



Università degli Studi di Salerno



**Dipartimento di Ingegneria dell'Informazione
ed Elettrica e Matematica Applicata**

Dottorato di Ricerca in Ingegneria dell'Informazione
Ciclo 38

TESI DI DOTTORATO / PH.D. THESIS

Towards Socially Intelligent Robots: Multi-User Engagement and Multimodal Multi-Task Biometric Recognition

GIUSEPPE DE SIMONE

SUPERVISOR: **PROF. SSA ALESSIA SAGGESE**

PHD PROGRAM DIRECTOR: **PROF. PASQUALE FOGGIA**

Anno 2026

Giuseppe De Simone

Abstract

In recent years, the integration of social robots into everyday environments has witnessed substantial growth, primarily driven by the increasing demand for more natural, intuitive, and effective forms of human-robot interaction (HRI). This trend reflects a broader shift toward developing robotic systems capable of understanding and responding to human social cues, thereby facilitating seamless collaboration and communication in real-world settings. In this context, this thesis addresses key challenges in social robotics by proposing a comprehensive framework that enhances HRI with a particular focus on multi-user interaction, robot’s proactive behavior generation, and soft-biometric recognition through multimodal and multi-task learning strategies.

First, we introduce a hardware-agnostic architecture based on the Robot Operating System (ROS), designed to support real-time interaction in dynamic, multi-user scenarios. The system integrates multimodal perception modules, including head pose estimation, active speaker detection, and direction-of-arrival analysis, together with reasoning components built upon Behavior Trees and Finite State Machines. This design effectively addresses the Multi-Engagement Problem, enabling robots to manage simultaneous interactions with multiple users. Furthermore, the framework was extended with an interactive game to support elderly assistance and enhance user engagement through entertainment.

Second, the thesis presents a novel Vision-Language Model (VLM)-based approach for generating context-aware sentences from video input. A dual-teacher Knowledge Distillation pipeline is proposed to automatically generate training data, combining the strengths of VLMs and Large Language Models (LLMs).

The resulting lightweight model is optimized for deployment on embedded platforms, enabling proactive and contextually relevant verbal behaviors in social robots.

Third, the MAGNET architecture is introduced: a Multi-Modal Multi-Task learning framework for soft-biometric estimation, capable of jointly recognizing gender and emotion from audio and video data. By employing soft parameter sharing, MAGNET achieves state-of-the-art performance while maintaining low computational overhead, making it suitable for real-time applications.

Extensive quantitative and qualitative evaluations confirm the effectiveness of the proposed methods in improving the naturalness, responsiveness, and social acceptability of human–robot interactions. This thesis contributes a scalable and efficient solution for deploying socially intelligent robots in real-world environments.

Contents

1	Introduction	3
1.1	Thesis overview	8
2	State Of The Art	11
2.1	The Multi-Engagement Problem	11
2.1.1	Classic Methods	12
2.1.2	Deep learning approaches	17
2.1.2.1	Multi-Modal Engagement Pre- diction	20
2.1.3	Discussion	29
2.2	Multi-Modal Multi-Task Soft-Biometrics analysis for social robots	31
2.2.1	From Unimodal to Multimodal Approaches in Gender and Emotion Recog- nition	34
2.3	Natural Language Generation	41
2.4	Socially Assistive Robots	47
3	Proposed Method	51
3.1	ROS-based architecture	52
3.2	Perception modules	53
3.2.1	Estimation of the User Engagement State	53
3.2.2	Speech Recognition	57
3.2.3	Hand gesture recognition	58
3.2.4	Soft-biometrics analysis	58
3.2.4.1	Audio encoder	59
3.2.4.2	Video encoder	61
3.2.4.3	Fusion layer	63
3.2.5	Multi-Task framework	63

3.3	Reasoning modules	66
3.3.1	Conversational Modality	69
3.3.2	Game Modality	74
3.3.3	Rule-based Dialogue Management	76
3.3.4	End-to-End Dialogue Management . . .	79
4	Results	87
4.1	Quantitative and Qualitative Evaluation of En- gagement and Interaction Capabilities	87
4.1.1	Quantitative Analysis	88
4.1.1.1	UE-HRI Dataset	88
4.1.1.2	MIVIA-Multi Engagement Dataset	89
4.1.1.3	Experimental Results	91
4.1.2	Qualitative Analysis	92
4.1.2.1	Laboratory Tests	92
4.1.2.2	In the Wild Tests	98
4.1.3	Discussion	100
4.2	Multi-Modal Multi-Task Soft-Biometrics estimation: Experimental Eval- uation	101
4.2.1	Dataset Description and Partitioning . .	102
4.2.2	Implementation details: Visual pre-processing	103
4.2.3	Implementation details: Audio pre-processing	104
4.2.4	Experimental protocol and setup	105
4.2.5	Experimental results	105
4.2.6	Ablation study	114
4.2.7	Inference Time Analysis	116
4.2.8	Discussion	116
4.3	Evaluation of VLM-Based Proactive Interaction Module	118
4.3.1	Quantitative Results	118
4.3.2	Qualitative Results	120
4.3.3	On Device Tests	123
4.3.4	Discussion	124
5	Conclusions	127

*To those who have supported and stood by me throughout this
journey, with patience and affection.*

Chapter 1

Introduction

Nowadays, there has been a noticeable increase in the deployment of robots across a wide range of sectors, including industrial, retail, and medical domains. However, most currently available robots, while proficient in executing predefined and repetitive tasks, generally lack cognitive capabilities. As a result, establishing meaningful collaboration and interaction between humans and robots remains a significant challenge. In particular, in recent years, there has been a growing demand for service robot applications in both domestic and industrial environments [1].

In light of these considerations, a growing number of robot manufacturers have started integrating human characteristics [2] so as to replicate human behaviors [3]. This specific category of robots is commonly referred to as *social robots* [4]. In recent years, the proliferation of social robots has extended into various fields, including retail [5, 6], education [7], and healthcare [8], where they have demonstrated remarkable performance. Specifically, Socially Assistive Robots (SARs) have emerged as a promising technology to support healthcare professionals in delivering targeted interventions for individuals with Autism Spectrum Disorder (ASD). Through collaborative interactions between robots and therapists, significant improvements in social, cognitive, and motor skills can be achieved. These enhancements contribute to the effectiveness of conventional therapeutic approaches, including both behavioral and physical rehabilitation programs for young patients [9, 10]. Within the

domain of SARs, another noteworthy contribution found in the scientific literature is the use of these robots as platforms for digital games, aimed at promoting engagement and learning in children with ASD [11]. In particular, various empirical studies [12, 13, 14] have highlighted the effectiveness of integrating social robots in mitigating feelings of isolation among the elderly. These interventions have been associated with favorable psychosocial outcomes, including reductions in anxiety, stress, and depressive symptoms. Beyond these psychosocial benefits, social robots also demonstrate the capability to monitor the well-being of elderly individuals and detect critical events, such as falls, thereby contributing to improved safety and emergency responsiveness. Furthermore, these technological systems can be employed to support physical health by facilitating rehabilitative exercises aimed at enhancing users' physical well-being. More recently, an emerging area of interest involves the deployment of social robots in public events and environments. For example, Aaltonen et al. [15] employed the Pepper robot to interact with visitors in a shopping mall setting. In another study, a social robot was utilized to assist customers in locating restaurants [16], while in [17], a robot was deployed as a hotel concierge, responsible for welcoming guests and providing assistance.

The analysis of the referenced studies highlights that a fundamental requirement for social robots is the development of perceptual and reasoning capabilities, which are crucial for enabling effective interaction with users in a cognitively informed manner. This includes the ability to track facial expressions, maintain eye contact, and engage in conversations that are sensitive to emotional cues. As a result, a primary objective in this field is to improve the quality of human-robot interaction (HRI), making it more natural, intuitive, and consistent with human social norms and conventions. In this context, one of the major challenges encountered by researchers and developers is the *Multi-Engagement Problem*. This issue pertains to the robot's ability to interact effectively with multiple individuals or groups within a shared social environment, often requiring the management of simultaneous attention and the implementation of adaptive communication strategies. The importance

of implementing an intelligent conversational system is underscored in the study by Moujahid et al. [18]. In their research, the authors designed and developed a robotic head intended to serve as a receptionist. The study defined a set of behavioral policies, each varying in the degree of emphasis placed on empathy and social interaction. These policies incorporated both verbal behaviors, such as the robot greeting users upon their arrival, and non-verbal behaviors, including the robot orienting its head toward the interlocutor to signal attentiveness. Moreover, to further enhance the conversational capabilities of social robots, it is essential to equip them with the ability to analyze user biometrics, such as gender, emotional state, and age, using deep learning approaches [19]. These features are critical for enabling robots to engage in interactions that are both empathetic and contextually appropriate. Nevertheless, beyond these aspects, an equally crucial factor in social robotics is the robot’s response time to user inputs. As demonstrated in [20], latency in the robot’s responses can significantly degrade the quality of human-robot interaction. Therefore, deploying AI models on embedded devices is essential to eliminate network-induced delays and to optimize inference performance. However, such systems are constrained by limited hardware resources, necessitating the optimization of models to reduce the number of parameters, memory footprint, and computational time. In this context, Multi-Task Learning (MTL) offers a promising solution [21]. Notably, Foggia et al. [22] proposed a Multi-Task framework for soft-biometric estimation from audio signals. Their approach demonstrated a reduced computational load on NVIDIA Jetson devices ¹ compared to a Single-Task Learning counterpart, making it more suitable for deployment in resource-constrained environments. Based on the considerations discussed above, Figure 1.1 illustrates the essential components required for a cognitive robot. The robot’s skill set is divided into three overarching domains: *Perception*, *Reasoning*, and *Action*.

The domain of perception encompasses the elements that empower the robot to analyze its surrounding environment,

¹<https://www.nvidia.com/de-de/autonomous-machines/embedded-systems/>

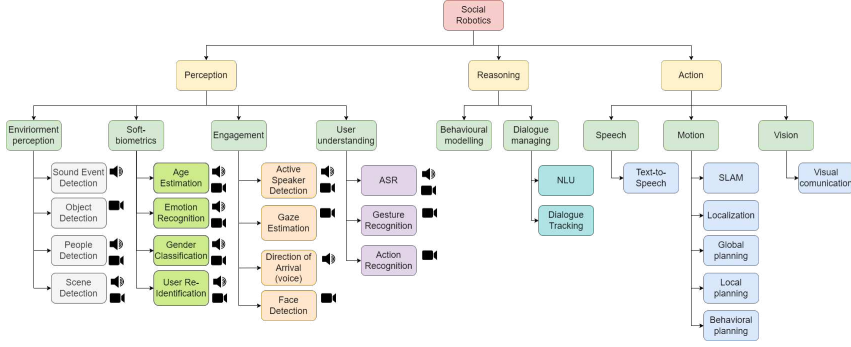


Figure 1.1: The illustration provides a clear depiction of the essential components necessary for the operational framework of a social robot. The presence of the video camera icon and the audio icon signifies that the tasks can be addressed using either video data, audio data, or both, as required.

evaluate the user’s state, and understand user-initiated queries. This capability allows for informed and contextually appropriate communication. In light of the three main domains delineated above, the following key components are described:

- **Environment Perception.** This category encompasses a collection of algorithms employed to analyze the surrounding environment, including functionalities such as *Object Recognition* [23], *People Detection* [23], *Sound Event Detection* [24], and other related tasks.
- **Soft Biometrics Analysis.** These algorithms facilitate the analysis of users’ soft biometric attributes, providing insights into critical aspects such as *Age* [25], *Gender* [26], and *Emotions* [27]. These insights inform the robot’s behavior, enabling personalized and context-aware interactions.
- **Engagement.** This module is dedicated to implementing algorithms that replicate human-like behavior in scenarios involving multiple individuals within a shared environment. Its primary objective is to manage conversational exchanges effectively through strategic mechanisms such as *Active Speaker Detection* [28], *Gaze Esti-*

mation [29], and *Direction of Arrival (DOA)* analysis of vocal cues.

- **User Understanding.** This category includes algorithms designed to interpret user intentions through both verbal and non-verbal cues. Key components include:
 - **ASR (Automatic Speech Recognition).** Responsible for the real-time transcription of auditory inputs into textual data [30].
 - **Gesture and Action Recognition.** Since communication is not limited to speech, this module focuses on interpreting gestures and actions, which are essential for conveying intentions, emotions, and commands. It involves the development of models capable of accurately recognizing gestures [31] and actions [32], thereby enhancing the robot’s ability to respond to diverse forms of human communication.

The reasoning domain encompasses the modules utilized by the system to process the information obtained from the perception components, facilitating the implementation of intricate and user-friendly behaviors. The principal components within this domain are as follows:

- **Behavioral Modeling.** This component involves methods for modeling complex behaviors by amalgamating information derived from the perception modules. Noteworthy methods in this category include the employment of *Finite State Machines (FSM)* and *Behavior Trees* [33] to represent and manage complex decision-making processes and behaviors.
- **Dialogue Management.** This category encompasses modules designed to facilitate human-robot interactions in conversations. It encompasses mechanisms to monitor the states of ongoing conversations, known as *Dialogue Tracking* [34]. Additionally, this component includes modules for understanding human intents, often referred to as *Natural Language Understanding (NLU)*.

NLU is responsible for parsing and comprehending textual inputs provided by users, and is often used in conjunction with ASR to enable a comprehensive understanding of user intent [35].

The action domain provides a comprehensive overview of the essential modules necessary to enable the robot’s interactions with both its environment and human counterparts. It is worth noting that these principal components are categorized into three main groups:

- **Speech** encompasses all the modules that may be employed for communication with the user through speech. One of the critical constituents within this category is the *TTS (Text-to-speech)* module. The TTS component assumes the responsibility of converting textual content into auditory speech data. This particular functionality is utilized by the robot for the purpose of engaging in verbal interactions with human users, utilizing a synthesized voice output.
- **Vision** encompasses a variety of methods that can be utilized for user interaction through visual tools. An illustrative example involves the use of a *Graphical User Interface (GUI)* for this purpose, such as tablets.

1.1 Thesis overview

In this chapter, we conducted a comprehensive exploration of the foundational elements essential for enabling effective human interaction within the field of cognitive robotics, with a particular emphasis on social robotics. Specifically, we examined various application domains of social robots, including retail and healthcare. Across these domains, the necessity of equipping social robots with perceptual and reasoning capabilities emerged as a critical requirement for facilitating meaningful human-robot interaction. Within this context, a particularly important aspect is the concept of *Multi-Engagement*, which denotes the robot’s ability to manage conversations with

multiple users in multi-party scenarios. Nevertheless, perceptual and engagement capabilities alone are not sufficient to guarantee effective interaction. Therefore, analyzing user characteristics, particularly soft-biometric traits, becomes essential for enabling empathetic and cognitively aware dialogues. Furthermore, our analysis indicates that integrating interactive games into social robots can enhance user engagement, particularly in assistance-oriented applications for elderly individuals. In light of these considerations, this thesis focuses on methodologies designed to strengthen the interaction between humans and robots.

Specifically, Chapter 2 provides a comprehensive review of state-of-the-art methods aimed at enhancing human-robot interaction. Specifically, Section 2.1 explores the techniques that can be employed to address the Multi-Engagement Problem in social robots, analyzing both *classical* approaches and modern *deep learning-based* methods. Subsequently, Section 2.2 discusses the estimation of soft-biometric attributes using both single-modal and multi-modal approaches, with the aim of enhancing performance and robustness. Given the importance of minimizing memory usage and inference time in embedded systems, this section also examines the application of *Multi-Task Learning* techniques as a means to optimize computational efficiency. Furthermore, Section 2.3 addresses the importance of implementing proactive behaviors in social robots. It reviews the techniques available in the literature that enable robots to initiate and sustain meaningful interactions, thereby improving user engagement. Finally, Section 2.4 analyzes the use of social robots in healthcare, particularly in supporting elderly individuals through entertainment and assistive functionalities. This section highlights the benefits of employing robots in this domain and identifies the key capabilities required for effective user engagement and assistance.

In Chapter 3, we present the methodologies developed to address the aforementioned challenges. Specifically, Section 3.1 introduces a hardware-agnostic ROS-based architecture designed for this purpose. Then, this section provides a formal definition of the Multi-Engagement Problem and outlines the proposed solution, which integrates multimodal techniques based

on video and audio data with *Behavior Trees* for behavior modeling. In Section 3.3.4, we propose a method based on *Vision-Language Models (VLMs)* to implement proactive behaviors in social robots. This approach enables the generation of contextually relevant sentences by analyzing visual input data, with the additional constraint of operating efficiently on embedded devices. Finally, Section 3.2.4 introduces a *Multi-Modal Multi-Task* learning framework for gender and emotion recognition. The proposed method integrates both video and audio modalities to address these tasks within a unified multi-task architecture, achieving real-time performance on resource-constrained embedded platforms.

In Chapter 4, we present the experimental results. Specifically, Section 4.1 analyzes the impact of the multi-engagement system using quantitative metrics such as precision, recall, and F1-score, together with qualitative evaluations based on user surveys assessing the perceived impact on human–robot interaction. Section 4.2 discusses the results of the proposed *Multi-Modal Multi-Task* method for gender and emotion recognition from video, comparing single-modal and single-task settings with the multimodal multitask configuration, and benchmarking against state-of-the-art approaches. Finally, Section 4.3 reports the evaluation of the VLM-based system for implementing proactive robot behaviors, including quantitative metrics such as mean-token accuracy, as well as qualitative assessments through human evaluations and performance analysis on embedded devices.

Chapter 2

State Of The Art

2.1 The Multi-Engagement Problem

In the field of social robotics, enabling robots to engage seamlessly with multiple individuals in a social context is a significant and challenging endeavor. While robots have excelled in one-on-one interactions, the transition to multi-party engagement is pivotal, offering a multitude of practical applications and transforming human-robot interactions. The importance of multi-engagement is evident across various domains, from education and healthcare to public spaces and homes. For instance, in a classroom, a robot must effectively manage questions and contributions from multiple students, ensuring each one feels valued. In healthcare, robots need to interact with patients and medical professionals concurrently, providing optimal care. As social robots become more integrated into our daily lives, their ability to engage with multiple users simultaneously becomes increasingly essential. In public spaces like airports and shopping centers, they must adapt to diverse user needs, offering information, assistance, or entertainment in real-time. Notably, in healthcare, multi-engagement capabilities are vital as the aging population grows. Social robots should interact with multiple individuals, providing companionship and support tailored to their unique needs. In sum, multi-engagement is a pivotal advancement in social robotics, enhancing practicality, versatility, and the quality of interactions. The pursuit of effective multi-engagement solutions is

fundamental to advancing social robots' capabilities and their seamless integration into our daily lives. In light of these considerations, in this section, we will outline the methods discussed in the state of the art that can be applied to address the problem under investigation. These methods can be broadly categorized into two main groups. The first group encompasses approaches that aim to tackle the issue using pre-deep-learning methods. The second group comprises methods that employ deep learning techniques to address the problem

2.1.1 Classic Methods

Sensor-based methods. One approach employed in human-robot interaction research involves the utilization of sensors, specifically laser scans, for the purpose of detecting the positions and orientations of individuals relative to the robot. This approach has been demonstrated in the work conducted by Yousuf et al. [36]. In the context of this study, a novel model has been implemented on a humanoid robot designed to serve as a guide within a museum and engage in conversations with visitors. The robot is equipped with the ability to move through wheels installed on its underside, as well as manipulate its head and arms by controlling its joints. The robot's head is equipped with eye cameras and an ear microphone, allowing it to move along three axes, thus replicating the movements of a human head. To estimate the location and orientation of visitors within its operational area, the experiment employs a system referred to as a "sensor pole." This system enables the simultaneous tracking of both the robot's and the visitors' positions and orientations independently. The strategic placement of the sensor pole is in front of the designated experimental area. It incorporates two laser range sensors, with one situated at the shoulder level of the visitors and the other positioned at ankle level, approximately 15 cm above the floor. The upper laser range sensor, situated at shoulder level, is responsible for tracking the positions and body orientations of the visitors. The lower laser range sensor, installed at ankle level, serves the purpose of tracking the position and orientation of the guide robot. Furthermore, two cameras are

integrated into the chest area of the robot to detect and track visitors' faces and calculate their directional information. To ensure accurate face tracking, the cameras are suitably tilted upwards, taking into consideration an average human height of 160 cm. In order to improve the human-robot interaction, it is important to note that the robot is developed in the way to respect the social norms with a particular attention to social distance. Notwithstanding the system's commendable performance, it is important to acknowledge certain limitations and considerations associated with this method. One notable aspect to consider is the cost-effectiveness of the system, which may be a concern due to its reliance on specialized hardware requirements. Specifically, the need for specific hardware components can potentially increase the overall cost of implementing and maintaining the system. Furthermore, the utilization of a sensor pole, while instrumental in tracking both the robot and visitors, may introduce certain challenges related to user-friendliness. The presence of such hardware in the environment can potentially detract from the overall user experience, and its use may result in a less user-friendly setting. This, in turn, could have implications for the quality of human-robot interaction, potentially reducing its effectiveness. Lastly, it is essential to consider the context in which the system has been evaluated. As indicated by the authors, the system's testing and evaluation took place in a controlled environment, typically a laboratory setting. Consequently, there exists uncertainty regarding the extent to which these results can be extrapolated to real-world, uncontrolled scenarios or "in the wild." The controlled nature of the testing environment may not fully capture the complexities and nuances of real-world interactions, thereby raising questions about the system's generalizability and real-world applicability.

Pre-deep learning approaches. A seminal contribution within the realm of multi-party interaction is encapsulated in the work by Foster et al. [37], wherein they conceptualized and constructed a robotic bartender as an integral facet of the JAMES project. The JAMES robot bartender (Figure 2.1) facilitates interactions exemplified in the depicted scenario, wherein two



A customer attracts the bartender's attention

ROBOT: [Looks at Customer 1] How can I help you?

CUSTOMER 1: A pint of cider, please.

Another customer attracts the bartender's attention

ROBOT: [Looks at Customer 2] One moment, please.

ROBOT: [Serves Customer 1]

ROBOT: [Looks at Customer 2]

Thanks for waiting. How can I help you?

CUSTOMER 2: I'd like a pint of beer.

ROBOT: [Serves Customer 2]

Figure 2.1: The JAMES socially aware robot bartender [37]

patrons enter the bar area and individually endeavor to place drink orders with the bartender. It is noteworthy that in instances where the bartender is already engaged with the first customer and the second customer arrives, the bartender responds by instructing the latter to wait, culminating the ongoing transaction with the initial customer, and subsequently attending to the second customer's request. Within the bartending context, the foundational stride toward effectuating a successful collaborative endeavor between the robot and the customer involves accurately categorizing the engagement status of all prospective customers present within the scene. This categorization pertains to both the inception of the interaction and its progression. In essence, for the bartender to effectively execute its interactive duty, it is imperative for it to comprehend the societal milieu in its immediate vicinity. This comprehension is essential to ensure that the bartender exclusively engages with potential customers who genuinely seek interaction with it.

The configuration of the system is explicated in Figure 2.2. Evidently, the system adopts a multimodal framework, amalgamating information derived from both auditory and visual sources. The preeminent constituents within the delineated

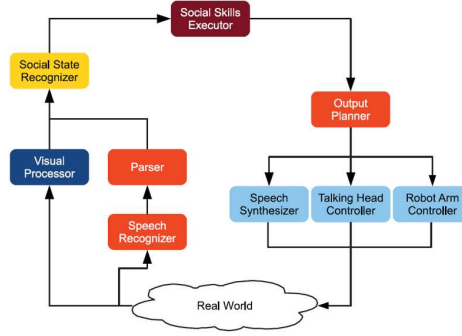


Figure 2.2: The figure below illustrates the system architecture proposed in [37], representing the core components of the "James Project". This architectural framework highlights the primary modules and their interactions within the system. The robot's interaction with the external environment commences with data acquisition through its various sensors. These sensory inputs are then processed by modules responsible for speech recognition and video analysis. The output from these modules is subsequently directed to the **Social State Recognizer**, a critical component that discerns which users from the environment wish to engage with the robot. Following this initial analysis, the **Social Skills Executor** comes into play, responsible for the implementation of the engagement policy. It processes the information from the "Social State Recognizer" and selects the specific user with whom the robot should engage. Finally, the robot's response and actions are executed through its actuators, translating the chosen interaction into tangible and perceptible behaviors.

structure encompass the **Social State Recognizer** and the **Social Skills Executor**. The Social State Recognizer (SSR) is endowed with audio and video data as inputs, leveraging them to extrapolate a social state for the system. Upon identification of pertinent shifts within the social state such as a customer's approach or departure from the bar, emergence of a new hypothesis from the speech recognizer, or completion of the robot's service to a customer, the SSR proclaims the updated state to the Social Skills Executor (SSE). Subsequently, the Social Skills Executor (SSE) ascertains the appropriate

course of action for the robot in light of the revised state. The implementation of the Social State Recognizer can be executed through various methodologies. For instance, a *Hand-Coded Rule* approach can be employed, which involves the analysis of human head pose orientation and proximity concerning the robot. Alternatively, a *Supervised Learning* strategy grounded in classification techniques can be utilized. The researchers of this study have opted for the latter, training the system using a multimodal dataset across three distinct "states":

- **NotSeekingEngagement:** In this state, the customer's visibility to the bartender is established, yet no specific requirements are being expressed.
- **SeekingEngagement:** This state involves the customer actively attempting to initiate interaction with the bartender, without having received any acknowledgment thus far.
- **Engaged:** The *Engaged* state signifies that both the customer and the bartender are engrossed in a transaction pertaining to the ordering of beverages.

An examination of the outcomes attained through the utilization of the two aforementioned strategies, namely the **rule-based** classifier and the **trained** classifiers, yields intriguing insights. In the context of the rule-based classifier, user classification as *seeking engagement* is contingent upon their proximity to the bar (within 30cm) and their orientation predominantly facing forwards (with an absolute torso angle not exceeding 10°). In contrast, within the realm of supervised techniques, including but not confined to *Logistic Regression*, *Support Vector Machine*, and *Nearest-Neighbor*, among others, have been employed.

In the initial series of experiments, the authors subjected both the trained classifiers and the rule-based classifiers to a rigorous 5-fold cross-validation procedure within an offline context. The findings of this assessment reveal that the *IB1* classifier (Nearest Neighbor) exhibited the highest level of accuracy, while conversely, the *rule-based* classifier performed the

least effectively in terms of accuracy. However, in a subsequent set of experiments, an intriguing deviation from the previous results was observed. The acquisition of new data yielded outcomes that contradicted the earlier findings, with the *rule-based* classifier emerging as the top-performing classifier, achieving an accuracy rate of 68%, whereas the *IB1* classifier underperformed. From these results, it becomes apparent that non-deep learning methodologies may not be well-suited for effectively modeling the inherent complexities of this intricate problem. Instead, hand-coded techniques, exemplified by the *rule-based* classifier, appear to offer superior accuracy and stability, underscoring their efficacy in addressing the challenges posed by the problem at hand.

2.1.2 Deep learning approaches

Gaze estimation. With the advent of deep learning, a series of methods have been developed to retrieve information based on user physiognomy. One strategy for enabling automated systems to comprehend a user’s engagement state involves leveraging visual cues [38]. Visual cues encompass implicit visual information that humans use to convey their intentions to their interlocutors. A popular technique consists of the estimation of gaze direction. In the study by Park et al. [39], an attempt is made to solve this problem by employing a cascaded neural network. Figure 2.3 provides a high-level overview of the architecture of the proposed method.

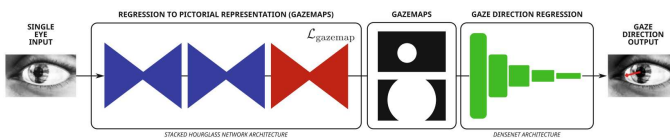


Figure 2.3: General architecture of the gaze estimation system proposed by [39]

As depicted in Figure 2.3, the system is composed of an initial module responsible for estimating two gaze maps. A gaze map is an abstract pictorial representation of the eyeball, with the iris and pupil at its center. These gaze maps are computed

using a series of stacked convolutional neural networks (CNNs). Once the gaze maps are obtained, the system computes the gaze direction directly using the DenseNet architecture for regression. This approach presents a promising avenue for the precise estimation of gaze direction, a crucial factor in comprehending a user's engagement state and promoting effective human-robot interaction. However, it may fall short in certain scenarios. For instance, a user might look at the robot without the intention of initiating a conversation, posing a challenge in accurately assessing their engagement. Additionally, complications arise in situations where multiple individuals within the scene are gazing at the robot simultaneously. Solely relying on gaze estimation becomes insufficient for determining which user to engage with, highlighting the need for more sophisticated approaches to address these complexities in multi-user interaction contexts.

Head pose estimation. Head pose estimation is another technique that can be used to understand the intention of the user regarding which person to talk to. This method consists of estimating the orientation of the user's head in terms of the angles *roll*, *pitch*, and *yaw*, with respect to the orientation system of the camera. A relevant method that can be employed to solve this task is presented in [40], where the authors take advantage of **multi-task learning** to build a system capable of estimating head orientation and detecting a set of facial landmarks.

