International Journal of Social Robotics (2026) 18:35
https://doi.org/10.1007/s12369-026-01367-1

Pepper Plays: *Guess Who?*

Juan Echevarria[1] · Modesto Castrillón-Santana[3] · Igor Rodriguez[1] · Javier Lorenzo-Navarro[3] · Elena Lazkano[1] · Unai Zabala[2]

Received: 7 March 2025 / Revised: 12 November 2025 / Accepted: 19 January 2026 / Published online: 23 February 2026

Abstract
Social games offer a rich and challenging environment to develop and evaluate complex robot behaviors. Social interactions with robots benefit from personalization, which requires person identification—a process that raises privacy concerns. Soft biometrics, however, offer a privacy-preserving alternative by enabling short-term identification without relying on private or sensitive features. In this work, we leverage a zero-shot Visual Question Answering (VQA) person identification system and adapt it to make Pepper play *Guess Who?* The identification module performance is firstly studied offline in the AveRobot dataset. Afterwards, it is integrated in a ROS based architecture that manages the game flow and enables natural verbal interaction. Experiments conducted with 29 subjects demonstrate that the system achieves identification performance nearly equivalent to that of hard biometric-based systems while also offering engaging and entertaining gameplay. Notably, user feedback reveals a high level of acceptance, highlighting that social gaming with robots is a promising avenue for developing and testing complex behaviors.

Keywords Social games · Robot-based gameplay · Visual question answering · Semantic soft biometrics · Person identification · Robot behavior development

1 Introduction

Social Robots (SR) should possess the capability for social interactions and are positioned as a potential solution to address their growing demand across various domains. From entertainment to assistance, providing companionship to the elderly, aiding children in learning, or facilitating meal delivery, their potential use is broadening [1]. As mentioned in [2], social games are an excellent scenario to explore complex social interactions. Gaming using robots has been already used for promoting exercise [3], or for rehabilitation for people with Parkinson's disease [4, 5], as a tool for memory stimulation [6], to treat the attention deficit hyperactivity disorder [7], and as edutainment platform [8], among others.

Viewed from a human-centric perspective, perception is also essential for robots to personalize their interactions with individual humans. Personalization not only enhances user trust but also significantly impacts the quality of the interaction [9]. Robust person identification is essential for SR personalization, either long or short-term. Facial information has traditionally been used to establish the identity of a person of interest such as the face identification

All authors contributed equally to this work.

✉ Elena Lazkano
e.lazkano@ehu.eus

Juan Echevarria
jechevarria@ikasle.ehu.eus

Modesto Castrillón-Santana
mcastrillon@iusiani.ulpgc.es

Igor Rodriguez
igor.rodriguez@ehu.eus

Javier Lorenzo-Navarro
javier.lorenzo@ulpgc.es

Unai Zabala
unai.zabalac@ehu.eus

[1] Computer Science and Artificial Intelligence, EHU, Manuel Lardizabal 1, Donostia, Gipuzkoa 20018, Spain

[2] Computer Architecture and Technology, EHU, Manuel Lardizabal 1, Donostia, Gipuzkoa 20018, Spain

[3] SIANI, Universidad de Las Palmas de Gran Canaria, Edificio Central del Parque Científico y Tecnológico, Las Palmas 35017, Spain

proposal in Rodriguez et al. [10]. However, the use of hard biometric traits, such as the face, even when resulting in higher accurate systems, raises privacy concerns because they retain sensitive personal data. An alternative is provided by soft biometrics traits, defined as human attributes (e.g. gender, age, ethnicity, etc.) which alone do not definitively distinguish an individual, but combined has been proposed as a non-intrusive and viable solution for extracting visual discrete features of a person such as gender, hair color, clothing, and accessories among others [11, 12]. This approximation is reliable when short-term identification is required.

Social games offer the perfect scenario for improving HRI. We claim that by deploying a social game in SRs the applicability and usability of complex robot behaviors can be analysed and evaluated. In this paper, we argue that a soft biometrics-based approach is well suited for short-term person identification, while preserving privacy. We propose leveraging Visual Question Answering (VQA) to get soft biometrics descriptors from an individual, and adopt them for person identification in a SR to carry out an interactive social game. Our motivation is to equip robots with explainable visual descriptions of the individuals they interact with for short-term person identification. In that vein, we propose a gaming setup to asses person recognition using a zero-shot VQA system. By developing an attribute-based person description and recognition module, Pepper is able to play the well known *Guess Who?* game with a group of people. Our contributions are threefold:

- Semantic Soft Biometrics Description: We employ VQA to generate a reliable soft biometrics description of the person interacting with the robot.
- Robust Identification Strategy: We define and assess a strategy for identifying individuals based on a collection of non-unique soft biometrics features.
- Robot Integration: The VQA-based identification strategy is deployed in Pepper to allow it to play *Guess Who?*

As far as we know, this is the first work that reports the use of VQA for SR perception. Moreover, we contribute our bit to the area of robotic games, an active research topic that aims to develop robots able to play as a mean to improve the life of children with disabilities and people at large [13, 14].

2 Attribute-Based Person Description and Recognition

The field of person recognition has gained significant interest due to its practical applications in security monitoring, authentication, and other scenarios. This technology relies on different human traits, including fingerprints, facial and voice features, and gait patterns.

When it comes to user recognition within HRI applications, facial recognition stands as the most widely adopted approach, with deep learning models currently predominating for this purpose [15]. Along these lines, Wang et al. [16] propose a method for multi-face re-identification using facial embeddings and unsupervised clustering, achieving an overall accuracy of 93.55% on the TERESA dataset and 90.41% on the YTF dataset. However, despite its promising results, this approach remains in the research phase and, as far as we know, has not been integrated into the TERESA robot. Conversely, the humanoid robot Nadine demonstrates a real-world instance of facial recognition integration [17]. Nadine employs an external camera and leverages OpenFace [18], to identify users and personalize interactions. Another real-world successful implementation of face recognition in service robotic applications is demonstrated by the BRILLO robot [19]. BRILLO utilizes a panoramic camera and a combination of a YOLO detector and OpenFace. This powerful combination ensures accurate face recognition even under challenging conditions. This capability allows BRILLO to not only identify users but also track them, gather information about their preferences, and personalize interactions accordingly. In [10], a face re-identification/verification module is evaluated offline, and integrated into the multirobot guide system GidaBot and, to avoid new visitors interfering with those attended.

While facial recognition is powerful, it has limitations. Facial changes, such as beard growth and makeup, coupled with common challenges like occlusions and varying lighting conditions, can significantly degrade performance. The latter affects particularly in short-term identification scenarios. Valid alternatives apply multimodal biometrics [20]. In the context of HRI scenarios, Freire et at. [21] combine facial and voice recognition modalities and incorporate a unique input data sampling method, they aim to enhance verification accuracy. This strategy increases intra-class variability, potentially improving the robustness of the biometric system, reporting on the AveRobot dataset an Equal Error Rate (EER) of 12.22%. In addition, the use of hard biometric traits such as the face requires managing sensitive personal information. In an interaction context where it is not necessary to uniquely identify the person, it would be ideal to use non-unique features.

