

Social Referencing Disambiguation with Robot Mental Imagery for Domestic Service Robots

Fan, Kevin; Jouaiti, Melanie; Noormohammadi-Asl, Ali; Dautenhahn, Kerstin; Nehaniv, Chrystopher L.

DOI:

[10.1007/s12369-025-01223-8](https://doi.org/10.1007/s12369-025-01223-8)

License:

Other (please specify with Rights Statement)

Document Version

Peer reviewed version

Citation for published version (Harvard):

Fan, K, Jouaiti, M, Noormohammadi-Asl, A, Dautenhahn, K & Nehaniv, CL 2025, 'Social Referencing Disambiguation with Robot Mental Imagery for Domestic Service Robots: System Implementation and Validation in an Object Selection Task', *International Journal of Social Robotics*. <https://doi.org/10.1007/s12369-025-01223-8>

[Link to publication on Research at Birmingham portal](#)

Publisher Rights Statement:

This version of the article has been accepted for publication, after peer review (when applicable) and is subject to Springer Nature's AM terms of use, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: <http://dx.doi.org/10.1007/s12369-025-01223-8>

General rights

Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

- Users may freely distribute the URL that is used to identify this publication.
- Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
- User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
- Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy

While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.

Social Referencing Disambiguation with Robot Mental Imagery for Domestic Service Robots - System Implementation and Validation in an Object Selection Task

Kevin Fan^{1*}, Melanie Jouaiti³, Ali Noormohammadi-Asl²,
Kerstin Dautenhahn¹, Chrystopher L. Nehaniv^{2,1}

^{1*}Dept. of Electrical & Computer Engineering, Organization, Waterloo, N2L 3G1, Ontario, Canada.

²Dept. of Systems Design Engineering, Organization, Waterloo, N2L 3G1, Ontario, Canada.

³Dept. of Mechanical Engineering, Imperial College London, London, SW7 2BX, Greater London, UK.

*Corresponding author(s). E-mail(s): kevin.fan1@uwaterloo.ca;
Contributing authors: m.jouaiti@imperial.ac.uk; ali.asl@uwaterloo.ca;
kerstin.dautenhahn@uwaterloo.ca; chrystopher.nehaniv@uwaterloo.ca;

Abstract

Domestic service robots need to have the cognitive ability to operate and succeed in complex, dynamic, and object-rich home environments. Inspired by human cognitive mechanisms combined with deep-learning techniques and soft computation, we developed a learning social referencing framework with mental imagery for domestic service robots and implemented the full architecture on a mobile manipulator robot, Fetch. This work demonstrates a comprehensive framework enabling service robots to address ambiguities in object selection tasks, and to continually learn under the guidance of a human interaction partner. We carried out a full system validation study with human participants to investigate user experience and attitudes towards the system, as well as the system's functional success. We experimentally evaluated the proposed cognitive framework in **four object selection** scenarios and found the framework effective at enabling the service robot to be adaptive and capable of handling various ambiguities in

interactions. Furthermore, participants perceived the robot positively in multiple dimensions, such as perceived intelligence, knowledge, sensibility and interactivity after interacting with the robot.

Keywords: Human-robot interaction, Intelligent robots, Robot learning

1 Introduction

Social referencing is an effective mechanism often employed by humans and some other social animals to resolve uncertainties and ambiguities when a more experienced or knowledgeable social partner is present [1–6]. There are two essential components of social referencing: 1- *information seeking*, where the subject looks to the social partner in uncertain situations to catch their attention, either verbally or non-verbally, signaling the need for additional information/clarification, 2- *behavior regulation*, whereby a subject that receives such feedback can alter behaviors accordingly [7]. Social referencing aids humans in gaining valuable information/knowledge according to the feedback of their social interaction partners. Importantly, humans can adaptively alter their own behaviors based upon the newly acquired knowledge. Prelinguistic infants, who regularly encounter uncertainty in their daily lives, due to their lack of life experience, employ non-verbal communication (e.g., gaze) and seek the attention of a caregiver to elicit disambiguation; social referencing, as a result, is often utilized by human infants to resolve ambiguities and to learn about their world [8], e.g. learning how to respond to a novel stimulus. Our computational social referencing simulates this human cognitive mechanism to empower service robots to behave naturally and effectively in human social environments.

Learning socially can work in tandem with the capacity for mental imagery. Think about a nice crisp red apple or the full moon on a starry night. You can almost “see” those in your mind based on these simple descriptive clues. The majority of humans have the ability to form mental imagery that allows them to visualize items, recall actions, relive experiences, and much more [9]. Humans are visual creatures that rely heavily on the visual system in discovering and understanding the environment; this is evidenced by the dominant proportion of cortical tissue allocated to visual information processing in comparison to all other senses [10]. We therefore incorporated robot mental visual imagery into our computational social referencing framework to better guide the robot in making logical, contextual decisions in ambiguous situations, as well as scaffolding its learning from human partners.

Let us consider the scenario depicted in Fig. 1. Imagine a human user is having a meal and instructs the robot to “pass the salt” from across the table. This instruction is clear to a non-expert user; however, the robot may be in an ambiguous state due to many possible reasons, such as multiple target detections, or the robot simply does not know which container contains the salt due to the limitation of its knowledge base. This type of ambiguity can frequently appear in a complex environment, such as a household environment, and a service robot can utilize our proposed social referencing framework to disambiguate the target reference naturally with human feedback and update its

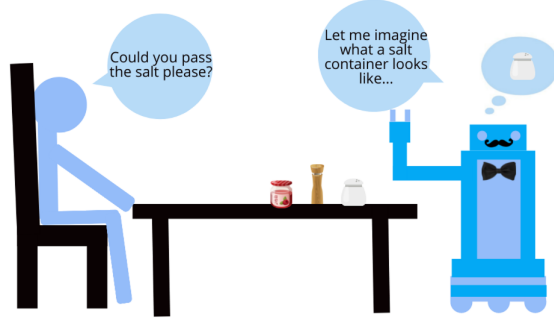


Fig. 1: The robot does not know which container contains “salt” due to its limited knowledge base; the robot invokes imagination and learning and performs disambiguation with human feedback if the target is still unclear.

knowledge base continuously, through interaction and imagination ([generation of an image through text](#)), to resolve ambiguous states in the future.

In this paper, we will describe a learning social referencing framework that allows the robot to resolve ambiguities through human interaction, utilize robot mental imagery to guide the robot’s learning and questioning process, and learn and update the robot’s knowledge base in a continuous fashion. Additionally, the framework incorporates the capability of mental imagery and an additional loop of social referencing, contributing further to the development of a more adaptive, “life-long learner” robot; moreover, we present a system validation experiment involving human participants interacting with the proposed framework implemented on a physical robot [in an object selection setting](#).

To clarify key concepts addressed in this paper, we define the following:

- **Social Referencing Disambiguation:** This refers to the robot’s ability to use social cues from human interaction—such as gaze direction, gestures, and verbal cues—to resolve ambiguities in understanding human intentions or instructions. In our framework, the robot leverages these social cues to interpret which object the human is referring to when encountering novelty or multiple possibilities exist.
- **Object Selection Task:** This is the practical application domain where the robot must select and interact with objects in a real-world environment. The object selection task presents scenarios where ambiguity can occur, and effective disambiguation strategies are essential for successful human-robot interaction.

Our research integrates social referencing disambiguation within the context of the object selection task. Specifically, we address how a domestic service robot can utilize social cues to resolve ambiguities when selecting objects and learn novel object labels through human-robot interactions.

Some specific technical parts of this work have been published previously [11, 12], but we present the final implementation and the enhanced comprehensive framework here (with robot imagination object learning). Moreover, the key contribution of this article is the addition of an extensive system validation user study.

This research is guided by the following research questions:

RQ1: How can we create a social referencing framework for service robots to handle object selection tasks?

RQ2: How feasible is it to develop a framework for service robots that allows the robot to learn objects through interaction and experience?

RQ3: How does the proposed framework affect users’ perceptions of the robot?

The contributions of this work are fourfold:

1. **Design and Implementation of a Social Referencing Object Disambiguation Framework:**

We developed a holistic framework that enables robots to disambiguate objects and learn new object labels through human-robot interaction. This framework leverages social referencing cues—such as gaze direction, gestures, and verbal communication—to interpret human intent in ambiguous object selection tasks. It integrates various deep learning modules for speech and gesture recognition, a stable diffusion model for robot mental imagery generation, and a parallel Siamese neural network for meta-learning, enabling the robot to disambiguate objects and continuously learn through interaction and imagination.

2. **Design of a Comprehensive System Validation Study:**

We designed and conducted a system validation study to evaluate the proposed framework within an object selection context. The study involved scenarios with different types of ambiguities, testing the robot’s ability to handle various disambiguation challenges and learn from interactions with users.

3. **Empirical Evaluation of the Framework’s Effectiveness:**

We performed an empirical evaluation of the framework’s effectiveness through both objective performance measurements and subjective human perception questionnaires. Our experimental results demonstrate that the robot can successfully disambiguate objects and learn new object labels, showing adaptability and competence in handling ambiguous situations.

4. **Investigation of the Framework’s Impact on User Perception:**

We examined the effects of the framework on participants’ perceptions of the robot in terms of perceived intelligence, efficiency, knowledgeability, sensibility, and interactivity. Our findings indicate that the framework positively influences user perceptions, enhancing the overall human-robot interaction experience.

The remainder of the article is structured as follows. Section 2 reviews related work in areas relevant to this article, while section 3 details the implementation of the proposed computational framework. Section 4 describes the system validation experiment. Results are presented in section 5 and discussed in section 6. Limitations and future work are outlined in section 7, followed by a conclusion (section 8).

2 Related Work

Our proposed framework is human cognition-inspired to handle the issue of object selection for domestic service robots. Related previous works have explored two aspects: 1- cognition-inspired robotic frameworks and 2- ambiguity resolution in object selection. Since our work is a combination of both aspects, we will discuss insights from prior work in these two areas.

2.1 Cognitively Inspired Robot Framework

2.1.1 Robot social referencing

The simulation of the social referencing mechanism has been attempted previously in robotics. Systems proposed in [13, 14] utilized emotion recognition and the robot’s internal states to explore the association of emotion to objects in robots. Specifically, they allowed the robot to observe the facial expression of the human interaction partner and an object presented by the interaction partner. Based on the facial expression of the interaction partner, the robot could assign internal emotion toward the presented object and invoke the appropriate motor actions when the object is shown again (i.e., the robot will try to avoid an object associated with negative emotions).