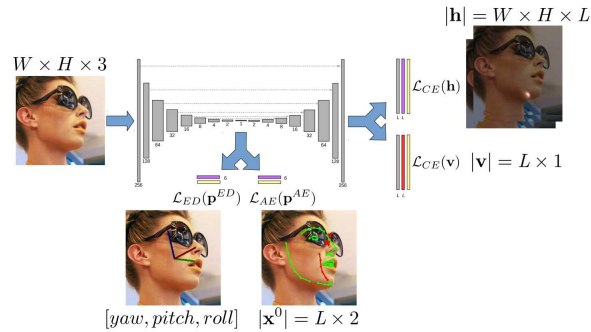


Figure 2.4: Architecture of the system proposed by [40]

The architecture of the method is described in Figure 2.4. It is composed of an encoder-decoder architecture. The goal of the encoder is to globally attribute the features of the subject’s face. Therefore, the authors compute the head pose from the bottleneck layer at the end of the encoder. Moreover, the output of the encoder is used to compute another task, which consists of localizing rigid landmarks on the face. This task requires a fine spatial resolution, and the feature maps at the end of the decoder are utilized to estimate these attributes. The final loss is a weighted sum of the four tasks described above. A significant result worth mentioning is that the system outperforms the state-of-the-art methods. The results obtained by the authors are impressive and demonstrate that training using a multi-task network can provide several advantages, surpassing the state-of-the-art in various aspects.

Body pose estimation. Communication with humans is not solely reliant on verbal cues, such as text or speech, but also encompasses non-verbal cues. Body language, in particular, can be harnessed to extract information about a user’s intent, specifically whether they are inclined to engage in conversation with a robot or not. Within the literature, a wide array of methods has been presented, with notable interest in those based on **2D-Multi-person Pose Estimation** (HPE - Human Pose Estimation). Multi-person HPE presents unique challenges because it necessitates determining the number of individuals present, their positions, and how to group keypoints to identify different people. To address these challenges, multi-person HPE methods can be categorized into **top-down** and **bottom-up** approaches [41]. Top-down methods employ pre-trained person detectors to identify individuals within input images, generating a set of bounding boxes (each corresponding to one person). Subsequently, single-person pose estimators are applied to each person’s bounding box to derive multi-person poses. In contrast, bottom-up methods initially locate all body joints in a single image and then group them into individual subjects. In the top-down pipeline, the number of people in the input image directly influences computation time, making it computationally intensive. Conversely, bottom-up methods

often exhibit faster processing speeds because they do not need to detect the pose for each person separately. The top-down pipeline consists of two key components: a human body detector to obtain person bounding boxes and a single-person pose estimator to predict the locations of keypoints within these bounding boxes. In contrast, the bottom-up pipeline involves two main steps: body joint detection, which includes extracting local features and predicting body joint candidates, and the assembly of joint candidates for individual bodies, entailing the grouping of joint candidates to construct pose representations using part association strategies. These multi-person pose estimation methods offer valuable tools for understanding and interpreting human body language, thereby facilitating improved human-robot interactions. In general, it is observed that numerous multi-person pose estimation methods can be conceptually linked to single-person pose estimation techniques. Within the state-of-the-art literature, two prominent approaches have been developed to address this challenge: *Regression methods* and *heatmap-based methods*. Regression methods leverage convolutional neural networks (CNNs) to predict the coordinates of individual joints directly from the input image. These methods employ a CNN architecture to establish a direct mapping between the image data and the respective joint coordinates [42]. This mapping enables the model to regress the precise 2D coordinates of the human joints based on the visual information extracted from the image. In contrast, heatmap-based methods in human pose estimation diverge from estimating the exact 2D coordinates of human joints. Instead, they focus on estimating 2D heatmaps. Each keypoint heatmap generated by these methods represents the probability distribution of the corresponding keypoint's presence at various spatial positions (x, y) within the image. By analyzing these heatmaps, the model infers the most likely location of each keypoint, enabling accurate pose estimation.

2.1.2.1 Multi-Modal Engagement Prediction

While the current state of the art presents various techniques, it is imperative to recognize the limitations of relying on a single

method exclusively. A case in point is the intricate challenges posed by gaze estimation and body pose analysis. Solely relying on gaze estimation may prove inadequate in determining which user to engage with, particularly in scenarios featuring multiple users simultaneously directing their attention toward the robot. Likewise, exclusive dependence on face or body estimation methods encounters analogous constraints. In light of these considerations, the study conducted by Abdelrahman et al. [43] seeks to address these challenges by introducing a novel model. This model employs rule-based policies to predict human engagement states during interactions involving collaborative robots. Notably, this user-friendly model capitalizes on accessible visual data obtained from a variety of robotic platforms. It demonstrates the capacity to effectively discern the continuum of human engagement, encompassing initiation, disengagement, and even the anticipation of re-engagement states. Importantly, the model leverages facial recognition features to enhance its predictive capabilities, underscoring its potential to enrich human-robot interactions. The model comprises three stages:

1. **Feature Extraction (Stage 1):** Crucial engagement-related features, including human head pose, gaze, and facial identity, are extracted using state-of-the-art convolutional neural networks. Body pose information is also extracted using Kinect v2 technology.
2. **Feature Processing (Stage 2):** Key engagement signals, such as the focus of attention and fused body signals, are inferred using a feed-forward neural network.
3. **Engagement Classification (Stage 3):** Subjects' engagement states are categorized using a straightforward rule-based policy based on the outcomes generated in stage two.

The Multi-Modal engagement and disengagement prediction model's architecture is illustrated in Figure 2.5. It determines whether individuals in the scene are focusing on the robot, primarily relying on a basic neural network model trained

using annotated data. The model employs a 5-dimensional feature vector derived from 3D gaze direction, 3D head pose angles, and facial identification, with a binary attention signal as output. Attention scores are updated based on individuals' frontal face alignment with the robot's gaze. The system exclusively focuses on users with valid skeletons and registered face IDs, which are combined into "fused bodies" for efficient interaction management.

The engagement process is initiated when a subject begins interacting with the system, with the individual exhibiting the highest focus of attention score selected in multi-user scenarios. The system responds by activating the monitor, orienting the robot, and extending personalized greetings. If the engaged individual is not recognized, the system guides them through the registration process. The disengagement evaluation continuously monitors engagement criteria. If prerequisites for interaction are not met, a reminder timer starts, and if the subject does not re-engage within the specified time, a message prompts their re-engagement. A logout timer is initiated during the message display, and upon expiry, the subject's ID is removed if engagement criteria remain unmet. After disengagement, the system transitions to an idle state or engages the next eligible user.

The model's validity is assessed through experiments involving 11 participants. Results show an average precision, recall, and F-score of 96%, 90%, and 93%, respectively. However, the model's applicability to social robots is limited as it processes only one person at a time, making it unsuitable for multi-user interactions and computationally demanding for embedded systems in social robots.

Active Speaker Detection. The analysis of conversational content within video media has garnered significant attention in recent years, with a growing emphasis on multi-modal approaches for audio-visual analysis. This research area holds both academic significance and practical relevance. One particular focus within audio-visual analysis is the examination of speech content within videos. This includes noteworthy subjects such as audio-visual speech recognition or lipread-

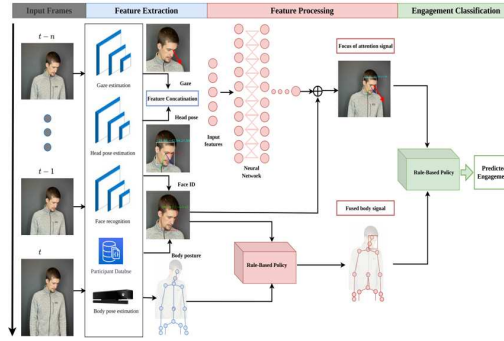


Figure 2.5: The overarching structure of our multi-modal engagement prediction model encompasses three distinct phases: feature extraction, feature processing, and engagement classification.

ing [44], as well as the separation and enhancement of audio-visual speech [45]. An issue of particular importance within this domain is the problem of active speaker detection, which involves identifying individuals speaking within a given scene [28] [46]. Current state-of-the-art solutions employ a multi-modal approach, wherein both audio and visual information are integrated into a unified representation. This integrated representation is subsequently used for classification and the detection of active speakers.

Multi-Task Learning for Audio-Visual Active Speaker Detection. In the work [28] it is introduced a new deep audio-visual model for the task and evaluate on the recently released AVA-ActiveSpeaker dataset [46].

The network, as illustrated in Figure 2.6, consists of two distinct subnetworks: a visual subnetwork and an audio subnetwork, both of which are responsible for extracting per-frame features. These extracted features are subsequently concatenated and processed through a two-layer bidirectional GRU. The primary objective of this bidirectional GRU is to operate on the concatenated features. Following this processing step, a final classifier is employed to generate speaker/non-speaker labels. Throughout the training and evaluation phases, the network receives input sequences comprising T frames along-

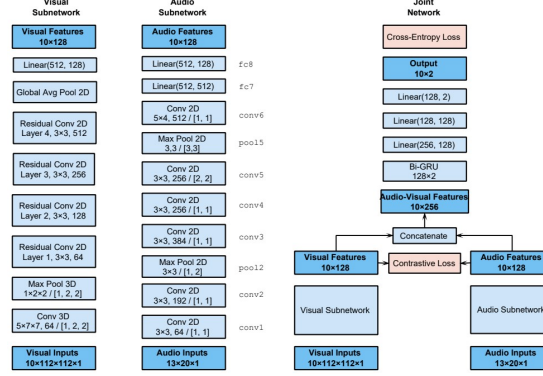


Figure 2.6: The architecture of the method proposed in the study by Zhang et al. [28] is structured around a neural network divided into two distinct subnetworks, each specialized in video and audio analysis, respectively. The video analysis subnetwork is designed to process input data in the form of a 4-dimensional tensor. This tensor encapsulates temporal information, height, width, and the channels of each frame. These components are integral to the analysis and understanding of video content. Conversely, the audio analysis subnetwork operates on a different type of input, specifically a 13-dimensional MFCC.

side corresponding audio representations. Notably, these audio representations vary in length due to discrepancies in frame rates. The network’s output consists of T speaker/non-speaker labels, with each label corresponding to a specific frame in the input sequence.

In the study conducted by the researchers, visual features were extracted using the 3D-ResNet18 architecture, as delineated in the work by Stafylakis et al. [47]. Conversely, audio features were obtained through a bespoke VGG-M network, which accepts 13-dimensional MFCC features as its input. The fusion of these features was executed through the concatenation operation, followed by the integration of the resultant joint features into a two-layer bidirectional GRU comprising 256 cells.

It is important to highlight that the model is trained using a contrastive loss approach. In details, the system minimizes the Euclidean distance among the visual and audio features

Table 2.1: Results of the method proposed in [28]

Method	Visual		Audio-Visual	
	Val	Test	Val	Test
Baseline [46]	/	0.711	/	0.821
Method proposed in [28] (w/o contrastive loss)	0.757	/	0.792	/
Method proposed in [28]	/	/	0.840	0.835

when the speaker is speaking and increase it when the speaker in the visual input is not speaking and the visual input does not match the audio input. The overall loss is the aggregated loss between the loss described before and the cross-entropy loss.

The outcomes, as depicted in Table 2.1, illustrate that this approach has yielded a 4.8% improvement in mean Average Precision (m-AP) compared to the baseline. An intriguing observation lies in the contrast of performance between scenarios utilizing solely visual information [46] and those incorporating audio-visual information. Evidently, the amalgamation of both audio and visual inputs yields a substantial enhancement in performance. Moreover, the incorporation of the contrastive loss has demonstrated a noteworthy enhancement in comparison to experiments conducted without its utilization.

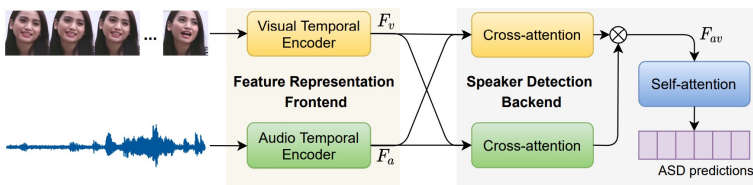
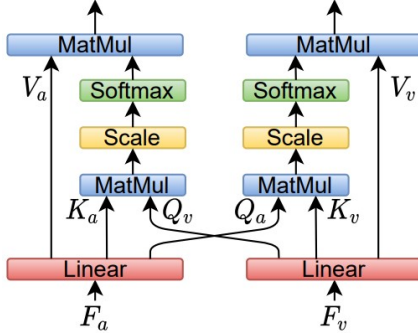


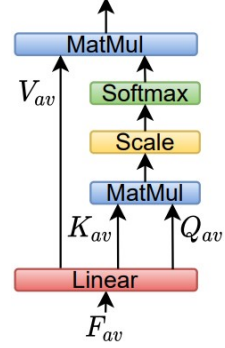
Figure 2.7: Architecture of TalkNet [48]

TalkNet ASD. The architecture proposed into [48] use a configuration similar to [28]. Also in this case we have two main modules dedicated to analysis to video and audio information. But in differently for previous work, the *features fusion* step is performed using the *cross-attention* Figure 2.8a on audio and visual features and then the fusion is performed towards the concatenation operation. The audio-visual vector is given to

a self-attention layer (Figure 2.8b) and then the classification is performed using a MLP. An explanation of the architecture described in the work is reported in the Figure 2.7.



(a) The audio and visual attention layers are designed to receive audio and visual features as input, respectively. Subsequently, they generate the key, query, and value components through the application of a linear layer. The attention values are computed by pairing the query vector with its counterpart (for instance, the query generated from the audio features is multiplied with the key vector derived from the visual features, and conversely)



(b) The self-attention layer bears resemblance to the cross-attention layer, albeit with a notable distinction: it accepts as input the fused audio-visual vector

Figure 2.8: The figures illustrate the steps involved in computing both the **cross-attention** and the **self-attention**, respectively.

The analysis of the *Visual Temporal Encoder* is of interest as it aims to acquire a long-term representation of facial expression dynamics. The objective pursued is the encoding of the visual stream into a sequence of visual embeddings with an equivalent time resolution. The visual frontend component of this approach delves into the spatial information encapsulated within each video frame. This frontend comprises a 3D convolutional layer succeeded by a ResNet18 block. Subsequently,

the visual temporal network is composed of a video temporal convolutional block, characterized by five interconnected residual units featuring rectified linear activation (ReLU), batch normalization (BN), and depth-wise separable convolutional layers. Following this, a Conv1D layer is applied to reduce the feature dimension. The audio temporal encoder is designed to acquire an audio content representation by capturing the temporal dynamics inherent in the audio data. Initially, each audio frame undergoes transformation into a vector comprising Mel-frequency cepstral coefficients (MFCCs). Subsequently, the audio temporal encoder processes the sequence of audio frames as its input and proceeds to generate a corresponding sequence of audio embeddings as its output. Utilizing these techniques, the authors have achieved a notable advancement beyond the current state-of-the-art. The experiments were conducted using the AVA-ActiveSpeaker Dataset [46], an audio-visual database developed by Google to address the active speaker detection challenge. In this experimental context, the *TalkNet* model has demonstrated remarkable performance, attaining a mean average precision (mAP) of 90.8%. This accomplishment solidifies its position as one of the preeminent methods for tackling this particular problem.

As observed, the performance achieved demonstrates a clear and deliberate focus on classification tasks. However, it is important to note that the incorporation of Conv3D layers substantially escalates the network’s parameter count, potentially rendering it ill-suited for deployment in embedded systems and potentially impeding its ability to meet real-time operational constraints.

A Light Weight Model for Active Speaker Detection.

The idea proposed in [49] is a lightweight end-to-end architecture designed to detect active speakers in **real time**. This is achieved thanks to a series of improvements:

- **Feature extraction:** splitting the 3D convolution of visual feature extraction into 2D convolution and 1D convolution to extract spatial and temporal information respectively, and splitting the 2D convolution for audio fea-

ture extraction into two 1D convolutions to extract the frequency and temporal information respectively

- **Cross-modal modelling:** using gated recurrent unit (GRU) with less calculation to replace the complex attention modules for the cross-modal modelling

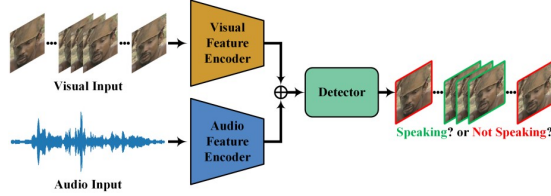


Figure 2.9: Architecture of the proposed method

An overview of the framework is depicted in Figure 2.9. The structure closely aligns with the contemporary state of the art, as evidenced by prior work [48], where two primary encoders are employed to extract features from audio and visual sources, respectively.

The encoder is structured into three visual blocks, each of which incorporates dual pathways designed for spatiotemporal feature extraction. One pathway involves convolution operations following a 3D convolution with a kernel size of 3, while the other pathway employs convolution operations following a 3D convolution with a kernel size of 5. This deliberate choice of design encompasses multiple pathways to facilitate feature extraction with distinct receptive fields, thereby encompassing a rich array of spatiotemporal information.

The features extracted from these distinct pathways are subsequently fused using 1D convolution operations. Each convolution operation within the visual block is complemented by batch normalization and ReLU activation. It is vital to note that all convolutions within the visual feature encoder employ a stride of 1, except for the 2D convolution within the first visual block, which utilizes a stride of 2. This strategic selection serves the purpose of reducing spatial dimensions during feature extraction, resulting in smaller feature maps as output from the visual feature encoder for subsequent processing. The

audio block, akin to the visual block, incorporates two paths with different receptive fields for feature extraction, and feature integration is achieved through a convolution with a kernel size of 1. It is noteworthy that the first two max-pooling layers in the audio feature encoder execute dimensionality reduction in the temporal dimension. Additionally, similar to the visual block, the 2D convolution is split into two 1D convolutions to extract information from the MFCCs. Following feature extraction, these extracted features are consolidated and provided as input to the classifier. The detector’s architecture comprises GRU models, which effectively capture the temporal contextual information within the multi-modal feature set. Subsequently, a fully connected layer is employed to predict whether the candidate is actively engaged in speech. The results demonstrate that the model is capable of achieving state-of-the-art performance while adhering to real-time constraints. Notably, this is achieved with a significant reduction in both the number of parameters and floating-point operations compared to alternative methods, all without any discernible drop in performance. Compared to *TalkNet* method, this one has achieved a reduction of the number of parameters of about the **93%** and an improvement in terms of accuracy of about **2%**. This underscores the model’s suitability for deployment on embedded systems. Additionally, the experiments were executed on a NVIDIA RTX 3090 GPU, where the model achieved an inference time of 96.04ms for 1000 frames. In conclusion, it can be asserted that this method has achieved state-of-the-art performance, demonstrating enhancements in both execution time and model size. Consequently, this method is deemed highly suitable for deployment on embedded systems.

A synthetic overview of the methods presented in the literature to address the problem under examination is depicted in Figure 2.10.

2.1.3 Discussion

From the preceding analysis, it has emerged that in single-user interaction scenarios, the multi-engagement problem is well-defined and relatively straightforward, as the system interacts

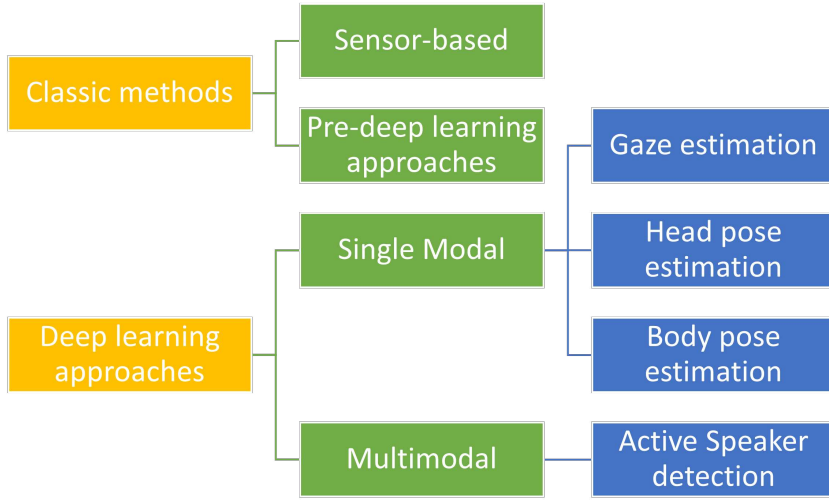


Figure 2.10: Taxonomy of methods presented in the literature for addressing the problem of multi-engagement.

solely with the user positioned in front of the robot. However, in multi-user scenarios, the complexity increases significantly, as the system must manage a variety of social situations. The first fundamental question becomes: *Which user should the robot engage with?* and more specifically, *How can the robot recognize that a user intends to initiate interaction?* This problem is particularly challenging because a user may be physically present in front of the robot without expressing any intention to communicate, or conversely, may attempt to initiate interaction while not being visible to the robot. In this context, classical methods are often inadequate due to their limited performance and flexibility. Deep learning-based approaches are more suitable for addressing these complexities. However, unimodal vision-based systems, such as those relying solely on gaze estimation as discussed in this section, are insufficient on their own. To achieve robust and context-aware interaction, a multimodal approach is required. In this context, combining *Active Speaker Detection* [46] with *Head Pose Estimation* [40] offers a viable solution for identifying individuals who are both speaking and directing their attention toward the

robot. Additionally, incorporating audio-based modules capable of estimating the sound source [50] of speech can help detect users who are speaking but are outside the robot’s field of view. Once the set of users intending to interact with the robot has been identified, the next challenge is: *How should conversational turns be managed?* A user may express the intention to initiate speech while the robot is already engaged with another individual. Therefore, it is essential to implement mechanisms for managing the dynamics of conversational interaction. This challenge, commonly referred to in the literature as *Turn-Taking* [38], can be effectively addressed through behavioral modeling. Engagement rules can be defined and implemented using formal structures such as Finite State Machines (FSMs), enabling the robot to manage multi-party interactions in a socially appropriate and structured manner.

Further details concerning the proposed solution to the addressed problem are discussed in Section 3.1.

2.2 Multi-Modal Multi-Task

Soft-Biometrics analysis for social robots

Advances in deep learning have significantly supported the ability of robotic systems to replicate human-like behaviors. To facilitate natural and socially acceptable interactions between humans and robots, accurate interpretation of soft biometric traits plays a fundamental role, especially in enabling robots to engage in cognitively and emotionally intelligent dialogues with users [20, 51], thus contributing to greater user comfort and acceptance [52, 53]. Within this field, the estimation of gender and emotion has traditionally relied on facial analysis from images to infer these soft biometric attributes [54]. While these methods have demonstrated high accuracy, they predominantly employ *single-task* networks, where separate models are trained independently of each other for each recognition task. This approach not only increases the computational burden but also introduces redundancy when multiple soft biometrics

have to be identified concurrently, since replicating the model for each task significantly increases both computational and memory requirements, which is not a viable solution on embedded devices aboard robotic platforms.[55]. Indeed, an important and non-negligible feature when talking about social robotics is the computational factor: in fact, as experimentally demonstrated in [20], during interactions long processing delays can lead to the perception of low intelligence or low responsiveness of the robotics, undermining users' trust and engagement. Within this context, *multi-task learning* (MTL) approaches have been investigated as a more resource-efficient alternative [56], leveraging the shared structure and inherent correlation between tasks [57, 58]. The core mechanism of MTL lies in shared representation learning, which enhances generalization by utilizing task-related information as an implicit form of regularization [59, 60, 61] while simultaneously reducing computational and memory requirements, primarily due to an overall reduction in the number of parameters compared to single-task approaches. In this context, several architectural paradigms have been proposed to implement MTL effectively. Specifically, *hard parameters sharing* architectures consist of a shared backbone model for feature extraction, to which task-specific classification heads are attached. In contrast, *soft parameter sharing* architectures involve separate feature extraction pipelines followed by task-specific branches. Each branch includes additional layers dedicated to learning task-specific features and culminates in independent classification heads. This design allows for greater flexibility and specialization, enabling the model to better capture the nuances of each task. While hard sharing is advantageous in terms of memory footprint and training speed, it often struggles to accommodate tasks with differing characteristics or learning dynamics. In fact, the rigid shared representation may fail to capture task-specific subtleties, leading to suboptimal performance. On the other hand, soft sharing, despite its slight higher computational cost, typically yields better results, especially in scenarios where tasks are only loosely related or exhibit different temporal or semantic properties, as is the case with emotion and gender recognition. However, despite its ad-

vantages, *Multi-Task Learning (MTL)* presents a central challenge: achieving balanced optimization across tasks that may differ significantly in complexity, loss scale, and convergence behavior. This issue, is commonly known as the *task balancing problem*. It necessitates the careful balance of different tasks to ensure that network parameters converge toward robust shared features that prove beneficial across all tasks. If a particular task dominates the training process excessively, it is likely to manifest its dominance by inducing gradients with relatively large magnitudes. In practical terms, the multitask loss function $\mathcal{L}$ is often assumed to be linear in relation to the single task losses $\mathcal{L}_i$. To elaborate further, the problem can be formally defined as follows:

$$\mathcal{L} = \sum_{i=1}^n \lambda_i \mathcal{L}_i$$

Where $\mathcal{L}_i$ represents the loss function for the i^{th} task, and $\mathcal{L}$ is the final loss, defined as the weighted sum of $\mathcal{L}_i$ contributions over the n tasks. Consequently, the selection of appropriate weights can significantly impact the system’s performance. To address this challenge, various methods have been proposed in the literature to adaptively adjust the weights during training [62]. Notable among these are *Uncertainty Weighting* and *Dynamic Task Prioritization*, which dynamically adjust the contribution of each task’s loss based on its inherent difficulty or predictive uncertainty. Such adaptive weighting mechanisms facilitate more stable and effective joint learning by ensuring that no single task dominates the optimization process. Another well-known method is GradNorm, which seeks to balance the training dynamics by dynamically adjusting the gradient magnitudes associated with each task [63].

In the context of computer vision and soft-biometrics analysis, MTL has been successfully applied to gender and emotion recognition; for instance, in [19] the authors employed convolutional neural networks (CNNs), specifically ResNet and MobileNet, to address these tasks using static face images. However, while image-based approaches are generally adequate for identifying stable attributes such as gender, they are substantially less effective in modeling emotional states, which typically require temporal context for effective characterization.

Indeed, emotions are inherently dynamic phenomena that require the interpretation of *temporal features*, including motion, velocity, and facial muscle activity [64]. Hence, video-based methods are better suited for emotion recognition, as they can capture both short- and long-term temporal patterns necessary for accurate affective analysis [65]. While a valid alternative for addressing gender and emotion classification may be the use of audio data [66], relying solely on audio signals presents significant limitations (compared to visual cues), largely due to the more subtle and ambiguous nature of vocal expressions. Audio-only emotion recognition approaches often suffer from challenges such as background noise, speaker variability, and a less direct mapping between acoustic features and specific emotional states, which can compromise their reliability. Empirical studies consistently demonstrate that integrating audio and visual modalities substantially improves emotion recognition performance, as these modalities offer complementary information and can compensate for each other's limitations [67]. Furthermore, in the context of social robotics, the simultaneous use of audio and video data closely mirrors the way humans naturally perceive and interpret social cues. Just as people intuitively combine facial expressions and vocal tones to assess gender and emotion, robots equipped with multimodal processing capabilities can achieve more accurate and human-like perception. However, when audio and video modalities are processed independently, they often produce separate outputs for the same task, thus requiring a decision-level fusion mechanism to reconcile potentially divergent predictions and generate a coherent interpretation of the observed behavior.

2.2.1 From Unimodal to Multimodal Approaches in Gender and Emotion Recognition

Conventional approaches to gender and emotion recognition have typically relied on unimodal representations, leveraging either audio or visual cues to extract relevant features.

Single Modality - Audio Analytics. In the domain of audio analysis, various neural architectures have been proposed

to extract meaningful information from speech signals and have been applied to a wide range of tasks, including speech recognition, emotion classification, and acoustic event detection.

Early studies primarily employed Convolutional Neural Network (CNN) based architectures to extract features from either spectrograms or raw audio waveforms. For example, Issa et al. [68] introduced a 1D CNN-based model for emotion classification from speech signals.

More recent approaches have incorporated the self-attention mechanism introduced by Transformers [69], leading to hybrid architectures that combine CNNs and multi-head attention blocks. Notably, the Conformer model [70] integrates convolutional modules with self-attention layers, achieving state-of-the-art performance in automatic speech recognition tasks by capturing both local and global dependencies in the audio signal.

In parallel, the emergence of self-supervised learning (SSL) has facilitated the use of large-scale unlabeled audio data to learn generalizable speech representations. In this framework, models are pre-trained on vast amounts of unlabeled data and subsequently fine-tuned on specific tasks, such as emotion or gender recognition. In this context, *HuBERT* (Hidden-Unit BERT) [71] is a Transformer-based SSL model that discretizes acoustic features using clustering and learns to predict masked segments of these sequences, thereby capturing rich contextual information. Similarly, *Wav2Vec2* [72] combines a convolutional feature encoder with a Transformer-based context network to learn robust speech representations from raw audio in a self-supervised manner, without requiring labeled transcriptions. *CoordViT* [73] extends the conventional Vision Transformer (ViT) framework to the audio domain by introducing coordinate-aware representations for spectrogram inputs. Instead of relying solely on positional embeddings, *CoordViT* concatenates normalized two-dimensional coordinate maps $(x, y) \in [-1, 1]$ to each spectrogram patch prior to linear projection. This design preserves the spatial layout of temporal–frequency features and enables the Transformer to capture local acoustic patterns and global dependencies concurrently.

Despite their impressive performance across a wide range of

tasks [74], Transformer-based architectures typically involve a large number of parameters and high computational costs. As a result, their applicability in social robotics remains limited, particularly in scenarios that require real-time processing.