One of the many applications of large language models (LLMs), more specifically visual language models (VLMs), is the development of VQA models, which are trained to answer questions about visual content. They combine computer vision and natural language processing to interpret and understand the visual information on an image and respond to questions about it, becoming a meeting point for

vision-language tasks [22, 23]. As described by Radford et al. [24], fine tuning an encoder with an LLM, aligns visual and linguistic representation spaces, allowing VQA models to achieve competitive results with zero-shot transfer, i.e. without requiring domain-specific training.

In the next sections, we firstly evaluate offline the reliability of the VQA-based identification approach in a HRI benchmark, later describe the game flow, to finally present the participants' experience evaluation.

3 Zero-Shot VQA Based Person Identification

As stated before, a preliminary experiment was conducted considering a short-term verification scenario (i.e., with no significant changes in clothes and accessories), using a privacy-preserving strategy that avoids hard biometrics and instead relies on appearance descriptors provided by a VQA model, to test offline the strength of combining those descriptors in person identification. First, we will provide a brief description of the dataset.

3.1 AveRobot Dataset

The AveRobot dataset[1] [25] was originally designed as a benchmark for person recognition in HRI within a multi-robot scenario. This dataset contains 2664 video clips captured from the robot's perspective of 111 different identities recorded across nine locations spanning three floors of a building, with three locations per floor. Each clip is associated with a unique person, floor, location, and camera. Given that different robots manage each floor, and they can integrate multiple sensors, the dataset encompasses a variety of capturing devices and lighting conditions.

3.2 VQA-Based Identification Pipeline

A zero-shot VQA approach outperformed all competitors in the PAR Contest - CAIP23 benchmark [26]. The benchmark focused on a surveillance scenario where pedestrians are captured at a distance, evaluating the competitors ability to estimate five different person features: gender, presence of bag or hat, and the main color of the upper and lower body clothes. The winner team achievement was attained without any VQA model specific adaptation to the scenario [27]. The results evidenced the great performance of the VQA-based proposal across diverse pedestrian datasets and tasks. After these results, in this proposal, we leverage VQA to extract each person soft biometric features, with the aim at identify individuals within a group during short-term interactions.

Given that a HRI session is not a single image but an AveRobot clip, the VQA model will return a list of answers (one per frame processed) for each particular question. VQA models are characterized for providing short answers, see the sample answers reported in Fig. 1. Given this VQA characteristic, a voting or consensus approach is used to select the final answer for each of the considered questions.

To validate the approach, AveRobot samples from any floor are compared with those samples captured in any of the other floors by any camera. Therefore, the evaluation considers as genuine any HRI clip VQA description pair of the same identity captured by any sensor in other floor, and as impostor any HRI clip VQA description pair of different identities captured by any sensor in other floor.

The first step when processing each AveRobot clip is to detect the person. For human detection purposes, we adopted a pretrained YOLOv8[2] model, specifically YOLOv8n, with the default enclosed tracker module (i.e., ByteTrack), to detect and track individuals during the HRI session. We configured the detection to focus solely on the person class.

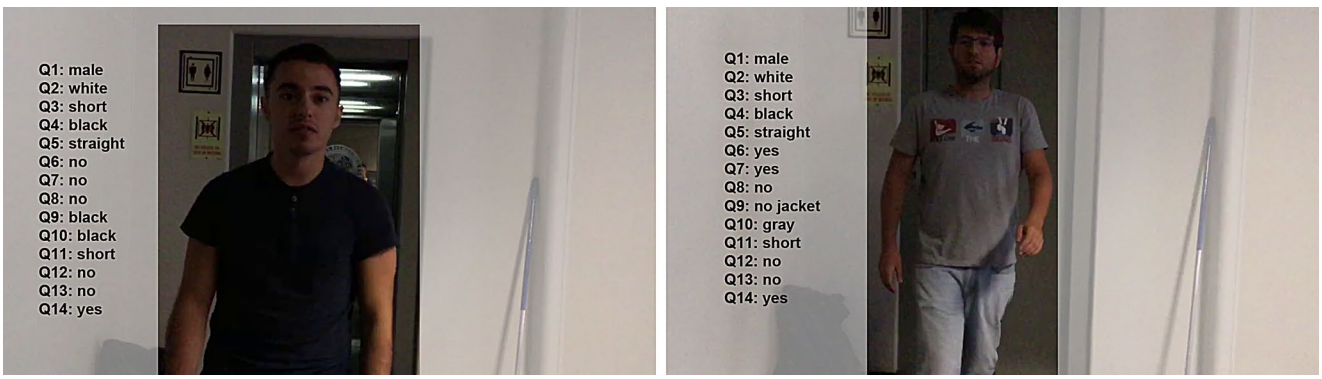


Fig. 1 Answers provided by the VQA model for two AveRobot clips. For those particular individuals, the matching score obtained is 10 for the set of 14 questions

[1] https://mozart.dis.ulpgc.es/averobot/.

[2] https://github.com/ultralytics/ultralytics.

Table 1 Full set of appearance-related questions used in visual analysis

Questions (1–7)	Questions (8–14)
Q1. Is the person male or female?	**Q8**. Does the person wear a jacket?
Q2. Which color is the person's skin?	**Q9**. What color is the person's jacket?
Q3. Has the person short or long hair?	**Q10**. What color is the person's shirt?
Q4. Which color is the person's hair?	**Q11**. What type of collar does the shirt have?
Q5. Is the person's hair curly or straight?	**Q12**. Has the person's shirt stripes?
Q6. Has the person a beard?	**Q13**. Has the person's shirt squares?
Q7. Does the person wear glasses?	**Q14**. Is the person's shirt plain?

If multiple individuals are detected during the interaction, only the tracked person with the largest bounding box, occupying at least one-third of the image height, is considered. For those individuals, we posed the set of VQA questions listed in Table 1, without enclosing any hard biometric data.

The chosen VQA model makes use of the BLIP-2 pre-training strategy [28], integrating off-the-shelf frozen pre-trained image encoders and a frozen Open Pre-trained Transformers (OPT) Language Model [29]. The model was fine-tuned for VQA with the Visual Transformer (ViT) base backbone [30]. This model was adopted following its outstanding performance in the pedestrian attribute recognition challenge. A more recent comparison with other models, as described in [27], showed that certain models, such as FLANT5XL, may achieve slightly higher accuracy in binary questions within the pedestrian attribute recognition scenario. However, this advantage comes with the drawback of increased computational cost and the need for additional processing, as its responses tend to be more verbose compared to the OPT model, which is known for providing concise answers.

The matching between two samples, i.e. the individual interacting in two AveRobot clips, is measured using the Jaccard similarity index, which is formally defined as the size of the intersection divided by the size of the union. In our case, if the answer obtained for a particular question is identical for both samples, the similarity increases by 1. We have introduced a slight modification for partial matching in those questions related to colors, e.g. "blue and white" and "blue and gray" increasing the similarity by 0.5. In Fig. 1, the reader may observe the final output for two AveRobot clips, showing a sample frame of each one.