Similar to our research direction, the above mentioned studies investigate computational models of social referencing in robotic applications. However, they focus on the emotional association aspect of the mechanism, while our framework concentrates on the utilitarian aspect of social referencing (object disambiguation and learning) to improve the practical utility of service robots.

2.1.2 Robot mental visual imagery/imagination

Visual mental imagery/imagination is beneficial in terms of increasing the general adaptability of a robot. In early work [15], a mental imagery system with spatial language understanding was proposed to allow a robot manipulator to maintain object permanence and understand perspective differences with a conversational agent. The mental model was constructed using simple 3D rigid body objects in a simulated world, and a physical dynamic predictor made predictions of the object’s location in the next image frame. More recent work [16] realized the robot’s mental representation of the workspace using a 3D game engine, enabling the robot to have a “general understanding” of the objects that the robot needs to interact with and allowing comparisons between imagined expectations and the actual sensory input. Furthermore, [17] proposed a mental imagery model to construct the 3D representation of the scene through multiple 2D images captured by the robot’s sensors.

Our research differs from the works mentioned above as we focus on the visual appearance of *novel* objects that were not known to the robot previously. We utilize visual imagery generation as an assistive mechanism to aid the robot in learning and locating *unfamiliar* objects rapidly and logically.

2.2 Object Selection Ambiguity Resolution

2.2.1 Ambiguity detection

The initial step in object selection tasks is to detect the existence and nature of ambiguity in the task. Existing literature predominantly relies on simplistic Boolean logic for ambiguity determination. In the work presented by Hatori et al. [18], the system assessed potential targets by establishing a hard threshold for the confidence level of the detected object. The system deemed the situation ambiguous when there were multiple potential targets in the workspace. Similarly, Sibirtseva et al. [19] introduced a system that computes ambiguity by identifying the number of objects the system

can detect, as per the human partner’s request. If multiple objects were detected in the final analysis, the system inferred ambiguity. Likewise, Haasch et al. [20] also utilized the number of objects that fulfill the user’s request as a basis for determining ambiguity.

These works oversimplify the ambiguities involved in the object selection task, and they all commonly address one single type of ambiguity: the detection of multiple potential candidates. In comparison, our system addresses various types of ambiguities that may occur in the common object selection task. In addition, our system can provide the robot with probabilistic reasoning of task situations that resembles human thought logic, which is more robust and natural than the harsh pure binary logic typically utilized.

2.2.2 Attention detection

The evaluation of human attention is critical for interactive robotic applications. Understanding the state of the human interaction partner can provide invaluable insights for robots to engage effectively. The ability to assess human attention can find relevance in various robotic tasks, including but not limited to object selection. Lemaignan et al. [21] presented a system that leveraged keypoint-based head pose estimation to approximate gaze direction, utilizing head pose to estimate attention focus. The system calculated the “with-me-ness” value, a normalized ratio metric reflecting human attention based on attention time devoted to the target task and the expected interaction time. Minguillon et al. [22] proposed the detection of human attention in scenarios involving multiple speakers, utilizing EEG data and fuzzy logic inference. Li et al. [23], on the other hand, employed fuzzy logic and gaze direction measurements to evaluate a surgeon’s attention during surgeries, directing a robotic laparoscope system to adaptively change viewpoints based on the user’s attention. Asteriadis et al. [24] employed convolutional neural networks (CNN) to classify head orientation. They approximated gaze direction by transforming eye images and computing the gaze vector through comparison with a front-facing eye image. The combination of head pose and gaze data was then integrated with a fuzzy inference system to determine attention levels.

Compared to the works mentioned above, our attention detection method utilizes multiple input sources, such as the head orientation and the gaze direction of the user, to estimate human attention levels. We also employ fuzzy inference systems to enable more robust assessment than traditional logics and to be more time-efficient when compared to CNN-based approaches.

2.2.3 Resolution strategies

Due to the essential nature and the common occurrence of object selection tasks for service robots, various ambiguity resolution methods have been proposed by exploiting a wide range of human signals. Gesture is one of the most well-utilized and effective methods of disambiguation in object selection [20, 25]. While the system in [20] relied heavily on human deictic (pointing) gestures to help the robot disambiguate similar items, the system in [25] combined human deictic gestures and speech queries with a partially observable Markov decision process (POMDP) to select the target object.

More recently, additional systems have oriented more toward language processing to resolve ambiguities [18, 19, 26–29]. Methods in [18, 19] analyzed unstructured language queries coupled with the robot’s visual input, and posed informative speech queries to the human partner for disambiguation. The system in [26] proposed a model with attention breaches to predict the object target-source pairs based on human linguistic input and the robot’s sensory information. The method presented in [27] suggested an interactive dialogue system with a POMDP model to track the agent’s observation history and provide more relevant, informative robot language queries for clarification. Similarly, [28, 29] provided an interactive dialogue system with enhanced language understanding and more appropriate robot clarification queries.

Our framework incorporates social referencing and mental imagery in disambiguation, and it utilizes and analyzes both verbal and nonverbal signals. Furthermore, the framework enables the robot to grow its knowledge through interactions and human guidance, allowing the robot to adapt to the user’s specific environment.

3 Framework

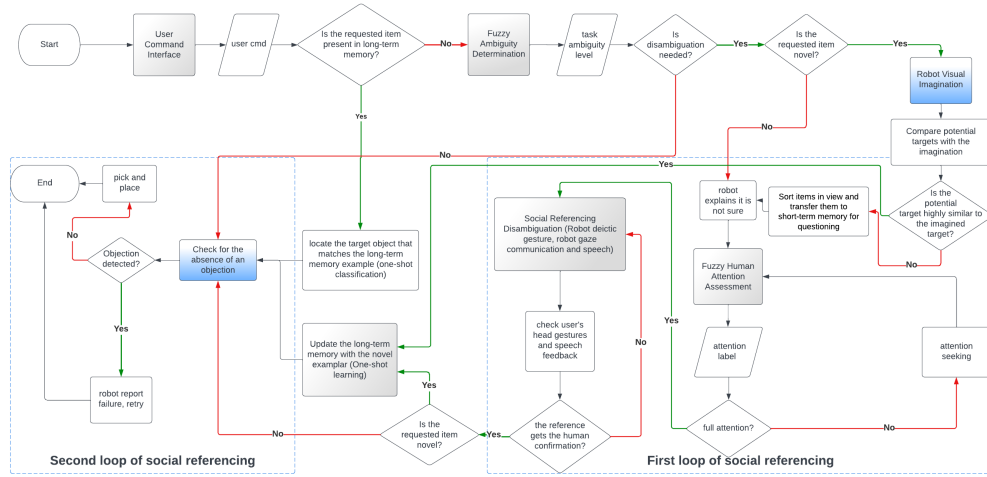


Fig. 2: High-level flow diagram of the social referencing disambiguation framework, including robot mental imagery and loops of social referencing interaction. In the context of this framework, a “loop” of social referencing allows the robot to change its behavior based on the feedback of its social interaction partner (human). The “first loop” allows the robot to locate the target object with human feedback. The “second loop” allows the robot to passively observe the human to avoid mistakes.

In this section, we will first give a general overview of the entire computational framework; specifically, the framework’s high-level behaviors are summarized, and the architectural components of the framework are introduced. Subsequently, we will

explain these architectural functional components in detail. We described the basic overall framework and some of the fundamental functional components in our previous publications in [11, 12]. Here, we present additional details, as well as an enhanced framework that is more robust and adapted to learning novel objects.

3.1 System Overview

With a robot running the proposed framework, the user command interface enables the user to issue commands either in natural language or text-based input. The command interface analyzes and extracts relevant information from the user instruction and passes the extracted label to the fuzzy task ambiguity level analysis, where this component instructs the robot to analyze the visual scene of the environment and perform fuzzy computation combined with the robot’s long-term memory to compute a continuous ambiguity level of the current selection task. The robot will either invoke motor actions to fetch the target item if the task ambiguity level is low, or plan to engage in social referencing disambiguation if the task ambiguity is high. However, before engaging in the social referencing disambiguation process, the fuzzy human attention assessment will be executed to determine whether the human user is paying attention to the robot. This is important as the social referencing mechanism is most effective if the social interaction partner is attentive enough to provide adequate feedback. If an insufficient attention level is detected (see details in section 3.4), the robot will perform attention-seeking behaviors such as gestures (e.g., waving, gaze) or speech. If the user instructs the robot with a novel class label that is not present in the base object detection model, the robot will “imagine” the appearance of the object based on the description from the user, and compare the imagined target with all the potential objects in the robot’s current view. It can then either perform contextual questioning, if necessary, or transfer the item that is highly similar (over 80% similarity) to the imagined target directly to the robot’s long-term memory for future retrieval and current task execution. The disambiguation process enables the robot to utilize deictic gestures, gaze communication, and simple verbal queries to ask questions to the human interaction partner and allows the robot to read human head gestures and speech feedback to clarify the target reference. Lastly, if the requested object is novel to the robot, the robot will transfer the requested item into its long-term memory for knowledge update.

To summarize, the robot’s social referencing framework is highly modular to facilitate better functional clarity and maintainability and can be divided into the following five fundamental functional blocks:

- User command interface,
- Fuzzy task ambiguity level analysis,
- Fuzzy human attention level assessment,
- Social referencing disambiguation,
- Short-term long-term memory object learning.

The above five fundamental functional blocks achieve the base framework that enables the robot to disambiguate and learn objects with human guidance. This system was outlined in [11]. We enhanced the previously described framework by adding

additional functionalities with two new modules that enhance the robot’s learning and interaction robustness, respectively:

- Robot mental imagery/imagination,
- Absence of objection detection.