Single Modality - Face Analytics. In the context of video analysis, state-of-the-art methods focus on extracting spatiotemporal features from frame sequences using a variety of architectural paradigms. A common strategy involves extending 2D convolutional neural networks with temporal modeling components, such as recurrent neural networks or temporal pooling operations, to capture dynamic patterns over time. For instance, Ma et al. [75] combine 2D CNNs to extract spatial features by video frames and then these are given as input to LSTM to elaborate temporal information achieving impressive results for the action recognition task.

Another noteworthy class of methods employs 3D convolutional neural networks (3D CNNs), which process spatial and temporal dimensions jointly, thereby enabling the direct modeling of motion patterns without requiring additional temporal modules.

More recently, Transformer-based architectures have been introduced to model spatiotemporal relationships by applying self-attention mechanisms over frame patches.

In this context, *Swin Transformer* [76] extend the classical ViT framework by introducing a hierarchical architecture specifically designed for the efficient and scalable processing of high-resolution images. This architecture achieves improved scalability and adaptability across a wide range of computer vision tasks, including image classification, object detection, and semantic segmentation. The Swin Transformer integrates key inductive biases from convolutional networks, such as locality and shift-invariance, through the use of shifted windows, effectively combining the strengths of CNNs and Transformers. In the same line, the work in [77] introduces *MTCAE-DFER* (Multi-Task Cascaded framEwork for Dynamic Facial Expression Recognition). Here, three stages, namely dynamic face detection, landmark localization, and expression recognition, are jointly trained, so as to progressively enrich feature granularity and promote shared representation learning. The model

employs a pre-trained VideoMAE encoder as a shared feature extractor to obtain global dynamic representations, followed by task-specific Cascaded ViT Decoders designed to enhance global-local feature interaction via a modified multi-head attention mechanism.

While these models have demonstrated strong performance, they are often associated with high computational demands. A viable solution for extracting spatiotemporal features while mitigating computational overhead is the use of 3D convolutional architectures, which jointly process spatial and temporal dimensions in a unified framework. Among early 3D convolutional approaches, *ResNet3D* [78] extends the standard residual network architecture by incorporating 3D convolutions, enabling direct temporal information propagation through stacked residual blocks without introducing architectural bottlenecks. Building on this, $R(2+1)D$ [78] decomposes 3D convolutions into separate spatial and temporal convolutions, simplifying optimization and enhancing the model’s representational capacity.

Multiple Modality. Unimodal systems can perform well in controlled settings, but they often struggle in real-world scenarios where key information is distributed across multiple sources. Leveraging multiple modalities generally improves robustness and overall performance compared to unimodal approaches [67]. However, moving from unimodal to multimodal learning introduces challenges, particularly the need to extract discriminative features from each modality and integrate them coherently. Transformer-based architectures address this by using self-attention and cross-attention to model intra- and inter-modal relationships [79]. Despite their strong generalization, their attention operations introduce substantial computational overhead, limiting their suitability for real-time or resource-constrained applications. CNN-based encoders therefore remain an appealing alternative due to their computational efficiency, though naïve CNN-based multimodal models [49, 45] still require explicit fusion strategies.

Fusion strategies for CNN-based approaches can be grouped into three categories [79]. *Data-level* (early) fusion combines raw modalities before feature extraction, as in RGB–depth fu-

sion. Although effective, it demands that modalities share compatible formats, making it unsuitable for heterogeneous data such as video (namely a 4D tensor) and audio (namely a 2D tensor) [80]. *Decision-level (late)* fusion, based on independent unimodal predictions aggregated through rules such as majority voting, is simple but less robust (e.g., when modalities fail) and offers limited cross-modal interaction modeling [79]. For instance, the authors in [81] propose MATER, modeling temporal audio–visual interactions through Transformer-based encoders and applying late fusion by averaging modality-specific predictions, thus enabling each modality to generalize at its own pace while preserving complementary cues. Finally, *feature-level* fusion avoids the constraint of data-level fusion by using modality-specific encoders whose outputs are combined before classification, enabling modularity and heterogeneous input handling, but requiring more sophisticated mechanisms to enhance cross-modal correlations. This is currently among the most adopted strategy in this field. For instance, the model in [82] employs modality-specific CNN encoders (ResNet-18 for audio spectrograms and 3D-ResNet101 for video) followed by hierarchical attention mechanisms, along with a cross-attention fusion module that captures inter-modal dependencies. A tailored emotional metric loss further enhances embedding discriminability by enforcing polarity-aware constraints inspired by Plutchik’s emotion wheel. Most recently, the VAVL framework [83] introduces a versatile architecture for both unimodal and multimodal emotion analysis that supports partially unpaired audio–visual data. It combines conformer-based acoustic, visual, and shared branches, with residual cross-modal connections and an auxiliary unimodal reconstruction task that strengthens shared representations. A training strategy alternating between unimodal and multimodal optimization further enables robust performance under missing-modality conditions.

Multi Task Learning (MTL). MTL has emerged as a powerful strategy to reduce the computational overhead associated with training and deploying separate models for individual tasks. It is a training paradigm in which a single model is optimized to perform multiple related tasks simultaneously, leveraging shared representations to improve both learning ef-

iciency and generalization. By jointly learning from multiple objectives, MTL enables the model to capture underlying shared structures or interrelated patterns across tasks, which can lead to improved performance, especially when tasks are complementary.

A central challenge in MTL lies in achieving balanced optimization across tasks that differ in complexity, loss scale, and convergence behavior. This issue, commonly known as the *task balancing problem*, has been the focus of various strategies proposed in the literature [84]. Notable among these are *Uncertainty Weighting* and *Dynamic Task Prioritization*, which dynamically adjust the contribution of each task’s loss based on its inherent difficulty or predictive uncertainty. Such adaptive weighting mechanisms facilitate more stable and effective joint learning by ensuring that no single task dominates the optimization process.

Another well-known method is GradNorm, which seeks to balance the training dynamics by dynamically adjusting the gradient magnitudes associated with each task [63].

In the context of MTL, several architectural paradigms have been proposed to implement MTL effectively.

Specifically, *hard parameters sharing* architectures consists of a shared backbone model for feature extraction, to which task-specific classification heads are attached. The work in [85] introduces MTL-BAM (Multitask Learning with Bidirectional Attention Mechanism), a framework enhancing multimodal emotion recognition by emphasizing effective fusion of text, speech, and visual features. It employs modality-specific encoders to extract context-rich representations, which are integrated through a cross-modal interactive attention mechanism that models fine-grained dependencies and highlights emotionally salient information across modalities. A self-supervised label generation unit produces unimodal pseudo-labels to support auxiliary unimodal emotion recognition tasks, which complement the main multimodal classification task. Through hard parameter sharing and joint optimization, MTL-BAM balances learning across modalities, enabling the auxiliary tasks to strengthen the shared representation and improve the overall quality of multimodal fusion and emotion recognition. In

contrast, *soft parameter sharing* architectures involve separate feature extraction pipelines followed by task-specific branches. Each branch includes additional layers dedicated to learning task-specific features and culminates in independent classification heads. This design allows for greater flexibility and specialization, enabling the model to better capture the nuances of each task.

While hard sharing is advantageous in terms of memory footprint and training speed, it often struggles to accommodate tasks with differing characteristics or learning dynamics. In fact, the rigid shared representation may fail to capture task-specific subtleties, leading to suboptimal performance. On the other hand, soft sharing, despite its slight higher computational cost, typically yields better results, especially in scenarios where tasks are only loosely related or exhibit different temporal or semantic properties, as is the case with emotion and gender recognition.

Discussion. Although the promising results achieved by single modality approaches for both gender and emotion, also in multitask settings, we can highlight that multimodal approaches are particularly well-suited to social robotics, as they align with the way humans naturally integrate auditory and visual cues to interpret their surroundings and make decisions. As previously described, among the available architectures, Transformer-based approaches have demonstrated strong performance in multimodal tasks, owing to their capacity to capture complex inter-modal dependencies. However, this advantage often comes at the expense of increased computational demands and inference latency, which can negatively impact the responsiveness required for human-robot interaction. Conversely, CNN-based architectures offer a more lightweight alternative, making them more appropriate for real-time applications on embedded devices. Nonetheless, to achieve effective multimodal integration, these architectures should be carefully designed with modality-specific feature extractors and dedicated fusion mechanisms to ensure efficient and meaningful combination of information across modalities.

In this context, feature-level fusion can offer a good balance between flexibility, performance, and computational efficiency.

Furthermore, deploying separate single-task models for each recognition task increases memory usage and redundancy. MTL offers an impressive alternative, by enabling shared representations across tasks, thereby reducing the overall memory footprint and potentially improving generalization through implicit regularization. In applications such as social robotics, where socially acceptance from the human requires a good tradeoff between accuracy and computational burden, soft sharing offers a balance by enabling richer representations without sacrificing the benefits of multitask learning.

2.3 Natural Language Generation

In the context of social robotics, the implementation of proactive behaviors, which are actions performed by a robot to actively initiate or participate in interactions with users, has been shown to produce positive outcomes, particularly by improving the quality of interaction [86]. These proactive behaviors can be broadly categorized into two main areas, based on verbal and non-verbal cues. Specifically, in Section 2.1, we discussed techniques that can be employed to direct the robot’s attention toward interlocutors, thereby enabling behaviors based on non-verbal cues. In this section, we address the problem of *Natural Language Generation (NLG)*, defined as “the subfield of artificial intelligence and computational linguistics that is concerned with the construction of computer systems that can produce understandable texts in English or other human languages from some underlying non-linguistic representation of information” [87]. Thereby, our focus lies on methods grounded in verbal cues, with particular attention to the generation of contextually appropriate sentences that enhance human–robot interaction through natural spoken language. Specifically, we begin with an analysis of state-of-the-art approaches, starting from early methods based on the recurrent neural network family and progressing toward modern transformer-based architectures.

Over the years, the NLG problem has been approached in various ways [88]. In particular, one of the most influen-

tial paradigms involves the use of Recurrent Neural Networks (RNNs) [89, 90], typically structured in an encoder-decoder architecture, where both components are implemented as RNNs. In this architecture, the encoder processes each token of the input sequence sequentially. At each time step, it takes as input the current token and the hidden state from the previous step, producing an updated hidden state. Commonly, only the final hidden state of the encoder is retained, as it is assumed to encapsulate the semantic representation of the entire input sequence. This final hidden state is then passed to the decoder as its initial hidden state. The decoder, also an RNN, generates the output sequence token by token. At each step, it takes as input the hidden state and the output token from the previous step. The generation process continues until a special End-of-Sentence (EOS) token is produced. Despite their past success, RNN-based models exhibit several limitations. Most notably, they compress the entire input sequence into a single fixed-size hidden state, referred to as the *encoding bottleneck*, which limits their capacity to represent long or complex sequences effectively [91, 92, 93]. Furthermore, due to their inherently sequential nature, RNNs require Backpropagation Through Time (BPTT) during training, which scales linearly with the length of the input sequence [94]. This characteristic makes them inefficient on modern parallel processing hardware such as GPUs [95]. In light of these considerations, unlike RNN-based approaches, which process sequences sequentially and thus suffer from limited ability to capture long-range dependencies, Transformers leverage the *self-attention* mechanism [69]. This mechanism enables direct interaction between all tokens in a sequence, thereby effectively modeling long-range relationships. Furthermore, the parallel processing of tokens significantly accelerates training, particularly on modern hardware architectures. Building on this innovation, numerous approaches have been developed, achieving state-of-the-art performance in text analysis and generation [96], while also exhibiting strong generalization capabilities in zero-shot settings [97].

Futhermore, recent advancements in Large Language Models (LLMs) [98] have substantially improved the cognitive and reasoning capabilities of robotic systems, facilitating their de-

ployment across a wide range of applications. In particular, ongoing research efforts are focusing on the integration of LLM-based reasoning with traditional control paradigms, such as reinforcement learning [99], to address complex decision-making and task execution challenges. A notable example is Say-Can [100], which represents a seminal contribution toward combining LLMs with robotic control. The framework leverages the high-level semantic reasoning abilities of language models alongside a library of pre-trained low-level robotic skills. Upon receiving a natural language instruction, the system generates a set of candidate actions and ranks them based on semantic relevance and feasibility, using learned affordance functions. These developments are increasingly shaping the field of socially interactive robotics by empowering *social robots* to engage in natural, context-aware, and human-like interactions [101]. Specifically, as discussed in Section 1, social robots must be capable of perceiving and analyzing both their surrounding environment and user behavior in order to perform context-dependent actions and interact effectively with humans. In this regard, implementing proactive behaviors [102, 103] has been shown to be particularly useful for capturing users' attention and initiating interactions [104]. Nevertheless, improving these capabilities requires enhancing the perceptual abilities of social robots [105]. To this end, Large Language Models (LLMs) [97], and in particular, Vision-Language Models (VLMs), offer promising avenues, due to their capacity to integrate multimodal data into a unified representation. VLMs such as CLIP [106] and SigLip [107] have demonstrated remarkable performance in tasks like image captioning, often requiring minimal task-specific fine-tuning. Indeed, their ability to align visual and linguistic information within a shared embedding space enables robust generalization across a wide range of real-world scenarios. Despite their strong performance on benchmark tasks, these models remain difficult to be effectively employed in social robotics due to their limited capacity for generating natural and fluent dialogue. This is primarily because they are trained to align image-text representations and are not natively equipped to produce novel textual content in a human-like manner. More recently, models such as BLIP [108] has ad-

vanced this line of research by integrating visual understanding with more sophisticated generative capabilities than those of previous models. In general, BLIP is capable of generating detailed image-based descriptions and answering visual questions. However, since it operates solely on static images, it lacks access to temporal information, which is crucial for producing contextually coherent descriptions and sustaining meaningful interactions over time. For example, a user may join a group that is laughing and naturally inquire about the humorous incident; as evident, this interaction requires an understanding of temporally evolving context. In such cases, video-based models are better suited to capture the nuances of short-term temporal dynamics. In light of this, recent multimodal architectures such as VideoLLaMA [109] and LLaVA-OneVision [110] confirm to be particularly promising: indeed, these models are capable of analyzing both static and dynamic visual input, with the latter showing improved performance in generating contextually accurate descriptions by modeling temporal evolution. A notable innovation in LLaVA-OneVision is its token compression technique, which applies bilinear interpolation to video frames, significantly reducing the number of tokens processed by the underlying LLM. This approach enables substantial reductions in inference time and memory usage, without a significant loss in performance [111]. This capability is particularly valuable in the domain of social robotics, where long computational times can negatively affect the quality of human-robot interaction. This has been experimentally evaluated in [20], where the authors developed a social robot system capable of conversing and playing with users during trade shows. The evaluation, conducted using 5-point Likert scale surveys [112], revealed that while users positively rated the robot’s cognitive and empathic conversational abilities, they reported lower satisfaction regarding the system’s response speed, which was perceived as too slow and consequently detracted from the overall interaction quality. Such latency is not only attributed to the model’s inference time but also to communication delays associated with cloud-based processing. For these reasons, it is necessary to deploy LLM- and VLM-based systems directly on embedded hardware [19], such as NVIDIA Jetson platforms,

thereby eliminating network-induced latency and enabling operation in environments without internet connectivity. However, given the limited computational resources of such embedded devices, it becomes essential to reduce both the number of model parameters and the dimensionality of input data. Achieving an effective trade-off between system performance and computational efficiency is therefore a key consideration for the deployment of advanced AI models in real-world social robotic applications. Nevertheless, although VLM-based approaches have achieved impressive results, particularly in video understanding tasks [113], they have demonstrated the ability to generate accurate descriptions [114, 115] while effectively capturing salient visual and temporal events within a video sequence [116]. These capabilities enable such models to identify and describe the most informative moments by leveraging spatiotemporal reasoning [117] and attention mechanisms [69]. However, their integration into the social robotics domain remains limited. This is primarily due to the unique challenges posed by real-time interactive settings, the demand for robust contextual reasoning, the risk of generating hallucinated or factually inconsistent content [118], and the necessity to produce proactive behaviors and human-like, contextually utterances in dynamic and unstructured environments to enable effective interaction with users. In light of these considerations, at the best of our knowledge VLMs have not yet been integrated into robotic platforms for the purpose of generating context-aware sentences aimed at enabling proactive behaviors for capturing users’ attention. This limitation is primarily due to the current inability of VLMs to generate sufficiently human-like conversational language necessary for natural and engaging interactions [119]. As previously discussed, VLM-based approaches are predominantly designed and optimized for video understanding tasks [118]. Specifically, after pre-training on large-scale corpora through self-supervised learning, LLMs and VLMs acquire strong general-purpose language capabilities. However, these models often lack task-specific knowledge or exhibit undesired behaviors, such as hallucinations [118, 120], when deployed in applied specific settings. Therefore, adapting these to the task at hand—namely, gen-

erating human-like context-aware utterances to enable proactive behaviors in social robots, requires specific training and fine-tuning strategies tailored to the requirements of Human-Robot Interaction (HRI) scenarios[121]. A viable approach is to apply Supervised Fine-Tuning (SFT) [122], a method that involves training the model over datasets composed of human-annotated input-output pairs. SFT is a crucial step in aligning VLM behavior with task-specific demands by optimizing model parameters using a supervised loss function. However, generating datasets for SFT can be time-consuming and resource-intensive [123]. To address this challenge, *Knowledge Elicitation* and *Knowledge Distillation* techniques [124] can be employed to streamline the data preparation process. These approaches involve prompting a large, pretrained language model to produce high-quality labels or responses by carefully crafting input prompts. For instance, Hsieh et al. [125] used a large language model to generate labels for training smaller models [126]. By leveraging the latent capabilities of LLMs, this form of knowledge distillation enables the automatic generation of training data without requiring additional gradient-based updates to the model, effectively treating the model as a zero or few-shot annotator. This technique is particularly advantageous in domains such as social robotics, where annotating multimodal and interaction-rich datasets is costly, labor-intensive, and difficult to scale. In particular, the aforementioned *Knowledge Distillation* technique enables the rapid creation of task-relevant examples that reflect realistic human-robot interaction scenarios, providing a scalable and cost-effective strategy for curating high-quality training data. This approach facilitates the adaptation of general-purpose models to the specific requirements of the social robotics domain.

In light of the aforementioned considerations and recognizing the importance of enabling proactive behaviors in social robots. In Section 3.3.4 is proposed a Vision-Language Model (VLM) specifically tailored for social robotics applications. The proposed system processes video input and generates contextually sentences that reflect visual attributes such as users' outfits, actions, and other environmental cues. These generated utterances are intended to attract users' attention

and foster more engaging and natural interactions with the robot.

2.4 Socially Assistive Robots

In recent times, one of the most significant challenges we are confronted with is the global aging population, which has led to an increasing number of individuals in need of assistance with their physical and mental health, particularly in cases of conditions such as Dementia. According to the projections of the World Health Organization (WHO), by the year 2050, approximately 2.1 billion people worldwide will be over the age of 60, signifying a substantial rise in requests for assistance. This demographic shift is certain to exacerbate the existing shortage of caregivers. In response to this growing problem, cutting-edge technology, including robotics, has emerged as a potential solution. The integration of socially assistive robots (SARs) into the daily lives of elderly individuals and caregivers holds the promise of enhancing their Quality of Life (QOL) and working conditions, respectively.

Elderly care is fraught with a multitude of challenges and routine issues [12]:

- **Loneliness and Social Isolation:** Loneliness and feelings of rejection are among the major concerns that contribute to depression and overall negative emotional states in seniors. SARs, with their social characteristics, present a viable solution to this problem. These systems have the capacity to engage with elderly individuals through repetitive interactions, playing games, and engaging in conversations on relevant topics.
- **Physical Decline and Mobility Issues:** Many seniors experience physical decay and mobility problems that often deter them from moving independently due to the risk of falls attributed to balance issues. This highlights the need for close supervision, which may not always be feasible for caregivers to provide, especially around the clock.

- **Mental Health and Memory Decline:** Mental health and memory decline are also significant concerns for the elderly population. SARs can assist by providing reminders for important events, daily medication schedules, mealtimes, and hydration needs.

In light of these challenges, the authors in [13] designed an application requiring users to recognize songs from the years when they were younger. The song recognition exercise is a task of retrograde memory (for known songs) and anterograde memory (for unknown songs) to which is potentially added also a component of autobiographical (retrograde) memory when the known songs evoke possible personal memories of the individuals. To investigate if this kind of application can be useful, the authors employed two technological tools: a tablet and the social robot Pepper.

Each time the user makes a reply to the quiz question, the game gives positive or negative feedback if the answer is correct or not. One difference between the two devices is the additional feedback that is provided by the robot after a user provides the answer to a specific question. Indeed, the humanoid robot also provides feedback in terms of additional modalities such as head, body, hand, and arm movements, as well as sounds and colored LEDs positioned on shoulders and eyes. The game was well appreciated by all users and resulted in a very competitive atmosphere. In addition, during the tests, some patients were observed as putting in place compensatory strategies for their memory lacks: for instance, one user tried to remember that the correct answer to a specific question was the one having the longest title among the possible options, instead of trying to memorize the correct answer. Caregivers reported that the robot significantly contributed to making the atmosphere in the training program very dynamic, activating, and stimulating, even during a part of the training period.

These remarkable outcomes have also been exemplified in other contemporary works, such as the study conducted by Savage et al. [127]. In their research, the Nao robot was utilized for both entertainment and physical rehabilitation purposes, further highlighting the potential and versatility of socially assistive robots in various domains.

The question at hand is: *What type of game should be developed?* Despite the gradual increase in the percentage of senior gamers, there remains a limited understanding of their gaming preferences and habits. It remains unclear which types of games seniors tend to engage with. In an investigation conducted by Nap et al. [128], the impact of a category known as "**casual games**" on elderly individuals was analyzed. Casual games are characterized by their straightforward in-game rules and suitability for short gameplay sessions. Their design primarily emphasizes gameplay rather than intricate storylines or high-end graphics. The study's outcomes indicated that all participants derived enjoyment from playing casual games. Notably, one participant particularly relished engaging in turn-based gaming with her (grand)children. In contrast, another participant expressed a disliking for multiplayer or turn-based gameplay with his grandchildren, as their enjoyment seemed to stem from his occasional failures. Two participants mentioned that they played digital games as a means to enhance their cognitive faculties and reflexes. An intriguing observation was made among participants whose partners were coping with moderate-stage dementia. These participants found that digital gaming provided a respite from their sorrows. Notably, one participant, whose spouse had dementia, ceased discussing their sorrow when the gaming session commenced, instead expressing delight and immersion in the gameplay experience.

In another study conducted by Groenewoud et al. [11], the authors sought to explore the potential impact of different types of games on the quality of life of individuals affected by dementia. The investigation yielded a spectrum of positive experiences associated with game engagement. These positive outcomes included the fulfillment of individuals' need for achievement, a boost in self-esteem, an elevated sense of connection and belonging, reinforcement of personal identity, engagement in meaningful activities, and a genuine appreciation for the games themselves. Conversely, negative experiences emerged, characterized by instances of failure, which led to diminished self-esteem, irritation, and an increased sense of insecurity. It is imperative, however, to underscore the critical importance of achieving a harmonious alignment between the

chosen game, the touchscreen skills of the individual with dementia, and their unique personal ambitions or interests. The process of identifying the ideal game tailored to the specific needs and capabilities of each individual can indeed pose a challenging endeavor.

The rationale for implementing SARs in elderly care is compelling, with a multitude of benefits. These systems can, in many cases, fully or partially restore the autonomy of the elderly population. By integrating telemedicine capabilities, SARs enable remote treatment of patients, assist caregivers in their roles, and provide continuous monitoring of seniors. Moreover, SARs possess empathic skills that can emotionally support patients, adding another dimension to their potential benefits in elderly care. The integration of socially assistive robots in elderly care has the potential to address the numerous challenges faced by both the elderly and caregivers. These robots can provide companionship, assist with physical tasks, and enhance the overall quality of life for the elderly population. As the world's elderly population continues to grow, the role of SARs in elderly care is expected to become increasingly important.

Chapter 3

Proposed Method

The extensive literature review presented in Chapter 2 highlights the growing importance of cognitive robotics, particularly within the domain of social robotics. Intelligent robotic systems have made substantial progress, enabling them to engage in human-like conversations, extract relevant information, and address a range of challenges, including but not limited to *Spoken Language Understanding and Generation* and *Soft-Biometrics Analysis*. In Chapter 1, we examined the fundamental characteristics that a social robot must possess. Subsequently, Chapter 2 explored state-of-the-art techniques proposed in the literature to enhance human-robot interaction, while identifying existing limitations and research gaps. Section 2.1 investigated techniques for addressing the Multi-Engagement Problem. The analysis revealed that conventional methods exhibit limited performance compared to deep learning-based approaches. Within this context, we compared unimodal and multimodal approaches, and found that multimodal techniques integrating both audio and visual data, demonstrated superior effectiveness. Subsequently, in Section 2.2, we explored both unimodal and multimodal techniques for estimating soft biometrics, such as gender and emotion, from audio and visual inputs within a Multi-Task Learning (MTL) framework. Once again, multimodal approaches outperformed unimodal counterparts. When combined with MTL, these methods proved suitable for deployment on embedded systems, offering reduced memory footprint and faster inference times.

Furthermore, Section 2.3 discussed strategies for implementing proactive verbal behaviors using transformer-based architectures, with a focus on deploying these models on embedded devices to ensure low-latency and real-time responsiveness. Finally, Section 2.4 examined a practical application of social robots in healthcare environments, where they are employed to support and entertain elderly individuals in their daily routines.

In light of these considerations, Section 3.1 introduces a hardware-agnostic architecture built upon the ROS framework, serving as the foundational software infrastructure for implementing the discussed capabilities of social robots. The modules responsible for the robot’s perception capabilities are described in Section 3.2, with particular emphasis on soft-biometrics analysis using the Multi-Modal Multi-Task Learning approach presented in Subsection 3.2.4. Subsequently, Section 3.3 details the modules responsible for processing the acquired information and coordinating human-robot interaction.

3.1 ROS-based architecture

In this section, we describe the proposed a *hardware-agnostic* ROS-based architecture specifically devised for social robots, featuring multi-engagement capabilities. The proposed social robot is designed to autonomously interpret user requests through voice activity detection and speech recognition, and to facilitate interaction with multiple users by implementing a multi-engagement policy based on multi-modal inputs (audio and video). The robot is also equipped with empathic conversational capabilities. Moreover, motivated by evidence that engaging in interactive games with the robot enhances the quality of human interaction [20], as discussed in Section 2.4, we have added a functionality that enables humans to play games with the robot.

The overall architecture of the proposed system is illustrated in Figure 3.1. Due to its complexity and the potential difficulty in interpretation, its components are described separately and in detail in the following sections. A high-level

representation is provided in Figure 3.2.

The proposed system can be integrated with multiple multimodal devices (audio and video) capable of gathering input data. These inputs are processed by the *Perception* modules, which are responsible for analyzing both the environment and the individuals within it. Then, *Reasoning* modules are instead responsible for analyzing data obtained by perception components, guiding the robot’s next steps according to sensory input, contextual cues, and prior interactions. Finally, *Action* modules enable user communication through visual (e.g., displaying information on a tablet), audio (e.g., speech output or audio playback) and motion (e.g., moving of the head or the arms) cues. Note that our system is hardware-agnostic: the only hardware dependent ROS nodes are the acquisition nodes (for audio and video streams) and the actuators nodes, the latter managed through specifically devised ROS action servers (e.g. allowing the movement of different parts of the robot through its SDK). This allows the core of the platform (both the reasoning and perception modules) to be reused when using different robotic platforms, as only the hardware-dependent nodes would need an adaptation. Furthermore, our solution can run in real time on the low-power embedded device installed on board the robot. For this purpose, we have carefully selected and customized the algorithms to guarantee the best trade-off between accuracy and computational burden, with timely responses and low memory footprint requirements.

3.2 Perception modules

3.2.1 Estimation of the User Engagement State

An overview of the proposed multi-engagement approach is shown in Figure 3.3. The video stream is given as input to a face detection and tracking module; face detection is performed using the UltraFace detector ¹, which provides high accuracy with low computational burden. As for the track-

¹<https://github.com/Linzaer/Ultra-Light-Fast-Generic-Face-Detector-1MB>

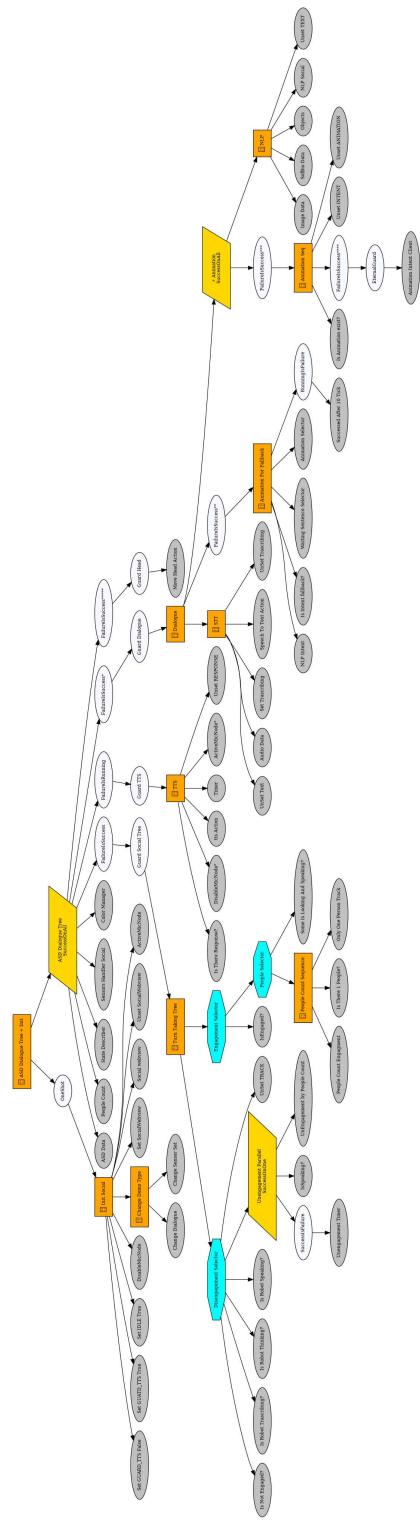


Figure 3.1: Complete Behavioral Tree of the proposed system, used to model the robot’s behavior during interaction with users.