A quantitative offline assessment was conducted to compare the results of the soft biometric approach with those reported by Freire et al. in [21] with a multi hard biometrics (audio and face) approach. To illustrate the feasibility of the approach, we adopted as metric the EER, that corresponds to the similarity threshold configuration where both the false acceptance and false rejection rates are equal. For computational efficiency, VQA questions were posed for every tenth frame of each AveRobot clip. The final clip feature for each considered question was determined through the aforementioned voting mechanism applied to the resulting VQA descriptions.

Table 2 EER comparison (lower is better)

Approach	EER
Freire et al. visual cue [21]	13.38%
Freire et al. multi cue [21]	12.22%
Ours (Q1-2)	28.16%
Ours (Q1, Q3-5)	32.28%
Ours (Q1–Q7)	21.06%
Ours (Q1-14)	15.01%

Table 2 summarizes the results, including for comparison those reported in [21]. The EER reported using the clip description based in the answer to the 14 questions is 15.01%, being almost three points higher, i.e. worse, than the best one reported in [21]. Observe that this results are achieved without any domain specific training nor hard biometrics information.

To investigate the impact of the different questions, we conducted some experiments by reducing the number of VQA questions used to describe individuals. When removing questions related to shirt and jacket (Q1-7), the EER increased slightly up to 21.06%, evidencing the importance of clothes to match correctly an identity in this context. After leaving just four questions related to gender and hair style (Q1, Q3-5), the EER increases significantly up to 32.28%. Leaving just the questions to Q1 and Q2 (gender and skin color), the EER reached 28.16% given the dataset distribution, with better behavior than including the hair-related questions. Fig. 2 reports the respective receiver operating characteristic (ROC) curves and EERs obtained for the VQA-based pipeline comprising a different number of VQA answers. The advantage of adding answers to the individual description clearly improves the Area Under the Curve (AUC), while reducing the EER. The reader may observe the significant drop in performance when less than 10 questions are adopted. Certainly, the selection of questions more adapted to the specific population included in AveRobot could help to decrease further the EER, but that is not the aim of this paper.

Certainly, the EER is slightly worse than the reported in the literature adopting a hard biometrics approach. However, no training was needed for a population of more than 100 individuals. This is achieved without the need to register and store personal data from the individuals during the HRI session. The identification approach seems suitable for

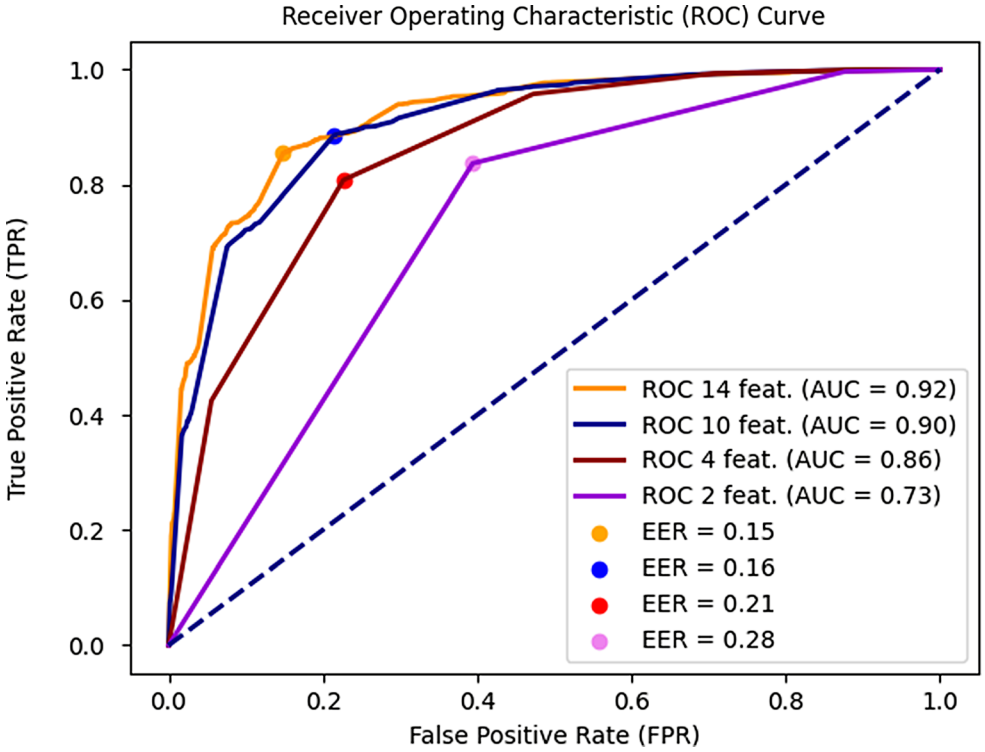


Fig. 2 ROC curves and EER reported for three different number of features

identifying individuals during the interaction of SR with a group of individuals.

4 Guess Who?

Once the validity of the VQA model for person identification has been tested in an offline setting using the AveRobot dataset, it is integrated into an interactive and real-time application, such as the *Guess Who* game.

In the classical *Guess Who?* game, two players compete to identify the other's secretly chosen character from a set of options. Each player selects a character from their character board, and then, players alternate asking each other yes or no questions about the appearance of the opponent's chosen character to narrow down the possibilities. The objective of the game is to guess who has the other player chosen, so, at any given time, a player can decide to take a guess instead of asking a question. At the end, whoever guesses first wins the game.

To adapt the game for human-robot interaction, where players compete against Pepper, we have modified the original rules as follows:

- Instead of a predefined set of characters included in a character board, we dynamically select the characters based on the people present around the robot.
- The game is going to be played in two different stages, in each one, a player will choose a person and the other will try to guess. Once finished, the roles will swap.
- Both the maximum number of questions and the number of permitted guesses are dynamically adjusted based on the number of participants, n, and determined by Eq 1.