To support these modules, our framework integrates various established deep learning-based techniques and incorporates custom models, enhancing the robot’s capabilities in perception, understanding, and adaptation. Below, we provide an overview of the deep learning techniques utilized in our system:

- *Object Detection and Recognition:*
 - YOLO (You Only Look Once): We employ the YOLO algorithm [30] for real-time object detection and recognition. YOLO is a convolutional neural network (CNN) that uses deep learning to detect objects within images quickly and accurately. This allows the robot to identify and locate objects in its environment based on visual input.
 - Caffe models: We also employ Caffe-based models for object detection. Caffe is a deep learning framework developed by the Berkeley Vision and Learning Center (BVLC) [31], known for its speed and modularity in deploying convolutional neural networks.
- *Speech Recognition and Natural Language Understanding:*
 - Custom LSTM Model for Speech Analysis: We utilize the Google Cloud Speech-to-Text API [32] to transcribe user speech in real-time. To interpret and analyze the transcribed speech, we developed a custom Long Short-Term Memory (LSTM) neural network. The LSTM model processes the sequence of words to understand the context and extract meaningful information, such as user verbal feedback to the robot.
- *Visual Mental Imagery Generation:*
 - Stable Diffusion Model: The robot’s ability to generate visual mental imagery is powered by the Stable Diffusion model [33], a type of latent diffusion model that utilizes deep learning for image synthesis. When the robot encounters a novel object label, it uses the Stable Diffusion model to generate a visual representation based on the user’s verbal description.
- *Meta-Learning for Continuous Adaptation:*
 - Siamese Neural Network: We implemented a Siamese neural network for meta-learning purposes. This network helps the robot to learn from limited examples and generalize knowledge to new, unseen objects. The Siamese network compares features from known objects to new inputs, facilitating the robot’s ability to adapt and learn continuously from user interactions.

The high-level flow diagram of the entire framework is illustrated in Figure 2, where the fundamental functional blocks are highlighted in grey, and the auxiliary sub-components are highlighted in blue.

3.2 User Command Interface

This functional component is the first contact point for the user, and it obtains the user instruction and pass the command to the subsequent functional components. The user can provide instructions to the robot with either text-based inputs or verbal

commands. To support text-based inputs and input verifications, we developed an interactive graphical user interface (GUI).

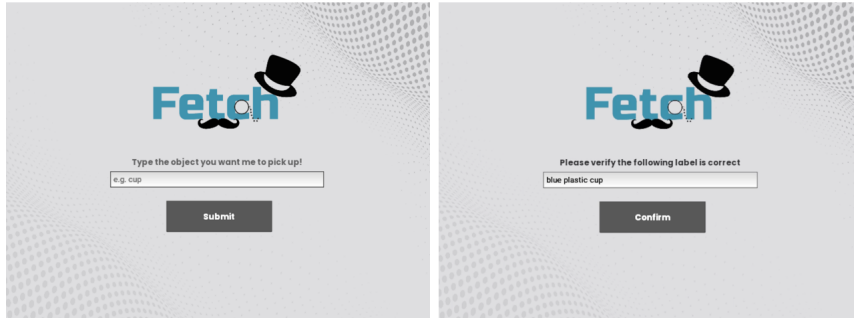


Fig. 3: Left: GUI waiting for user input. Right: input verification.

If the user chooses to directly issue verbal commands, we utilize the Google Speech Recognition API [32] to transcribe speech to text and perform the label extraction process. However, since human language is creative and productive in nature, there are various ways to express an idea. For example, if the user wants the robot to pass a cup, one may say, “please pass me the cup”, or “the cup please”, or “hand me the cup”. It is also natural for users to add descriptors to the target object such as “give me the red plastic cup”. Since we do not restrict how the user issues verbal commands, the possible variations in verbal expressions are huge. Instead of building complex language models to extract the relevant information, we tokenize the entire instruction into word tokens, clean the query by removing stopwords and punctuation, and apply the HunPos tagger [34] to classify tokens’ word classes. We then combine all adjectives followed by nouns to construct the target label. For example, for the command “can you please give me the red plastic cup” the extracted label will be “red plastic cup”. Additionally, we recognize that speech-to-text transcription can be faulty due to many factors, such as accents, signal strength, and audio noise. To address those issues, the GUI displays what the robot heard automatically, if a valid instruction is recognized, prompts the user, and then the robot verbally asks for confirmation before processing the (potentially faulty) command. The inclusion of this additional user confirmation step intends to address the unreliability of natural language interactions with robots due to e.g. noise, model bias, and user accents.

3.3 Ambiguity Analysis

After receiving the user command through the user command interface. The robot examines the workspace with the YOLO object detection model [30] to store all potential targets into the temporary, short-term memory. Subsequently, the framework utilizes a fuzzy inference system to reason about the state of the task and compute a task ambiguity level. The number of potential targets that satisfy the user requirements and their associated confidence levels are explicitly recorded as inputs for the

fuzzy inference system for task ambiguity analysis. Implicitly, the fuzzy Ambiguity analysis block also considers the novelty of object labels; this is to account for the important ambiguity type where the robot cannot locate the correct reference because of a lack of granular knowledge of the object (i.e., the robot is asked for a “red water bottle”, but the robot can only recognize “bottle”). A numeric task ambiguity level can be computed as the final output of this functional block, the ambiguity level can drive the robot’s subsequent behaviors accordingly. An overview of the ambiguity analysis functional block is shown in Figure 4:

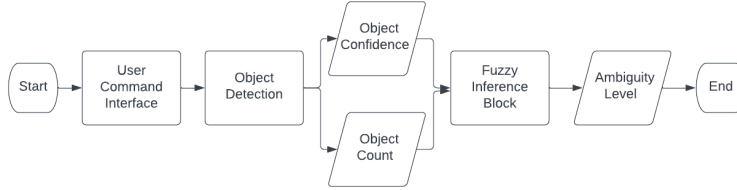


Fig. 4: Ambiguity Analysis functional block overview. [12]

As outlined in Zadeh’s work on fuzzy set theory and fuzzy inference system [35–37], the fuzzy logic system represents data’s imprecision by accepting linguistic variables and modeling them using membership functions. The relationships between terms in the fuzzy set are determined by a network of if-else rules. The membership function is activated based on the inputs and the if-else network, and finally the system yields a numerical output through a defuzzification method.

3.3.1 Membership Function

As described previously, the fuzzy ambiguity analysis block explicitly takes in the number of potential targets and the associated object confidence levels as inputs. These inputs form the antecedents of the system and the consequent is the task ambiguity level. The antecedent linguistic variable ‘object confidence’ is segmented into five categories: very low (vlow), low, medium (med), high, and very high (vhigh); its universe of discourse is normalized, spanning a numerical range from zero to one. Similarly, the ‘ambiguity level’ variable, akin to the confidence level, is framed as a percentage, sharing the same linguistic labels and universe of discourse as the ‘object confidence’. The ‘object count’ input is characterized by four terms: no object (noObj), single object (one), two potential targets (two), and multiple objects (more). The universe of discourse range for ‘object count’ is defined from zero to seven.

The membership functions for both the object confidence level and ambiguity level are characterized by Gaussian functions, having means at 0, 0.25, 0.5, 0.75, 1 and a standard deviation set at 0.1.

$$f(x) = \exp\left(-\left(\frac{x - \mu}{\sigma}\right)^2\right) \quad (1)$$

Ambiguity can emerge when there is low confidence in detecting objects. This diminished confidence can result from multiple reasons, including inadequate lighting, atypical appearances of specific custom objects, viewing angles, among other factors.

The membership function for the object count is determined by a mix of triangular functions, a Kronecker delta function, and a trapezoidal function.

The triangular function can be specified in the general form of:

$$f(x) = \begin{cases} \frac{x-a}{b-a}, & a < x < b \\ \frac{c-x}{c-b}, & b < x < c \\ 0, & x \leq a \text{ or } x \geq c, \end{cases} \quad (2)$$

the translatable Kronecker delta function:

$$\delta(x - T) = \begin{cases} 1, & x = T \\ 0, & \text{otherwise,} \end{cases} \quad (3)$$

and the half trapezoidal function:

$$f(x) = \begin{cases} \frac{x-a}{b-a}, & a < x < b \\ 1, & b \leq x. \end{cases} \quad (4)$$

together the aggregated membership functions:

$$f(x) = \begin{cases} -x + 1, & 0 \leq x < 1 \\ \delta(x - 1), & x = 1 \\ x - 1, & 1 < x \leq 2 \\ -x + 3 \text{ and } (x - 2), & 2 < x \leq 3 \\ 1, & x \geq 3 \end{cases} \quad (5)$$

Incorporating the object count data is crucial since it introduces an important ambiguity type that suggests multiple objects might be potential targets. Even though the fuzziness of the object count variable is confined, the membership function in equation 5 allows it to be woven seamlessly into the broader fuzzy inference system. A graphical illustration of the fuzzy membership functions described is shown in Figure 5:

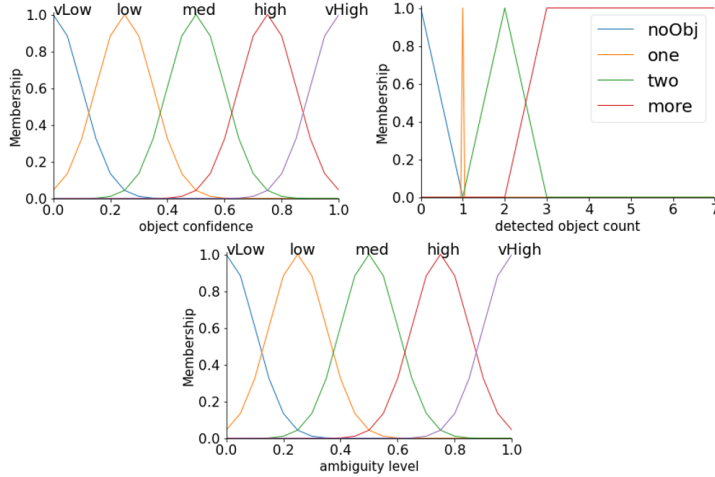


Fig. 5: Ambiguity determination membership functions; top panels show input membership functions (Antecedents) and the bottom panel shows the output membership function (Consequent)

3.3.2 Ambiguity analysis knowledge rules

The inference rules outline the connections between terms of linguistic variables. To gauge the task's ambiguity level, the ambiguity analysis system employs eight intuitive inference rules. These rules are detailed as follows:

1. **IF** (*object confidence* is very low **OR** low) **AND** (*object count* has no object **OR** multiple potential targets) **THEN** *ambiguity level* is very high
2. **IF** (*object confidence* is very low **OR** low) **AND** (*object count* is one) **THEN** *ambiguity level* is high
3. **IF** (*object confidence* is low **OR** medium) **AND** (*object count* has multiple potential targets) **THEN** *ambiguity level* is high
4. **IF** (*object confidence* is low **OR** medium) **AND** (*object count* is one) **THEN** *ambiguity level* is high **OR** medium
5. **IF** (*object confidence* is medium **OR** high) **AND** (*object count* has multiple potential targets) **THEN** *ambiguity level* is medium
6. **IF** (*object confidence* is medium **OR** high) **AND** (*object count* is one) **THEN** *ambiguity level* is low
7. **IF** (*object confidence* is high **OR** very high) **AND** (*object count* has multiple potential targets) **THEN** *ambiguity level* is medium
8. **IF** (*object confidence* is high **OR** very high) **AND** (*object count* is one) **THEN** *ambiguity level* is very low

If multiple items meet the user's criteria, the system retrieves the highest confidence score from all possible targets. For defuzzification, the centroid method is used, which returns the centroid of the area covered by the activated fuzzy membership functions as the precise value [38].