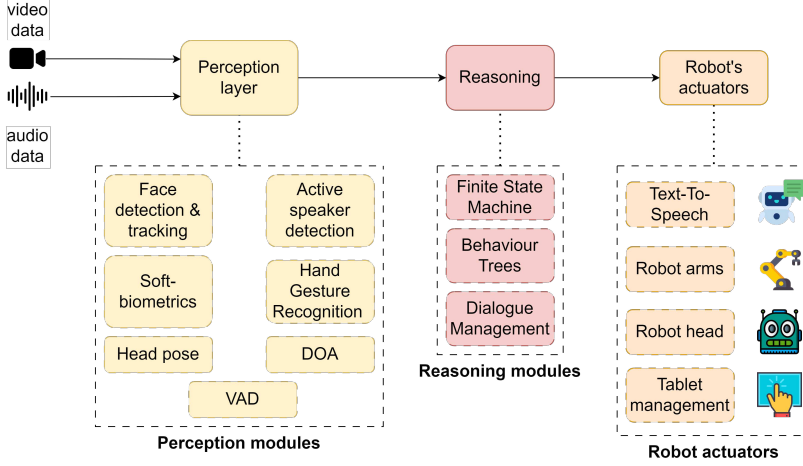


Figure 3.2: High-level architecture of the proposed system. Video and audio data gathered by sensors are processed by the *perception* layer, which includes modules for detecting and tracking users' faces in the scene, estimating head pose for each user, detecting speakers, recognizing hand gestures, estimating the direction of arrival of sound (DOA), and detecting voice activity (VAD). The processed data is then passed to the *reasoning* layer, which comprises a Finite State Machine (FSM) and a Behavior Trees (BTs), together with a dialogue management module. This layer models the system's behavior by considering both perceptual data and interaction history. Finally, the output is transmitted to the robot's actuators to perform actions such as arm movements, head rotations and robot speech.

ing algorithm, we use the well known *SORT* [129] algorithm, since facial movements during conversations are gradual, allowing SORT to reliably track and distinguish face identities over time. Next, for each face image, head pose estimation is performed. Specifically, the proposed system employs the 6DRepNet [130] model to estimate head positions, which has achieved state-of-the-art performance while reducing the number of model parameters and thus the computational burden of the network. In particular, the RepVGG convolutional neural network [131] is used to extract features from face images. Subsequently, the 6DRepNet model uses these features to pre-

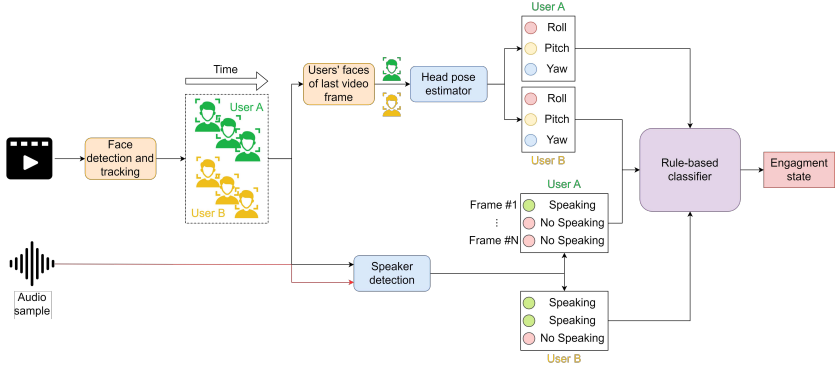


Figure 3.3: Proposed architecture for solving the multi-engagement problem. The system takes both video and audio data as input. Face detection and tracking are applied to all users within the scene, followed by two deep neural networks that perform detection of *active speakers* and head pose estimation. This information is then given as input to a rule-based classifier to determine which user to engage with, effectively managing the conversation between multiple users and robot.

dict a six-parameter rotation matrix, which can be converted into Euler angles (yaw, pitch, and roll).

Anyway, head pose estimation alone may not be sufficient: indeed, users may look at the robot without intent to engage, and multiple users may look at it at the same time. Therefore, it is essential to detect whether a person is speaking. To achieve this goal, the system simultaneously analyzes audio and video and integrates an *active speaker* detection module, using the lightweight architecture proposed by Liao et al. [49]. This method employs a video encoder with (2+1)D convolutional layers for spatial and temporal analysis. In a more formal way, this module processes input tensors with dimensions $T \times H \times W \times C$, where T represents the temporal dimension, H and W denote the height and width of the face image, respectively, and C is the channel axis. It is imperative that the input tensor contains faces belonging to the same user; thus, the previously mentioned tracking algorithm plays a vital role in managing user entry and exit within the scene. Subsequently, the audio data is processed using an audio encoder that employs

(1+1)D convolutional layers applied to the spectrogram-based representation of the audio samples. The resulting embeddings from both audio and video streams are fused and passed to a bidirectional GRU, followed by a Multi-Layer Perceptron (MLP) for the classification task of determining whether the user is speaking. This approach achieves high accuracy while maintaining a reduced network weight, making it suitable for real-time operation on devices with constrained computational resources. Moreover, to handle situations where the individual intending to engage with the robot is beyond the field of view (FOV) of the camera, the system incorporates a direction of arrival (DOA) estimation module. This module enables the user to capture the robot’s attention by vocally addressing it. The system then uses DOA analysis to rotate the robot’s head to the estimated angle. While positioning may not be perfectly accurate due to potential errors in angle estimation, the critical aspect is ensuring that the user enters the field of view of the robot’s camera. This allows the system to use visual data to refine the orientation of the robot’s head and focus its attention more precisely on the user. The estimated head poses, active speaker and DOA data feed a *rule-based* classifier based on behavior tree (detailed in Section 3.3.1), which uses such information to select the user to engage and address the *turn-taking* problem.

3.2.2 Speech Recognition

Given that spoken language is the most intuitive form of communication for humans, we have integrated speech recognition features into our system. First, the audio data is provided as input to a Voice Activity Detection (VAD) module, implemented using the built-in algorithm of the ReSpeaker MicArray V2 ². This module is responsible for detecting and extracting speech frames from the audio sample. By filtering out noise, the module ensures that errors in subsequent components are reduced and allows it to precisely identify when a user initiates a request to the robot. Subsequently, the processed audio data is fed into the Automatic Speech Recognition (ASR) mod-

²https://wiki.seeedstudio.com/ReSpeaker_Mic_Array_v2.0/

ule which is based on Whisper [132] - a multi-language and multi-task network capable of transcribing speech in various languages. This feature provides flexibility and adaptability to different contexts.

3.2.3 Hand gesture recognition

This module is used to detect the user’s hand and generate a skeleton representation using the *mediapipe* [133] framework. Mediapipe is an open-source library developed by Google that provides ready-to-use solutions for real-time processing of multimedia data, such as images and videos. It offers optimized machine learning pipelines for various computer vision applications, including hand skeleton recognition. The user’s gesture is subsequently recognized by enumerating the extended fingers.

3.2.4 Soft-biometrics analysis

In Sections 1 and 2.2, we discussed the importance of estimating soft biometrics to enable empathic and cognitively aware interactions between users and robots. Moreover, based on the previous assessments, such estimations must emulate human capabilities by jointly analyzing video and audio data within a short time frame.

For these reasons, in this section we propose a Multi-Modal Multi-Task architecture, referred to as *MAGNET*, illustrated in Figure 3.4. As shown, it is designed to jointly estimate gender and emotion by leveraging soft parameter sharing across tasks. The framework integrates dedicated audio and visual encoders, specifically devised for each modality. This choice allows to effectively capture the unique characteristics of each modality, so that each encoder specializes in extracting the most relevant features from its respective input domain, thereby enhancing the overall representational capacity.

Furthermore, a concatenation-based fusion strategy is employed to integrate modality-specific features, preserving the distinct representational properties of each modality within the joint embedding space. The resulting multimodal representa-

tion is subsequently processed by convolutional modules that follow a *progressive reduction* strategy, which incrementally reduces the feature dimensionality before reaching the classification head. This approach ensures that essential information from both audio and visual streams is retained, allowing the model to exploit complementary cues without degrading the integrity of the individual modality contributions.

Let's see in more details both the design of encoders (Subsections 3.2.4.1 and 3.2.4.2), the fusion strategy (Subsection 3.2.4.3) and the multi task learning paradigm design (Subsection 3.2.5).

3.2.4.1 Audio encoder

The proposed audio encoder, named MAGNET-A, extracts audio embeddings from a time-frequency representation based on Mel-Frequency Cepstral Coefficients (MFCC). The use of MFCC is motivated by its alignment with the human auditory system, as it organizes pitch information in a way that closely mirrors human perceptual sensitivity to sound. MFCC computation involves segmenting the audio signal into short overlapping frames, each windowed to minimize spectral leakage. The Short-Time Fourier Transform (STFT) is then applied to obtain the frequency representation of each frame. Subsequently, the power spectrum is mapped onto the Mel scale using a filterbank of triangular filters that emulate the human auditory system's sensitivity to frequency. The resulting Mel-spectrogram undergoes a logarithmic transformation to compress the dynamic range and enhance perceptual relevance. Finally, a Discrete Cosine Transform (DCT) is applied to the log-Mel spectrum to yield a compact set of decorrelated coefficients.

The audio encoder thus receives as input an MFCC tensor with shape $B \times C_a (= 1) \times MFCC \times T$, where B denotes the batch size, C_a is a dummy channel dimension that accumulates features after each convolutional operation, $MFCC$ is the number of coefficients, and T represents the temporal dimension. As shown in the Audio Encoder block of the architecture illustrated in Figure 3.4, the input tensor is passed

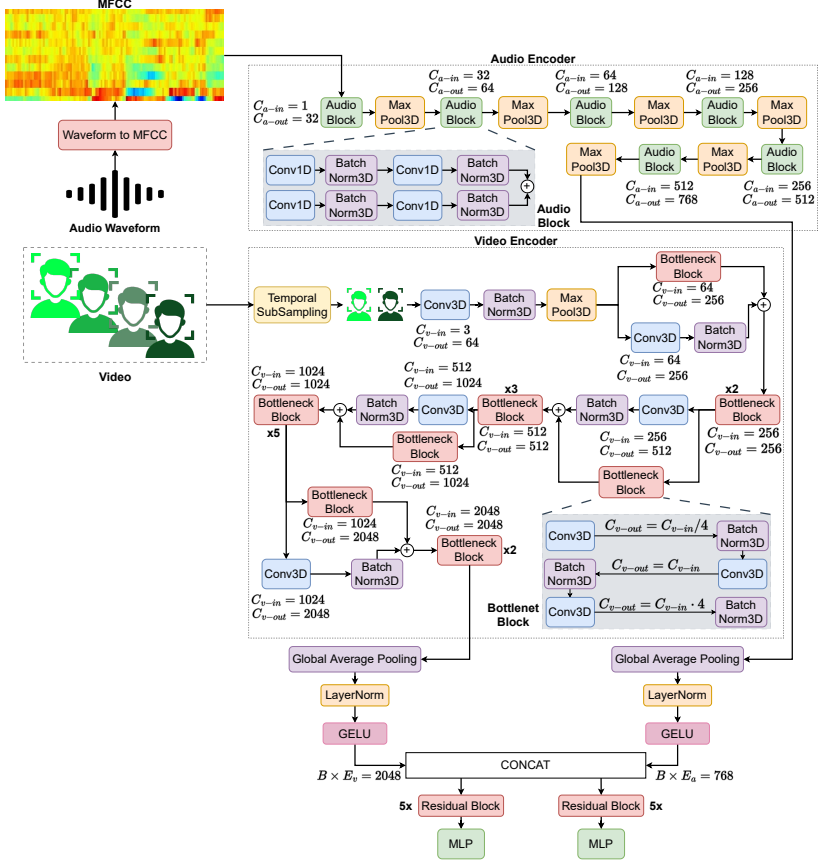


Figure 3.4: Overall architecture of the proposed MAG-NET. The audio encoder processes MFCC features using Conv(1+1)D operations for feature extraction. In parallel, the video encoder receives a video clip as input, applies uniform temporal subsampling along the temporal axis, and extracts features using 3D convolutions and bottleneck layers. For each convolutional block, the number of input and output channels for the video and audio encoders are denoted as C_{v-in} , C_{a-in} and C_{v-out} , C_{a-out} , respectively. The embeddings generated by both encoders are temporally aggregated using pooling operations and then normalized. The resulting feature vectors are concatenated along the embedding dimension and progressively reduced using a series of 1D Residual Blocks. The final representation is passed to the classification head.

through a sequence of stacked *Audio Blocks*, each followed by a *3D Max Pooling* operation to reduce the temporal resolution. Each *Audio Block* comprises sequential Conv(1+1)D operations, applied first along the feature dimension and then along the temporal axis. Normalization layers are inserted between convolutional modules to stabilize training and enhance feature extraction. The final output of the audio encoder is an embedding tensor with shape $B \times E_a$, where $E_a = 768$ denotes the dimensionality of the audio embedding. The design of this module is inspired by the architecture proposed in [49], which demonstrated effective audio feature extraction. Specifically, the authors reported a state of the art 94% accuracy on the task of active speaker detection. Even if effective for that specific task, the original architecture was too simple to address the complexity of the tasks under investigation. Therefore, we decided to increase the number of *Audio Blocks*, which in turn increased the dimensionality of the final embedding. Notably, even if increasing the size of the network with respect to the original version, this architecture exhibits a reduction of approximately 91% in the number of parameters compared to the well known ResNet-based model, thus confirming to be a good tradeoff between accuracy and computational burden required. A detailed, layer-wise specification of the proposed MAGNET-A audio encoder is provided in Table 3.1.

3.2.4.2 Video encoder

One of the most promising architecture in the literature in *Video Action Recognition* task is *SlowFast* [134], which introduces a dual-pathway architecture: a *slow* branch processes input at a low frame rate with higher channel capacity to capture rich semantic features, while a *fast* branch operates at a high frame rate with reduced channel capacity to capture fine-grained motion dynamics. This complementary design enables effective modeling of both static and dynamic aspects of video content.

In the context of emotion and gender recognition, motion plays a critical role, particularly for emotion, which is expressed through dynamic facial movements such as changes

Table 3.1: MAGNET-A architecture. The model takes as input an MFCC representation of the audio, where B denotes the batch size, C_a the number of input channels (initialized to 1), $MFCC$ the number of MFC Coefficients, and T_a the temporal dimension. For each layer, the number of input channels C_{a-in} and output channels C_{a-out} are reported. For pooling operations, kernel size, stride, and padding parameters are specified.

Layer	CNN Channels		Pooling Parameters		
	C_{a-in}	C_{a-out}	Kernel	Stride	Padding
<i>Input: $B \times C_a (= 1) \times MFCC \times T$</i>					
<i>Audio Block 0</i>	1	32			
<i>Max Pooling3D 0</i>			(1,1,3)	(1,1,2)	(0,0,1)
$\left(\begin{array}{c} \text{Audio Block } i \\ \text{Max Pooling3D } i \end{array} \right)_{i=1}^4$	$32 \cdot 2^{(i-1)}$	$32 \cdot 2^i$	(1,1,3)	(1,1,2)	(0,0,1)
<i>Audio Block 5</i>	512	768			

in muscle tension, velocity of expressions, and temporal transitions. However, unlike action recognition tasks that require high temporal resolution to capture rapid and complex movements, emotional expressions typically unfold over longer time scales and involve macro-level dynamics. Therefore, the high temporal sensitivity provided by the fast branch is not strictly necessary. Moreover, gender is a temporally invariant attribute, relying more on static visual features such as facial structure and appearance. This is why we decide to adopt a simplified yet effective variant of the SlowFast architecture by retaining only the slow branch, referred to as MAGNET-V. This choice is motivated by the need to balance accuracy and computational efficiency: in fact, the slow branch alone is sufficient to model the temporal patterns relevant to emotion recognition while also capturing the static features necessary for gender classification. Furthermore, it significantly reduces the number of parameters and computational cost compared to the full *SlowFast* configuration, making it more suitable for real-time applications and resource-constrained environments such as social robotics.

Specifically, video frames are uniformly subsampled along the temporal dimension. The resulting input tensor, with shape

(B, C_v, T_v, H, W) , where B is the batch size, $C_v = 3$ denotes the RGB channels, T_v the number of frames, and H, W the spatial resolution, is fed into the model for feature extraction. The encoder, illustrated in Figure 3.4, consists of a sequence of 3D convolutional layers, each followed by normalization and max pooling operations. At the end, the encoder produces an output embedding of dimension $E_v = 2048$.

3.2.4.3 Fusion layer

As described above, the audio encoder outputs an embedding of shape (B, E_a) , where B denotes the batch size and E_a represents the dimensionality of the audio embedding. Similarly, the video encoder produces an embedding of shape (B, E_v) , where E_v is the dimensionality of the video embedding. Due to the potential disparity in scale and distribution across these two modalities, each embedding is independently normalized before fusion, employing *Layer Normalization*, as follows:

$$y = \frac{x - \mathbb{E}[x]}{\sqrt{\text{Var}[x] + \epsilon}} \cdot \gamma + \beta \quad (3.1)$$

Here, $\mathbb{E}[x]$ and $\text{Var}[x]$ refer to the mean and variance of the input features for each individual sample, γ and β are learnable parameters, and ϵ is a small constant to ensure numerical stability.

Following normalization, the embeddings from both modalities are concatenated along the feature dimension, resulting in a unified representation. This fusion operation is formally defined as:

$$\text{Concat}[(B, E_a), (B, E_v)] = (B, E_a \| E_v) = (B, E_{av}), \quad (3.2)$$

where E_{av} denotes the dimensionality of the concatenated audio-visual embedding, and $\|$ represents the concatenation operator along the feature dimension.

3.2.5 Multi-Task framework

Inspired by the approach proposed in [19] for still image face analytics, we propose to use a soft-sharing multi task configura-

tion, employing task-specific convolutional blocks to enable the extraction of features that are more relevant and specialized for each individual task. In more details, each branch integrates a *ResidualBlock* stacked $N(= 6)$ times, as originally used in ResNet architectures. The residual blocks progressively reduce [135] the embedding dimension E_{av} while concurrently enriching the feature representation along a new axis E'_{av} , resulting in an intermediate tensor of shape (B, E'_{av}, E_{av}) . Each residual block reduces the E_{av} dimensionality by a factor of two through strided convolutions, while progressively increasing the feature representation along the E'_{av} axis. This process accumulates up to 768 channels, resulting in a compact output tensor of shape $(B, 768)$. Unlike traditional versions that use 2D convolutions,

Table 3.3: Design of the task-specific branches. For each layer, the transformation of features is reported, specifically illustrating how the input embedding size E_{av} is progressively reduced and how the task-specific representation E'_{av} is subsequently expanded.

Layer	E'_{av}	E_{av}
<i>Input: $B \times E'_{av} \times E_{av}$</i>	1	2816
<i>ResidualBlock 0</i>	32	1024
<i>ResidualBlock 1</i>	64	512
<i>ResidualBlock 2</i>	128	256
<i>ResidualBlock 3</i>	256	128
<i>ResidualBlock 4</i>	512	64
<i>ResidualBlock 5</i>	768	32
<i>Average Pooling</i>	768	1
<i>MLP 5</i>	512	
<i>MLP 6</i>	256	
<i>MLP 7</i>	128	

this implementation employs *1D convolutions* with a kernel size of 5. Two separate convolutions are used with strides of 2 and 1, respectively. The downsampling effect is achieved through the stride-2 convolution. Furthermore, the *GELU* activation function is employed in place of *ReLU*, as it has been shown to improve generalization and provide smoother gradient propagation during training. After the pooling stage, the resulting

tensor of shape $(B, 768)$ is passed to a dedicated MLP for each task. The design of the task-specific branches is detailed in Table 3.2, while a graphical illustration of the architecture is provided in Figure 3.4.

As discussed in Section 2.2, the overall loss function $\mathcal{L}$, used to train the model in the multi-task setting, is defined as the weighted sum of the individual task losses $\mathcal{L}_i$, as shown in Eq. 3.3:

$$\mathcal{L} = \sum_{i=1}^n \lambda_i \mathcal{L}_i \quad (3.3)$$

where λ_i represents the weight associated with the i^{th} task.

In our case, we consider two classification tasks ($n = 2$): gender and emotion recognition. Therefore, we adopt the Categorical Cross-Entropy loss function for both tasks, as defined in Eq. 3.4:

$$\mathcal{L}_i = -\frac{1}{n} \sum_{j=1}^n y_{i,j} \cdot \log(\tilde{y}_{i,j}) \quad (3.4)$$

where $y_{i,j}$ denotes the ground truth for the j^{th} class in task i , and $\tilde{y}_{i,j}$ represents the corresponding model prediction.

This raises a critical question: *How should the task-specific loss weights λ_i be set to ensure effective training in a multitask setting?* One widely adopted approach in the literature is the use of the *GradNorm* algorithm [63], which dynamically adjusts task weights during training to balance gradient magnitudes.

Formally, let $G_i = \|\nabla_{\theta}(\lambda_i \mathcal{L}_i)\|_2$ denote the gradient norm of task i with respect to the shared parameters θ . GradNorm aligns G_i with a target gradient norm $\hat{G}_i$, computed as:

$$\hat{G}_i(t) = \bar{G}(t) \left(\frac{r_i(t)}{\bar{r}(t)} \right)^{\alpha}$$

where:

- $r_i(t) = \frac{\mathcal{L}_i(t)}{\mathcal{L}_i(0)}$ is the relative inverse training rate of task i ,
- $\bar{G}(t)$ is the average gradient norm across all tasks at step t ,
- $\bar{r}(t)$ is the average relative training rate at step t ,

- α is a tunable hyperparameter controlling the strength of rebalancing.

The GradNorm loss, which is used to update the task weights λ_i , is defined as:

$$\mathcal{L}_{\text{GradNorm}} = \sum_{i=1}^n \left| G_i(t) - \hat{G}_i(t) \right|$$

By minimizing $\mathcal{L}_{\text{GradNorm}}$, the algorithm adaptively tunes the task weights λ_i to ensure that all tasks progress at comparable rates, thereby mitigating the risk of one task dominating the training process. In our implementation, we set the GradNorm hyperparameter $\alpha = 1.5$ to account for differences in task difficulty and learning dynamics.

3.3 Reasoning modules

In this section, we will explore how the high level features extracted by perception modules are processed by *controllers*, with the aim to select the action to perform based on the user requests, interaction history and environmental data. The proposed system can exhibit two main interaction modalities - *Conversational* and *Game*: in the former, the system is configured to engage in a conversation with users, responding to their questions; in the latter, the system is set up to play an interactive game (*rock, paper, scissors*) with the user.

A state of the art solution to model inanimate agents and solve behavioral modeling tasks is the employing of Finite State Machines (FSMs); even if very simple and widely used, their use in handling very complex tasks, such as social robots' engagement, can lead to an explosion in the number state, resulting in a loss of maintainability and scalability [136]. This is why *Behavior Trees* (BTs) [136] emerged, and have been used in several robotics applications [137, 138, 33]. BTs are a formal and structured way to model the behavior of autonomous agents and robots. They consist of nodes representing different behaviors or actions, connected in a tree structure. The root node is the starting point and the leaf nodes represent individual actions or tasks. BTs help organize and control task

execution, making complex behaviors easier to design, understand and maintain. These characteristics make them suitable for modeling the behaviors and tasks of social robots. However, incorporating various unrelated behaviors, such as the robot's behavior when playing with the user versus when it is holding a conversation, could significantly increase the number of states in the tree, thus increasing its depth and reducing modularity and human readability. Therefore, it is preferable to separate different behaviors into distinct and independent behavioral manager modules, preserving the modularity of the system, and in turn improving maintainability and debugging efficiency.

For these reasons, in this thesis we propose a hierarchical architecture that combines both a FSM and BTs. At a high level, the transition between the two different modalities is handled by means of a properly devised Finite State Machine (FSM). Vice-versa, the specific engagements in each modality (conversational and game, respectively) are modeled using the *Behavior Trees* (BTs).

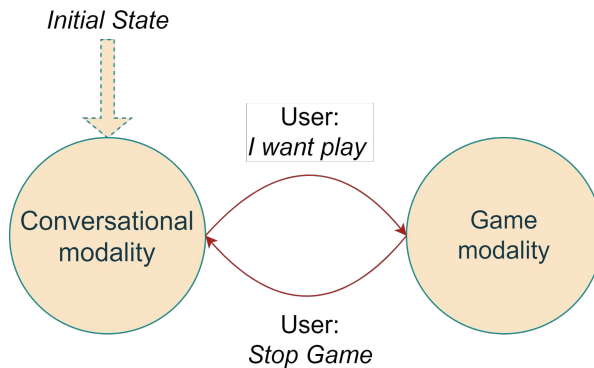


Figure 3.5: Finite State Machine (FSM): the high-level controller designed to manage the system's modalities. Each state within the FSM is associated with a specific Behavior Tree (BT). The conversation modality serves as the initial state of the FSM. Transitions between modalities can be triggered by user requests, which can be performed either through voice commands or by pressing the physical robot buttons (bumpers).

The FSM controls the macro-state (Figure 3.5), specifically managing the robot’s *modality* that the system is able to implement, such as conversational or game mode. Each FSM state is associated with a BT that controls, in detail, the behavior and interaction flow of that specific modality. In this way, it is possible to easily add additional behaviors such as a new modality, by just adding a new state to FSM and implementing the corresponding BT and transitions. Moreover, this solution can assist in managing hardware resources. By switching from one state to another, it is possible to deallocate some resources, such as variables or threads, that are no longer needed for the current system state. This enables better resource management, particularly beneficial for systems with low computational capabilities.

Another important component of the *reasoning layer* is the *dialogue management* module. The primary goal of this module is to classify the user’s intent using Natural Language Processing (NLP) algorithms and to generate appropriate responses. Furthermore, it keeps track of user requests and system responses to maintain conversation history. Additionally, this module integrates data from the *Perception layer*, such as soft biometrics, to facilitate empathic interactions. It also leverages video data to generate more sophisticated responses using Visual Question Answering (VQA) algorithms.

In light of these considerations, our system starts in the *conversational* modality, which is set as the initial state, and switches between modalities based on user requests. Transitions may be issued either via voice commands or by pressing the physical robot buttons (bumpers). The two BTs are explained in detail in Section 3.3.1 for the *Conversational* modality and in Section 3.3.2 for the *Game* modality. For the dialogue management module, we propose two versions: the first adopts a rule-based approach, described in detail in Section 3.3.3, while the second employs an end-to-end strategy based on Visual Language Models, presented in Section 3.3.4.

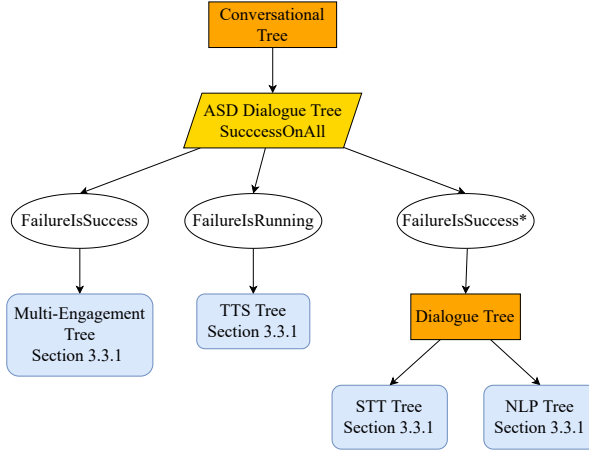


Figure 3.6: Behavior Tree for conversational modality. This BT models the robot’s behavior during interactions with the user. It specifically implements multi-engagement strategies and formulates responses to user requests.

3.3.1 Conversational Modality

In Figure 3.6 the overall architecture of the proposed conversational BT is shown. In general, during the conversational modality, the BT implements the engagement algorithm (*Multi-Engagement Tree*), processes the user’s speech requests, and generates appropriate responses (*Dialogue Tree*), which are communicated to the user using the Text-to-Speech modules (*TTS Tree*). Furthermore, given the complexity of the proposed controller, a detailed analysis will be conducted by examining each subtree, categorized according to its specific task.