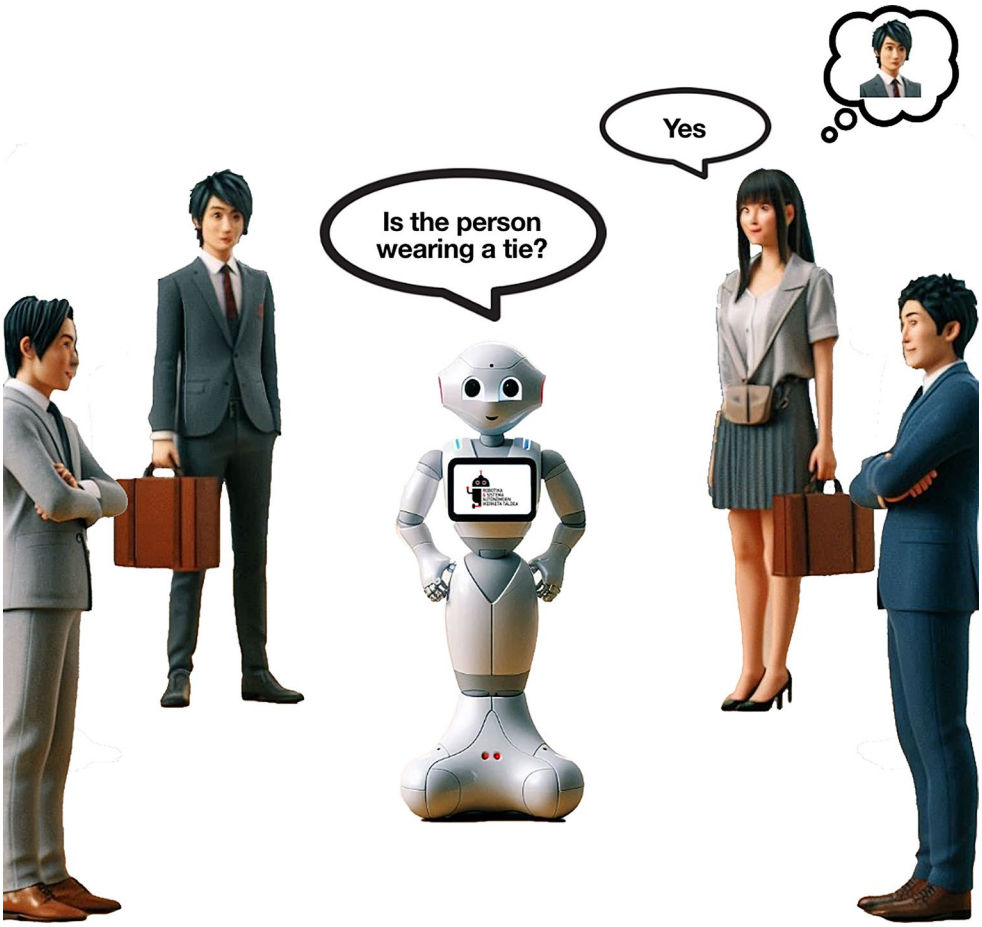
$$max_{questions} = \max\left(1, \lfloor 2\log_2(n)) \rfloor\right) \tag{1a}$$

$$max_{guesses} = \max(1, \min(3, \left\lfloor \frac{n}{2} \right\rfloor)) \tag{1b}$$

Fig. 3 shows the interaction setup, where Pepper would be acts as the guesser, the human player answers the questions and the rest of the people complete the game board.

A ROS-based architecture has been specifically designed to support this game interaction. It follows a modular structure, with individual components handling perception, dialogue, decision-making, and behavior coordination, as described in the following subsections. Due to Pepper's limited onboard computational capabilities, all intensive processing (including people detection, visual feature extraction and VQA reasoning) is executed on an external computer equipped with an NVIDIA RTX 4070 Super 12GB GPU.

Fig. 3 *Gues Who?* scenario



4.1 Character Categorization

Accurate character categorization is essential for gameplay, where the extraction of the visual features of each participant is fundamental. To achieve that, both Pepper's frontal cameras captures are combined into a single image, so that the whole body of the people playing is visible. After that, the aforementioned VQA based person identification module is applied to get the features of each participant.

When the robot looks at the different people surrounding it, a random name is assigned to each person based on their detected gender. This is made for an easier identification of each person in the later interaction with the player.

To improve the game experience, the questions set was expanded beyond those introduced in section 3. The additional questions are related to the person's lower body (not always visible in AveRobot), such as the garment they are wearing and their shoes, as well as a question about other accessories such as hats, scarfs, watches or similar items.

Given that the VQA model always provides an answer, i.e., it hallucinates, even when the answer cannot be inferred from the provided image, "*control questions*" are introduced to mitigate this issue. For example, when asking "Which color are the person's shoes?", the model may still output a color even if no shoes are visible. To prevent this, a control question such as "Do you see the person's shoes?" is asked first, requiring a positive answer before proceeding. For more details on the full set of control questions, see Appendix A (Table 5).

4.2 Game Initiative and Flow

As mentioned before, two primary game flows are considered: 1) robot-initiated guessing and 2) user-initiated guessing. The flow is decided in the first interaction with the robot, and varies depending on who initiates the interaction. If the user explicitly expresses a desire to play, the robot assumes the guessing role. Conversely, when the robot initiates the game, triggered by a set time elapsing before any interaction or a user greeting, the user becomes the guesser. At the end of each round, the roles of the game will swap.

Throughout the game, there are a few phases: firstly, the robot will spin around itself, storing the assigned fictitious name, the features obtained and the odometry pose for each. Then, it will enter the questioning phase, where one player will ask questions and the other will answer them. To finish,

the guessing player will try and figure out who the other player was thinking about. If they manage to guess correctly in the predefined amount of tries, they will win the round. If the robot is the one who guessed correctly, it will go back to the chosen person using the stored odometry pose to show they know who that person is. All the different stages of the game and the activation or deactivation of the different tools and behaviours of the robot are managed by a main ROS node, as well as the assignation of the roles and making the decisions needed in the game.

4.3 Person Discrimination

Whenever it is the robot's turn to guess, it has to be able to choose questions that will give him relevant information. To achieve that, we use the cross entropy measure with all the people's features and select the question that will give us the most information. Once the user answers, the robot has to select all the people that match the answer given by the user, and stop taking into account all the others for the search. When facing unknown information, the robot plays it safe and does not discard any person based on information it may not know.

4.4 Game Dialogue

In the game, interaction is required between the robot and the user, usually in the form of verbal communication. To achieve a smooth interaction, the user is always indicated what they are meant to do by the robot. The robot generates questions and answers according to the interaction with the user and the information gathered throughout the game.

Vosk [3] tool handles automatic speech recognition, ensuring accurate understanding of the user's spoken requests, while Nuance's [4] text-to-speech engine generates the audio.

4.5 Body Expressiveness

The robot employs automatic gesticulation during speech, generating synchronized beats as described in [31]. Additionally, the robot reacts with two different emotions: happiness to exhibit success and frustration/sadness, expressed when it fails to guess the identity of the goal person.

4.6 Information Display

For a more open interaction with the player, and to display the thinking process of the robot, we have utilized the tablet in Pepper's chest. We display information like the snapshots and names assigned to the people detected. In case one person is selected by the user, the information the robot knows about that user will be displayed too. Once the robot decides a person is not the one the user has chosen, it will draw a red X on the person snapshot, displaying to the user the decisions it is making. An example of the display is shown in Fig 4.

Fig. 4 People list displayed on Pepper's tablet

[3] https://alphacephei.com/vosk/.

[4] https://docs.nuance.com/mix/apis/tts-grpc/.