3.4 Attention Assessment

The fuzzy attention assessment block leverages facial key points to monitor the user’s head position and where they are looking. This system was created during the COVID-19 pandemic, a time when facial masks often obstructed significant portions of users’ faces. Considering this challenge, our system adopts a “diamond shape” key point layout on the forehead of the user for head orientation tracking. We categorize head poses into four main directions: ‘head up’, ‘head down’, ‘head left’, and ‘head right’. These directions are ascertained using straightforward thresholding logic relative to the center of the ”diamond” key point structure in both the x and y planes. For gaze tracking, the iris locations and the outer corners of the eyes are selected as key points. The layout of facial keypoints can be seen in Figure 6.

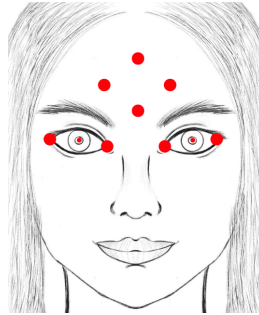


Fig. 6: The facial key points’ layout. [12]

Eyes are subtle facial features; thus, gaze direction necessitates meticulous processing for accurate and reliable assessment. As a result, we utilize a fuzzy inference system that integrates head orientation labels to determine the direction of eye gaze. Both the left and right eye positions serve as the fuzzy system’s antecedents and are categorized into three levels: low, medium, and high. The universe of discourse of each eye is based on the coordinates of its inner and outer corners, allowing the fuzzy system to be locally calibrated to accommodate variations in eye shapes and sizes. The positions of both irises are then used as input data for inferential analysis within the fuzzy system.

The overview of the attention assessment functional block is shown in Figure 7

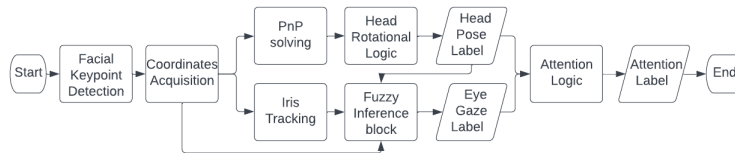


Fig. 7: Attention Assessment functional block overview. [12]

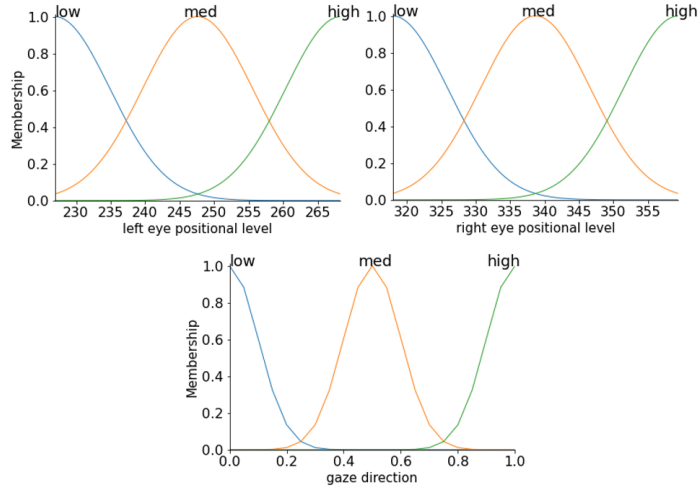


Fig. 8: Eye gaze membership functions, colors refer to respective membership functions; the top panels show input membership functions (Antecedents) and the bottom panel shows the output membership function (Consequent) [12]

The ‘gaze direction’ serves as the resultant linguistic variable within the fuzzy system. This variable is defined by three linguistic terms that characterize the general direction of the gaze: eye left, eye middle, and eye right. The universe of discourse of this variable is gauged on a normalized scale, ranging from zero to one, to represent gaze direction.

3.4.1 Gaze membership functions

The membership functions for the left and right eye position levels adopt a Gaussian shape, as depicted in equation 1. These membership functions are dynamically constructed and tailored to each specific eye. Consequently, the means of the functions are continually fine-tuned to align with the outermost, innermost, and central points of the respective eyes, utilizing a predetermined standard deviation of 8 to facilitate optimal overlap among membership segments. Moreover, by accounting for the head orientation label, the fuzzy system dynamically tweaks the Gaussian membership function’s means by adding offsets to the eye key points. This refinement allows the system to more effectively discern if a user is looking at the subject (in this case, the robot), even when their head is not directly facing the camera.

The gaze direction variable utilizes Gaussian membership functions as well. As previously stated, the universe of discourse of this consequent variable remains static. Hence, the means of the three Gaussian functions that correspond to the three terms in the fuzzy set are fixed at 0, 0.5, and 1, all sharing a uniform standard deviation of 0.1.

3.4.2 Gaze and attention rules

For effective clarification, it is crucial for the robot to discern if the human collaborator is focused on it before commencing any disambiguation actions. Therefore, obtaining insights into the human’s gaze contact becomes essential. If the user isn’t directing their gaze at the robot, it likely indicates a lack of sufficient attention towards the robot. Conversely, if the user’s general gaze direction is toward the robot, it is likely an indication of sufficient human attention. The logic governing gaze contact inference is encapsulated in the following rules:

1. **IF** (*left eye position level* is low) **OR** (*right eye position level* is low) **THEN** *overall gaze direction* is left
2. **IF** (*left eye position level* is medium) **OR** (*right eye position level* is medium) **THEN** *overall gaze direction* is middle
3. **IF** (*left eye position level* is high) **OR** (*right eye position level* is high) **THEN** *overall gaze direction* is right

The final numerical output is determined using the centroid method, as referenced in [38]. This value is subsequently subjected to thresholding functions, allowing the robot to recognize when human eye contact is established or discern when the user’s gaze is directed elsewhere.

Once we determine the label for the human’s gaze focus, we integrate it with the label for human head orientation to assess the overall human attention:

1. **IF** (*head orientation* is center) **AND** (*gaze focus* is eye contact) **THEN** *attention* is full attention
2. **IF** (*head orientation* is center) **AND** (*gaze focus* is eye away) **THEN** *attention* is semi-attention
3. **IF** (*head orientation* is left **OR** right) **AND** (*gaze focus* is eye contact) **THEN** *attention* is semi-attention
4. **IF** (*head orientation* is left **OR** right) **AND** (*gaze focus* is eye away) **THEN** *attention* is no attention
5. **IF** (*head orientation* is down) **THEN** *attention* is no attention
6. **IF** (*head orientation* is up) **THEN** *attention* is no attention

3.5 Disambiguation

The framework allows the robot to establish bidirectional communications between the robot and the user to direct the robot to the correct reference. Specifically, the framework provides the robot with the ability to analyze the human head gesture with a Caffe model [31] for head detection along with the Lucas-Kanade optical flow [39] to analyze the motion of the head. Additionally, the system employs a multi-layer, bi-directional Long Short-Term Memory (LSTM) recurrent neural network trained on the Twitter US Airline Sentiment dataset [40] to analyze speech feedback. After confirming the correct object through interaction, the robot will slowly approach the target, and shifts its gaze (head) between the object and the user to give hints about which target is scheduled to be picked up. As shown in Figure 2, before the robot proceeds to fetch the target item, the robot will be passively observing the human user to seek the ‘absence of objection’. If no objection is detected while the robot

is preparing to fetch the object, the robot will proceed to execute the task. If an objection from the user (through speech or gesture) is detected, then the robot will halt task execution, acknowledge mistakes and retry. This behavior of seeking the *absence of objection* before task execution forms the second loop of social referencing. This simulates the passive, indirect social referencing behavior in humans [41], and can increase the overall robustness of the system. This second loop can increase robustness in safety or otherwise critical object selection tasks. For example, in the case of a home companion or health support robot, making the correct selections (e.g., in terms of medication delivery) can be of life-saving or life-threatening importance, while in the context of robotic co-workers (co-bots), mistakes made during assembly tasks might have additional severe financial and safety critical consequences.

3.6 Learning and Object Selection

When ambiguity arises due to object novelty, the system needs to disambiguate and update its schema. As described in [11], a short-term/long-term memory scheme works in parallel with the robot’s base knowledge schema. If the robot determines the target label is novel, relevant data such as the object’s name, visual representation (in image forms), and map/robot frame coordinates of potential candidates are stored in the short-term memory of the robot. Additionally, the robot would engage in a “mental imagination” process based on the latent diffusion method proposed in [33]. The latent diffusion model with T latents $x_1 \dots x_T$ is in the form of [42]:

$$p_\theta(x_0) := \int p_\theta(x_0 \dots x_T) dx_1 \dots dx_T, \quad (6)$$

where x_0 is the original image, $p_\theta(x_0)$ is the probability the generative model assigns, and the joint probability $p_\theta(x_0 \dots x_T)$ represents the denoising reverse process.

The latent diffusion model denoises noise-filled samples by iteratively removing the normally distributed noise in the reverse process. This reverse process corresponds to the learning of Gaussian transitions of a fixed Markov chain. There is a deterministic forward process that is fixed to a Markov chain to gradually add Gaussian noise to the sample. More details on the training and tuning of the model can be found in [33]. In our system, we configure the unconditional guidance scale parameter to be 18 to make the model follow the user’s speech prompt more precisely. Typically this value is in the range between 5 to 10 to allow creativity of the model, i.e., model can generate contents content without strictly sticking to the prompt. We set the speech prompt of the model input to be “a singular, complete, full front view, ‘*userinput*’ in the center, grey background, photo-realistic”, where ‘*userinput*’ is the extracted object label from the user’s instruction.

