Sciences Institute (ICMAT-CSIC), Madrid, 28049, Spain
{name.surname}@icmat.es

Abstract. The EMOROBCARE project aims at developing a low-cost, expressive social robot designed to support therapy for children with Autism Spectrum Disorder. This paper presents the system software design and development tools. We illustrate an example use case game that combines perception, speech, reasoning and expressive capabilities of the robot. Additional games will be developed through co-design with therapists in order to foster communication and emotional expression.

Keywords: ASD · Child–Robot Interaction · Architecture

1 Introduction

Autism Spectrum Disorder (ASD) is characterized by challenges in social communication, interaction, and behaviour. Children with ASD Level 2, in particular, often exhibit limited verbal communication and require structured, personalized therapeutic interventions. Social robots have emerged as promising tools in this domain, offering consistent, engaging, and non-judgmental interaction partners that can support therapy goals [3].

Robots such as QTrobot, NAO, and Kaspar have demonstrated effectiveness in supporting imitation, emotion recognition, and joint attention through structured and predictable interactions [8] [7]. These robots are often used in clinical or educational settings, where their simplified appearance and expressive capabilities help reduce sensory overload and sustain engagement. Platforms such as Jibo and MARIA T21 have been adapted for long-term interventions, integrating serious games and real-time feedback to enhance communication and learning outcomes [11] [9]. Less sophisticated robots, such as OPT and parrot-like companions, have also shown promise in emotion recognition and turn-taking activities, contributing to the development of specific social skills [12]. Despite these advances, many existing systems are limited by high costs, proprietary ecosystems.

The EMOROBCARE project is a multidisciplinary initiative aimed at designing and validating a low-cost, expressive social robot to assist therapists working with children with ASD Level 2. The project emphasizes affordability, modularity, and open-source development, leveraging technologies such as ROS2, Jetson-based hardware, and lightweight speech and vision models. Its primary goal is to evaluate the feasibility and therapeutic value of the robot in real-world therapy sessions. These sessions are structured around a triadic interaction model involving the therapist, the child, and the robot. The robot engages the child through interactive games, expressive speech, and facial animations, while the therapist guides the session and adapts the interaction based on the child's needs.

This paper presents the software architecture of EMOROBCARE, currently tested on a virtual model before deployment on the physical robot.

2 System Architecture

The 30 cm-tall EMOROBCARE robot has a spherical base, a 3-DOF head powered by affordable high-torque servos, a camera, chest-mounted microphone, side speakers, and a screen-based expressive face. A Jetson Nano handles perception, while a separate tablet displays game content, keeping the hardware simple and cost-effective.

The system architecture (Fig. 1) builds upon ROS 2-based components and a previously developed open-source architecture for situated social robots [5], which itself is based on the ROS4HRI framework [10]. This architecture provides multimodal social perception, symbolic reasoning, an LLM-based dialogue manager, intent-based controller and an interactive GUI. Additional ROS 2-based components have been incorporated: *Head controller*, which manages joint trajectories across its three joints (pan, tilt, roll); *Play motion2*, which executes predefined expressive gestures such as nodding or head shaking; *Attention manager*, to manage the robot's gaze behavior, enabling it to track objects or humans or perform random gaze shifts; and *Face interface*, which displays cartoon-like facial expressions (e.g., happy, sad, surprised).

We next outline the framework's main pipelines for autism therapy.