Multi-Engagement Tree The multi-engagement algorithm, depicted in Figure 3.7, begins by gathering data from the face detector, tracker, head pose estimator, and active speaker detection modules (Section 3.2.1). Thus, for each visible face inside the scene it is possible to know the track of that face (by means of the tracking algorithm), the head orientation in degrees (by the head pose estimator) and if the user is speaking (by the active speaker detection). Once such data are obtained, the first branch (*Disengagement Selector*) is executed. The

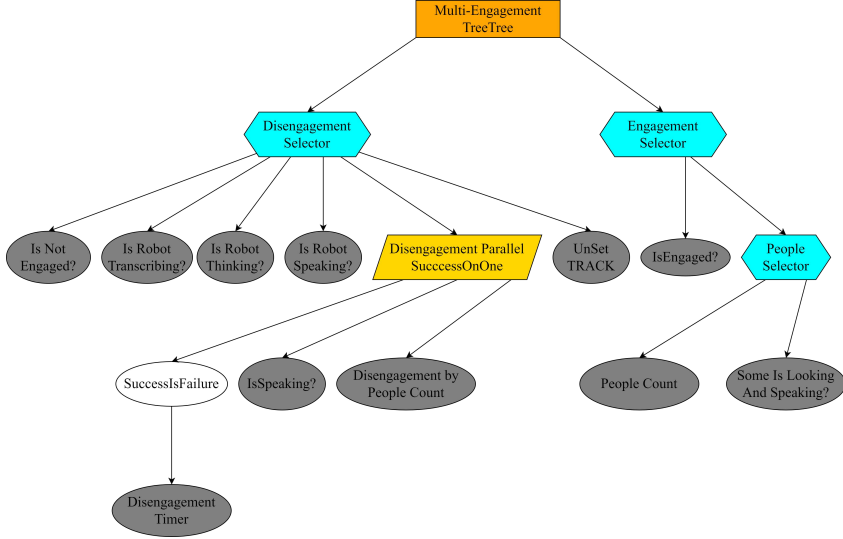


Figure 3.7: Multi-Engagement Tree. This BT manages the system’s ability to handle multiple engagements. Specifically, it uses data obtained from the perceptive layer to select the appropriate user to engage in the scene. Additionally, it determines whether to disengage the currently engaged user, taking into account the context and the state of the conversation.

purpose of this branch is to determine whether a user within the scene is engaged and decide whether to maintain the engagement. Initially, the system checks if no user is engaged (*Is Not Engaged?* node). Subsequently, if someone is engaged, the algorithm verifies if the robot is interacting with the user by evaluating whether the robot is transcribing speech into text (*Is Robot Transcribing?* node), processing the user’s request (*Is Robot Thinking?* node), or speaking (*Is Robot Speaking?* node). In such scenarios, no disengagement operations are performed. If these nodes return *Failure*, it indicates that a user is engaged and the robot is not performing operations, prompting a timer to start (*Disengagement Timer* node). This timer measures the duration during which the engaged user is not speaking (*IsSpeaking?* node) with the robot. If the engaged user resumes speaking before the timeout, the *IsSpeaking?* node, and consequently the *Disengagement Parallel* node, return *Success*, preventing any further actions from being ex-

ecuted. Conversely, if a timeout occurs or if the engaged user leaves the scene (*Disengagement By People Count* node), the *Disengagement Parallel* node returns *Failure*. Consequently, the BT removes the track of the engaged user from the blackboard (*UnSetTRACK* node) and subsequently returns *Success*. In this context, it is useful to recall the concept of the *blackboard*: a blackboard refers to a data structure where variables are stored during interaction. Since these variables are shared among all nodes in the tree, each node can write to and read from the blackboard, utilizing such information to make decisions.

After this, the *Engagement Selector* branch is executed. The purpose of this branch is to determine whether any user is already engaged (*IsEngaged?* node). If no user is engaged, the system first evaluates the number of users present in the scene. If there is only one user, it is engaged by default (*People Count* node). In a multi-user scenario, the algorithm checks if any individual within the scene is both looking at and speaking to the robot (*Some Is Looking And Speaking?* node). If such a user is identified, their ID, as provided by the face tracker, is stored on the blackboard as the engaged user. Subsequently, the user's track and associated data can be utilized by additional modules to focus the robot's attention on the selected engaged user in the scene. For instance, the robot can use bounding box data to orient its head toward the engaged user and initiate a conversation with them.

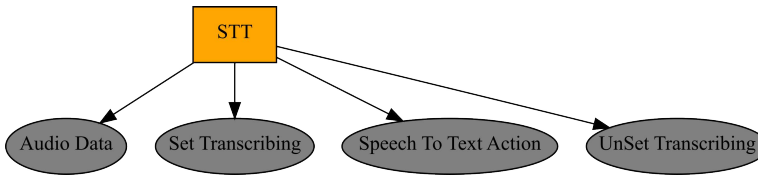


Figure 3.8: Speech To Text Tree. This Behavior Tree is designed to await a speech request from the user and subsequently invokes the necessary services to transcribe the input audio into text.

Speech To Text Tree This subtree, illustrated in Figure 3.8, is responsible for executing the Speech-to-Text (STT) task. The *Audio Data* node retrieves the audio sample filtered by the VAD module, the state variable is set on the blackboard (*Set Transcribing* node), and calls the ASR service (*Speech To Text Action* node), which takes the utterance as input and transcribes it. The output is then stored on the blackboard. At the end, the state variable is unset on the blackboard (*Unset Transcribing* node).

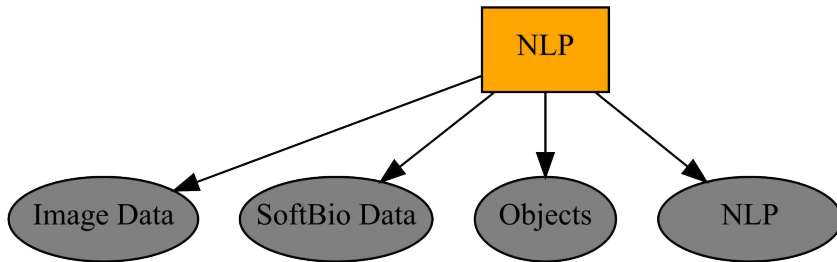


Figure 3.9: Natural Language Processing Tree: This BT gathers contextual data from the scene, including the image of the environment, soft biometrics of users, and identified objects, along with the transcribed user request. It then calls the dialogue server, passing this information, and subsequently manages the server’s response.

Natural Language Processing Tree The NLP sub-tree is responsible for invoking the NLP service to process the text transcribed by the ASR and recognize user requests. Specifically, it collects the image of the current scene (*Image Data* node), the soft biometric information of the users present (*Soft-Bio Data* node), the objects in the environment (*Objects* node), and the transcribed text (*NLP* node). All this information is used by the *NLP* node to call the *Dialogue Management* server. The dialogue management process is described in detail in Section 3.3.3. Finally, the server’s response is recorded by the node into the blackboard for further processing.

Text To Speech Tree This subtree is responsible for the Text to Speech task. If a RESPONSE exists (*Is There Re-*

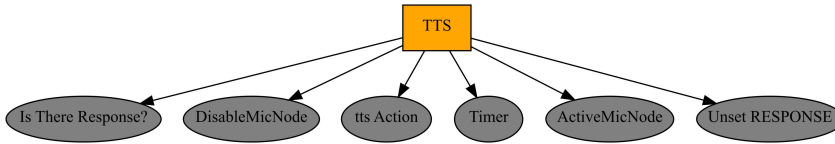


Figure 3.10: Text to Speech Tree: This Behavior Tree receives textual input and invokes the TTS service to convert the text into audio, which is then played through the robotic platform.

sponse? node), the microphone is disabled (*DisableMicNode* leaf); this action, along with the subsequent reactivation of the microphone, is necessary to prevent the robot from hearing itself. This ensures that the system does not listen to its own output. The TTS server is then called (*tts Action* node) to play the sent text using audio output. The Behavior Tree will remain in the RUNNING state until the robot finishes speaking. Finally, the microphone is reactivated (*ActiveMicNode* leaf), and the RESPONSE is unset (*Unset RESPONSE* node).

Asynchronous events

A Behavior Tree (BT) determines the next decision based on the current state of the system. Specifically, a *control node* can activate only its child nodes and not other parts of the tree. However, in practical applications, certain events may occur asynchronously and independently of the system's state, such as a button press on the robot. These events must be managed appropriately to prevent the system from transitioning into an incoherent state. We refer to these category of events as *asynchronous events*. An asynchronous event can be either a hardware or software event that can be triggered at any moment during the system's execution and interaction. Asynchronous events can arise from user interactions with the robot's sensors, such as buttons, or from the system's modules themselves under specific conditions. Examples of interactions include when a user touches the robot's bumper to asynchronous event an ongoing action, such as suspending the interaction, or when a user verbally indicates a desire to start a game. Each time

an asynchronous event occurs, it is added to a queue. Subsequently, each asynchronous event in the queue is handled during the *post-tick* of the tree, which occurs between the end of the current tree’s tick and the beginning of the new one. The *tick* in a Behavior Tree is a periodic signal that traverses the tree, activating nodes and updating their states (*Success*, *Failure*, or *Running*). It is used to dynamically evaluate and manage the agent’s behavior based on conditions and priorities.

3.3.2 Game Modality

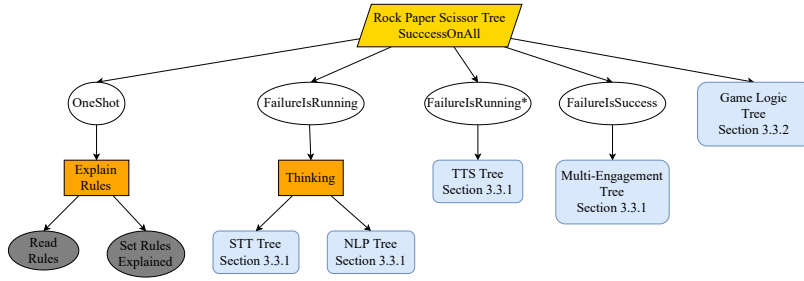


Figure 3.11: **Game Modality Tree:** This Behavior Tree (BT) models the system’s behavior during the ”rock, paper, scissors” game with the user. During the initialization step, the robot explains the game rules to the user. The Speech to Text, Text to Speech, and multi-engagement tasks are handled as previously discussed in the reported section. The ”Game Logic” subtree manages the gameplay by detecting the user’s hand gesture, determining the winner, and communicating the result.

Given the advantages of incorporating interactive games with social robots [13, 20, 11], we integrated a ”rock, paper, scissors” game [139] into our system. Specifically, this game mode operates alongside the conversational mode and is reachable when the corresponding node of the FSM (Figure 3.5) is activated. A Behavior Tree (BT) has been designed to coordinate user-robot interactions throughout the game. The BT structure is illustrated in Figure 3.11. Initially, the robot enters into the initialization phase, where it communicates the game rules through the *Explain Rules* subtree. In this phase, the

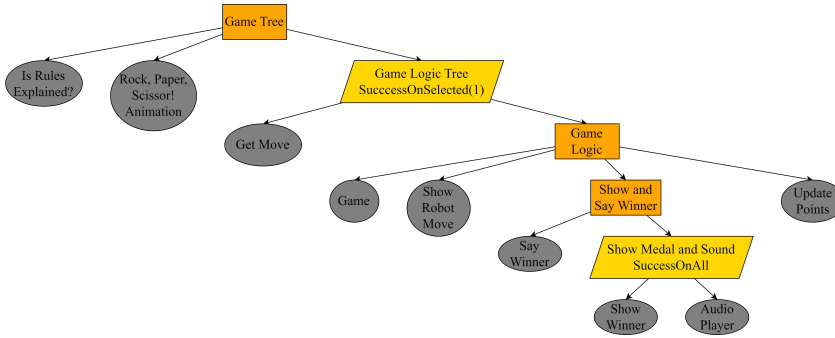


Figure 3.12: Behavior Tree (BT) managing the gameplay. Initially, it verifies whether the rules have been explained, then starts the *rock, paper, scissors* animation by moving the robot’s arms. Subsequently, it recognizes the user’s hand gesture and randomly selects the robot’s move. Finally, the winner is determined and communicated to the user using both visual cues (icons displayed on the tablet) and auditory feedback (speech and specific sounds).

robot’s speech is managed by the *TTS* subtree, as described in Section 3.3.1. Throughout the gameplay, the system continuously monitors the user’s face using the *Multi-Engagement* subtree. During the interaction, the user can communicate with the robot, either by pressing the robot’s bumpers or verbally notifying their intent to restart the game or terminate the match, transitioning back to the conversational modality. In such cases, the user’s intent is processed through the previously discussed *STT* and *NLP* subtrees (Section 3.3.1).

Once the initialization phase is complete, the game begins. The gameplay is controlled by the *Game Tree* subtree, depicted in Figure 3.12. Initially, the algorithm verifies whether the rules have been explained (*Is Rules Explained?* node). If this condition is satisfied, the robot engages with the user by initiating the game animation, simulating the typical game gestures by moving its left arm in a circular motion (*“Rock, Paper, Scissors! Animation”* node). This sequence is repeated at the end of each match. Subsequently, the algorithm processes the user’s move (*Get Move* node), randomly selects the robot’s move, and determines the winner (*Game* node). Using data

from the perception layer, particularly the *gesture recognition* module (see Section 3.2.3), the system identifies the user’s chosen option: rock, paper, or scissors (an example for each move in Figure 3.13). Upon determining the winner, the robot communicates the result to the user through the *Show and Say Winner* subtree. The interaction with the player is further facilitated through a custom graphical interface displayed on the robot’s tablet. This interface shows each player’s (robot and user) score and two interactive panels (see Figure 3.14). The video feed captured by the robot is displayed on the right side of the interface, while the left side shows the game action selected by the system (rock, paper, or scissors). Based on the result (win, loss, or tie), it announces the outcome (*Show Medal and Sound* subtree) through speech (*Say Winner** node), distinct sounds (*Audio Player* node), and icons displayed on the tablet, providing a rewarding feedback mechanism [140] for the player.

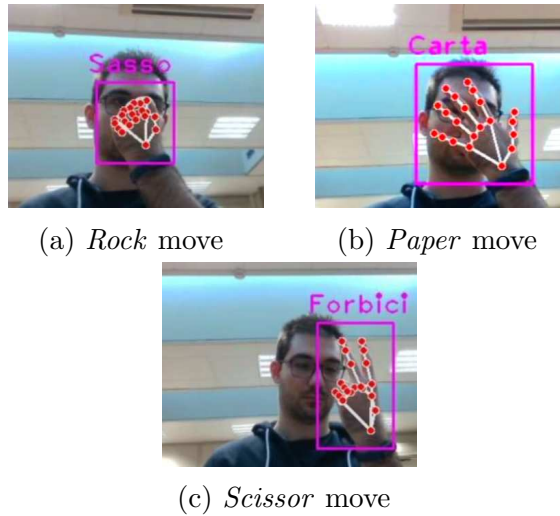


Figure 3.13: List of the allowed game moves: *rock*, *paper*, *scissor*.

3.3.3 Rule-based Dialogue Management

For a social robot, engaging in conversation with users is fundamental to human-robot interaction; thus, it is essential to man-



Figure 3.14: Tablet view during the game session. The right side displays the video feed captured by the camera, along with the detected game move chosen by the user (in this specific case it is paper). On the left side, the option selected by the proposed system is shown, which is scissors in this case.

age the dialogue effectively to achieve high interaction performance. To address these needs, the proposed system integrates a dialogue server built using the widely adopted *RASA* [141] framework. *RASA* provides robust NLP capabilities, enabling the analysis of input text through a hybrid approach that combines advanced transformer based techniques with traditional methods, such as *regex*-based analysis, for effective intent classification. Furthermore, the framework facilitates conversation tracking and decision-making by considering both the analyzed input and the evolving state of the dialogue. We have thus defined three macro-intent that can be used to classify and understand user requests.

- **Soft-biometrics intent:** The user might inquire about personal attributes such as gender, emotion, or age. Example requests include: "What is my gender?" or "What is my age?" This is useful for demonstrating the robot's perceptive capabilities, establishing an initial connection with the user. Additionally, this information can be used by the dialogue server itself to implement cognitive and empathetic responses. For instance, the system might

adapt its behavior and dialogue based on whether the user is a young woman or an older man.

- **General Purpose intent:** Users may ask general questions, such as "What is the capital of Italy?" or "How can I cook a pizza?" Given the open-ended nature of such inquiries, we employ a Large Language Model [142] to generate an appropriate response. Specifically, we use the widely adopted ChatGPT [97, 143], connected via network APIs. Furthermore, considering advancements in the field of *Visual Question Answering* [144] (VQA), we divide this macro-intent into two sub-intents: VQA-based and text-only intents. Once a user request is classified as a *general purpose* intent, it is further analyzed to determine if scene analysis is required. For instance, the user might ask, "Can you describe the scene?" requiring the system to capture the current scene's image and send both the image and user request to the VQA system. Alternatively, if the inquiry is purely textual, such as "What is the typical dish in Naples?", a text-only response suffices, bypassing scene analysis to improve computational efficiency and response time. For both VQA and text-only responses, we use the *gpt4o* model.
- **Game intent:** This intent is used to detect if the user wishes to change the robot's mode and initiate an interactive game, such as by saying, "Let's play." When this intent is recognized, the dialogue manager raises an *asynchronous event* (see Section 3.3.1), signaling the FSM (see Section 3.3) to switch to a new state.
- **DOA intent:** This intent is used to detect users outside the FOV of the robot who want to interact with it, as described in Section 3.2.1. In this context, the user could call the robot using specific keywords or sentences (for example, "Look at me"). When the dialogue server detects this intent, an *asynchronous event* is raised. During the handling of this, a temporary behavior tree is executed. This tree uses the angle estimated by the DOA module (Section 3.2) to rotate the robot's head towards the user;

thus, using the TTS tree, the user is invited to interact. At the end of this interaction, the temporary behavior tree is terminated, and the control returns to the main behavior tree controller.

The call to the dialogue server is performed by the *NLP* tree (Figure 3.9) when a transcription is ready. Subsequently, the text response from the server is set on the blackboard, and the *TTS* tree (Figure 3.10) will use this response to engage in conversation with the user.

3.3.4 End-to-End Dialogue Management

In light of the considerations discussed in Section 2.3, and recognizing the importance of enabling proactive behaviors in social robots, in this section we propose a Vision-Language Model (VLM) specifically tailored for applications in social robotics. Indeed, VLMs are particularly well-suited for supporting the implementation of proactive behaviors informed by the robot’s perceived context, due to their inherent ability to integrate and interpret multimodal data. However, several challenges must be addressed: (i) these models require task-specific fine-tuning to generate natural and fluent human-like sentences that are appropriate for the intended application; and (ii) they must be optimized for deployment on resource-constrained devices, requiring a careful balance between performance and computational efficiency to ensure acceptable inference times. Regarding (i), it is worth to highlight that LLMs typically require substantial amounts of training data, but generating labeled data for a specific task, such as proactive human-robot interaction, is time-consuming and labor-intensive. To mitigate this, a possible solution is to use *Knowledge Distillation*, so that a pre-trained model automatically labels data for the target task, significantly reducing the manual annotation effort. However, for the specific task at hand, namely generating context-aware sentences from video input to engage users, there are currently no pre-trained VLMs designed for this purpose [119]. Vice-versa, LLMs are more suitable for producing human-like text due to their strong generalization capabilities, but they do not

natively support video input processing. To exploit the pros of both VLMs and LLMs, overcoming their cons, we propose a hybrid approach that leverages the strengths of both modalities through a hierarchical dual-teacher data labeling pipeline. The first teacher, a VLM, processes the video input and generates a set of contextual scene descriptions by configuring the *top-k* and *temperature* parameters [145] (the prompt used is reported in Listing 3.1). These descriptions are subsequently used to feed to a second teacher, namely a LLM, which receives a task-specific prompt along with the scene descriptions and generates contextually appropriate sentences using the *Diverse Beam Search* algorithm [146]. This strategy is employed to promote diversity across the generated sentences, producing outputs that are semantically or structurally distinct. By appropriately tuning its parameters, it is possible to generate multiple sentence variations conditioned on the same video context, thereby enhancing the data augmentation process. The prompt used for generation is provided in Listing 3.2. The generation of multiple descriptions and corresponding context-aware sentences serves as a form of data augmentation, mitigating the risk of overfitting and enhancing the diversity of the training samples. This pipeline enables the automatic creation of high-quality, task-specific training data. An overview of the proposed pipeline is shown in Figure 3.15.

Listing 3.1: Prompt for VLM teacher

```
Describe the video in detail. Include the
actions of the people, their emotions,
and any other relevant information. Use
a formal tone and avoid using emojis or
informal language. The description
should be suitable for a professional
setting.
```

Listing 3.2: Prompt for LLM teacher

```
You are a social robot that has to assist
and play with people.
```

```
1.Generate a greeting sentence in english:
```

- a. take into account the contest that i will provide you
 - b. be polite and kind
 - c. make compliments
 - d. write sentence in the format tts::"sentence"
 - e. Don't say good morning, good afternoon or good evening.
- Give me the responses directly, I don't want to know you reasoning
-

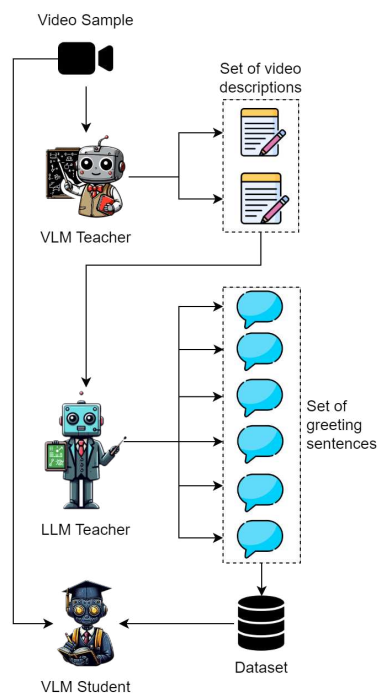


Figure 3.15: Label generation pipeline. A video sample from the dataset is provided as input to the VLM teacher to generate contextual captions. These captions are then passed to the LLM teacher, which produces an augmented set of context-aware sentences. The resulting dataset consists of video samples paired with multiple associated sentences.

After defining the dataset generation strategy, the subsequent step involved selecting suitable pre-trained models to serve as

teachers. For the VLM teacher, we identified high-performing models for video captioning based on recent literature. The selection process was informed by a comprehensive analysis of performance metrics and memory consumption. In addition to quantitative evaluation, a qualitative analysis was conducted by a group of domain experts. In this context, we selected the *LLaVA-OneVision* model [110], which has demonstrated competitive performance on both image and video benchmarks. Specifically, the LLaVA-OV-7B variant, where 7B denotes a model with 7 billion parameters, achieved a zero-shot accuracy of 88.9% on the Image Detail Description (ImageDC) dataset [147], and an accuracy of 89.5% on NLVR2, a benchmark for natural language visual reasoning [148]. These results surpass those of comparable models, such as VPG-C-7B [149], which achieved an accuracy of 73.2% on the same NLVR2 task.

For the LLM teacher, we considered state-of-the-art open-weight models. Due to the absence of benchmarks specifically tailored to the target task, an initial screening was performed using general-purpose evaluation metrics. Subsequently, a qualitative assessment was conducted via the LM-Arena³ platform, which allows for direct comparison of model outputs given identical prompts. This two-step evaluation strategy facilitated the selection of a model capable of generating contextually relevant and fluent sentences. The candidate models included *DeepSeek-V3* [150], *Qwen2-72B* [151], and *Mistral-24B* [152]. According to LM-Arena scores, *DeepSeek-V3* obtained the highest score (1396), followed by *Mistral-24B* (1268) and *Qwen2-72B* (1258). *Qwen2-72B* was excluded from further consideration due to its lower score and substantially higher parameter count compared to *Mistral-24B*. Although *DeepSeek-V3* achieved the highest LM-Arena score, it also has the largest model size (671B parameters). Furthermore, our qualitative evaluation did not reveal substantial differences in the quality of generated sentences between *DeepSeek-V3* and *Mistral-24B*. Therefore, we selected *Mistral-24B* as the LLM teacher, as it provides an optimal trade-off between performance and computational efficiency. Nevertheless, it is important to emphasize

³<https://lmarena.ai/>

that the proposed *Knowledge Distillation* protocol is model-agnostic and can be flexibly adapted by selecting alternative VLM and LLM architectures according to the specific requirements of the target application.

Implementation Details

Considering the analysis presented in Section 3.3.4 and the advantages highlighted in [110, 111], the LLaVA-OneVision-7B model was selected as the Vision-Language Model (VLM) teacher, while Mistral-24B [152] was adopted as the Large Language Model (LLM) teacher. To serve as student model, a lightweight version of LLaVA-OneVision with a reduced parameter count of 0.5B was employed. This decision is supported by the strong performance reported in [110], particularly due to its token compression mechanism, which enables substantial reductions in inference time and memory consumption while maintaining output quality. Empirically, the selected student model achieves a memory usage of 1964.80 MB and generates 48 tokens in approximately 3 seconds. In contrast, the teacher model requires 94 seconds and 15319.23 MB to generate the same number of tokens. A detailed analysis of the deployment on embedded hardware is provided in Section 4.3.3.

Data Processing. The initial dataset consists of video recordings from the MuMMER [153] and AVDIAR [154] corpora. These datasets were merged into a unified database and subsequently augmented using the aforementioned *Knowledge Distillation* protocol. Specifically, the MuMMER dataset [153] is intended to support research in natural human-robot interaction (HRI). It contains approximately 1 hour and 29 minutes of recorded interaction sessions between humans and a humanoid robot deployed in real-world public environments, such as shopping malls. The dataset includes multimodal data streams such as audio and video capturing spontaneous, task-oriented, and socially driven behaviors. Owing to its ecological validity and multimodal richness, MuMMER is particularly well-suited for investigating perceptual processing, conversational modeling, and interaction strategies in social robotics. The

AVDIAR dataset [154] (Audio-Visual Diarization Dataset) was created to facilitate research in multimodal person diarization and tracking. It comprises approximately 23 sessions, totaling around 27 minutes of real-world interactions involving two or more participants in indoor settings. The recordings feature synchronized high-resolution RGB video and multi-channel audio streams, capturing natural conversations with overlapping speech, head movements, and spontaneous gestures. AVDIAR is particularly well-suited for developing and evaluating models in speaker localization, audio-visual alignment, and social signal processing, making it valuable for human-robot interaction (HRI) applications that require robust multimodal understanding. The MuMMER and AVDIAR datasets were merged into a unified corpus and processed according to the *Knowledge Distillation* framework described in this Section 3.3.4. The resulting dataset was first partitioned into training, validation, and test sets, with 88% of the samples allocated for training and 6% each for validation and testing. Care was taken to ensure that each video appeared exclusively in one subset to avoid data leakage. Given that social robots typically operate over fixed temporal windows, all videos were segmented into 4-second clips. Each clip was then used to generate three descriptive captions, and for each caption, three context-aware sentences were produced. As a result, every 4-second clip is associated with a total of nine sentence-level annotations. The final dataset comprises 3,762 samples for training and 261 samples each for validation and testing. Prior to training, each video clip was preprocessed using *Uniform Temporal Sampling* to standardize the number of frames per clip, denoted as T . To assess the influence of temporal context on model performance, three model variants were trained with different values of T , specifically $T \in \{1, 4, 8, 16\}$. Finally, the dataset was formatted to comply with the input requirements of the LLaVA-OneVision framework [110], ensuring seamless integration into the training pipeline.

Training. The training procedure was conducted using the HuggingFace framework⁴, starting from a pretrained version of LLaVA-OneVision. Training was performed with mixed precision [155] for a total of 1,500 steps, using a batch size of 4. A cosine learning rate schedule was adopted, with an initial learning rate of 2×10^{-5} decaying to a final value of 6.5×10^{-8} . The AdamW optimizer was used for optimization, and the training objective was defined by the *Cross Entropy Loss* function.

In this section, we introduced a hardware-agnostic architecture built upon the ROS framework, designed to enhance human-robot interaction by addressing two key challenges: the Multi-Engagement Problem, which includes both the estimation of the user’s engagement state (Section 3.2.1) and turn-taking management (Section 3.3). Moreover, we addressed the problem of soft-biometrics estimation by employing a Multi-Modal Multi-Task Learning approach named MAGNET (see Section 3.2.4). Finally, we presented how this information is utilized by a rule-based dialogue module (Section 3.3.3) to implement empathic dialogues. Subsequently, an end-to-end dialogue module based on Visual Language Models is introduced in Section 3.3.4, designed to generate context-aware sentences that enhance user engagement.

Conclusions In the next chapter, we begin presenting the results of the proposed methods. Specifically, Section 4.2 reports the performance of the soft-biometrics analysis network, MAGNET, evaluated on the CREMA-D dataset. Furthermore, Section 3.3.4 introduces a method based on Vision-Language Models (VLMs), aimed at further enhancing interaction and engagement between users and robots. This approach enables robots to exhibit proactive behaviors by analyzing video data and generating contextually appropriate sentences. Finally, in Section 4.1, we present a comprehensive evaluation of the overall system on a robotic platform. First, quantitative results on user engagement prediction are provided in Section 4.1.1, using both a state-of-the-art dataset [156] and a custom dataset.

⁴<https://huggingface.co/>

Then, in Section 4.1.2, we assess the impact on users through 5-point Likert-scale surveys conducted in real-world scenarios.

Chapter 4

Results

4.1 Quantitative and Qualitative Evaluation of Engagement and Interaction Capabilities

In Section 3.1, we presented a hardware-agnostic architecture designed to address the Multi-Engagement problem, with a specific focus on estimating users' engagement states and managing *turn-taking*. Additionally, we integrated entertainment capabilities by implementing the *Rock, Paper, Scissors* game.

In this section, we evaluate the engagement module as well as the conversational and entertainment capabilities of the proposed social robotic framework through a combination of quantitative (Section 4.1.1) and qualitative (Section 4.1.2) analyses.

For the quantitative evaluation, we employ widely adopted metrics such as *precision*, *recall*, and *F1-score*. The qualitative evaluation focuses on assessing the quality of user experience in terms of system perception and interaction capabilities. To this end, participants completed structured surveys using a 5-point Likert scale to measure various aspects of user satisfaction and engagement. Furthermore, a statistical analysis of the survey results was conducted to better understand user trends and experiences during interaction with the robot.