Given a novel object label, the robot “imagines” one possible visual appearance of the target item based on the speech instruction of the user. More detailed user descriptions give the latent diffusion model a more precise speech prompt to generate more concrete robot mental visual imagery. Examples of the robot’s visual mental imagery can be seen in Figure 9. Due to speed consideration, only one image is generated from the speech prompt. **Before performing the imagination process, the robot verbally communicates to the participant that it will think about how the target object**



Fig. 9: Examples of the robot’s mental imagery: objects’ visual appearances are generated based on the description of the user.

should look. This communication of intent helps set participant expectations by indicating that the robot is processing information. Following this, the robot appears to be idle, remaining stationary while it generates the image. In addition, the generated image is cropped with MobileNet object detection[43], where the most salient (confident) object is selected for comparison (Figure 10). Essentially, this is analogous to the robot ‘paying attention’ to the most salient object in an imagined scene. This process of locating the most prominent imagined object provides the siamese network with the expected data form (as the network is designed to recognize objects, not entire visual scenes), thereby enhancing network performance. In this process, the recognition label is unimportant; only the bounding box is necessary as the robot needs to isolate the most salient object in the imagined scene. However, if the object detection fails in this process, i.e., no bounding box can be formed, the robot would keep the entire imagined visual scene for later analysis instead of the isolated visual object. This behavior ensures that the system can still proceed to subsequent phases in case of sub optimal data. In the case where the imagination fails to generate or the imagination quality is highly distorted, the robot will proceed with the human-guided disambiguation process as described in Section 3.5 so that the correct reference can still be identified.

After the formation of the robot’s mental visual imagery, the robot would compare potential candidates in its short-term memory (what the robot ‘sees’) with the imagined target candidate (what the robot ‘thinks’) via a custom siamese neural network [44]. Our siamese network is a twin head network with VGG-19 [45] as the two feature extraction embeddings. The siamese neural network computes the similarity score between the query input and the anchor input. We empirically tested different distance metrics such as the L1, L2 distance and the cosine similarity, and various loss metrics such as mean squared error, cross-entropy loss, and contrastive loss, the results are shown in Table 1,

and we determined that the L1 distance of the feature embedding:

$$L1 = \sum_{i=1}^n |anchor_i - query_i|, \quad (7)$$



Fig. 10: An example of an extracted object from the robot’s imagination.



Fig. 11: A comparison of the real vs. imagined object, top: actual robot camera input; bottom: extracted imagined object.

and the contrastive loss (with Margin, M , empirically set to the value of 1, Y represents the true label, and Y_{pred} is the prediction) [11]:

$$L = \text{mean}((1 - Y) \times (Y_{pred})^2 + Y \times (\max(M - Y_{pred}, 0)^2)), \quad (8)$$

perform most satisfactorily for the given task. In the case where the robot’s generated mental image is highly similar to one of the potential candidates (more than 80% similar), the ambiguity of the reference selection task is eliminated, and the robot can execute the object fetching and learning without asking any additional questions. However, if the mental image does not closely resemble any of the potential targets in

Table 1: Loss and distance function experiment results

Distance Metrics	Loss Function	Validation Accuracy (out of 1)
L1	Mean Squared Error	0.86
L2	Mean Squared Error	0.89
Cosine Distance	Mean Squared Error	0.70
L1	Cross Entropy Loss	0.85
L2	Cross Entropy Loss	0.87
Cosine Distance	Cross Entropy Loss	0.74
L1	Contrastive Loss	0.92
L2	Contrastive Loss	0.89
Cosine Distance	Contrastive Loss	0.72

the short-term memory, the memory items will be sorted in descending order based on their similarity to the mental image. When query items are organized this way, the most likely object can be inquired about first. The robot will inquire about the object reference as described in 3.2. Here, the robot can disambiguate in a contextual fashion. Query items are asked based on the visual resemblance of the reference and the robot’s mental imagery. Once the target reference is clarified, either directly with highly similar robot mental imagery or with social feedback from the user, the associated short-term memory item of the target reference is transferred into the long-term memory, with the updated granular information learned through the interaction. Note that the learning is only performed when the system determines that the ambiguity mainly originates from object novelty. It is inefficient and unnecessary to learn object labels that the base knowledge schema already contains.

In later interactions, if the robot recognizes that the requested item has been learned previously, the long-term memory items can be retrieved through the same siamese network. In the retrieval process, the robot stores all potential candidates in its short-term memory and invokes the associated long-term memory item. In this process, the robot compares the short-term memory candidates (what the robot sees) with the long-term memory item (what the robot remembers). The correct reference can be identified by the short-term memory candidate that has the highest similarity with the long-term memory item. As we mentioned, our learning framework is working in parallel with the base object detection model; thus, the original weights of the base object detection model are intact, and a separate knowledge base (object appearance/image, name, characteristics, etc.) grows alongside the base model. Our system enables the robot to continuously update its knowledge base through natural interaction and imagination, yet does not suffer from serious issues in continual learning agents such as catastrophic forgetting. The learning network is illustrated in the left part of Figure 12.

4 System Validation Experiment

Participants interacted with the Fetch robot in four scenarios designed to validate components of the framework. They completed a pre- and post-questionnaire to assess the user’s perception of the robot.

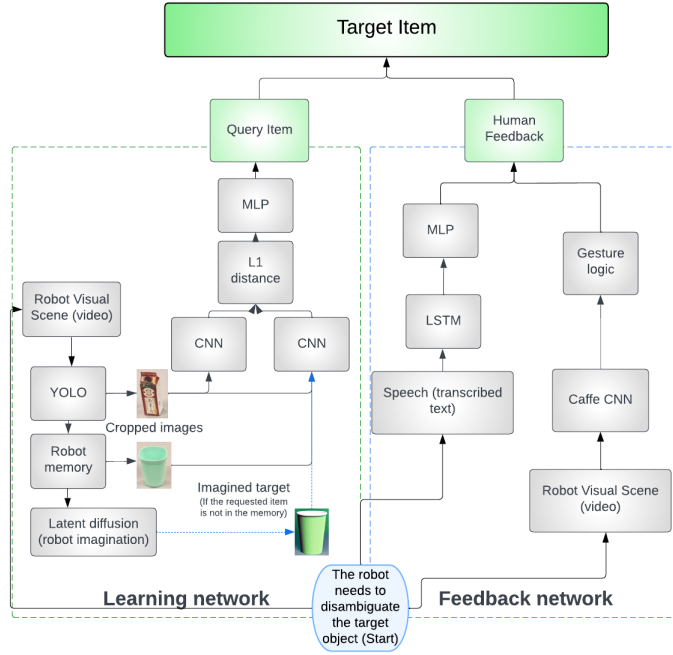


Fig. 12: Network overview of the social referencing disambiguation framework. Learning network with robot imagination (left), feedback network with speech analysis and head gesture analysis (right).

4.1 Procedure

To validate our system, we designed four interaction scenarios.

- **Scenario 1:** The target item is *unique* among the objects. The participant instructs the robot to select an item *known* to the robot. This scenario shows whether the system can analyze an unambiguous scene.
- **Scenario 2:** There are *multiple* potential targets among the objects. The participant instructs the robot to select an item *known* to the robot. **The experimenter verbally provided the participant with a list of labels for objects known to the robot prior to the interaction.** This scenario shows whether the system can analyze an ambiguous scene with multiple potential candidates and/or low detection confidence in base object detection and perform the task with the social referencing disambiguation process.
- **Scenario 3:** There are *multiple* potential targets among the objects. The participant instructs the robot to select an item *unknown* to the robot. **The experimenter verbally instructed the participant to freely describe the object of interest without being limited to predefined labels.** This scenario shows whether the system can analyze an ambiguous scene due to novel object instructions and perform robot mental imagination, learning, and social referencing disambiguation.

- **Scenario 4:** Items in *Scenario 3* are shuffled. The participant instructs the robot to select the item mentioned in *Scenario 3*. This scenario shows the interaction learning outcome. We can see whether the robot learnt the novel object labels through interaction. Can the robot retrieve the appropriate memory item and locate the target reference without disambiguation?



Fig. 13: Experiment setup and the Fetch Robot.

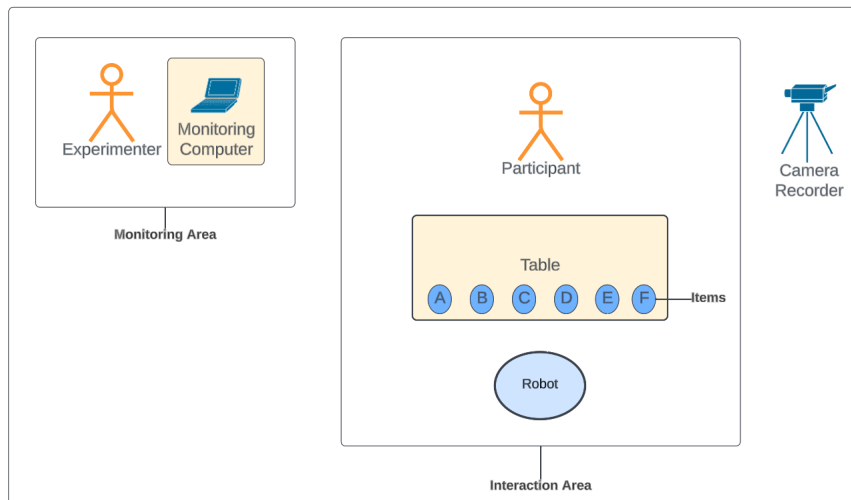


Fig. 14: High level Experimental layout (not to scale).

To ensure better generalization and test the system’s robustness, we randomized the positions of objects and added/removed objects between trials, specifically:

- Rearrangements or additions/removals of objects were applied before the start of each trial with the experimenter verbally informing the participant to ensure clarity. The length adjustment operation was performed first, then shuffling was applied. All random operations utilized a standard pseudo random number generator to ensure unbiased selection.
- The experimenter physically shuffled the positions of the objects on the table to create a new arrangement (swapping existing items on the table). The logic of the shuffling process is defined in Algorithm 1.
- The number of objects on the table was varied between 5 to 8 for each trial, and objects were either added or removed manually by the experimenter to achieve this variation. The selection of which objects to add or remove was done randomly through the logic defined in Algorithm 2.

The random shuffling logic described in Algorithm 1 takes in a list which contains the current layout of objects on the table, and generates a new list with the same objects in a randomly shuffled sequence. This process ensures that the relative positions of the objects are randomized, which helps eliminate any positional bias in the experiment.

Algorithm 1: Random Shuffling Algorithm

Input: List of items L (existing items on the table)
Output: Shuffled list of items R
Initialize empty list $R \leftarrow []$;
while *length of R < length of L* **do**
 Randomly select an item x from L ;
 if x *not in* R **then**
 | Add x to R ;
 end
end
return R ;

The object number adjustment logic described in Algorithm 2 accepts the current sequence of objects on the table as a list and optionally takes a ‘*retain_label*’ parameter, which guarantees that the object with this label will remain in the list after the operation. This optional parameter is important because, in Scenario 3, the robot associates a novel label with an object taught by the participant. Ensuring that this object is not removed allows us to study the recall result in Scenario 4. Overall, the algorithm randomly determines a target length and then adjusts the current object list to match this target length by either appending or removing elements from the list.