2.1 Vision

The goal of the vision module is to enable the robot to perceive and understand its surrounding environment in real-time during interaction with children in order to respond appropriately to children's actions and to support therapeutic games. The module builds on the ROS4HRI framework. While body detection, face detection and recognition, and gesture recognition are reused, the pipeline is extended with the following new modules:

– **Pointing gesture detection:** when a pointing gesture is detected, this module estimates the pointing direction by combining the orientation of the

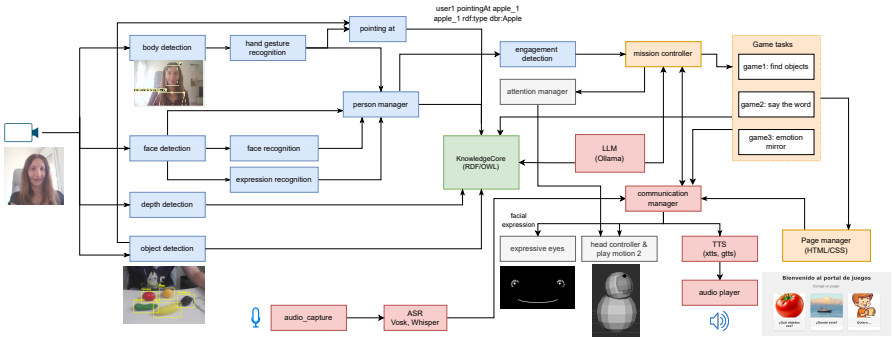

Fig. 1: EMOROBCARE software architecture: blue—perception; green—reasoning; red—speech; gray—execution; orange—supervision.

finger and the forearm. The resulting vector is intersected with detected objects or persons. This enables the robot to semantically interpret deictic gestures (e.g. john is pointing at book), allowing children to indicate objects during gameplay.

– **Object detection:** the module integrates two object-detection methods: YOLOv8 [4] and YOLO-world [4] to identify objects in the world. YOLO performs best in controlled environments where objects are known in advance. YOLO-world supports open-vocabulary detection, enabling the recognition of unseen object categories, suitable for uncontrolled environments. As an example, Table 1 presents a minimal performance comparison of YOLOv8-small and YOLO-world-small on a test dataset consisting of 53 images featuring five fruit toys. The YOLO model was trained on 378 images. For YOLO-world, we evaluate two prompts: one containing only the object name (prompt 1, e.g, *pear toy*), and another including both the object and its color (prompt 2, e.g, *yellow pear toy*). As expected, YOLO-world performs worse than YOLOv8 due to the lack of task-specific fine-tuning. Interestingly, adding color to the prompt improves detection accuracy.
– **Depth estimation:** the module integrates MiDaS [2] to infer relative depth information. The aim is to symbolically represent spatial information of the world, e.g. the child is far from the robot, or the cup is in front of the child.

2.2 Communication

In this project, we extend the original pipeline in the following way:

– **ASR:** the robot supports both Vosk[5] and Whisper[6] at this point. To compare their performance, the Word Error Rate (WER)[1] was used on two

[4] https://github.com/ultralytics/ultralytics
[5] https://github.com/alphacep/vosk-api
[6] https://github.com/openai/whisper

Model Comparison	Precision	Recall	F1-score
YOLO	0.910	0.879	0.894
YOLO-world prompt 1	0.615	0.537	0.573
YOLO-world prompt 2	0.662	0.604	0.632

Table 1: YOLO vs YOLO-world comparison. Precision shows the proportion of correct positive predictions, while recall shows the proportion of actual positives correctly predicted. F1-score provides the harmonic mean of precision and recall.

Model Comparison	Audio 1	Audio 2
Whisper (best) vs Whisper (small)	0.394	0.485
Whisper (best) vs Vosk	0.535	0.781

Table 2: Word Error Rate (WER) comparison for Whisper and Vosk on two child speech recordings. Lower values indicate better transcription quality.

different child audio recordings collected during therapy sessions, using Whisper's best-performing model as a reference. The results, presented in Table 2, highlight the differences in accuracy across smaller models from both Vosk and Whisper.