4.1.1 Quantitative Analysis

To assess the effectiveness of our method in analyzing users within the scene and identifying which ones to engage, we conducted tests using the publicly available UE-HRI dataset [156]. In addition, we collected a dataset in the MIVIA Laboratory of the University of Salerno, referred hereinafter as MIVIA-MultiEngagement (MIVIA-ME), to perform an extended evaluation and validation.

4.1.1.1 UE-HRI Dataset

The UE-HRI dataset comprises 54 video recordings of interactions between individuals and the robot Pepper, capturing both single-user and multi-user conversations in authentic, uncontrolled settings. The UE-HRI dataset offers a valuable basis for assessment as it represents real-world interaction dynamics, including a variety of engagement levels.

Inspired by Youssef et al. [157], we formulate our engagement detection task as a *binary classification problem* and we decided to label the dataset accordingly. The labels are thus defined as follows:

- **Negative Class, NENG** includes situations in which no users are visible within the scene, indicating a complete disengagement from interaction with the robot and also situations in which users are present in the scene but have temporarily shifted focus away from the robot, evidenced by behaviors such as looking away or conversing with one another.
- **Positive Class, ENG** includes situations in which at least one user in the scene is engaged in interaction with the robot.

Evaluation metrics To assess the system’s performance, we utilized the *F1 score*, a widely recognized metric in classification tasks. Calculating the F1 score involved determining the counts for True Positives (TP), False Positives (FP), and False Negatives (FN), defined as follows:

- **True Positive (TP)**: In the i^{th} frame of the ground truth, where the label is ENG (indicating engagement), the system correctly predicts ENG.
- **False Positive (FP)**: In the i^{th} frame of the ground truth, where the label is NENG (indicating disengagement), the system wrongly predicts ENG.
- **False Negative (FN)**: In the i^{th} frame of the ground truth, where the label is ENG (indicating engagement), the system incorrectly predicts a negative label (NENG).

Based on these definitions, we calculated the *precision*, *recall*, and *F1 score* metrics as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4.1)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4.2)$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4.3)$$

4.1.1.2 MIVIA-Multi Engagement Dataset

Our solution has also been evaluated on the MIVIA-Multi Engagement (MIVIA-ME) dataset (Figure 4.1), which comprises three hours of video recordings capturing interactions between individuals (at least two people in each sequence) and the Pepper robot, involving a total of 51 participants. The labeling has been performed event-based and the following events have been identified:

- UL^i : The i -th user is looking at the camera.
- US^i : The i -th user is speaking.
- UNL^i : The i -th user is not looking at the camera.
- UNS^i : The i -th is not speaking.

In this context, a *True Positive* occurs when the system correctly engages the user who is both looking at and speaking to the robot, i.e., the system selects the i -th user labeled with *UL* and *US*. A *False Positive* is observed when the system engages a user who is neither looking at nor speaking to the robot, i.e., the system selects the i -th user labeled with *UNL* or *UNS*. Conversely, a *False Negative* arises when the system fails to engage a user who is both looking at and speaking to the robot, i.e., the system does *not* select the i -th user labeled with *UL* and *US*. Subsequently, the evaluation was conducted by computing the *Precision*, *Recall*, and *F1 score*, as described in Section 4.1.1.1. The ground truth annotations comprised a total of 96 events, corresponding all of them to clips in which at least two users were simultaneously present within the robot’s field of view.



Figure 4.1: Image from the MIVIA-ME dataset illustrating two users interacting with the robot. The user on the left is gazing at the robot but not speaking, as highlighted by a yellow rectangle, while the user on the right is both gazing at and speaking to the robot, indicated by a green rectangle.

4.1.1.3 Experimental Results

UE-HRI dataset. The results obtained by the proposed multi-engagement algorithm are summarized in Table 4.1. In particular we report Precision (P), Recall (R) and F1-Score (F1) obtained by exclusively applying the head pose module or the active speaker module; a combination of both modules (head pose + active speaker) is then evaluated. In the first configuration (head pose predictor only) the system achieved high precision, minimizing the occurrence of false positives. However, this configuration led to a reduced recall rate, increasing the number of false negatives and resulting in an overall F1 score of 78%. An improvement was achieved by using the active speaker detection module (thus by combining audio and video streams). This adjustment slightly reduced precision but significantly boosted recall, leading to an F1 score of 80%. The best overall performance was obtained by combining both the head pose and active speaker detection modules: this resulted in the highest recall value and yielded a robust F1 score of 82%, reflecting an increased ability to accurately identify engaged users among several interaction scenarios.

Table 4.1: Results obtained for Precision (P), Recall (R), and F1 score (F1), using different model settings: (1) only the *head pose* module, (2) only the *active speaker* module, and (3) both the *head pose* and the *active speaker* modules.

	P	R	F1
Head Pose	0.92	0.67	0.78
Active Speaker	0.88	0.74	0.80
Proposed: Head Pose + + Active Speaker	0.88	0.77	0.82

MIVIA-ME dataset. On the MIVIA-ME dataset the system achieved a high *recall* rate of 86%, although a drop in *precision* was observed (reaching only 74%) due to increase in *false positives*. This resulted in an overall *F1-score* of 79%, as summarized in Table 4.2.

Table 4.2: System performance on MIVIA-ME dataset in terms of *Precision*, *Recall*, and *F1-score*

Precision	Recall	F1-score
0.74	0.86	0.79

The observed differences in precision and recall across the two test sets can be attributed to two main factors. During the analysis, it was found that the system correctly detected the interlocutor in the majority of instances, achieving an F1 score of 0.79 for the laboratory setting tests and an F1 score of 0.82 for the UE-HRI dataset. In the UE-HRI dataset, there is a delay in the transition between negative labels and the positive label. This delay increases the number of False Negatives, leading to a reduction in recall. In contrast, in the MIVIA-ME dataset, the system has shown sensitivity to lip movements, which increases the number of False Positives, resulting in a reduction in precision.

4.1.2 Qualitative Analysis

A qualitative evaluation involves examining the perceived quality of the interaction between humans and robots from the human perspective: a series of structured interviews were thus carried out in two distinct settings, namely a controlled laboratory environment and a real-world, uncontrolled, setting. The purpose of these interviews was to offer a thorough understanding of how the system performs under different conditions.

4.1.2.1 Laboratory Tests

Experimental Setup. A group of two users was invited to interact with the same robot (same hardware setup but different software versions and engagement modes). The first mode is based on the approach presented in [20], where the robot engages with the closest person; this has been considered as our baseline solution. The second mode leverages the solution proposed in this thesis. The interaction scenario consisted of users having a dialogue with the robot and asking questions freely.

The experiment employed a *within-subjects* design [158], where each participant interacted with both engagement modes. To maintain the integrity of the test, participants were unaware of the active engagement mode during each interaction, and the initial engagement mode was randomly assigned at the beginning of each session. During the sessions, researchers monitored the interaction and could switch the robot’s engagement mode at runtime without disclosing which mode was active at any given time. At the end of the session, participants were asked to complete a 5-point Likert-based questionnaire. Two types of questionnaires were employed. The first one consisted of three items assessing the system’s engagement capabilities:

1. *How would you evaluate the engagement system (i.e., the robot’s ability to focus on the interlocutor)?*
2. *How involved did you feel during the conversation?*
3. *How would you rate the robot’s ability to engage with multiple users simultaneously?*

For the second questionnaire, we used the widely adopted *Godspeed Questionnaire* [159], which evaluates human-robot interaction across five scales: *Anthropomorphism*, *Animacy*, *Likeability*, *Perceived Intelligence*, and *Perceived Safety*. In this study, only the *Likeability* and *Perceived Intelligence* scales were selected, as they are the most relevant to assessing the performance of the proposed system. The survey questions are presented in Figure 4.2.

Data Validation. A total of 51 surveys were collected. The population involved in the study consisted of students aged between 19 and 27 years, with a mix of participants having both IT knowledge and those without any technical background. In order to assess the validity of the data and the results obtained, we computed Cohen’s d effect size using the **G*Power** software [160], as recommended by Hoffman et al. [158]. Cohen’s d (Eq. 4.4) is a measure of effect size that quantifies the difference between two groups relative to the variability within

Likeability						Perceived Intelligence							
Dislike	1	2	3	4	5	Like	Incompetent	1	2	3	4	5	Competent
Unfriendly	1	2	3	4	5	Friendly	Ignorant	1	2	3	4	5	Knowledgeable
Unkind	1	2	3	4	5	Kind	Irresponsible	1	2	3	4	5	Responsible
Unpleasant	1	2	3	4	5	Pleasant	Unintelligent	1	2	3	4	5	Intelligent
Awful	1	2	3	4	5	Nice	Foolish	1	2	3	4	5	Sensible

(a) Godspeed Questionnaire questions regarding *Likeability*.

(b) Godspeed Questionnaire questions regarding *Perceived Intelligence*.

Figure 4.2: Questions from the Godspeed Questionnaire assessing user perceptions of *Likeability* and *Perceived Intelligence*.

those groups. It is used to determine how practically significant the difference between two means is. The Cohen's d can be computed as follows:

$$d = \frac{M_1 - M_2}{SD_{\text{pooled}}} \quad (4.4)$$

where:

- M_1 and M_2 are the means of the two groups,
- SD_{pooled} is the pooled standard deviation, calculated as:

$$SD_{\text{pooled}} = \sqrt{\frac{(n_1 - 1) \cdot SD_1^2 + (n_2 - 1) \cdot SD_2^2}{n_1 + n_2 - 2}}$$

where:

- SD_1 and SD_2 are the standard deviations of the two groups,
- n_1 and n_2 are the sample sizes of the two groups.

A common convention is that $d = 0.2$ represents a small effect, $d = 0.5$ represents a medium effect, and $d \geq 0.8$ represents a large effect. Another important indicator that is used to measure if the obtained result are statistically significant is the so called *statistical power* [158]. Statistical power is the probability that a statistical test will detect an effect if it truly

exists. In other words, it is the ability of a test to correctly identify a difference or effect when the null hypothesis is false. Low power (e.g., below 50%) indicates that the test has a low probability of detecting a true effect, which may lead to Type II errors (false negatives). High power is desirable, as it reduces the likelihood of Type II errors and increases the probability of making a correct detection. The statistical power (β) of a test is calculated as the probability of correctly rejecting the null hypothesis when it is false. The general formula for statistical power is:

$$\text{Power} = 1 - \beta$$

Where: β is the probability of committing a Type II error (false negative), i.e., failing to detect an effect that truly exists. The calculation of power depends on several factors, including:

- **Effect size** (d): the difference between the groups or the strength of the effect being tested.
- **Sample size** (N): the number of observations in the sample.
- **Significance level** (α): the probability of committing a Type I error (false positive), i.e., rejecting the null hypothesis when it is true.
- **Test distribution**: typically, a normal distribution or a Student's t-distribution, depending on the type of test used.

After collecting data, the first step was to assess the validity of the proposed survey test [161] using Cronbach's alpha [162] (Eq. 4.5), which is a statistical measure used to assess the reliability or internal consistency of a set of items within a group.

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum_{i=1}^k \sigma_i^2}{\sigma_T^2} \right) \quad (4.5)$$

where:

- k is the number of items
- σ_i^2 is the variance of the i -th item
- σ_T^2 is the total variance of the sum of all items

It is primarily employed to evaluate questionnaires, rating scales, or psychometric tests, ensuring that the items are well-correlated with each other. The alpha value ranges from 0 to 1, with higher values indicating greater consistency among the items, while lower values suggest lack of correlation [163]. In our case, a Cronbach's alpha of 0.88 was obtained for *engagement questionnaire*, confirming the validity of the proposed test [158].

Questionnaires setup. A first test was performed to demonstrate the effectiveness of our engagement system compared to the baseline. The *Engagement Questionnaire* has been formulated to examine the following hypothesis:

H1: The users consider the proposed engagement system more engaging than the baseline system.

We used the procedure proposed by O'Brien et al. [164], which involves summing the scores of each item per participant, then computing the sum of the scores on questions, and finally dividing the total by the number of questions. In our case:

$$\text{final_score}_j = \frac{1}{3} \sum_{i=1}^3 \text{score}_{ij} \quad (4.6)$$

where j is the j -th participant and i is the i -th question. This procedure was applied to each participant, and the data were then split by modality, resulting in the score distributions X and Y (for the proposed system and the baseline respectively). A second set of tests was conducted using the *Godspeed Questionnaire* to analyze the *Likeability* and *Perceived Intelligence* constructs. For both questionnaires, given the distributions of the two groups, the *Wilcoxon* [158, 164, 165] test was employed at a **95%** confidence level and a critical value for a one-tailed

Table 4.3: Summary of the results obtained for the constructs *Engagement (ENG)*, *Likeability (LI)*, and *Perceived Intelligence (PI)*. The table reports the mean and standard deviation (SD) for Group 1 (Baseline system) and Group 2 (Proposed system), along with the p-values, z-scores obtained from the statistical tests, and the corresponding effect size and statistical power.

	ENG		LI		PI	
	<i>Group1</i>	<i>Group2</i>	<i>Group1</i>	<i>Group2</i>	<i>Group1</i>	<i>Group2</i>
Mean	2.96	4.34	3.97	4.44	4.11	4.48
SD	1.00	0.56	0.76	0.55	0.79	0.48
p-value	1.21e-8		7.7e-4		2.2e-2	
z-score	5.69		3.36		2.28	
effect size	1.70		0.70		0.57	
power	0.99		0.99		0.99	

z-test (right-tail test) of **1.65**. The results are reported in Table 4.3, which presents in detail the Wilcoxon test outcomes, including p-values and z-scores, for the constructs of *Engagement*, *Likeability*, and *Perceived Intelligence*.

Discussion. A total of 51 surveys were collected. Analyzing the engagement-related questionnaire, users positively evaluated the system on a 5-point Likert scale, with a mean score of 4.34 and a standard deviation of 0.56 for the proposed system, compared to a mean score of 2.96 and a standard deviation of 1.0 for the baseline. The Wilcoxon test yielded a p-value of 1.21e-8, a z-score of 5.69, an effect size of 1.7, and a statistical power greater than 0.95, considering a significance level of 95%. These results confirm the significance of the reported statistical outcomes and demonstrate that the proposed system is more engaging than the baseline. Moreover, the results of the *Godspeed Questionnaire* also demonstrate positive feedback. For the *Likeability* item, the system achieved a p-value of 7.7e-4 and a z-score of 3.66, with an effect size of 0.7 and a

statistical power of 99%. For the *Perceived Intelligence* item, the system achieved a p-value of $2.2e-2$ and a z-score of 2.28, with an effect size of 0.57 and a statistical power of 99%. In both cases, the p-values were below the threshold α of 0.05, that confirms the statistical significance of the results at a 95% confidence level.

4.1.2.2 In the Wild Tests

Experimental Setup Regarding tests conducted in real-world scenarios, it was not feasible to allow users to try both system modalities due to the chaotic and dynamic nature of these environments. Nevertheless, 62 Likert-based surveys were collected, yielding an overall positive evaluation of the system. These tests were conducted through surveys gathered during the *Maker Faire* exhibition in Rome and the *Futuro Remoto* event in Naples, both held in Italy in 2024. The data collection involved participants ranging in age from 10 to 80 years, resulting in a total of 62 surveys: 41 from *Futuro Remoto* and 21 from *Maker Faire*. Figure 4.3 depicts the robotic platform proposed in this thesis in operation during the exhibition in both the events.

During the event, participants were encouraged to interact freely with the Pepper robot, and 62 users completed the survey form. Following the interaction, participants were invited to respond to a 5-point Likert scale questionnaire assessing various aspects of their experience [20]:

- (Q1) *How engaging was the conversation with the robot?*
- (Q2) *How do you rate the robot's ability to understand questions?*
- (Q3) *How would you evaluate the speed of the robot's responses?*
- (Q4) *How appropriate were the robot's responses?*

Furthermore, users who played the "rock, paper, scissors" game with Pepper were asked two additional questions:



(a) Pepper robot interacting with people at *Maker Faire* in Rome, Italy.



(b) Pepper robot interacting with people at *Futuro Remoto* in Naples, Italy.

Figure 4.3: Pepper robot in action during the *Maker Faire* and *Futuro Remoto* exhibition.

- (QG1) *How do you rate the social robot's ability to play with humans?*
- (QG2) *Did you have fun playing with Pepper?*

Finally, participants were asked to provide an overall assessment of the human-robot interaction experience:

- (QF) *How would you rate the social robot overall (reliability, performance)?*

Considering the difficulty of conducting experimentation in real-world scenarios using a *within-subject* procedure, we opted to test the proposed system directly by analyzing user feedback. The summarized results are presented in Table 4.4, which provides an analysis based on mean, standard deviation, quartiles, and minimum/maximum values.

Results. The overall system performance (QF) received a strong evaluation, with a mean rating of 4.3 and a standard deviation of 0.72. Notably, question Q1, which focuses on the

Table 4.4: Summary of the survey responses for each question, including statistics such as mean, standard deviation (std), minimum (min), maximum (max), and quartiles (25%, 50%, and 75%).

	Q1	Q2	Q3	Q4	QG1	QG2	QF
mean	4.40	4.10	3.73	4.53	4.55	4.47	4.30
std	0.69	0.74	0.89	0.72	0.79	0.71	0.72
25%	4	4	3	4	4	4	4
50%	5	4	4	5	5	5	4
75%	5	5	4	5	5	5	5
min	3	2	1	2	2	2	2
max	5	5	5	5	5	5	5

robot’s engagement abilities, attained a mean score of 4.4 (close to the maximum score of 5) and a standard deviation of 0.69, indicating a positive impact from the engagement strategies implemented. Question Q2, reflecting the robot’s comprehension abilities, received an average rating of 4.1 with a standard deviation of 0.74. Also in this case, the result is very close to the maximum value, despite the presence of background noise. Moreover, Q4, which evaluates the appropriateness of the robot’s responses, achieved particularly high ratings, with a mean of 4.53 and a standard deviation of 0.72. The lower score observed in Q3, relating to response speed, is attributed to network latency caused by the use of large language models. Furthermore, analyzing the game-related questions, QG1 and QG2, we observe a high level of appreciation, with very high mean scores of 4.55 and 4.47, respectively, and low standard deviations of 0.79 and 0.71 for QG1 and QG2.

4.1.3 Discussion

In this work, the challenge of improving the social acceptability of social robots by equipping them with Multi-Engagement capabilities is addressed. In particular, we propose a multi-modal architecture which integrates audio and video inputs to determine if, when, and which individuals in a scene intend to engage with the robot. The system combines a deep learning-

based head pose estimator with an active speaker detector and utilizes a Behavior Tree model to manage engagement states and guide the robot’s behavior based on interaction history.

Extensive validation demonstrated the system’s effectiveness, achieving an F1 score of **0.79** in laboratory tests and **0.82** using the UE-HRI dataset. User evaluations revealed significant improvements in engagement perception, with scores of **4.34** for the proposed system compared to **2.89** for the baseline on a 5-point Likert scale. Statistical analysis confirmed the significance of these results, showing strong effect sizes and high statistical power across multiple metrics. Moreover, results obtained from the *Godspeed Questionnaire* further validated the system’s effectiveness.

This approach, introduced in Section 3.1, can be deployed across diverse robotic platforms without the need for specialized equipment, offering a versatile and cost-effective solution. By enhancing robots’ ability to perceive and respond to human engagement cues, this method improves the naturalness and effectiveness of human-robot interaction, thereby promoting broader acceptance of social robots in dynamic, multi-user environments.

4.2 Multi-Modal Multi-Task Soft-Biometrics estimation: Experimental Evaluation

In Sections 1 and 2.2, we discussed the importance of estimating soft-biometric information using multimodal strategies to better emulate user behavior and enhance system performance. Moreover, we explored Multi-Task Learning techniques as a means to reduce memory consumption and improve inference time, thereby enabling real-time interaction between users and robots. Subsequently, in Section 3.2.4, we introduced *MAGNET*, a Multi-Modal Multi-Task learning method based on soft parameter sharing.

In this section, to validate the effectiveness of the proposed

MAGNET architecture, we compare our model against uni-modal baselines—referred to as MAGNET-A and MAGNET-V, based on audio and visual encoders respectively—as well as several state-of-the-art audio and video encoders, evaluated in both multi-task and single-task settings. For the audio modality, we evaluate state-of-the-art encoders such as *HuBERT* and *Wav2Vec2*, which operate on raw waveforms, together with the *Conformer*, which processes audio spectrograms. For the visual modality, we evaluate again state of the art video encoders, including *SlowFast*, *R(2+1)D*, *Swin*, and *ConvNeXt*, which differ in architectural complexity and capacity to model spatial and temporal information. To isolate the contributions of architectural choices, we further compare our soft parameter sharing strategy with a hard sharing counterpart. This multi-factors comparison enables us to explore the benefits of multitask learning and modality fusion while also considering model efficiency.

4.2.1 Dataset Description and Partitioning

The dataset used to both train and test the proposed approach is the well known CREMA-D dataset [166]. It features recordings from 91 actors of diverse ethnic backgrounds, including African-American, Asian, Caucasian, Hispanic, and unspecified ethnicities, comprising 48 males and 43 females, with ages ranging from 20 to 74 years. This diversity contributes to the robustness and generalizability of the dataset. The actors simulate six distinct emotions, namely Neutral, Happy, Angry, Sad, Fear, and Disgusted, while reciting scripted phrases in a controlled environment with a green screen background. Each actor utters one or more of a set of 12 predefined sentences ¹.

To ensure a fair evaluation and avoid *data leakage*, the CREMA-D dataset is partitioned based on speakers rather than samples. This approach guarantees that no speaker ap-

¹It’s eleven o’clock; that is exactly what happened; I’m on my way to the meeting; I wonder what this is about; the airplane is almost full; maybe tomorrow it will be cold; I would like a new alarm clock; I think I have a doctor’s appointment; don’t forget a jacket; I think I’ve seen this before; the surface is slick; we’ll stop in a couple of minutes

pears in more than one subset (training, validation, or test) thus preserving the integrity of the evaluation process. A 90-10 split strategy is employed, with 80% of the speakers allocated to the training set and 10% on the validation set. The remaining 10% is allocated to the test set.

Furthermore, the splitting process is designed to maintain a balanced distribution of gender across the subsets. Given that speakers contribute an unequal number of samples across emotional categories, the split also takes emotional distribution into account to ensure a more uniform representation of both gender and emotion classes within each partition. To evaluate the generalization ability of our proposed method we additionally use the RAVDESS (Ryerson Audio-Visual

Database of Emotional Speech and Song) dataset [167].

RAVDESS is a widely used benchmark for emotion recognition, containing both *audio* and *visual* modalities, and consisting of approximately 7,356 recordings performed by 24 professional actors (12 male and 12 female). Each actor utters two lexically matched statements expressed in eight emotional categories: *neutral*, *calm*, *happy*, *sad*, *angry*, *fearful*, *disgust*, and *surprised*. All emotions, except for “neutral”, are expressed at two intensity levels (normal and strong), thus providing variability in emotional strength. Owing to its balanced and structured design, RAVDESS serves as an effective benchmark for assessing cross-dataset robustness and the transferability of models trained on other corpora such as CREMA-D.

4.2.2 Implementation details:

Visual pre-processing

The first step in the video processing pipeline involves face detection on each frame to ensure that the encoder focuses on the most informative regions, primarily the face. This is performed using the S3FD detector, which provides high-precision detections. Since all models are pre-trained, their respective preprocessing pipelines are strictly followed to maintain compatibility with the original training conditions. This typically involves selecting a fixed number of frames T_v , uniformly sampled across the original video’s temporal dimension, and apply-

ing a sequence of image transformations, including normalization (scaling pixel values using pre-defined means and standard deviations), resizing (either to a fixed resolution or by preserving the aspect ratio), and central cropping. For the proposed video encoder MAGNET-V, the preprocessing pipeline of *Slow-Fast* is employed: 8 frames are sampled, resized to maintain a shorter side of 256 pixels, normalized, and center-cropped to 256 pixels. The $R(2+1)D$ architecture processes 16 frames, resized to 256×256 , followed by center cropping. Both *ConvNeXt* and *Swin* operate on 32 frames, with *ConvNeXt* resizing to 230×230 followed by a 224-pixel center crop, while *Swin* resizes to 256×256 before cropping to 224 pixels. To improve model generalization, *RandAugment* is used for data augmentation. This method applies three random image transformations with low intensity, each with a 25% probability, directly to the image tensors during training.

Finally, to enhance realism and better reflect real-world human-robot interaction scenarios, the original green screen backgrounds in the videos are replaced with indoor scenes sampled from the MIT Indoor Scene Recognition dataset [168]. This dataset provides a wide range of high-resolution indoor environments, enabling the creation of contextually rich training samples.

4.2.3 Implementation details: Audio pre-processing

The proposed MAGNET-A model uses MFCCs as input features, with the number of coefficients set to 128, a sampling rate of 16 kHz, 128 Mel filterbanks, and a *Hann* window function. In contrast, the *HuBERT* and *Wav2Vec2* models operate directly on raw audio waveforms, represented as one-dimensional tensors whose lengths correspond to the number of samples. The *Conformer* model, instead, takes as input Mel-spectrograms computed using the same parameters adopted for our MAGNET-A.

4.2.4 Experimental protocol and setup

Metric. In order to validate the proposed approach we consider the widely adopted Accuracy metric:

$$\text{Accuracy} = \frac{\sum_{i=1}^C TP_i}{\sum_{i=1}^C N_i} \quad (4.7)$$

where TP_i denotes the true positive instances for the i -th class, while C represents the total number of classes, and N_i is the number of samples belonging to the i -th class.

Training Configuration. In the single-task setup, all models are trained using the AdamW optimizer. The learning rate is set to 1×10^{-4} for gender classification and 5×10^{-5} for emotion classification. To adapt the learning rate during training, the `ReduceLROnPlateau` scheduler is employed, thus lowering the learning rate by a factor of 0.1 when no improvement in validation loss is observed over four consecutive epochs. Training continues for up to 200 epochs, but to avoid overfitting, `EarlyStopping` halts training if the validation loss does not improve after 10 epochs. The batch size is set to 32 for gender recognition and 128 for emotion recognition.

In the multitask configuration, the training procedure mirrors the single-task setup with a few adjustments to the hyperparameters. Specifically, the learning rate is uniformly set to 5×10^{-5} , and a batch size of 16 is used. Due to the multitask nature of the framework, the loss function being optimized is a weighted combination of the individual task losses, regulated using the GradNorm algorithm.

4.2.5 Experimental results

The achieved results are summarized in Tables 4.5 - 4.8. In more details, Table 4.5 reports the results obtained by the single task architectures, both with single modality and multi modality. Tables 4.6 and 4.7 show the results of both emotion and gender tasks of the multi task framework, with both the soft and hard parameters sharing, respectively. Finally, a comparison of the proposed approach with state-of-the-art algorithms is shown in Table 4.8.

Table 4.5: Results for the emotion recognition and gender recognition tasks in a single-task framework, both single modality and multi modality. For each metric, the values corresponding to the best-performing models are highlighted in bold. Note that M stays for Million, K stays for Kilo.

Audio Encoder	Video Encoder	#Param ↓	Acc. on Emotion ↑	Acc. on Gender ↑
SINGLE MODALITY: AUDIO				
MAGNET-A	-	14.2M	0.595	0.957
Conformer	-	774.0K	0.429	0.927
Wav2Vec2	-	95.5M	0.624	0.986
HuBERT	-	95.5M	0.616	0.984
SINGLE MODALITY: VIDEO				
-	R(2+1)D	34.3M	0.777	1.000
-	SlowFast	42.0M	0.773	0.997
-	MAGNET-V	38.6M	0.805	0.995
-	ConvNeXt	53.0M	0.705	1.000
-	Swin	50.7M	0.773	0.996
MULTI MODALITY: AUDIO-VIDEO				
MAGNET-A	R(2+1)D	51.8M	0.857	1.000
MAGNET-A	SlowFast	65.2M	0.851	1.000
MAGNET-A	Swin	67.2M	0.789	0.998
MAGNET-A	ConvNeXt	69.5M	0.775	1.000
Wav2Vec2	R(2+1)D	133.0M	0.850	0.993
Wav2Vec2	MAGNET-V	137.0M	0.829	0.998
Wav2Vec2	SlowFast	146.0M	0.819	0.998
Wav2Vec2	Swin	148.0M	0.826	0.998
Wav2Vec2	ConvNeXt	150.0M	0.803	1.000
HuBERT	R(2+1)D	133.0M	0.836	1.000
HuBERT	MAGNET-V	137.0M	0.814	0.995
HuBERT	SlowFast	146.0M	0.810	1.000
HuBERT	Swin	148.0M	0.825	1.000
HuBERT	ConvNeXt	150.0M	0.798	1.000
Conformer	R(2+1)D	35.7M	0.798	0.998
Conformer	MAGNET-V	40.1M	0.796	0.984
Conformer	SlowFast	43.5M	0.802	0.998
Conformer	Swin	51.7M	0.759	1.000
Conformer	ConvNeXt	54.0M	0.782	1.000
MAGNET (Single Task Variant)		56.2M	0.882	1.000

Single Task Architecture. The results of the single task architectures for both emotion and gender recognition tasks are reported in Table 4.5. The table is horizontally partitioned into three sections, referring to both single modalities (audio and video) and multi modalities. The proposed encoders (namely MAGNET-A for audio and MAGNET-V for video) are compared with the considered state of the art encoders, widely adopted in the literature.

Regarding emotion recognition, the results indicate a relatively consistent performance across models. The highest accuracy is achieved by *Wav2Vec2*, closely followed by *Hu-*

BERT, which performs nearly on par with the best-performing method. In contrast, *Conformer* exhibits a notable drop in accuracy, trailing by approximately 20% compared to *Wav2Vec2*. Notably, the MAGNET-A model demonstrates competitive accuracy while significantly reducing the parameter count to 81.3M fewer parameters (14.2M vs. 95.5M), highlighting its effectiveness as a balanced solution between predictive performance and computational efficiency.