Algorithm 2: Adjusting the Number of Objects with Retention Option

```
Input: Current list of items curr;  
Optional label to retain retain_label (default: None)  
Output: Adjusted list curr with size target_size  
Randomly select target_size  $\leftarrow$  integer between 5 and 8 (inclusive);  
Global variable: List of all items all_items;  
if length of curr = target_size then  
| Do nothing;  
else  
| if length of curr > target_size then  
| | while length of curr > target_size do  
| | | Randomly select an item x from curr;  
| | | if retain_label is set and x = retain_label then  
| | | | Continue to next iteration;  
| | | end  
| | | Remove x from curr;  
| | end  
| else  
| | while length of curr < target_size do  
| | | Randomly select an item x from all_items;  
| | | if x not in curr then  
| | | | Add x to curr;  
| | | end  
| | end  
| end  
end  
return curr;
```

We considered a trial to be successful if the robot could select the correct item (the success of object pickup is out of the scope of this paper). The experimental setup is shown in Figure 13. One experimenter was present in the experimental room to monitor the outputs of the framework and recorded the scenario outcome as failure (0), success (1), or task aborted due to user objection detection (0.5) after each interaction. The experimenter was not directly facing the participant but was not hidden. The robot was fully autonomous during the interaction, and the experimenter did not intervene with any decision making of the framework. However, the experimenter had a stop button for safety considerations. The stop button could remotely shut down the robot in case of emergencies.

The participant provided informed consent prior to the experiment. At the start of the experimental session, the participant was welcomed by the experimenter, and a short introduction of the Fetch robot was given. The introduction was conducted in front of the robot to help reinforce the experimenter’s verbal explanations. The introduction involved Fetch robot’s manufacturer, applications, and minor technical details of the robot’s hardware, we also demonstrated some of the robot’s motor motions (robot base movement and head motions). The introduction was designed to

last approximately 3 minutes. The participant was given opportunities to ask questions about the robot; the brief Q&A session lasted at most 1 minute. After the introduction, the participant completed the pre-questionnaire, and then the participant proceeded to interact with the robot scenario by scenario, [as described earlier in this subsection](#). After engaging with all four scenarios, the participant completed the post-questionnaire. The individual study sessions lasted about 40 minutes each.

4.2 Questionnaires

The pre-questionnaire was based on an adapted Technology-Specific Expectations Scale (TSES) [46], the adapted Robotic Social Attributes Scale (ROSAS) on robot competence [47], the Godspeed questionnaire subscale on Perceived intelligence [48], the adapted Perceived Safety with Emotional State (PSE) questionnaire [49], as well as general demographics questions. The pre-questionnaire helped assess participants' expectations and familiarity with technology and robotics. The post-questionnaire was based on the adapted short user experience questionnaire (UEQ-S) [50], as well as the same Godspeed, ROSAS, and PSE questions that were asked in the pre-questionnaire, in addition to some questions on robot predictability, confidence in the robot, robot's dependability, plus an open-ended question on possible points of improvement (*"Which abilities would you improve or add to Fetch if you were to employ it in a home service setting?"*). The ROSAS questionnaire employed a 5-point Likert scale, and the Godspeed and PSE utilized a 5-point semantic differential scale. Additionally, we had 10-point scaled questions regarding the robot's predictability, dependability, and participant confidence in the robot.

4.3 Participants

We recruited 45 participants (17 males and 28 females). The age of our participants ranged from 18 to 41 years old, with an average of 23.36 ± 5.18 years. The majority of the participants were university students (43 out of 45 participants, 2 other participants are outside tourists) from various academic backgrounds, including the Department of Mathematics, different disciplines of engineering (mechanical, civil, electrical and computer engineering), computer science, medicine and health science. The recruitment took place at the University of Waterloo's main campus. This study was approved by the University of Waterloo Human Research Ethics Board.

4.4 Robotic System and Hardware

The system was implemented on a Fetch mobile manipulator robot [51], as illustrated in Figure 13. A Microsoft Surface laptop provided a keyboard input channel and reflected back the robot states as described in 3.2. A Google Pixel 4a camera was mounted onto the 'head' of the robot to provide an additional high-resolution video stream for the attention analysis module. The framework was integrated with the robot operating system (ROS). An Ubuntu 18.04 PC with an NVIDIA RTX 2060 GPU and an i7 Intel Core CPU (11th Gen Intel i7-11700 @ 2.5GHz x 16) hosted all data processing pipelines, deep learning models, and robot planning and control models.

5 Results

5.1 System Performance

We made objective measurements of the outcome of each interaction described in 4.1. We computed the success rate (out of 1) of the framework in each interaction as follows: scenario 1: 0.92 ± 0.26 , scenario 2: 0.84 ± 0.37 , scenario 3: 0.89 ± 0.24 and scenario 4: 0.9 ± 0.27 . Each scenario verifies the different abilities of the framework, and the success rate objectively shows the performance of the framework in resolving the tested ambiguities in object section tasks.

5.2 Perceived Robot Attributes

We performed Cronbach’s alpha test to ensure data reliability and consistency. Our pre-questionnaire data has a Cronbach’s alpha of 0.88, which indicates good data consistency, and our post-questionnaire data has a Cronbach’s alpha of 0.93, which indicates excellent internal data consistency. On average, participants had previously interacted with 1.80 ± 2.84 robots, controlled (remote/wired) 1.07 ± 1.88 robots, and built/programmed 0.51 ± 1.10 robots. The technology/robotic expectation score was computed as the average of all questions presented in the Technology-Specific Expectations Scale. The maximum expectation score is 5 representing the highest expectation, and the lowest is 1. From our participant sample, the overall average expectation score is 3.10 ± 1.33 indicating a moderate amount of expectation toward technology and robotics.

To investigate whether there are significant differences in the perceived intelligence, robot competence, and user-reported emotional state before and after the interaction. We performed the two sample paired t-test on the adapted Godspeed, ROSAS, and the PSE questionnaire:

Table 2: Analysis of participants’ responses before and after the interaction with the system

Construct	Category	Before	After	P-value
Godspeed	Competent	4.04 ± 0.71	4.20 ± 0.94	0.37
	Knowledgeable	3.73 ± 0.81	4.09 ± 0.92	0.0079†
	Responsible	3.8 ± 0.82	4.16 ± 0.95	0.056
	Intelligent	3.76 ± 0.98	4.29 ± 0.79	0.0012†
	Sensible	3.56 ± 0.78	4.13 ± 0.76	0.00023†
ROSAS	Reliable	3.80 ± 0.79	3.80 ± 1.01	1.00
	Interactive	4.00 ± 0.98	4.40 ± 0.84	0.030
	Responsive	4.02 ± 0.78	4.27 ± 1.05	0.19
	Capable	3.98 ± 0.81	4.11 ± 0.91	0.39
PSE	Relaxed	3.75 ± 1.11	4.40 ± 0.84	0.0020†
	Calm	4.16 ± 0.77	4.42 ± 0.92	0.10
	Engaged	3.33 ± 1.00	4.04 ± 0.90	3.53e-5†

Note, mean scores are reported, significant differences, $p < 0.05$ are bolded, items that remain significant after the Benjamini and Hochberg False Discovery Rate correction are labeled with †.

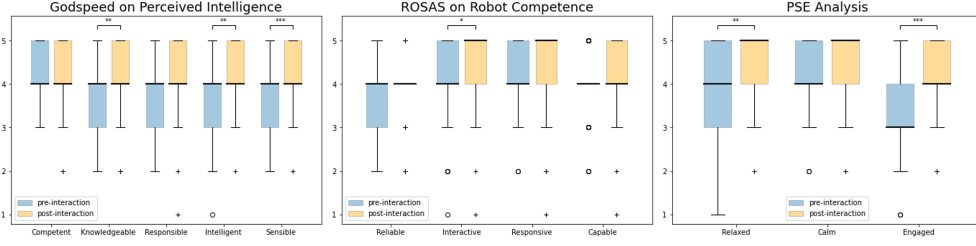


Fig. 15: Participants perception of the robot before and after the interaction. Pre-interaction data are represented by light blue boxes with circle outliers (o), while post-interaction data are represented by light orange boxes with cross outliers (x).

We also asked about the degree of the robot’s predictability and confidence in the robot, as well as the robot’s dependability on a 10-point scale (1 means the lowest degree, and 10 indicates the highest). The averages of these item are computed as 7.20 ± 1.55 , 8.18 ± 1.71 , and 7.36 ± 1.79 , respectively.

Additionally, we computed the mean scores of the **UEQ-S** after the interaction to better understand the general user experience, as shown in Table 3 and Figure 16.

Table 3: User experience after the interaction

Construct	Category	Mean score (out of 5)
UEQ-S	Supportive	4.29 ± 0.79
	Easy	4.06 ± 0.96
	Efficient	3.67 ± 0.95
	Clear	4.09 ± 1.00
	Exciting	4.40 ± 0.78
	Interesting	4.62 ± 0.61
	Inventive	4.11 ± 0.86
	Leading edge	4.00 ± 0.80

Average scores are from 1 to 5, with 5 the highest rating for each attribute.

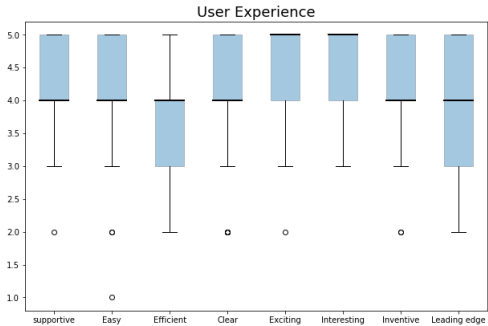


Fig. 16: User experience (**UEQ-S** results).

Given that the “Efficiency” category of the UEQ-S received the lowest scores among all dimensions, we investigated whether there is a correlation between participants’ perceptions of the robot’s efficiency and the robot’s success rate per participant. We calculated the robot’s success rate per participant by averaging the robot’s success outcomes across the four scenarios for each participant. **We then computed the Spearman’s correlation coefficient between the efficiency scores and the robot’s success rates, which was $r = -0.13$ with a p-value of 0.38.**

6 Discussion

The focus of this research lies in understanding the dynamics and practicality of integrating a social referencing framework for service robots, primarily focusing on object selection tasks. Delving deeper into the intricacies of interactive robotic systems offers insights not only into their technical feasibility but also into the human perception shaped by these interactions. By analyzing our findings in the subsequent sections, we seek to provide answers, driving discussions that reveal the strengths, challenges, and potential implications of our developed framework.