- **TTS:** Coqui XTTS[7]enables expressive speech with emotional tones (e.g., happy, surprised, afraid, among others). Emotional speech is achieved by performing inference on the XTTS cloning model using emotional speech samples from professional voice actors. In therapeutic games, therapists often use whispering to support children's responses. The system can replicate such whispers through voice samples based on Autonomous Sensory Meridian Response (ASMR) voices. Alternatively, we have also integrated gTTS[8] as a lightweight fallback in case of limited resources in the system (e.g, if no GPU is available).
- **LLM Integration:** this module is in charge of generating a dialogue response to user inputs parsed through the communication manager. The original pipeline, based on Ollama models[9] (e.g., LLama 3.2) has been adjusted to ensure responses are predictable, supportive, and aligned with therapeutic goals. Prompts are tuned so that the robot generates friendly, peer—encouraging, patient, and age-appropriate utterances –rather than an adult or instructor outputs. Additionally, utterances include markup tags to allow expressive speech (aligning facial expression, prosodics and motions).
- **Expressive face**: consists of a mouth and two eyes, possibly adding eyebrows and color, to display different expressions (see Figure 2).
- **Communication Manager**: it is in charge of handling the communicative acts between the robot and user processing the input from two sources: 1) from the user input (ASR), and sends it to the mission controller to

[7] https://github.com/coqui-ai/TTS
[8] https://github.com/pndurette/gTTS
[9] https://github.com/ollama/ollama

process dialogue through the LLM. 2) from the mission controller, either in response to the user request, or to initiate a conversation. In this case, the response includes markup tags to allow multimodal communication based on: prosodics, facial expression and motions. For example,

```
 <expression(surprise)> Such a surprise! </expression>
<do motion(shaking)> <expression(happy)> I can't believe you
are here.</expression>
```

The communication manager parsers the tags and triggers the respective modules accordingly to output the specified data. In the example above, the TTS would generate a surprised voice speaking the utterance "Such a surprise" while simultaneously displaying a surprised facial expression. Next, it triggers a shaking gesture through the head controller while speaking "I can't believe you are here" with a happy voice and facial expression. This emotional alignment across modalities enhances the robot's social presence and supports therapeutic goals by modeling expressive, context-sensitive behaviour.

2.3 Behavior coordination

The intent-based *mission controller* processes user intents by selecting which task (in this case, games) to start. It coordinates the flow of the game, triggering the different robot skills (e.g., show game page, ask questions, provide feedback).

The symbolic *Knowledge Core*, in this context, allows for storing symbolic information of the environment, their properties and interaction events. At the moment, it keeps track of detected objects, humans, and tuples indicating which object the user is pointing to. The different components of the system can then use this information within the interaction to reason and make decisions.

2.4 Interactive GUI and Robot Model

The GUI interface[10] (Fig. 2) provides the means to visualize robot internal state and trigger actions. We have expanded its original version with the following components:

- *Perception*: a visualization tool that displays vision-based detections. It extends the existing human detection to object detection and pointing gestures.
- *Radar*: a representation of the environment with spatial information of detected humans. It was also extended to include objects present in the environment and whether they are being pointed at or not.
- *Robot face*: displays the face of the robot along with transitions of facial expressions in real-time.
- *Chat*: exposes the verbal input/output of the robot. It includes a button to trigger the voice pipeline of the system whenever a user wants to verbally interact with the robot.

[10] The GUI is meant to be used for research and development of the system only.