Table 3 Questionnaire contents. Questions are grouped into 4 categories

VQA implementation performance
QQ1: Are the questions proposed by the robot relevant to achieve the identification?
QQ2: Are the questions proposed by the robot varied enough and include enough features?
QQ3: Does Pepper succeed in identifying the goal person?
QQ4: Does the robot employ too much time in getting the feature information (effectiveness)?
Understanding and Expressing
QQ5: How well did Pepper understand your questions and respond in a clear way?
QQ6: At times, did you find Pepper's behavior to be frustrating or difficult to deal with?
QQ7: How informative was Pepper's body language in understanding its reactions?
QQ8: Did you find the information displayed on the robot's tablet useful?
Gameplay experience
QQ9: How well did Pepper keep you engaged throughout the game?
QQ10: Did you find the game entertaining and amusing?
QQ11: Did you find the game easy to understand and to play?
Global satisfaction
QQ12: Evaluate your global experience playing with Pepper

Table 4 Population details

Age	Male	Female	Experience with robots	Game knowledge
10–15	40%	60%	2.4 ($\pm$0.55)	4 ($\pm$0.71)
20–40	53.33%	46.66%	3 ($\pm$1.41)	4.44 ($\pm$0.73)
50+	33.33%	66.66%	2.33 ($\pm$1.32)	4.66 ($\pm$0.5)
total	44.8%	55.18%	2.76 ($\pm$1.21)	4.48 ($\pm$0.63)

5 Experimental Evaluation

Before starting, participants received simple instructions about the game flow, especially those unfamiliar with the original *Guess Who?*. The explanation covered the game dynamics and the role the robot would take during each round. They were also informed that the robot would try to identify people using visible traits like clothing or hair color, based on what it could see through its cameras. It was made clear that the robot might occasionally make mistakes, and that if it did not understand a response, it would ask the user to repeat it to keep on the conversation.

The game scenario was completed with up to 7 people to make the game more entertaining as well as to increase the variety of guessing options. Each game round, including interaction and response time, typically lasted between 3 to 5 minutes. This duration excludes the initial setup and reset between rounds.

After each session, the participants completed a form to asses the quality of the developmental state of the robot's behavior. The questionnaire contained 12 questions divided in four categories and an additional text box to add optional comments (see Table 3).

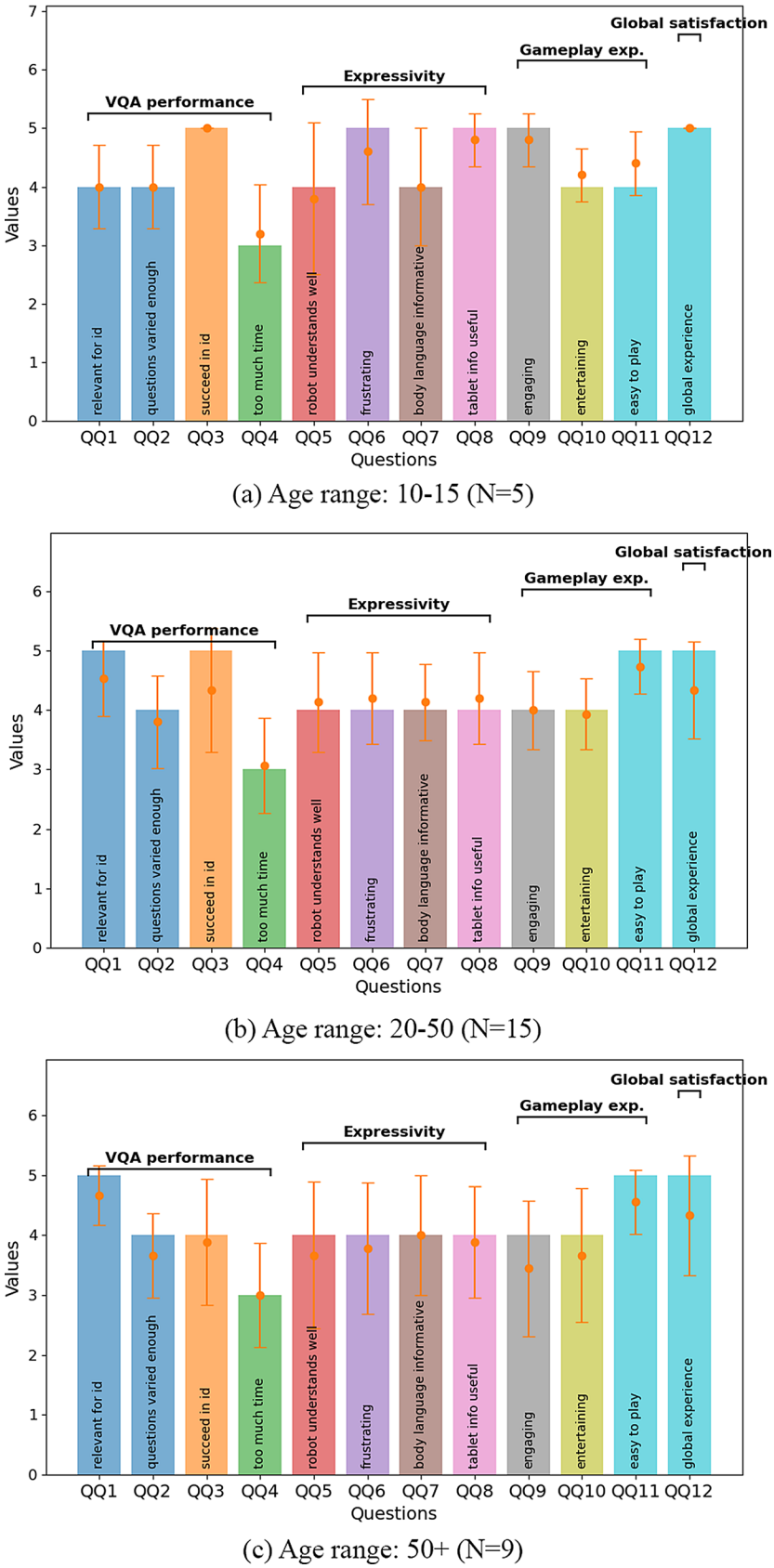
In total, 29 participants were engaged, details are shown in Table 4. 58 rounds were played (Pepper acted as the guesser half of the times). Three age ranges were covered, not balanced sets though. We were able to gather five responses from children of age between 10 and 14, nine responses from people above 50 (up to 64) and the remaining corresponded to people of age in range 20–40. Globally, 46% were female and, two children and two persons above 50 did not have any previous experience with the game itself. Its worth mentioning that not many of the participants had previous experience with robots as the average value of 2.76 over Likert 5-scale reflects. Collected data also showed that the population more confident with robots was concentrated in the 20–40 age range.

The robot made a successful guess 83% of the times and only two times needed more than one chance. Thus, the incorrect guesses were mainly due to lack of features. Besides, the robot made 3.03 questions per game on average. A slightly higher number compared to the 2.96 questions per game needed by the human users. Finally, we detected that the YOLO tracker loses the human rather frequently. Consequently, the same person was captured twice (17%) with the side effect that a player was given two different identities.

Regarding the questionnaire, Fig. 5 shows the median values of each 5-scale question, together with their means and standard deviations for the three different age clusters.

At a glance, there are no significant differences among the three age ranges. Children show to be more passionate in their responses, probably due to the excitement and the novelty of playing with the robot. Although the three populations are not balanced, the similarity of the results allow us to generalize and to state that age did not make difference in the answers. This is corroborated by the non-parametrical Kruskal-Wallis test with the exception of question QQ9 that showed a p_value (0.018329) shorter than 0.05. Surprisingly, the population over the age of 50 reported higher engagement with the game, which might be attributed to their comparatively lower familiarity with robots and technology overall – though this remains speculative.