RQ1: How can we create a social referencing framework for service robots to handle object selection tasks?

We constructed the framework with modularity in mind, encompassing five core functional components: *User Command Interface*, *Fuzzy Ambiguity Determination*, *Fuzzy Human Attention Assessment*, *Social Referencing Disambiguation*, and *Short-term Long-term Memory Object Learning*. Together, these components establish a foundational social referencing framework, enabling capabilities like input capture, scene comprehension, human state estimation, bidirectional object disambiguation, and interactive learning. Subsequent enhancements to the framework introduced *Robot Mental Imagery/Imagination* and *Objection Absence Detection*, further strengthening the system’s ability to handle unfamiliar objects and fostering more robust interactions than the foundational framework.

RQ2: How feasible is it to develop a framework for service robots that allows the robot to learn objects through interaction and experience?

Our framework’s design and its subsequent deployment on a physical mobile robot demonstrated the viability of developing an interactive system that empowers a service robot to clarify and learn about objects through human guidance and the robot’s own imaginative capabilities. Moreover, our objective success rate metric offers insights into the robot’s ability to manage different ambiguities during object selection tasks, in our experimental settings.

While we consider that all interaction success rates are sufficiently high, we can observe that scenario 2 (where there are multiple potential targets that satisfy the user requirement) has the lowest success rate. This is within our expectation because scenario 2 involves most of the interactions with the user, so that there are more opportunities for the system to make mistakes. Note, scenario 3’s success rate (where the user command involves a novel object label that is not known to the robot) was higher than scenario 2’s. We hypothesize this is because scenario 3 is assisted by the robot’s imagination process, where ambiguity is eliminated directly or after a small number

of clarification questions that the robot needs to ask, thus minimizing the chance for failure. However, we acknowledge that without further data analysis or participant feedback, this remains a hypothesis. The framework’s performance was influenced by multiple factors. Firstly, the framework relies heavily on visual data, robustness of the object detection is critical. Additionally, speech transcription quality has great impact on the fluency of interaction. Improving speech recognition can enhance the framework’s interactivity and create smoother interactions. Furthermore, the diffusion model occasionally generates images that are faulty (incomplete images, out of bound objects ,etc.). This is likely due to the limited number of inference iterations (20 iterations) required to maintain the speed necessary for real-time applications given our computational constraints. Despite these occasional inaccuracies in the model’s output, the framework can still proceed to disambiguation when the imagination is poor as the robot will clarify the reference with human guidance instead.

RQ3: How does the proposed framework affect users’ perceptions of the robot?

From our user perception questionnaire data 2, we can observe that the participants in our sample perceived the robot as significantly more knowledgeable, intelligent, sensible, and interactive after interacting with the system. We believe multiple factors can influence the increase in positive perceptions of the robot: The robot’s ability to learn through social interactions and imagination, the robot’s understanding of human gestures and speech, and the robot’s own gestures, speech, and gaze interaction can all positively impact the user’s perception of the robot. It is important to note that participants completed the pre-questionnaire after the experimenter’s verbal introduction to the robot. This introduction might have influenced their initial perceptions of the robot, potentially leading to higher pre-interaction scores. Therefore, some of the observed improvements in perceptions could be partially attributed to both the introduction and the actual interaction with the robot. Indeed, as shown in Table 2, participants’ pre-interaction scores were generally high, with means above 3 on a 5-point scale. This indicates that participants already had positive perceptions of the robot before the interaction, potentially due to the brief pre-introduction of the robot and more likely due to our participant demographics (University STEM students who have some level of technical familiarity). The high initial scores indicate a potential ceiling effect, which may have limited the observable impact of our framework on measured dimensions. Despite this, significant improvements were still observed post-interaction in areas such as perceived intelligence and interactivity. This indicates that our framework can enhance user perceptions even among individuals who are already familiar with technology and may have positive predispositions toward robots. However, not all categories of perceived intelligence and competence yielded significant differences (specifically, in the categories of reliability, responsiveness, and capability). We speculate that the robot’s navigation and object manipulation ability, the speed of the mental imagination process (on average 36 seconds processing time per instance of the imagination process), and the occasional delay in speech processing and gesture detection due to device networking can negatively influence how the participants perceived the robot, compared to initial expectations they might have had after being first introduced to the robot.

As seen in Table 3, all user experience categories exceed the neutral 3-point mark, and the overall user experience score is 4.14 ± 0.28 . We can also observe the excitement level is rated highly; thus, it is in accordance with the significant differences observed (Table 2) for the ‘Engaged’ item in the PSE questionnaire after the interaction. However, we observe that the efficiency category is rated at the lowest among the user experience categories. This agrees with our responses to the open-ended question, where the most commonly mentioned point (11 responses) is related to the robot’s motor motions and response time. *Although we hypothesized that participants’ perceptions of efficiency would correlate directly with the robot’s task success rate per participant, our analysis did not find a significant correlation (Spearman’s $\rho = -0.13$, $p = 0.38$). This suggests that participants’ perceptions of efficiency were not directly influenced by the robot’s success in completing tasks. Instead, other factors—such as processing delays during the mental imagery generation or interaction smoothness—may have had a more substantial impact on efficiency perceptions. Addressing these factors, particularly by reducing processing times and improving the fluidity of interactions, could potentially enhance users’ perceptions of the robot’s efficiency in future implementations.* Lastly, the clarity aspect can be improved with more speech production from the robot side. Six open-ended question responses mentioned giving more speech capability to the robot and that utilizing robot speech more frequently can be beneficial. The remaining open-ended question responses are not directly relevant to the system’s development (i.e., improving the physical appearance of the robot, mechanical characteristics of the robot etc.).

7 Limitations and Future Work

The proposed framework is not without its limitations. Firstly, there are areas for improvement, such as in language understanding, production, and gesture recognition and production. However, the framework is designed to be modular and extendable, so that, for example, large language models and more sophisticated pose estimation and motion production methods could potentially be integrated to achieve richer and more natural interaction experiences.

In the future, we also plan to enhance the framework’s interactivity and extend the framework to perform other general service robot tasks. We are also interested in incorporating the emotion association [13] aspect of social referencing, and perhaps utilizing object emotional learning to provide more practical utility to the system, such as object preference estimation.

Furthermore, the current framework only focuses on a relatively simplistic single-step task (the object selection task). Currently, the framework stands as a high-level conceptual prototype for service robots, tackling the pragmatic challenge of object selection ambiguity by incorporating human-assisted robot disambiguation and learning. While the current focus is honed on a single-step task, primarily due to time and resource considerations, the framework holds the potential to evolve. With additional time and resources dedicated towards further development, the framework should be expanded and generalized to accommodate multi-step tasks.

Moreover, although our system validation study highlighted the effectiveness of the framework and favourable participants’ views of the robot, it primarily reflected the robot framework’s short-term successes in single session interactions with each participant. As the framework evolves in future, to handle broader multi-step tasks, conducting a long-term study to assess its long-term impact would be a promising next step.

As mentioned in the previous section, participants made suggestions on how to improve the system, including issues regarding the robot’s hardware, movements and embodiment; many of those issues could be addressed in future with a more agile or humanoid robotic platform. Following this point, due to time and resource constraint, we only asked a single open-ended question for participants to make suggestions on the robot after the experiment. In future studies, we should conduct follow-up semi-structured interviews to collect additional qualitative data to reveal more insights on participants’ perceptions on the robot.

Finally, our participant sample consisted mainly of university students studying STEM subjects. The homogeneity of the participants’ demographics is not ideal for extensively assessing the framework’s impact on human perception of the robot, as participants were familiar with technology and more likely to have positive predispositions toward robots. A more diverse sample—including individuals from various educational backgrounds, age groups, and levels of technological familiarity—could provide deeper insights into how participants with little or no technical expertise would perceive and use our system. Future work should aim to recruit a broader participant pool to assess the generalizability of our findings and determine whether similar positive impacts are observed across different demographics.

8 Conclusions

We presented an interactive learning social referencing disambiguation framework with robot mental imagery to enhance service robots’ adaptability to unfamiliar situations.

We demonstrated our framework’s feasibility, competency, and positive impact on participants’ perception of the Fetch robot in the specific chosen experimental setting through our design, implementation, and experimental results.

We detailed the framework’s overall architecture, technical information and background, and implementation details. We also explained different kinds of ambiguity that a robot can encounter in an object selection task, and the framework’s corresponding disambiguation strategies. In addition, we described the framework’s adaptability in handling sub-optimal data. The framework is inspired by human cognitive mechanisms and exploits modern deep learning methods to resolve many ambiguities through natural interactions and robot mental imagery.

We implemented the proposed framework on a physical mobile manipulator robot and validated our system’s performance through a system validation experiment involving 45 human participants. We demonstrated that the robot can competently perform object selection tasks in different object selection scenarios. We analyzed our system through both objective performance measurements and subjective human perception questionnaires.

Results showed significant positive impacts of interacting with the system on user perceptions across multiple dimensions. Specifically, participants reported significant improvements in their perceptions of the robot’s knowledgeability, intelligence, sensibility, and interactivity after the interaction. These findings indicate that our framework not only achieves the robot’s operational capabilities but also positively influences user attitudes.

The strengths of our approach include the robot’s adaptability to ambiguous object selection instructions through social referencing and mental imagery, and its ability to continually learn novel object labels from human interactions, leading to improved user experiences. However, we identified limitations such as the processing time during mental imagery generation, averaging 36 seconds per instance, which affected participants’ perceptions of efficiency. Furthermore, our participant sample consisted primarily of university STEM students, potentially limiting the generalizability of our findings.

This framework holds the future potential to empower service robots to continuously learn through tasks and become “life-long” learners [52] that can continually adapt to their users and their changing needs and preferences.

Declarations

8.1 Funding

This work was supported, in part, thanks to funding from the Canada 150 Research Chairs Program, the Canada Foundation for Innovation John R. Evans Leaders Fund, and the Ontario Research Fund.

8.2 Conflict of interest

The authors have no Conflict of interest to declare that are relevant in the context of this article.

8.3 Ethical approval

The study obtained approval from the University of Waterloo Human Research Ethics Board.

8.4 Consent of participation

Informed consent was obtained from all individual participants involved in the study.

8.5 Consent for publication

Participants gave informed consent to publish anonymized results from this study.