Among the video-based models, accuracy is generally comparable, with the exception of *ConvNeXt*, which under-performs by approximately 10% compared to the best-performing model, namely our MAGNET-V. Notably, MAGNET-V not only achieves the highest accuracy among video encoders, but also maintains a competitive parameter count, with just 38.6M parameters (versus the best video modality, namely 34.3M of R(2+1)D), making it an efficient and effective choice.

In the multimodal setting, all the possible combinations of the considered video and audio encoders have been evaluated. We can notably see that the proposed solution, named MAGNET, achieves the highest accuracy of 88.2% on the test set, representing gains of 7.7% and 28.7% over the single modality counterpart MAGNET-V and MAGNET-A, respectively. These results underscore the effectiveness of multimodal fusion and the complementarity of the proposed representations. Across the various configurations, multimodal variants based on MAGNET-A or MAGNET-V consistently outperform alternative designs.

As for gender recognition, the accuracy is consistently high, indicating the relative ease of this task when compared with emotion recognition. The proposed multimodal solution, combining MAGNET-A and MAGNET-V still provides a 4.3% accuracy improvement over its unimodal audio counterpart MAGNET-A, confirming the benefits of modality integration, albeit to a lesser extent.

Multitask Architecture. Table 4.6 presents the results of the proposed multitask, multimodal MAGNET, compared with alternative multitask architectures employing soft parameter sharing. As before, the comparison includes both single-modal

Table 4.6: Results for the emotion recognition and gender recognition tasks in a multitask soft-sharing framework, both single modality and multi modality. For each metric, the values corresponding to the best performing models are highlighted in bold. Note that M stays for Million.

Audio Encoder	Video Encoder	#Param ↓	Acc. on Emotion ↑	Acc. on Gender ↑
SINGLE MODALITY: AUDIO				
MAGNET-A	-	30.9M	0.609	0.950
Conformer	-	18.1M	0.358	0.891
Wav2Vec2	-	112M	0.631	0.991
HuBERT	-	112M	0.624	0.991
SINGLE MODALITY: VIDEO				
-	R(2+1)D	48.8M	0.786	1.000
-	SlowFast	56.3M	0.800	1.000
-	MAGNET-V	53.2M	0.756	1.000
-	ConvNeXt	69.7M	0.431	0.995
-	Swin	67.4M	0.718	1.000
MULTI MODALITY: AUDIO-VIDEO				
MAGNET-A	R(2+1)D	65.6M	0.861	0.996
MAGNET-A	SlowFast	73.4M	0.861	1.000
MAGNET-A	Swin	82.2M	0.798	1.000
MAGNET-A	ConvNeXt	84.5M	0.694	1.000
Wav2Vec2	R(2+1)D	146M	0.851	0.998
Wav2Vec2	MAGNET-V	151M	0.847	1.000
Wav2Vec2	SlowFast	154M	0.858	1.000
Wav2Vec2	Swin	163M	0.820	1.000
Wav2Vec2	ConvNeXt	165M	0.673	1.000
HuBERT	R(2+1)D	146M	0.868	0.990
HuBERT	MAGNET-V	151M	0.848	1.000
HuBERT	SlowFast	154M	0.867	1.000
HuBERT	Swin	163M	0.813	1.000
HuBERT	ConvNeXt	165M	0.667	1.000
Conformer	R(2+1)D	50.1M	0.804	0.992
Conformer	MAGNET-V	54.5M	0.799	1.000
Conformer	SlowFast	57.7M	0.798	0.998
Conformer	Swin	68.4M	0.776	0.995
Conformer	ConvNeXt	70.7M	0.705	1.000
MAGNET		69.9M	0.869	0.998

and multimodal configurations.

The experimental results indicate consistent emotion recognition accuracy across the evaluated models. Multitask approaches based exclusively on the video modality, namely *R(2+1)D*, *SlowFast*, and MAGNET-V, demonstrate comparable performance. In contrast, the *Swin* model exhibits a moderate performance drop of 8.2% relative to the best-performing method. The most significant decline is observed with the *ConvNeXt* model, which underperforms by 36.9% compared to *SlowFast*.

The audio-based models also exhibit consistent performance,

with the exception of the *Conformer*, which lags behind *Wav2Vec2*, the best-performing model in the audio-only modality, by 27.3%. Notably, the MAGNET-A model demonstrates strong efficiency, achieving competitive results while reducing the parameter count by 81.1 million compared to *Wav2Vec2*.

The top-performing multimodal configuration is confirmed to be the proposed MAGNET, which achieves an accuracy of 86.9%, outperforming its unimodal counterparts, MAGNET-V and MAGNET-A, by 6.9% and 26%, respectively.

A noteworthy finding is that comparable performance to the proposed solution can be achieved using *HuBERT* in combination with either *SlowFast* or $R(2+1)D$. However, this improvement comes at the expense of significantly larger model sizes (146M and 154M parameters, respectively) compared to fewer than 70M parameters in the proposed approach. In the gender recognition task, all models achieve consistently high accuracy, which can be attributed to the relative simplicity of the task compared to emotion recognition. The video modality alone appears sufficient to reach near-perfect performance, with all video-based models achieving accuracy values close to or equal to 1.000. Nevertheless, the audio modality also demonstrates strong performance, with results closely matching those of the video-based models. The proposed multimodal solution continues to perform competitively in this task, achieving an accuracy of 0.998, only 0.002 below the best-performing configuration. This further highlights the robustness and efficiency of the proposed architecture across both tasks, maintaining strong performance while using significantly fewer parameters than many alternative approaches.

Comparing the performance achieved in the multitask framework with its single-task counterparts, we can note that the results show only a slight decrease in accuracy for both emotion recognition (0.869 vs. 0.882) and gender recognition (0.998 vs. 1.000). However, this minor performance drop is offset by a substantial gain in model efficiency. In fact, the proposed multitask model MAGNET requires only 69.9M parameters, compared to the combined 112.4M parameters needed when using two separate single-task models (56.2M each), representing nearly a 50% reduction in model size. Moreover, a

significant reduction in the number of Floating Point Operations (FLOPs) can be also observed: a single-task network requires approximately 5612.344 GFLOPs (2806.172 GFLOPs per task), whereas the proposed multitask network achieves the same functionality with only 2807.347 GFLOPs, effectively halving the computational load.

Considering that inference time is a critical constraint in social robotics applications, we evaluated system performance in real-time conditions by deploying MAGNET, properly optimized using the TensorRT framework, on an NVIDIA Jetson Orin AGX 64GB: MAGNET achieved an inference time of 0.5 seconds per video clip (2fps) and 0.25s (4fps) using FP16 representation, compared to 1 second (1fps) for the single-task models. Therefore, given that the system requires only 0.5 seconds to process a video clip of approximately 2 seconds, this confirms the viability of the proposed solution for real-time applications in social robotics.

In order to also verify the effectiveness of the proposed multitask framework, we compare MAGNET with its hard-sharing counterpart; the results obtained are summarized in Table 4.7. While the hard-sharing variant has the advantage of a reduced parameter count (59.3M vs. 69.9M for MAGNET), this comes at the expense of predictive performance, particularly on the emotion recognition task. MAGNET achieves an accuracy of 86.9% in this task, whereas the best-performing hard-sharing configuration (using the *ASD* audio encoder and *SlowFast* visual backbone) reaches only 76.7%, marking a drop of more than 10 percentage points. This discrepancy highlights a fundamental limitation of hard parameter sharing: it enforces a shared feature representation across tasks, which can lead to gradient conflicts when the tasks are not fully aligned. In multitask settings involving heterogeneous objectives, such as emotion and gender recognition, this constraint can hinder the model’s ability to capture task-specific nuances. In contrast, MAGNET employs a soft parameter sharing mechanism, which enables each task to retain its own dedicated set of parameters while promoting information sharing through regularization. This approach facilitates task-specific specialization without sacrificing the benefits of shared learning. Ad-

Table 4.7: Results for the emotion recognition and gender recognition tasks in a multitask hard-sharing framework, across single modality and multi modality. For each metric, the values corresponding to the best performing models are highlighted in bold. Note that M stays for Million.

Audio Encoder	Video Encoder	#Param ↓	Acc. on Emotion ↑	Acc. on Gender ↑
SINGLE MODALITY: AUDIO				
MAGNET-A	-	14.7M	0.615	0.943
Conformer	-	783K	0.412	0.902
Wav2Vec2	-	96.1M	0.578	0.991
HuBERT	-	96.1M	0.590	0.984
SINGLE MODALITY: VIDEO				
-	R(2+1)D	37.1M	0.756	0.998
-	SlowFast	45.1M	0.743	0.987
-	MAGNET-V	38.6M	0.805	1.000
-	ConvNeXt	53.5M	0.649	1.000
-	Swin	51.2M	0.701	0.998
MULTI MODALITY: AUDIO-VIDEO				
MAGNET-A	R(2+1)D	55.4M	0.850	1.000
MAGNET-A	SlowFast	74.3M	0.858	1.000
MAGNET-A	Swin	69.4M	0.767	1.000
MAGNET-A	ConvNeXt	71.7M	0.685	0.996
Wav2Vec2	R(2+1)D	136M	0.839	1.000
Wav2Vec2	MAGNET-V	137.0M	0.829	1.000
Wav2Vec2	SlowFast	155M	0.832	0.998
Wav2Vec2	Swin	150M	0.710	0.998
Wav2Vec2	ConvNeXt	153M	0.718	1.000
HuBERT	R(2+1)D	136M	0.815	0.992
HuBERT	MAGNET-V	137.0M	0.814	0.998
HuBERT	SlowFast	155M	0.815	0.998
HuBERT	Swin	150M	0.729	0.998
HuBERT	ConvNeXt	153M	0.641	0.998
Conformer	R(2+1)D	38.6M	0.783	1.000
Conformer	MAGNET-V	40.1M	0.803	1.000
Conformer	SlowFast	46.7M	0.695	0.990
Conformer	Swin	52.3M	0.724	1.000
Conformer	ConvNeXt	54.6M	0.687	1.000
MAGNET (Hard Sharing Variant)		59.3M	0.767	1.000

ditionally, soft sharing allows the model to adaptively weight the contribution of each modality (audio and video) for each task. Rather than imposing a uniform representation, MAGNET learns which modalities are most informative for emotion versus gender recognition. This flexibility improves robustness and reduces the risk of negative transfer, where optimizing one task degrades performance on another. Overall, the experimental results confirm that soft parameter sharing, as implemented in MAGNET, offers a better trade-off between accuracy and model complexity, particularly in scenarios involving related but distinct tasks.

Table 4.8: Comparison of the proposed MAGNET and its single task variant with state of the art approaches on CREMA-D dataset. Several metrics are reported, depending on the ones available for that specific method.

Model	Modality	ACC	$F1_{macro}$	UAR	WAR
STATE-OF-THE-ART					
CoordViT [73]	Audio	83.0%	-	-	-
MTCAE-DFER [77]	Video	-	-	84.7%	85.0%
VAVL [83]	Audio-Video	-	77.9%	-	-
Mocanu et al. [82]	Audio-Video	84.57%	-	-	-
MATER [81]	Audio-Video	67.2%	-	-	-
Single task variant					
MAGNET-V	Video	80.5%	80.6%	-	-
MAGNET-A	Audio	59.5%	59.6%	-	-
MAGNET	Audio-Video	88.2%	88.0%	87.9%	88.2%
Multi task - Hard Sharing					
MAGNET-V	Video	71.5%	70.5%	71.8%	71.5%
MAGNET-A	Audio	61.5%	61.7%	61.7%	61.6%
MAGNET	Audio-Video	76.7%	75.3%	76.5%	76.7%
Multi task - Soft Sharing					
MAGNET-V	Video	75.6%	76.0%	75.4%	75.7%
MAGNET-A	Audio	60.9%	60.5%	60.7%	60.9%
MAGNET	Audio-Video	86.9%	86.8%	86.8%	86.9%

Comparison with State-of-the-Art Methods. In order to also validate the proposed approach with respect to the state of the art, we compare our results with both unimodal and multimodal approaches, well-known in the literature and reaching state of the art performance, with particular reference to emotion recognition task. The results are reported in Table 4.8. Note that, since the Accuracy is not always reported in the considered papers as metric, we also include $F1_{macro}$, Unweighted Average Recall (UAR) and Weighted Average Recall (WAR), which are accuracy-related metrics. For the sake of completeness, they can be computed as follows:

$$\text{recall}_i = \frac{TP_i}{TP_i + FN_i} \quad (4.8)$$

$$F1_{macro} = \frac{1}{C} \sum_{i=1}^C F1_i \quad (4.9)$$

$$\text{WAR} = \frac{\sum_{i=1}^C (\text{recall}_i \cdot N_i)}{\sum_{i=1}^C N_i} \quad \text{UAR} = \frac{1}{C} \sum_{i=1}^C \text{recall}_i \quad (4.10)$$

Here, TP_i , and FN_i denote the true positive, false positive, and false negative instances for the i -th class, respectively. C represents the total number of classes, and $F1_i$ corresponds to the F1-Score of the i -th class. N_i is the number of samples belonging to class i . Furthermore, recall_i denotes the recall for class i .

In particular, we select state-of-the-art models on the CREMA-D benchmark across modalities for comparative evaluation: *CoordViT* [73] for audio-based classification, *MTCAE-DFER* [77] for visual analysis, and the multimodal approaches by Mocanu et al. [82], *MATER* [81] and *VAVL* [83], which represents the current state of the art for joint audio-visual learning. Note that the same experimental setup has been used for evaluating all the considered algorithms.

MAGNET proves consistently to outperform all unimodal baselines. Specifically, it exceeds the performance of *CoordViT*, (+3.9%), and surpasses *MTCAE-DFER* (+2.1% in terms of UAR and +1.9% in terms of WAR). Moreover, it outperforms the multimodal approaches by Mocanu et al. [82] (+2.4% in terms of accuracy), *MATER* [81] (+19.7% in terms of accuracy) and the *VAVL* model (+8.9% in terms of F1-score).

In order to verify the effectiveness of the proposed MAGNET on a different dataset, we also compare it with state-of-the-art methods on the RAVDESS dataset. The results are reported in Table 4.9. Note that the model trained on CREMA-D has been only fine-tuned for domain adaptation. This cross-dataset evaluation highlights the model’s strong generalization capability, as it effectively transfers multimodal emotional representations learned from CREMA-D to RAVDESS. In fact, MAGNET attains the highest accuracy (92.0%), UAR (93.0%) and WAR (91.6%) values, with impressive macro-F1 score (91.9%), surpassing existing multimodal approaches such as Mocanu et al. [82] (+2.7%) and *MATER* [81] (+15.7%) and single modality approaches such as *MTCAE-DFER* [77] (+10.3% in terms of UAR and +7.9% in terms of WAR). Overall, the

consistent improvements across all evaluation metrics highlight that the combination of multitask learning and cross-modal attention fusion enables MAGNET to capture complementary emotional cues from audio and visual modalities while maintaining robust performance across datasets.

Table 4.9: Comparison of the proposed MAGNET with state of the art approaches on RAVDESS dataset. Several metrics are reported, depending on the ones available for that specific method.

Model	Modality	ACC	$F1_{macro}$	UAR	WAR
MTCAE-DFER [77]	Video	-	-	82.7%	83.7%
Mocanu et al. [82]	Audio-Video	89.3%	-	-	-
MATER [81]	Audio-Video	76.3%	-	-	-
MAGNET	Audio-Video	92.0%	91.9%	93.0%	91.6%

4.2.6 Ablation study

In this section we evaluate the robustness of the proposed approach by varying the possible configuration parameters of MAGNET: in particular, (i) the number of frames to be used for feeding the video encoder, (ii) the α parameter of Grad-Norm, controlling the strength of rebalancing of the two tasks.

Figure 4.4 reports the accuracy of MAGNET by varying the number of frames in the range $T = \{4, 8, 16, 32\}$. In the same figure we also report the accuracy of the different backbone configurations, still by varying the T value. We can note that overall, MAGNET consistently outperforms alternative configurations such as *Conformer*, *HuBERT*, and *Wav2Vec2*-based variants across most frame rates, confirming the effectiveness of its multimodal fusion strategy. In particular, the standard MAGNET configuration achieves the highest performance at $T = 16$ (0.88), indicating that this frame rate provides an optimal trade-off between temporal resolution and redundancy. Performance slightly decreases at higher T (32), suggesting potential oversampling effects and reduced discriminative power. These results demonstrate that MAGNET not only generalizes better across modalities but also maintains robust performance

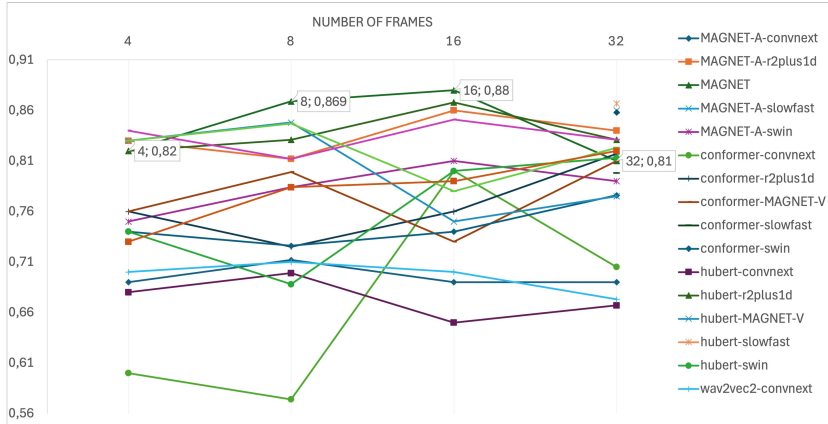


Figure 4.4: Accuracy of the proposed MAGNET solution, by varying the number of frames (in the range 4-32), and compared with the different backbones architectures and multi-modal baselines.

across different temporal granularities, with $T = 16$ emerging as the best operational setting.

We also evaluate the impact of α value in GradNorm, in the range $[0.5, 2.5]$ with $T = 16$. Table 4.10 reports the corresponding performance on both emotion and gender classification tasks.

Table 4.10: Emotion and Gender accuracy by varying the α value in GradNorm.

α	Emotion Accuracy	Gender Accuracy
0.5	0.864	0.990
1.0	0.862	1.000
1.5	0.880	0.990
2.0	0.866	0.990
2.5	0.812	0.990

As shown in Table 4.10, increasing α initially enhances emotion recognition performance, reaching its peak at $\alpha = 1.5$ (0.88), before declining for higher values. This indicates that an intermediate balance between the emotion and gender learning objectives leads to optimal feature disentanglement. Excessively low or high α values either underemphasize or overemphasize one modality, degrading the emotion classification ac-

curacy. Therefore, $\alpha = 1.5$ is adopted as the best trade-off, providing the highest emotion accuracy while maintaining stable gender recognition performance.

4.2.7 Inference Time Analysis

Table 4.11 reports the inference times of the proposed MAGNET, compared with the other considered backbone combinations, in case all the methods are optimized to FP16 representation. The reported inference times have been computed over an NVIDIA Jetson Orin AGX, a low power and low cost embedded device. While Conformer-based models exhibit significantly higher inference times (around 1.5 seconds), MAGNET variants maintain substantially lower times (approximately 0.24–0.27 seconds), achieving nearly a sixfold reduction in computational cost. Despite its lightweight design, MAGNET preserves multimodal representational richness through its integrated attention and multitask learning mechanisms, making it adaptable to different visual and audio backbones. Compared with faster yet unimodal-oriented architectures such as *HuBERT* or *Wav2Vec2*, MAGNET offers still a better trade-off between inference speed and cross-modal learning capacity.

It is important to highlight that both the encoders take as input 2 seconds of audio and video, respectively. Note that the video processes 16 frames in 2 seconds (8 frames per seconds). The time required by MAGNET for the elaboration is 0.24 seconds, which is strongly lower, even with overlapping windows, than the 2 seconds required for the elaboration.

These characteristics make MAGNET a good solution for real-time or large-scale multimodal emotion recognition tasks, where both responsiveness and representational depth are critical.

4.2.8 Discussion

The experimental results demonstrate that the multimodal architectures consistently outperform their unimodal counterparts, namely MAGNET-V and MAGNET-A, across all configurations, thereby confirming the effectiveness of the proposed

Table 4.11: Inference time comparison among different models, ordered by average time (expressed in second).

Model	Mean (s)	Std
conformer-swin	1.66	0.0008
conformer-convnext	1.55	0.3044
conformer-r2plus1d	1.45	0.0092
conformer-MAGNET-V	1.45	0.1013
MAGNET-A-convnext	0.27	0.0020
MAGNET-A-swin	0.26	0.0048
MAGNET	0.24	0.0009
MAGNET-A-r2plus1d	0.24	0.0007
wav2vec2-swin	0.09	0.0020
hubert-swin	0.09	0.0020
wav2vec2-convnext	0.08	0.0039
hubert-convnext	0.08	0.0038
hubert-MAGNET-V	0.03	0.0002
wav2vec2-r2plus1d	0.03	0.0061
wav2vec2-MAGNET-V	0.03	0.0058

multi-modality approach. Furthermore, in the multi-task setting, MAGNET exhibits performance comparable to that of its single-task counterparts in both gender and emotion recognition tasks, benefiting from the cross-task learning dynamics. In this context MAGNET, employing soft parameter sharing, consistently outperforms the hard-sharing variant in terms of accuracy, primarily due to the integration of task-specific branches that facilitate task specialization. Although soft sharing introduces a moderate increase in architectural complexity, it also leads to a significant improvement in model performance. Notably, both the memory footprint and inference time are approximately halved compared to the hard-sharing configuration. These improvements underscore the suitability of the proposed architecture for real-time applications in social robotics, where computational efficiency and low-latency response are critical requirements. Finally, a comparison on two widely adopted datasets, namely CREMA-D and RAVDESS, with several state-of-the-art methods namely, *CoordViT* for audio, *MTCAE-DFER* for video, and the multimodal *VAVL*, *MATER* and *Mocanu et al.*, shows that MAGNET consistently achieves better performance. Notably, competitive results are also obtained by the single-task variants, further supporting

the robustness of the proposed design.

4.3 Evaluation of VLM-Based Proactive Interaction Module

In Sections 1 and 2.3, we discussed the importance of implementing proactive behaviors in social robots. Subsequently, in Section 3.3.4, we introduced a VLM-based method for context-aware sentence generation from video data, trained using *Knowledge Distillation* and optimized for deployment on embedded devices. In this section, we first assessed the model’s performance by analyzing token prediction accuracy on the validation set (Section 4.3.1). Given the complexity and inherently social nature of the task, a *Human Evaluation* was also conducted, following the methodology proposed in [169], to evaluate the model’s effectiveness in real-world human-robot interaction scenarios (Section 4.3.2). In addition to these qualitative and quantitative evaluations, a technical assessment was performed to measure inference time and memory footprint. These experiments were conducted on an NVIDIA Jetson Orin AGX platform to verify the system’s suitability for deployment on embedded hardware integrated into the robot (Section 4.3.3).

4.3.1 Quantitative Results

In order to validate the system’s performance, an initial analysis was conducted to examine model convergence by evaluating the *mean token accuracy* on the validation set for different input temporal resolutions (see Section 3). In particular, the analysis was conducted on two versions of the LLaVA-OneVision model, differing in the number of parameters: 0.5B and 7B. Due to the substantial computational demands of the 7B model, evaluation was limited to the $T = 1$ configuration, as discussed in Section 4.3.3. The results on the validation set are presented in Figure 4.5, while the outcomes on the test set are summarized in Table 4.12. We analyzed the Mean Token Accuracy (MTA) for different values of $T \in \{1, 4, 8, 16\}$. Notably, all model configurations achieved an MTA of approximately

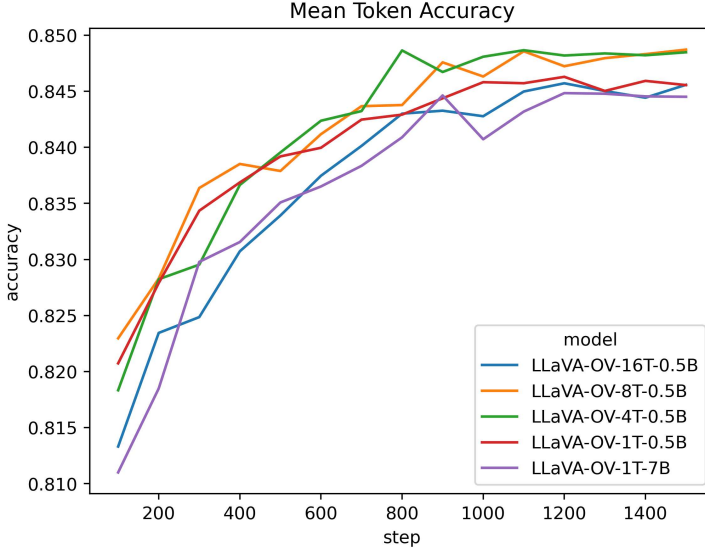


Figure 4.5: Mean token accuracy on the validation set.

84%. This result suggests that, for the task of generating contextually appropriate utterances, a single frame ($T = 1$) is sufficient. This observation can be attributed to the strong generative and representational capabilities of the underlying Vision-Language Model, **LLaVA-OneVision**. As demonstrated in [110], **LLaVA-OneVision** exhibits robust video understanding through task transfer from image-based learning, allowing it to effectively manage both image and video modalities. Furthermore, selecting $T = 1$ provides significant advantages for deployment on embedded systems by reducing both memory footprint and inference time, while maintaining high performance, as further discussed in Section 4.3.3. It is important to emphasize that this is a preliminary analysis aimed at assessing whether the model can generate syntactically correct and contextually meaningful sentences. A more in-depth evaluation of sentence quality is presented in Section 4.3.2.

Table 4.12: Mean Token Accuracy (MTA) achieved by the evaluated models on test set at varying levels of temporal video context T .

Model	T	MTA
llava-onevision-qwen2-0.5b	16	0.8391
	8	0.8419
	4	0.8414
	1	0.8441
llava-onevision-qwen2-7b	1	0.8433

4.3.2 Qualitative Results

Nevertheless, while the quantitative analysis (see Section 4.3.1) reported positive results regarding the model’s generalization and generation capabilities for the task, this alone is not sufficient. Given the inherently social nature of the problem, a human evaluation is necessary [169] to assess the impact and perceived quality of the model’s outputs from a user perspective. To this end, a questionnaire based on a 5-point Likert scale was designed and administered to a group of both technical and non-technical participants. The survey included the following five questions: *Q1: Do you think the generated sentence is pertinent to the context of the video?*, *Q2: Is the generated sentence syntactically correct?*, *Q3: Is the generated sentence semantically correct?*, *Q4: Do you think the sentence is appropriate to attract the user’s attention?*, and *QF: How do you rate the overall quality of the sentence?*. The questionnaire was distributed through a web interface developed using the *Label Studio* tool², allowing broad and accessible participation. The evaluation interface, illustrated in Figure 4.6, consisted of a set of evaluation guidelines, a video player, the generated context-aware sentence (i.e., the model output), and the aforementioned questions. Users were instructed to watch the video, read the sentence, and provide their evaluations using the 5-point Likert scale. Participants were presented with

²<https://labelstud.io/>

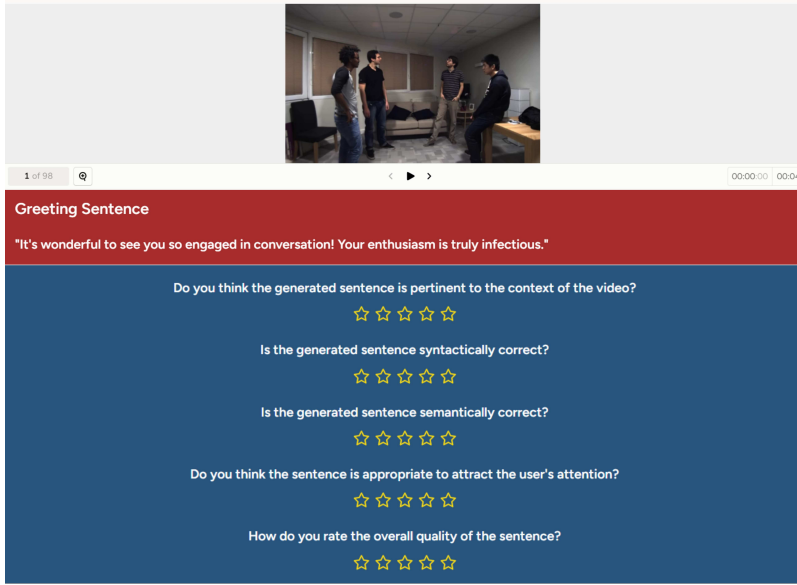


Figure 4.6: Web interface of the evaluation framework. Evaluation instructions are displayed in the blue box at the top. Below, a video player presents the video sample to be assessed. The sentence to be evaluated appears in the red box, followed by a 5-point Likert scale for rating. At the top of the web interface, the following labeling instructions are displayed: "Evaluate the sentence generated by a social robot to greet user(s). The system uses the video below to generate a greeting sentence. Press play and watch the video. Then, read the sentence shown in the red box titled *Greeting Sentence*, and provide your evaluation using the Likert scale below."

video samples that were excluded from the training phase of the network, comprising a total of 55 unique 4-second video clips. It is important to highlight that the size of this test dataset differs from that reported in Section 3, as the latter was augmented using the proposed *Knowledge Distillation* protocol. During the evaluation, each video could be evaluated multiple times by different users; however, the same user was not allowed to evaluate the same video more than once. Additionally, video samples were uniformly and randomly selected by the evaluation environment to ensure unbiased sampling across the dataset. Furthermore, to assess the internal consistency of

Table 4.13: Summary of the survey responses for each question, including statistics such as mean, standard deviation (std), and quartiles (25%, 50%, and 75%).