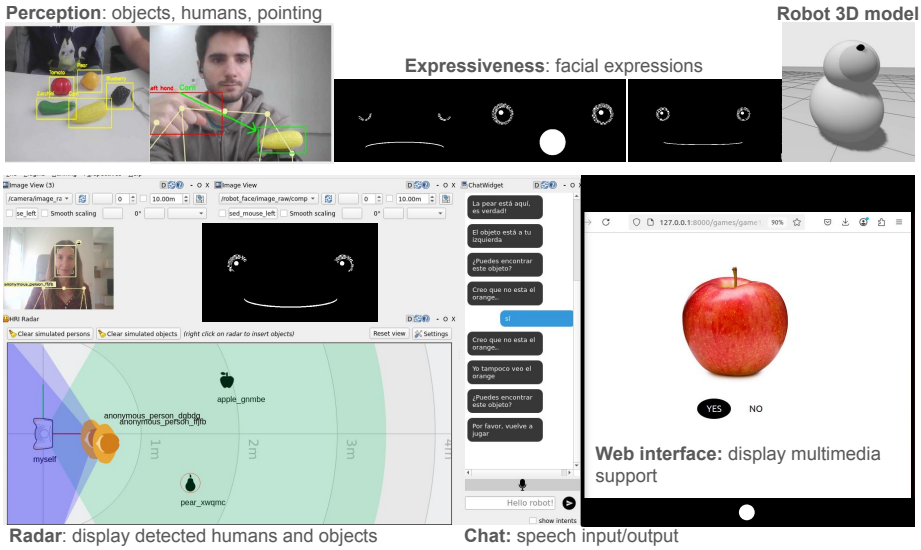

Fig. 2: Interactive GUI with perception and radar for object/human detection, facial expressions, chat for speech, web-based game content, and Gazebo robot model for testing head gestures.

The system also uses a *Page manager* to show HTML/JS multimedia content on a web browser, such as images or yes/no buttons. It interfaces with ROS 2 through rosbridge to provide user input data back to the mission controller. Finally, the robot model can be loaded into Gazebo with the interactive GUI for virtual testing. The physical robot will feature the same face as the GUI and include a tablet for displaying content.

3 Use case demonstrator

We developed a prototype game on the virtual system, called *Finding Objects* game, to promote joint attention and vocabulary acquisition, based on input from a therapist.

As illustrated in Figure 3, the game proceeds as follows. The interaction begins with the therapist initializing the session (1) by selecting a game on the touch-screen, which triggers the `find_objects` task via the mission controller. Next, the robot detects both the user (2) and visible objects (3) such as a pear or a tomato. The system updates its knowledge base (4) and displays the objects' locations on the radar tool (5). A first round of the game starts displaying an image of one of the detected fruits (e.g., a pear) on the tablet (6). The robot asks the child "Can you find this object?" using expressive TTS, motions and facial expressions (6). The child may respond verbally, pressing the YES/NO button, or pointing at the object (7). Either input is interpreted by the communication manager and passed to the mission controller.

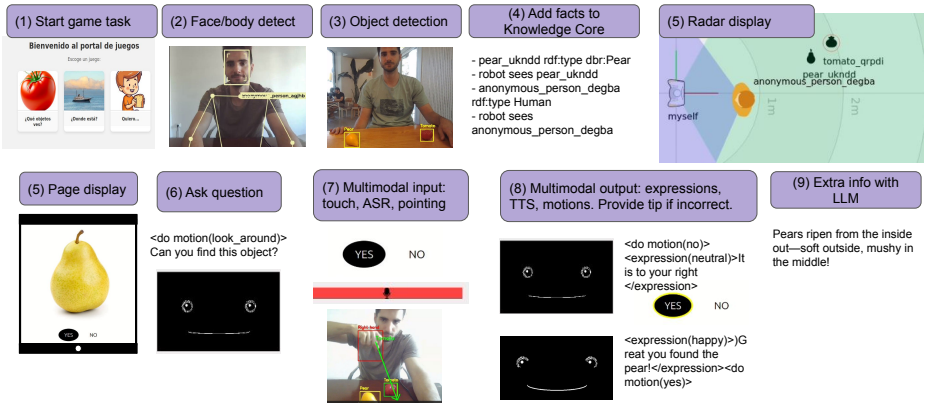

Fig. 3: Step-by-step flow of the "Finding Objects" game. The robot detects the user and objects, prompts the child with a fruit image, processes multimodal responses, provides expressive feedback and additional information using an LLM.