To evaluate the consistency of participant responses across conceptual categories of questions – Implementation performance, Understanding and experience, and Gameplay experience–, we computed the Intraclass Correlation Coefficient (ICC) using ICC(3,k) – which assumes fixed raters and evaluates mean ratings. The results indicate good reliability according to conventional interpretation guidelines, ICC(3,k)=0.848, 95% CI[0.72, 0.92]. The associated F-test ($F = 6.58, p = 1.266267e-09$) confirmed that the reliability is statistically greater than zero. Overall, these results



(a) Age range: 10-15 (N=5)

(b) Age range: 20-50 (N=15)

(c) Age range: 50+ (N=9)

Fig. 5 Questionnaire results. Bars show the mean and standard deviation values for each question (QQ1 to QQ12). Labels inside the bars are short descriptions of the questions and the bracket lines group questions in the 4 categories described in Table 4

suggests that the average ratings provided by the fixed raters show a consistent and reliable level of agreement.

Besides, the three clusters agree that the robot employs too much time (QQ4) in extracting user's features and this process should be faster.

Concerning the suggestions, a couple of them suggested that the robot should be allowed to fail or to miss the first attempt, that this make the robot more natural although the aim of the player of course is to win.

Two videos have been recorded during the experiments to visualize the standard flow of the game. In the first video [5], Pepper takes on the role of the guesser, examining the humans around and asking questions to narrow down the possibilities. In the second video[6], the roles swap, and the user becomes the guesser.

It must be mentioned that the game has been exhibited at the European Robotics Week (ERW-2024) in the Magic Space of the *Ibercaja Espacio Joven* in Zaragoza, event organized by *Hisparob*[7], where hundreds of kids and adults participated playing with the robot in groups of 8 to 10 during a whole afternoon (see Fig. 6).

6 Discussion

The current system demonstrates that a soft biometrics-based approach using zero-shot VQA is viable for enabling social robots to interact naturally in game-like scenarios. However, some limitations and ethical considerations emerged during the evaluation that deserve further reflection.

During the user sessions, an issue observed in a few instances was the misclassification of gender by the employed VQA model, particularly in the case of female participants with short hair. Although not critical to the game's mechanics, this misclassification affected how fictitious names and references were assigned during interaction, as gender is used for both, verbal and visual referencing. This raises ethical considerations regarding gender identity in public interactions and the potential discomfort caused by misidentification, which becomes relevant not only in terms of classifier accuracy but also in designing inclusive robotic behavior. However, in the cases observed, participants generally reacted with humour rather than discomfort, and no negative or awkward reactions were noted during or after the game.

Participants also noticed occasional delays in feature extraction, which impacted the fluidity of the interaction. Although the VQA module was configured with inference time limits, the overall delay came from the sequential process of the pipeline, where vision analysis, response generation, and robot motion occur one after the other, each

Fig. 6 *Guess who?* session at ERW-2024

[5] https://youtu.be/TFDDkXVd50I.

[6] https://youtu.be/7aZRBd8jzFQ.

[7] https://www.hisparob.es.

depending on the completion of the previous step. Because each step depended on the previous one, the overall process introduced delays during gameplay, which affected how participants perceived the robot's intelligence and responsiveness.

In addition to these timing issues, as mentioned in section 5, the tracking module occasionally duplicated the same individual (an error that occurred in around 17% of the cases). When this happened, the duplicated person was treated as two separate players, which could influence gameplay outcomes or user experience. While not critical to the interaction, it revealed limitations in the robustness of the current tracking and identification setup under real-world conditions.

Another issue observed during the experiments relates to the VQA models's tendency to produce answers without visual evidence, which could confuse users. To reduce this effect, we added control questions that verify the presence of relevant features. While effective, this also increased the overall inference time, and thus, the fluidity of the interaction.

7 Conclusions and Further Work

This work presents a robot system capable of playing the game *Guess Who?* using soft biometrics for short-term person identification. By leveraging a VQA-based pipeline integrated with a ROS-based architecture, the robot can identify individuals based on visual descriptors such as clothing, accessories, and hairstyle, without relying on sensitive hard biometric data. The system supports both verbal interaction and visual display, enabling a natural and entertaining game experience.

User evaluations confirmed that the robot's behavior was generally perceived as understandable, engaging, and effective, with high success rates in guessing the selected individuals. The system was especially well-received in public exhibitions, where it interacted with large and diverse groups of users.

There are several areas we plan to improve in future work. Reducing latency in feature extraction and interaction is one of our key priority. Although visual processing and response generation are offloaded to a GPU equipped external machine, there is still room to improve coordination between modules. Running some steps in parallel could further reduce delays.

From a technical point of view, latency reduction remains a central goal for upcoming iterations. Current delays mainly occurs from the sequential execution of perception, reasoning, and actuation modules. To mitigate this, we plan to parallelize the pipeline so that visual detection, VQA inference, and dialogue generation can run concurrently, improving temporal alignment between perception and speech. Further optimizations include adaptive frame-skipping, caching of visual features to avoid redundant queries, and the use of lighter trackers or multimodal re-identification methods for real-time operation. ROS2 intra-process communication and asynchronous callbacks will also be explored to reduce overhead, with the aim of keeping response times below the two-second threshold that users perceived as hesitation, thus improving the natural flow of interaction.

Another point to improve is the occasional duplication of the same person by the tracking system. When this happens, the user is treated as two separate players. Comparing visual features could help detect duplicates, though it may not work well when people look alike. Reducing latency could also help mitigate this issue, as tracking errors are more likely when the system falls behind real-time updates.

The choice of model also plays a key role in the system's performance. While BLIP-2 OPT was selected for its relatively low computational cost and speed, it introduces certain biases, particularly in attributes like gender. Exploring alternative models that offer better accuracy and fairness (such as PaliGemma, LLaVa or MiniGPT-4) is part of our future plans.

Beyond technical improvements, design choices can also help mitigate the effects of such biases. Using gender-neutral references, such as assigning names based on non-gendered animated characters (e.g., Avery, Riley, Maxie), would promote more inclusive behavior. Additionally, allowing users to self-identify or reframing questions to avoid categorical gender distinctions could enhance the experience in sensitive contexts.

Finally, although the current system was developed for a game-based interaction, the same approach can be extended to a broader range of social robotics scenarios. Beyond entertainment, a privacy-preserving VQA-based framework could support adaptive storytelling or cooperative learning activities in educational contexts, where the robot recognises participants' roles without storing personal identifiers. In eldercare or rehabilitation, similar mechanisms could assist daily routines by detecting familiar clothing or accessories, or by monitoring engagement in physical or cognitive exercises. Following recent trends in robot-mediated exergames and cognitive support platforms [3–8], the proposed architecture provides a general foundation for multimodal, privacy-conscious interaction. Thus, while the Guess Who? scenario served as an engaging proof of concept, the underlying modules hold potential for a wide spectrum of socially assistive and educational applications.