References

- [1] Salamon, A., Száraz, J., Miklósi, Á., Gácsi, M.: Movement and vocal intonation together evoke social referencing in companion dogs when confronted with a suspicious stranger. *Animal Cognition* **23**, 913–924 (2020)
- [2] Gácsi, M., Vas, J., Topál, J., Miklósi, Á.: Wolves do not join the dance: Sophisticated aggression control by adjusting to human social signals in dogs. *Applied Animal Behaviour Science* **145**(3-4), 109–122 (2013)
- [3] Walle, E.A., Reschke, P.J., Knothe, J.M.: Social referencing: Defining and delineating a basic process of emotion. *Emotion Review* **9**(3), 245–252 (2017)
- [4] Merola, I., Lazzaroni, M., Marshall-Pescini, S., Prato-Previde, E.: Social referencing and cat–human communication. *Animal cognition* **18**, 639–648 (2015)
- [5] Schrimpf, A., Single, M.-S., Nawroth, C.: Social referencing in the domestic horse. *Animals* **10**(1), 164 (2020)
- [6] Parkinson, B.: Social referencing in adults and children: Extending social appraisal approaches. *The social nature of emotion expression: What emotions can tell us about the world*, 119–140 (2019)
- [7] Campos, J.J., Thein, S., Owen, D.: A Darwinian legacy to understanding human infancy: Emotional expressions as behavior regulators. *Annals of the New York Academy of Sciences* **1000**(1), 110–134 (2003)
- [8] Bandura, A.: Social cognitive theory of social referencing. In: *Social Referencing and the Social Construction of Reality in Infancy*, pp. 175–208. Springer, ??? (1992)
- [9] Kosslyn, S.M., Thompson, W.L., Ganis, G.: *The Case for Mental Imagery*. Oxford University Press, Oxford (2006)
- [10] Pearson, J.: The human imagination: the cognitive neuroscience of visual mental imagery. *Nature Reviews Neuroscience* **20**(10), 624–634 (2019)
- [11] Fan, K., Jouaiti, M., Noormohammadi-Asl, A., Dautenhahn, K., Nehaniv, C.L.: Social referencing disambiguation framework for domestic service robots. In: *The 40th IEEE Conference on Robotics and Automation (ICRA 2023)* (2023). IEEE
- [12] Fan, K., Jouaiti, M., Dautenhahn, K., L.Nehaniv, C.: Fuzzy object ambiguity determination and human attention assessment for domestic service robots. In: *2022 9th International Conference on Soft Computing & Machine Intelligence (ISCFMI)* (2022). IEEE
- [13] Thomaz, A.L., Berlin, M., Breazeal, C.: An embodied computational model of social referencing. In: *ROMAN 2005. IEEE International Workshop on Robot*

- and Human Interactive Communication, 2005., pp. 591–598 (2005). IEEE
- [14] Boucenna, S., Gaussier, P., Hafemeister, L.: Development of first social referencing skills: Emotional interaction as a way to regulate robot behavior. *IEEE Transactions on Autonomous Mental Development* **6**(1), 42–55 (2013)
- [15] Roy, D., Hsiao, K.-Y., Mavridis, N.: Mental imagery for a conversational robot. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **34**(3), 1374–1383 (2004)
- [16] Mania, P., Kenfack, F.K., Neumann, M., Beetz, M.: Imagination-enabled robot perception. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 936–943 (2021). IEEE
- [17] Lin, Y.-C., Florence, P., Zeng, A., Barron, J.T., Du, Y., Ma, W.-C., Simeonov, A., Garcia, A.R., Isola, P.: Mira: Mental imagery for robotic affordances. In: 6th Annual Conference on Robot Learning (2022)
- [18] Hatori, J., Kikuchi, Y., Kobayashi, S., Takahashi, K., Tsuboi, Y., Unno, Y., Ko, W., Tan, J.: Interactively picking real-world objects with unconstrained spoken language instructions. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 3774–3781 (2018). IEEE
- [19] Sibirtseva, E., Kontogiorgos, D., Nykvist, O., Karaoguz, H., Leite, I., Gustafson, J., Kragic, D.: A comparison of visualisation methods for disambiguating verbal requests in human-robot interaction. In: 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 43–50 (2018). IEEE
- [20] Haasch, A., Hofemann, N., Fritsch, J., Sagerer, G.: A multi-modal object attention system for a mobile robot. In: 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2712–2717 (2005). IEEE
- [21] Lemaignan, S., Garcia, F., Jacq, A., Dillenbourg, P.: From real-time attention assessment to “with-me-ness” in human-robot interaction. In: 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 157–164 (2016). <https://doi.org/10.1109/HRI.2016.7451747> . ISSN: 2167-2148
- [22] Minguillon, J., Lopez-Gordo, M.A., Pelayo, F.: Detection of attention in multi-talker scenarios: A fuzzy approach. *Expert Systems with Applications* **64**, 261–268 (2016) <https://doi.org/10.1016/j.eswa.2016.07.042> . Accessed 2022-04-21
- [23] Li, S., Zhang, X., Kim, F.J., Silva, R., Gustafson, D., Molina, W.R.: Attention-Aware Robotic Laparoscope Based on Fuzzy Interpretation of Eye-Gaze Patterns. *Journal of Medical Devices* **9**(4), 041007 (2015) <https://doi.org/10.1115/1.4030608> . Accessed 2022-04-21

- [24] Asteriadis, S., Karpouzis, K., Kollias, S.: Robust validation of Visual Focus of Attention using adaptive fusion of head and eye gaze patterns. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 414–421 (2011). <https://doi.org/10.1109/ICCVW.2011.6130271>
- [25] Whitney, D., Rosen, E., MacGlashan, J., Wong, L.L., Tellex, S.: Reducing errors in object-fetching interactions through social feedback. In: 2017 IEEE International Conference on Robotics and Automation (ICRA), pp. 1006–1013 (2017). IEEE
- [26] Magassouba, A., Sugiura, K., Kawai, H.: A multimodal target-source classifier with attention branches to understand ambiguous instructions for fetching daily objects. *IEEE Robotics and Automation Letters* **5**(2), 532–539 (2020)
- [27] Zhang, H., Lu, Y., Yu, C., Hsu, D., La, X., Zheng, N.: Invigorate: Interactive visual grounding and grasping in clutter. *arXiv preprint arXiv:2108.11092* (2021)
- [28] Pramanick, P., Sarkar, C., Paul, S., Roychoudhury, R., Bhowmick, B.: Doro: Disambiguation of referred object for embodied agents. *IEEE Robotics and Automation Letters* (2022)
- [29] Pramanick, P., Sarkar, C., Banerjee, S., Bhowmick, B.: Talk-to-resolve: Combining scene understanding and spatial dialogue to resolve granular task ambiguity for a collocated robot. *Robotics and Autonomous Systems* **155**, 104183 (2022)
- [30] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788 (2016)
- [31] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 675–678 (2014)
- [32] André, N., Glen, S., Philip, J., Hans, W.: Web Speech API (2020). <https://wicg.github.io/speech-api/> Accessed 2023-10-12
- [33] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695 (2022)
- [34] Halácsy, P., Kornai, A., Oravecz, C.: Hunpos-an open source trigram tagger (2007)
- [35] Zadeh, L.A.: Fuzzy sets. *Information and Control* **8**(3), 338–353 (1965) [https://doi.org/10.1016/S0019-9958\(65\)90241-X](https://doi.org/10.1016/S0019-9958(65)90241-X) . Accessed 2022-04-26

- [36] Zadeh, L..A.: Fuzzy logic = computing with words. *IEEE Transactions on Fuzzy Systems* **4**(2), 103–111 (1996) <https://doi.org/10.1109/91.493904>
- [37] Zadeh, L.A.: Fuzzy logic. *Computer* **21**(4), 83–93 (1988) <https://doi.org/10.1109/2.53>
- [38] JDWarner/scikit-fuzzy: Scikit-Fuzzy version 0.4.2. Zenodo (2019). <https://doi.org/10.5281/zenodo.3541386> . <https://zenodo.org/record/3541386> Accessed 2022-04-25
- [39] Lucas, B.D., Kanade, T., *et al.*: An Iterative Image Registration Technique with an Application to Stereo Vision vol. 81. Vancouver, ??? (1981)
- [40] Eight, F.: Twitter us airline sentiment (2019). <https://www.kaggle.com/datasets/crowdfLOWER/twitter-airline-sentiment>
- [41] Feinman, S.: Social referencing in infancy. *Merrill-Palmer Quarterly* (1982-), 445–470 (1982)
- [42] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* **33**, 6840–6851 (2020)
- [43] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520 (2018)
- [44] Koch, G., Zemel, R., Salakhutdinov, R., *et al.*: Siamese neural networks for one-shot image recognition. In: *ICML Deep Learning Workshop*, vol. 2, p. 0 (2015). Lille
- [45] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition, pp. 1–14. *Computational and Biological Learning Society*, ??? (2015)
- [46] Alves-Oliveira, P., Ribeiro, T., Petisca, S., Di Tullio, E., Melo, F.S., Paiva, A.: An empathic robotic tutor for school classrooms: Considering expectation and satisfaction of children as end-users. In: *Social Robotics: 7th International Conference, ICSR 2015, Paris, France, October 26-30, 2015, Proceedings 7*, pp. 21–30 (2015). Springer
- [47] Carpinella, C.M., Wyman, A.B., Perez, M.A., Stroessner, S.J.: The robotic social attributes scale (rosas) development and validation. In: *Proceedings of the 2017 ACM/IEEE International Conference on Human-robot Interaction*, pp. 254–262 (2017)
- [48] Bartneck, C., Kulić, D., Croft, E., Zoghbi, S.: Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived

- safety of robots. *International Journal of Social Robotics* **1**, 71–81 (2009)
- [49] Zoghbi, S., Croft, E., Kulić, D., Loos, M.: Evaluation of affective state estimations using an on-line reporting device during human-robot interactions. In: 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3742–3749 (2009). IEEE
- [50] Schrepp, M., Hinderks, A., Thomaschewski, J.: Design and evaluation of a short version of the user experience questionnaire (UEQ-S). *Int. J. Interact. Multimed. Artif. Intell.* **4**(6), 103 (2017)
- [51] Freight Base Research – Fetch Robotics. <https://fetchrobotics.com/freight-base-research/> Accessed 2023-10-13
- [52] Dautenhahn, K.: Robots we like to live with?!-a developmental perspective on a personalized, life-long robot companion. In: RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No. 04TH8759), pp. 17–22 (2004). IEEE