Based on the child's response, the robot offers multimodal feedback (8): if the answer is correct, it provides positive reaction "`<expression(happy)> Great! You found the pear!</expression> <do motion(yes)>`" with a happy expression and a confirming motion; if incorrect, it might respond with a neutral expression "`<expression(neutral)> It is on your right</expression>` ", shift its gaze towards the object, and highlight the correct object on the screen. Finally it shares a fun fact using its LLM (9) i.e. "Pears ripen from the inside out –soft outside, mushy in the middle!". It then shows the next fruit.

4 Conclusions and future work

This work presents the software architecture and development tools for a low-cost social robot designed to support autism therapy in children, along with a simulated game. With core functionalities and infrastructure in place, the next steps involve refining the system and conducting a real-world pilot evaluation.

For speech, we plan to implement task-specific voices (e.g., motivational tones during games) and integrate voice emotion detection through the ASR to detect frustration, stress, or joy, enabling more adaptive responses. Advanced prompt engineering will tailor answers to therapy needs, avoiding unsuitable outputs like emoticons, offensive language, or biases. We also envision multi-robot conversations to further support therapy.

We plan to expand vision with color detection (for color-naming games), visual speech detection (to identify speakers), and head gesture recognition (for "yes" or "no" signals) to improve interaction. All vision features will be integrated into ReMap [6], a voxel-based spatial-semantic framework, enabling the robot to build 3D environment models and reason about objects and pointing directions beyond its view.

Once these tools are developed, we will build on them to enable engagement assessment by detecting emotional states (shouting, laughing), attention (e.g., child looking at the robot).Additional games will be developed with therapists on this architecture. The robot is scheduled for construction by summer 2025, with a pilot planned for autumn. User feedback will guide prototype refinements.

Acknowledgments. This research work was funded by the Ministry for Digital Transformation and the Civil Service, financed by the Recovery, Transformation and Resilience Plan through the European Union's Next Generation funds. EMOROBCARE. IASOMM24002.

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

References

1. Ali, A., Renals, S.: Word error rate estimation for speech recognition: e-wer. In: Proc. Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 20–24 (2018)
2. Birkl, R., Wofk, D., Müller, M.: Midas v3. 1–a model zoo for robust monocular relative depth estimation. arXiv preprint arXiv:2307.14460 (2023)
3. Cabibihan, et al.: Why robots? a survey on the roles and benefits of social robots in the therapy of children with autism. International Journal of Social Robotics **5**(4), 593–618 (2013)
4. Cheng, et al.: Yolo-world: Real-time open-vocabulary object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16901–16911 (2024)
5. Cooper, et al.: Demonstration of an open-source ros 2 framework and simulator for situated interactive social robots. In: 2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI). pp. 1770–1772 (2025)
6. Ferrini, et al.: remap: Spatially-grounded and queryable semantics for interactive robots. In: International Conference on Social Robotics. pp. 383–396. Springer (2024)
7. González-González, C.S., et al.: A long-term engagement with a social robot for autism therapy: A pilot study using user-centered design and icf-based assessment. Frontiers in Robotics and AI **8**, 669972 (2021)
8. LuxAI: Qtrobot for special needs education. https://luxai.com/assistive-tech-robot-for-special-needs-education/ (2021), accessed: 2025-06-09
9. Meza-Kubo, V.E., et al.: A new socially assistive robot with integrated serious games for therapies with children with autism spectrum disorder and down syndrome: A pilot study. Sensors **21**(24), 8414 (2021)
10. Mohamed, Y., Lemaignan, S.: Ros for human-robot interaction. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3020–3027. IEEE (2021)
11. Scassellati, B., et al.: Improving social skills in children with asd using a long-term, in-home social robot. Science Robotics **3**(21), eaat7544 (2022)
12. Silva, V., et al.: Social stories for promoting social communication with children with autism spectrum disorder using a humanoid robot: Step-by-step study. Technology, Knowledge and Learning **29**, 735–756 (2024)