Table 5 Full set of visual questions asked by the VQA model, including control questions designed to mitigate hallucinations

name	question	prec. by	prec.	ask times	ask till
gender	Is the person male or female?	—	—	9	—
adultornot	Is the person an adult?	face	yes	3	—
hair_color	What color is the person's hair?	—	—	3	—
hair_length	Does the person have short or long hair?	—	—	3	—
eyes	Do you see the person's eyes?	—	—	3	yes
face	Do you see the person's face?	—	—	3	yes
eyes_color	Which color are the person's eyes?	eyes	yes	3	—
glasses	Does the person wear glasses?	face	yes	3	allways
sunglasses	Is the person wearing sunglasses?	glasses	yes	3	allways
beard	Does the person have a beard?	gender	male	3	—
hair_face	Where on the face does the person have hair?	gender	male	3	—
arms	Do you see the person's arms?	—	—	3	yes
sleeves	Are the sleeves short or long?	arms	yes	3	—
pulleover_etc	What is the person wearing in their upper body?	sleeves	long	3	—
collar_type	What type of collar does the shirt have?	—	—	3	—
shirt_color	What color is the person's shirt?	—	—	3	—
shirt_stripes	Does the person's shirt have stripes?	—	—	3	—
stripes_orient	Are the person's shirt stripes vertical?	shirt_stripes	yes	3	—
jacket_color	What color is the person's jacket?	pulleover_etc	jacket	3	—
lower	Do you see the person's lower body?	—	—	3	yes
pants_color	What color are the person pants?	lower	yes	3	—
pants	What is the person wearing in their lower body?	lower	yes	3	—
shoes	Do you see the person's shoes?	—	—	3	yes
shoe_type	What type of shoes is the person wearing?	shoes	yes	3	—
shoe_color	Which color are the persons shoes?	shoes	yes	3	—
accessories	Is the person wearing any noticeable accessories?	—	—	3	—
accessory	What accessory is the person wearing?	accessories	yes	3	allways

Appendix

Acknowledgements This work is partially funded by the Spanish Ministry of Science and Innovation under projects PID2021-122402OB-C22, TED2021-131019B-10, and by the ACIISI-Gobierno de Canarias and European FEDER funds under projects ProID2021010012, ULPGC Facilities Net, and Grant EIS 2021 04. Author Zabala has received a PREDOKBERRI research Grant from Basque Government.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data Availability All the data used are properly referenced in the text. For further questions on data and analysis, please refer to the corresponding author.

Declarations

Compliance with Ethical Standards Not applicable.

Consent to Participate Informed consent was obtained from all individual participants included in the study.

Consent to Publish The authors affirm that all human persons provided informed consent for publications of the images.

Conflict of Interest The authors have no relevant financial of non financial Conflict of Interest to declare.

References

1. Yang J, Vindolet C, Olvera JRG, Cheng G (2024) On the impact of robot personalization on human-robot interaction: a review

2. Rato D, Correia F, Pereira A, Prada R (2023) Robots in games. Int J Soc Robot 15:37–57. https://doi.org/10.1007/s12369-022-00944-4
3. Fitter N, Mohan M, Preston R, Johnson M, Kuchenbecker K (2024) How should robots exercise with people? robot-mediated exergames win with music, social analogues, and gameplay clarity. Front Robot AI. https://doi.org/10.3389/frobt.2023.1155837
4. Allen J, McKay J, Sawers A, Hackney M, Lh T (2017) Increased neuromuscular consistency in gait and balance after partnered, dance-based rehabilitation in parkinson's disease. Neurophysiology 118(1):363–373. https://doi.org/10.1152/jn.00813.2016
5. Bevilacqua R, Benadduci M, Bonfigli A, Riccardi G, Melone G, Forgia AL, Macchiarulo N, Rossetti L, Marzorati M, Rizzo G, Di Bitonto P, Potenza A, Fiorini L, Loizzo FC, Viola CL, Cavallo F, Leone A, Rescio G, Caroppo A, Manni A, Cesta A, Cortellessa G, Fracasso F, Orlandini A, Umbrico A, Rossi L, Maranesi E (2021) Dancing with parkinson's disease: the SI-ROBOTICS study protocol. Front Public Health. https://doi.org/10.3389/fpubh.2021.780098
6. Louie WYG, McColl D, Nejat G (2012) Playing a memory game with a socially assistive robot: a case study at a long-term care facility. In 2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication, pp 345–350). https://doi.org/10.1109/ROMAN.2012.6343777
7. Cervantes J, López S, Cervantes S, Hernández A, Duarte H (2023) Social robots and brain-computer interface video games for dealing with attention deficit hyperactivity disorder: a systematic review. Brain Sci 13(8):1172. https://doi.org/10.3390/brainsci13081172
8. Martín FA, Gonzalez-Pacheco V, Castro-González Á, Ramey A, Yébenes M, Salichs M (2010) Using a social robot as a gaming platform, in International Conference on Social Robotics (ICSR) (30–39. https://doi.org/10.1007/978-3-642-17248-9_4
9. Lacroix D, Wullenkord R, Eyssel F (2023), HRI 'Who's in charge? Using personalization vs. Customization distinction to inform HRI research on adaptation to users. Companion The 2023 ACM/IEEE Int Conf Hum Rob Interact (Assoc Comput Machinery), New York, NY, USA, 23:580–586. 10.1145/3568294.3580152. https://doi.org/10.1145/3568294.3580152
10. Rodriguez I, Zabala U, Marín-Reyes PA, Jauregi E, Lorenzo-Navarro J, Lazkano E, Castrillón-Santana M (2020) Personal guides: heterogeneous robots sharing personal tours in multi-floor environments. Sensors 20(9):10.3390/s20092480. https://www.mdpi.com/1424-8220/20/9/2480
11. Kumar N, Berg AC, Belhumeur PN, Nayar SK (2011) Describable visual attributes for face verification and image search. IEEE Trans Pattern Anal Mach Intel 1962–1977
12. Gonzalez-Sosa E, Fierrez J, Vera-Rodriguez R, Alonso-Fernandez F (2018) Facial soft biometrics for recognition in the wild: recent works, annotation, and cots evaluation. IEEE Trans Inf Forensics Secur 13(8):2001–2014. https://doi.org/10.1109/TIFS.2018.2807791
13. Bonarini A, Besio S (2022) Robot play for all: developing toys and games for disability. Springer. https://doi.org/10.1007/978-3-031-05042-8
14. Rato D, Correia F, Pereira A, Prada R (2022) Robots in games. Int J Soc Robot 15:37–57. https://api.semanticscholar.org/CorpusID:253857239
15. Wang M, Deng W (2021) Deep face recognition: a survey. Neurocomputing 429:215–244
16. Wang Y, Shen J, Petridis S, Pantic M (2019) A real-time and unsupervised face re-identification system for human-robot interaction. Pattern Recognit Lett 128:559–568. https://doi.org/10.1016/j.patrec.2018.04.009. https://www.sciencedirect.com/science/article/pii/S0167865518301296
17. Tulsulkar G, Mishra N, Thalmann NM, Lim HE, Lee MP, Cheng SK (2021) Can a humanoid social robot stimulate the interactivity of cognitively impaired elderly? a thorough study based on computer vision methods. Visual Comput 37(12):3019–3038
18. Amos B, Ludwiczuk B, Satyanarayanan M et al. (2016) Openface: a general-purpose face recognition library with mobile applications. CMU Sch Comput Sci 6(2):20
19. John NE, Rossi A, Rossi S (2022) Personalized human-robot interaction with a robot bartender. In Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, pp 155–159
20. Oloyede MO, Hancke GP (2016) Unimodal and multimodal biometric sensing systems: a review. IEEE Access 4:7532–7555. https://doi.org/10.1109/ACCESS.2016.2614720
21. Freire-Obregón D, Rosales-Santana K, Marín-Reyes PA, Penate-Sanchez A, Lorenzo-Navarro J, Castrillón-Santana M (2021) Improving user verification in human-robot interaction from audio or image inputs through sample quality assessment. Pattern Recognit Lett 149:179–184
22. Agrawal A, Lu J, Antol S, Mitchell M, Zitnick CL, Parikh D, Batra D (2015) Vqa: visual question answering. Int J Comput Vision 123:4–31
23. Barra S, Bisogni C, De Marsico M, Ricciardi S (2021) Visual question answering: which investigated applications? Pattern Recognit Lett 151:325–331. https://doi.org/10.1016/j.patrec.2021.09.008. https://www.sciencedirect.com/science/article/pii/S0167865521003147
24. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I (2021) Learning transferable visual models from natural language supervision by Meila M, Zhang T (eds) Proceedings of the 38th International Conference on Machine Learning (ICML), vol 139. pp 8748–8763). http://proceedings.mlr.press/v139/radford21a.html
25. Marras M, Marín-Reyes P, Lorenzo-Navarro. J, Castrillón-Santana. M, Fenu G (2019) AveRobot: an audio-visual dataset for people Re-identification and verification in human-robot interaction. In Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods - Volume 1: ICPRAM, INSTICC (SciTePress), pp. 255–265). https://doi.org/10.5220/0007690902550265
26. Greco A, Vento B (2023) PAR contest 2023: pedestrian attributes recognition with multi-task learning. In: Tsapatsoulis N, Lanitis A, Pattichis M, Pattichis C, Kyrkou C, Kyriacou E, Theodosiou Z, Panayides A (Springer Nature Switzerland, Cham, eds) Computer analysis of images and patterns. pp 3–12
27. Castrillón-Santana M, Sánchez-Nielsen E, Freire-Obregón D, Santana OJ, Hernández-Sosa D, Lorenzo-Navarro J (2024). 10.1007/s42979-024-02985-0. Visual question answering models for zero-shot pedestrian attribute recognition: a comparative study. SN Comput Sci 5(6):680. https://doi.org/10.1007/s42979-024-02985-0
28. Li J, Li D, Savarese S, Hoi S (2023) BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. https://doi.org/10.48550/arXiv.2301.12597
29. Zhang S, Roller S, Goyal N, Artetxe M, Chen M, Chen S, Dewan C, Diab M, Li X, Lin XV, Mihaylov T, Ott M, Shleifer S, Shuster K, Simig D, Koura PS, Sridhar A, Wang T, Zettlemoyer L (2022) OPT: open pre-trained transformer language models
30. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2021) An image is worth 16x16. In Words: Transformers for Image Recognition at Scale, in 9th International Conference on Learning Representations, ICLR. https://openreview.net/forum?id=YicbFdNTTy

31. Zabala U, Rodriguez I, Martínez-Otzeta J, Irigoien I, Lazkano E (2021) Quantitative analysis of robot gesticulation behavior. Auton Robots 45(1):175–189. https://doi.org/10.1007/s10514-020-09958-1

Juan Echevarria is an M.Sc. student in Cognitive Neuroscience of Language at the University of the Basque Country and the Basque Center on Cognition, Brain and Language (BCBL). He received the B.Sc. degree in Artificial Intelligence from the University of the Basque Country in 2025. His research interests include the application of artificial intelligence to cognitive neuroscience and fMRI signal denoising.

Modesto Castrillón-Santana is a Full Professor with the Department of Computer Science and Systems, ULPGC. He received the M.Sc. and Ph.D. degrees in computer science from the Las Palmas de Gran Canaria University (ULPGC), in 1992 and 2003, respectively. His main research activities focus particularly on the automatic facial analysis, covering also different topics related to image processing, perceptual interaction, human-machine interaction, biometrics, and computer graphics. He is a member of the AEPIA and AERFAI, having coauthored around two hundred articles, including peer-reviewed international journals, book chapters, and conference proceedings. He has acted as an external expert for the Chilean, Italian, Qatar Research Agencies and the Spanish Accreditation Agency. He has served to the community as general co-chair in CAIP25, ICPRAM25, ICPRAM24, ICPRAM23 and SITIS18, also taking part of several conference programme and technical committees. He is currently an Associate Editor of Pattern Recognition Letters and past board member of Image and Vision Computing, and the IEEE Biometrics Newsletter.

Igor Rodriguez is an Assistant Professor in the Computer Sciences and Artificial Intelligence department (EHU). He received the B.Sc., M.Sc., and Ph.D. degrees in Computer Science from the University of the Basque Country in 2012, 2014, and 2019, respectively. He is currently an Associate Professor at the University of the Basque Country (EHU) and a member of its Robotics and Autonomous Systems Group (RSAIT). His research focuses on the development of advanced deep learning models applied to perception and decision-making in social robots, with a special interest in enhancing natural communication between humans and machines through gestures, body language, and emotional expression in robots.

Javier Lorenzo-Navarro is an Associate Professor of Computer Science and a researcher at the Intelligent Systems and Numerical Applications in Engineering Institute (SIANI). He received the M.S. degree in Computer Science in 1992 and the Ph.D. degree in Computer Science, cum laude, in 2001, both from the University of Las Palmas de Gran Canaria. His research interests include computer vision, human–machine interaction, machine learning for computer vision, soft biometrics, and person re-identification, with current work focused on emotion recognition. He has led several funded research projects and has coauthored more than 100 papers in international journals and conference proceedings. He serves as a reviewer for several scientific journals and as a program committee member of international conferences.

Elena Lazkano is an Associate Professor in the Computer Sciences and Artificial Intelligence department (EHU). She received her B.Sc. in Computer Sciences in 1992 (EHU), M.Sc. in Artificial Intelligence (Katholieke Universiteit Leuven, Belgium, 1993) and Ph.D. in Computer Sciences in 2004 (EHU). She is codirector of the Robotics and Autonomous Systems Group (RSAIT) in Donostia-San Sebastian and her research interest focuses in developing social aptitudes and proactivity in social robotics.

Unai Zabala is an Assistant Professor in the Computer Architecture and Technology department (EHU). He received the B.Sc., M.Sc., and Ph.D. degrees in Computer Science from the University of the Basque Country in 2019, 2020, and 2025, respectively. He is currently an Assistant Professor at the University of the Basque Country (EHU) and a member of its Robotics and Autonomous Systems Group (RSAIT). His research interests include human–robot interaction, gesture generation, and the expressive behavior of social robots.