	Q1	Q2	Q3	Q4	QF
mean	3.79	4.56	4.48	3.90	4.02
std	1.25	0.76	0.80	1.12	1.02
25%	3	4	4	3	3
50%	4	5	5	4	4
75%	5	5	5	5	5

the questionnaire, *Cronbach's alpha* (α) [170] was computed, as previously done in Section 4.1.2, yielding a value of 0.83. Cronbach's alpha is a widely adopted metric for evaluating scale reliability, measuring the degree to which a set of items are interrelated as a group. A value of $\alpha \geq 0.70$ is generally considered acceptable in psychological and social science research [158].

The Cronbach's alpha can be computed as follow:

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum_{i=1}^k \sigma_{Y_i}^2}{\sigma_X^2} \right) \quad (4.11)$$

where:

- k denotes the number of items,
- σ_X^2 represents the variance of the total score,
- $\sigma_{Y_i}^2$ indicates the variance of item i across the sample of individuals.

Higher values of α indicate greater internal consistency among the questionnaire items. The questionnaire was completed by 19 participants, resulting in a total of 298 evaluations. For each question, we computed the mean, standard deviation, and quartiles. The results are summarized in Table 4.13. Questions *Q2* and *Q3* received mean scores close to 4.5, supporting the quantitative results discussed in Section 4.3.1 and confirming

the model’s ability to generate syntactically and semantically sound sentences. Furthermore, questions $Q1$ and $Q4$ yielded mean scores close to 4, suggesting that while the model demonstrates a good understanding of the scene, there is still room for improvement in terms of contextual appropriateness and engagement. Finally, the results for QF indicate that users positively rated the overall quality of the generated sentences, supporting the model’s suitability for generating context-aware sentences to implement proactive behaviors in social human-robot interaction contexts.

4.3.3 On Device Tests

Building on the quantitative results presented in Section 4.3.1 and the need for efficient inference on embedded systems discussed in Sections 1 and 2.3, this section evaluates the computational performance of the trained models in the sentence generation task. Two versions of the **LLaVA-OneVision** model were analyzed, differing in the number of parameters: 0.5B and 7B. The memory footprint was assessed as a function of the number of input frames T , while inference time was measured by varying the maximum number of tokens generated. Although a maximum of 48 tokens is generally sufficient to produce meaningful sentences, additional experiments were conducted with limits of 32 and 64 tokens to evaluate the effect of output length on runtime performance. All tests were performed on an NVIDIA Jetson Orin AGX device using `float16` representation. The results, summarized in Table 4.14, show that **LLaVA-OneVision-Qwen2-0.5B** is the most lightweight configuration, achieving high throughput when generating 48 tokens while maintaining a low memory footprint of only 1964.80 MB across different values of T . In contrast, the **LLaVA-OneVision-Qwen2-7B** model requires substantially more memory, as expected, at least 15319.23 MB and incurs significantly longer inference times. Due to these computational demands, evaluations for this model were conducted only with $T = 1$, as inference time is expected to increase with higher values of T . In light of these considerations, and supported by the quantitative findings in Section 4.3.1,

the LLaVA-OneVision-Qwen2-0.5B model was selected for the human evaluation described in Section 4.3.2.

Table 4.14: Memory usage and inference time (mean and standard deviation) for different configurations of the LLaVA-OneVision models on the NVIDIA Jetson Orin AGX device. Results are reported for varying numbers of input frames T and generation lengths (#Toks).

# Toks	T	Model	Memory Used (MB)	Inference Time (seconds)
32	16	llava-onevision-qwen2-0.5b	1990.12	2.70 ± 0.09
	8	llava-onevision-qwen2-0.5b	1976.62	2.33 ± 0.07
	4	llava-onevision-qwen2-0.5b	1969.87	2.14 ± 0.08
	1	llava-onevision-qwen2-0.5b	1964.80	2.06 ± 0.06
		llava-onevision-qwen2-7b	15319.23	89.23 ± 0.07
48	16	llava-onevision-qwen2-0.5b	1990.12	3.66 ± 0.08
	8	llava-onevision-qwen2-0.5b	1976.62	3.25 ± 0.08
	4	llava-onevision-qwen2-0.5b	1969.87	3.14 ± 0.07
	1	llava-onevision-qwen2-0.5b	1964.80	3.10 ± 0.06
		llava-onevision-qwen2-7b	15319.23	93.51 ± 0.09
64	16	llava-onevision-qwen2-0.5b	1990.12	4.59 ± 0.07
	8	llava-onevision-qwen2-0.5b	1976.62	4.25 ± 0.07
	4	llava-onevision-qwen2-0.5b	1969.87	4.20 ± 0.09
	1	llava-onevision-qwen2-0.5b	1964.80	4.10 ± 0.14
		llava-onevision-qwen2-7b	15319.23	98.32 ± 0.07

4.3.4 Discussion

The experimental evaluation of the proposed Vision-Language Model (VLM)-based proactive interaction module demonstrates its effectiveness in generating contextually appropriate, sentence-level utterances for social robots. The results obtained from both quantitative and qualitative analyses provide compelling evidence of the model’s suitability for real-world deployment in Human-Robot Interaction (HRI) scenarios.

From a quantitative perspective, the model achieved a consistent Mean Token Accuracy (MTA) of approximately 84% across different temporal resolutions, indicating that even a single video frame ($T = 1$) is sufficient to generate syntactically and semantically coherent sentences. This finding aligns

with the hypothesis that the underlying LLaVA-OneVision architecture, pre-trained on large-scale image-text corpora, can effectively transfer its representational capabilities to video-based tasks. The ability to maintain high performance with minimal temporal input is particularly advantageous for embedded systems, where computational resources are limited.

The qualitative evaluation, conducted through a structured Likert-scale survey, further supports the model’s efficacy. Participants rated the generated sentences highly in terms of syntactic correctness (mean = 4.56) and semantic relevance (mean = 4.48), with overall sentence quality receiving a mean score of 4.02. These results validate the model’s capacity to produce human-like utterances that are not only grammatically accurate but also contextually meaningful and socially engaging. Notably, the internal consistency of the questionnaire (Cronbach’s $\alpha = 0.83$) confirms the reliability of the evaluation instrument.

In terms of deployment feasibility, the lightweight variant of the model (0.5B parameters) demonstrated excellent performance on the NVIDIA Jetson Orin AGX platform, achieving an inference time of approximately 3 seconds for 48-token outputs and maintaining a memory footprint below 2 GB. This confirms the model’s suitability for real-time applications in social robotics, where low latency and efficient resource utilization are critical.

Despite these promising results, some limitations remain. While the model performs well in generating context-aware sentences, the evaluation also revealed areas for improvement in terms of engagement effectiveness (Q4 mean = 3.90). This suggests that future work should explore more sophisticated prompting strategies or reinforcement learning techniques to further enhance the model’s ability to attract and sustain user attention.

In conclusion, the proposed VLM-based module represents a significant advancement in enabling proactive behaviors in social robots. By combining knowledge distillation, efficient architecture design, and rigorous evaluation, this work lays a strong foundation for scalable, real-time, and socially intelligent human-robot interactions.

Chapter 5

Conclusions

This thesis has addressed several key challenges in the field of social robotics, with a particular focus on enhancing human-robot interaction (HRI) through multimodal perception, proactive behavior generation, and efficient deployment on embedded systems. The proposed contributions span three main areas: multi-engagement management, context-aware sentence generation, and soft-biometric estimation via multi-task learning.

First, we introduced a hardware-agnostic architecture based on the ROS framework, designed to support real-time interaction in dynamic, multi-user environments. This architecture integrates multimodal perception modules, including head pose estimation, active speaker detection, and direction-of-arrival analysis, together with a reasoning layer built upon Behavior Trees and Finite State Machines. The system effectively addresses the Multi-Engagement Problem, enabling the robot to identify, select, and manage interactions with multiple users simultaneously. Quantitative evaluations on both public and custom datasets demonstrated robust performance, while qualitative assessments confirmed enhanced user engagement and perceived intelligence. Furthermore, the proposed framework was extended to integrate a casual game, thereby enabling the deployment of a complete social robotics application aimed at supporting elderly individuals by providing entertainment and promoting engagement.

Second, we proposed a novel method for generating context-

aware sentences from video data using Vision-Language Models (VLMs). To overcome the limitations of large-scale models in embedded settings, we introduced a dual-teacher *Knowledge Distillation* pipeline that leverages both VLMs and LLMs to generate high-quality training data. The resulting lightweight model, based on LLaVA-OneVision, was optimized for deployment on NVIDIA Jetson platforms. Experimental results confirmed the model’s ability to produce syntactically and semantically coherent utterances, with high user satisfaction in real-world evaluations.

Third, we presented MAGNET, a Multi-Modal Multi-Task architecture for soft-biometric estimation, capable of jointly recognizing gender and emotion from audio and video inputs. The proposed framework employs soft parameter sharing to balance task specialization and shared representation learning. Extensive experiments on the CREMA-D dataset demonstrated that MAGNET outperforms both unimodal and hard-sharing baselines, achieving state-of-the-art performance while maintaining a low computational footprint suitable for real-time applications.

From a broader perspective, the contributions of this thesis extend beyond purely technical advancements and address relevant societal challenges. In particular, socially intelligent robots have the potential to support populations that increasingly require assistance, such as elderly individuals or people living alone. The proposed framework demonstrates how a robot can interact naturally with multiple users, interpret emotional cues, and proactively engage in dialogue, capabilities that are essential for real-world socially assistive robotics applications. By integrating entertainment functionalities such as casual games, the system also contributes to improving user engagement and well-being, which is particularly relevant in contexts such as geriatric care, rehabilitation centers, and assisted living facilities.

Although the work presented in this thesis has primarily been validated through controlled experiments and real-world interaction tests, one of the long-term goals is the deployment of such systems in institutional environments. Potential application scenarios include hospitals, geriatric centers, and public

service environments where social robots could assist staff by providing companionship, entertainment, or basic interaction services to users. In this sense, the technologies developed in this thesis represent a step toward bridging the gap between academic research and practical deployment. Future collaborations with healthcare institutions and research centers could enable large-scale validation studies and facilitate the integration of socially assistive robots into daily care practices.

With respect to the objectives outlined at the beginning of the thesis, the proposed research has successfully addressed the main challenges identified in the introduction. Specifically, the work has demonstrated that multimodal perception, combined with efficient deep learning models and structured behavioral reasoning, can significantly improve the quality of human–robot interaction in multi-user environments. The results obtained across multiple experimental evaluations confirm that the proposed approaches are both technically effective and computationally feasible for deployment on embedded robotic platforms. While the proposed models provide a useful and deployable solution, it is important to emphasize that social robotics remains a complex and evolving field, and the models introduced in this thesis represent an important step rather than a final solution.

From a technological transfer perspective, the modular architecture proposed in this thesis, together with the lightweight models designed for embedded deployment, makes the system suitable for integration into commercial robotic platforms. Companies operating in the fields of service robotics, assistive robotics, and smart environments may benefit from these developments, particularly in applications requiring natural interaction with multiple users. While the primary focus of this research has been academic, the methodologies and models developed throughout this thesis provide a solid foundation for future collaborations with industry partners interested in advancing socially intelligent robotic systems.

In conclusion, this thesis contributes to the advancement of socially intelligent robotics by proposing methods that improve robots’ ability to perceive, understand, and interact with humans in natural social settings. By combining multimodal

perception, proactive dialogue generation, and efficient multi-task learning, the presented framework moves one step closer to the realization of socially aware robotic systems capable of operating in real-world environments and providing meaningful support to society.

Chapter 6

Appendix A

Publications Related to This Thesis

1. De Simone, Giuseppe, Alessia Saggese, and Mario Vento. "Empowering human interaction: A socially assistive robot for support in trade shows." 2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN). IEEE, 2024. [20]
2. De Simone, G., Greco, L., Saggese, A., & Vento, M. (2025, April). Robot to Human Interaction with multi-modal conversational engagement. In 2025 IEEE International Conference on Simulation, Modeling, and Programming for Autonomous Robots (SIMPAN) (pp. 1-8). IEEE. [171]
3. De Simone, G., Greco, L., Saggese, A., & Vento, M. (2025). Increasing realism of robot to human interaction with a multi-modal approach. International Journal of Social Robotics. [under-review]
4. De Simone, G., Greco, L., Saggese, A., & Vento, M. (2025). Multimodal audio-visual emotion recognition for social robotics. In Proceedings of the 21st International Conference on Computer Analysis of Images and Patterns (CAIP25). [172]
5. De Simone, G., Greco, L., Saggese, A., & Vento, M. (2025). Integrating visual and audio cues for emotion

and gender recognition in social robotics: A multimodal and multitask approach. *Information Fusion*. [173]

6. De Simone, G., Roveda, L., Saggese, A., & Vento, M. (2025). Towards proactive social robots: Distilling visual knowledge from large vision-language models. In *Proceedings of the Thirteenth International Workshop on Assistive Computer Vision and Robotics (ACVR25)*. [174]
7. De Simone, G., Foggia, P., Saggese, A., & Vento, M. (2023). Autonomous mobile robot for automatic out of stock detection in a supermarket. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1829-1838). [5]
8. De Simone, G., Saggese, A., Foggia, P., & Vento, M. (2025). Deep learning based empty shelf detection based on autonomous mobile robot. *Computer Vision and Image Understanding*. [175]
9. De Simone, G., Greco, A., Rosa, F., Saggese, A., & Vento, M. (2025). Context-aware data augmentation for enhanced speech command recognition in industrial environments. *Scientific Reports*, 15(1), 17445. [176]
10. Solfrizzi, V., Chiapparino, C., Castellano, G., De Carolis, B., Lofrese, D., Palestra, G., Toma, A., Perla, L., Massaro, S., Santacroce, M.T. and De Simone, G., 2025. *SISTER: Social robotS to support biopsychosocial frailTy*. In *Ital-IA 2025: 5th National Conference on Artificial Intelligence*, organized by CINI (pp. 1-6). [177]
11. De Simone, Giuseppe, et al. "Evolving Roles of Intelligent Robots in Social and Industrial Domains." *Ital-IA 2025: 5th National Conference on Artificial Intelligence*, organized by CINI. 2025. [178]

Bibliography

- [1] A. Ajoudani, A. M. Zanchettin, S. Ivaldi, A. Albu-Schäffer, K. Kosuge, and O. Khatib, “Progress and prospects of the human–robot collaboration,” *Autonomous Robots*, no. 5, Jun. 2018.
- [2] M. Mori, K. F. MacDorman, and N. Kageki, “The uncanny valley,” *IEEE Robotics & Automation Magazine*, vol. 19, no. 2, pp. 98–100, 2012.
- [3] A. Henschel, G. Laban, and E. S. Cross, “What Makes a Robot Social? A Review of Social Robots from Science Fiction to a Home or Hospital Near You,” *Current Robotics Reports*, vol. 2, no. 1, pp. 9–19, Mar. 2021. [Online]. Available: <https://doi.org/10.1007/s43154-020-00035-0>
- [4] C. Breazeal, K. Dautenhahn, and T. Kanda, “Social Robotics,” ser. Springer Handbooks, B. Siciliano and O. Khatib, Eds. Cham: Springer International Publishing, 2016, pp. 1935–1972.
- [5] G. De Simone, P. Foggia, A. Saggese, and M. Vento, “Autonomous Mobile Robot for Automatic out of Stock Detection in a Supermarket,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1829–1838. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2023W/ACVR/html/De_Simone_Autonomous_Mobile_Robot_for_Automatic_out_of_Stock_Detection_in_ICCVW_2023_paper.html
- [6] M. Niemelä, P. Heikkilä, and H. Lammi, “A Social Service Robot in a Shopping Mall: Expectations of the Management, Retailers and Consumers,” in *Proceedings of the Companion of the 2017 ACM/IEEE International*

- Conference on Human-Robot Interaction*, ser. HRI '17. New York, NY, USA: Association for Computing Machinery, Mar. 2017, pp. 227–228. [Online]. Available: <https://dl.acm.org/doi/10.1145/3029798.3038301>
- [7] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, “Social robots for education: A review,” *Science robotics*, vol. 3, no. 21, p. eaat5954, 2018.
- [8] P. Pennisi, A. Tonacci, G. Tartarisco, L. Billeci, L. Ruta, S. Gangemi, and G. Pioggia, “Autism and social robotics: A systematic review,” *Autism Research*, vol. 9, no. 2, pp. 165–183, 2016.
- [9] W.-Y. G. Louie, J. Korneder, I. Abbas, and C. Pawluk, “A study on an applied behavior analysis-based robot-mediated listening comprehension intervention for ASD,” *Paladyn, Journal of Behavioral Robotics*, vol. 12, no. 1, pp. 31–46, Jan. 2021. [Online]. Available: <https://www.degruyter.com/document/doi/10.1515/pjbr-2021-0005/html>
- [10] J. A. C. Panceri, É. Freitas, J. C. de Souza, S. da Luz Schreider, E. Caldeira, and T. F. Bastos, “A new socially assistive robot with integrated serious games for therapies with children with autism spectrum disorder and down syndrome: a pilot study,” *Sensors*, vol. 21, no. 24, p. 8414, 2021.
- [11] H. Groenewoud, J. de Lange, Y. Schikhof, A. Astell, P. Jodrell, and M. Goumans, “People with dementia playing casual games on a tablet,” *Gerontechnology*, vol. 16, no. 1, pp. 37–47, Mar. 2017. [Online]. Available: <https://centaur.reading.ac.uk/75404/>
- [12] D. Macis, S. Perilli, and C. Gena, “Employing Socially Assistive Robots in Elderly Care,” in *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, ser. UMAP '22 Adjunct. New York, NY, USA: Association for Computing Machinery, Jul. 2022, pp. 130–138. [Online]. Available: <https://dl.acm.org/doi/10.1145/3511047.3537687>
- [13] M. Manca, F. Paternò, C. Santoro, E. Zedda, C. Braschi, R. Franco, and A. Sale, “The impact of serious games with humanoid robots on mild cognitive impairment older adults,”

- International Journal of Human-Computer Studies*, vol. 145, p. 102509, Jan. 2021.
- [14] C. Huisman and H. Kort, “Two-Year Use of Care Robot Zora in Dutch Nursing Homes: An Evaluation Study,” *Healthcare*, vol. 7, no. 1, p. 31, Mar. 2019. [Online]. Available: <https://www.mdpi.com/2227-9032/7/1/31>
- [15] I. Aaltonen, A. Arvola, P. Heikkilä, and H. Lammi, “Hello Pepper, May I Tickle You? Children’s and Adults’ Responses to an Entertainment Robot at a Shopping Mall,” in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI ’17. New York, NY, USA: Association for Computing Machinery, Mar. 2017, pp. 53–54. [Online]. Available: <https://doi.org/10.1145/3029798.3038362>
- [16] A. Greco, A. Roberto, A. Saggese, and M. Vento, “Efficient transformers for on-robot natural language understanding,” in *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*, 2022, pp. 823–828.
- [17] S. C. Steinhäusser, M. Lein, M. Donnermann, and B. Lugin, “Designing Social Robots’ Speech in the Hotel Context - A Series of Online Studies,” in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, Aug. 2022.
- [18] M. Moujahid, H. Hastie, and O. Lemon, “Multi-party Interaction with a Robot Receptionist,” in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Mar. 2022, pp. 927–931.
- [19] P. Foggia, A. Greco, A. Saggese, and M. Vento, “Multi-task learning on the edge for effective gender, age, ethnicity and emotion recognition,” *Engineering Applications of Artificial Intelligence*, vol. 118, p. 105651, Feb. 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0952197622006418>
- [20] G. De Simone, A. Saggese, and M. Vento, “Empowering human interaction: A socially assistive robot for support in trade shows,” in *2024 33rd IEEE International Conference*

on Robot and Human Interactive Communication (ROMAN), 2024, pp. 901–908.

- [21] M. Crawshaw, “Multi-Task Learning with Deep Neural Networks: A Survey,” *ArXiv*, Sep. 2020. [Online]. Available: <https://www.semanticscholar.org/paper/Multi-Task-Learning-with-Deep-Neural-Networks%3A-A-Crawshaw/74f23063ca77f5b1caa3770a5957ae5fc565843e>
- [22] P. Foggia, A. Greco, A. Roberto, A. Saggese, and M. Vento, “Identity, gender, age, and emotion recognition from speaker voice with multi-task deep networks for cognitive robotics,” *Cognitive Computation*, vol. 16, no. 5, pp. 2713–2723, 2024.
- [23] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, “Object detection in 20 years: A survey,” *Proceedings of the IEEE*, 2023.
- [24] A. Greco, A. Saggese, M. Vento, and V. Vigilante, “A convolutional neural network for gender recognition optimizing the accuracy/speed tradeoff,” *IEEE Access*, vol. 8, pp. 130 771–130 781, 2020.
- [25] Y. Fu, G. Guo, and T. S. Huang, “Age synthesis and estimation via faces: a survey,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 11, pp. 1955–1976, Nov. 2010.
- [26] C.-B. Ng, Y.-H. Tay, and B.-M. Goi, “A review of facial gender recognition,” *Pattern Analysis and Applications*, vol. 18, no. 4, pp. 739–755, Nov. 2015. [Online]. Available: <https://doi.org/10.1007/s10044-015-0499-6>
- [27] A. Greco, A. Roberto, A. Saggese, M. Vento, and V. Vigilante, “Emotion analysis from faces for social robotics,” in *2019 IEEE international conference on systems, man and cybernetics (SMC)*. IEEE, 2019, pp. 358–364. [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8914039>
- [28] Y.-H. Zhang, J. Xiao, S. Yang, and S. Shan, “Multi-task learning for audio-visual active speaker detection,” *The ActivityNet Large-Scale Activity Recognition Challenge*, vol. 4, 2019.

- [29] S. O. Ba and J.-M. Odobez, “Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues,” in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, Mar. 2008, pp. 2221–2224, iSSN: 2379-190X.
- [30] J. Li *et al.*, “Recent advances in end-to-end automatic speech recognition,” *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.
- [31] P. Schneider, R. Memmesheimer, I. Kramer, and D. Paulus, “Gesture recognition in rgb videos using human body keypoints and dynamic time warping,” in *RoboCup 2019: Robot World Cup XXIII 23*. Springer, 2019, pp. 281–293.
- [32] A. Piergiovanni, C. Fan, and M. Ryoo, “Learning latent subevents in activity videos using temporal attention filters,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.
- [33] M. Iovino, E. Scukins, J. Styruud, P. Ögren, and C. Smith, “A survey of behavior trees in robotics and ai,” *Robotics and Autonomous Systems*, vol. 154, p. 104096, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0921889022000513>
- [34] J. Ni, T. Young, V. Pandelea, F. Xue, and E. Cambria, “Recent advances in deep learning based dialogue systems: a systematic survey,” *Artificial Intelligence Review*, vol. 56, no. 4, pp. 3055–3155, Apr. 2023. [Online]. Available: <https://doi.org/10.1007/s10462-022-10248-8>
- [35] H. Weld, X. Huang, S. Long, J. Poon, and S. C. Han, “A survey of joint intent detection and slot filling models in natural language understanding,” *ACM Computing Surveys*, vol. 55, no. 8, pp. 1–38, 2022.
- [36] M. A. Yousuf, Y. Kobayashi, Y. Kuno, A. Yamazaki, and K. Yamazaki, “How to move towards visitors: A model for museum guide robots to initiate conversation,” in *2013 IEEE RO-MAN*, Aug. 2013, pp. 587–592, iSSN: 1944-9437.
- [37] M. E. Foster, A. Gaschler, and M. Giuliani, “Automatically Classifying User Engagement for Dynamic Multi-party

- Human–Robot Interaction,” *International Journal of Social Robotics*, vol. 9, no. 5, pp. 659–674, Nov. 2017. [Online]. Available: <http://link.springer.com/10.1007/s12369-017-0414-y>
- [38] G. Skantze, “Turn-taking in Conversational Systems and Human-Robot Interaction: A Review,” *Computer Speech & Language*, vol. 67, p. 101178, May 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S088523082030111X>
- [39] S. Park, A. Spurr, and O. Hilliges, “Deep Pictorial Gaze Estimation,” in *Computer Vision – ECCV 2018*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 741–757.
- [40] R. Valle, J. M. Buenaposada, and L. Baumela, “Multi-Task Head Pose Estimation in-the-Wild,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 8, pp. 2874–2881, Aug. 2021.
- [41] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, “Deep Learning-based Human Pose Estimation: A Survey,” *ACM Computing Surveys*, vol. 56, no. 1, pp. 11:1–11:37, Aug. 2023. [Online]. Available: <https://dl.acm.org/doi/10.1145/3603618>
- [42] A. Toshev and C. Szegedy, “DeepPose: Human Pose Estimation via Deep Neural Networks,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 1653–1660, iSSN: 1063-6919.
- [43] A. A. Abdelrahman, D. Strazdas, A. Khalifa, J. Hintz, T. Hempel, and A. Al-Hamadi, “Multimodal Engagement Prediction in Multiperson Human–Robot Interaction,” *IEEE Access*, vol. 10, pp. 61 980–61 991, 2022.
- [44] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic, “Auto-avsr: Audio-visual speech recognition with automatic labels,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

- [45] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, “Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation,” *ACM Transactions on Graphics*, vol. 37, no. 4, p. 1–11, Jul. 2018. [Online]. Available: <http://dx.doi.org/10.1145/3197517.3201357>
- [46] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi, and C. Pantofaru, “Ava Active Speaker: An Audio-Visual Dataset for Active Speaker Detection,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 4492–4496, iSSN: 2379-190X.
- [47] T. Stafylakis and G. Tzimiropoulos, “Combining residual networks with lstms for lipreading,” 2017. [Online]. Available: <https://arxiv.org/pdf/1703.04105.pdf>
- [48] R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li, “Is Someone Speaking? Exploring Long-term Temporal Features for Audio-visual Active Speaker Detection,” in *Proceedings of the 29th ACM International Conference on Multimedia*, ser. MM ’21. New York, NY, USA: Association for Computing Machinery, Oct. 2021, pp. 3927–3935. [Online]. Available: <https://dl.acm.org/doi/10.1145/3474085.3475587>
- [49] J. Liao, H. Duan, K. Feng, W. Zhao, Y. Yang, and L. Chen, “A light weight model for active speaker detection,” in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 22 932–22 941.
- [50] J.-M. Valin, F. Michaud, J. Rouat, and D. Letourneau, “Robust sound source localization using a microphone array on a mobile robot,” in *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003) (Cat. No.03CH37453)*, vol. 2, Oct. 2003, pp. 1228–1233 vol.2.
- [51] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, “Emotion recognition in conversation: Research challenges, datasets, and recent advances,” *IEEE Access*, vol. 7, pp. 100 943–100 953, 2019.

- [52] J. Heredia, E. Lopes-Silva, Y. Cardinale, J. Diaz-Amado, I. Dongo, W. Graterol, and A. Aguilera, “Adaptive multimodal emotion detection architecture for social robots,” *IEEE Access*, vol. 10, pp. 20 727–20 744, 2022.
- [53] A. Ruiz-Garcia, M. Elshaw, A. Altahhan, and V. Palade, “A hybrid deep learning neural approach for emotion recognition from facial expressions for socially assistive robots,” *Neural Computing and Applications*, vol. 29, pp. 359 – 373, 2018.
- [54] G. Azzopardi, A. Greco, A. Saggese, and M. Vento, “Fusion of Domain-Specific and Trainable Features for Gender Recognition From Face Images,” *IEEE Access*, vol. 6, pp. 24 171–24 183, 2018. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8331979>
- [55] P. Foggia, F. Rosa, and M. Vento, “Improving learning from visual demonstration methods by target localization,” in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, 2024, pp. 740–747.
- [56] Y. Zhang and Q. Yang, “A survey on multi-task learning,” *IEEE transactions on knowledge and data engineering*, vol. 34, no. 12, pp. 5586–5609, 2022.
- [57] Y. F. Alharbi and Y. A. Alotaibi, “The Correlate of Emotion and Gender Classification Using EEG Signals,” in *2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP)*, Oct. 2021, pp. 790–794. [Online]. Available: <https://ieeexplore.ieee.org/document/9688884>
- [58] M. H. Sedaaghi, “Gender classification in emotional speech,” *Speech Recognition: Technologies and Applications*, pp. 363–376, 2008.
- [59] L. Deng and X. Li, “Machine Learning Paradigms for Speech Recognition: An Overview,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 1060–1089, May 2013. [Online]. Available: <https://ieeexplore.ieee.org/document/6423821>
- [60] R. Caruana, “Multitask learning,” *Machine Learning*, vol. 28, pp. 41–75, 1997.

- [61] Y. Zhang and D. Yeung, “A regularization approach to learning task relationships in multitask learning,” *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 8, pp. 1 – 31, 2014.
- [62] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool, “Multi-Task Learning for Dense Prediction Tasks: A Survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3614–3633, Jul. 2022.
- [63] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, “Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks,” in *International conference on machine learning*. PMLR, 2018, pp. 794–803.
- [64] S. Sowden, B. A. Schuster, C. T. Keating, D. S. Fraser, and J. L. Cook, “The role of movement kinematics in facial emotion expression production and recognition.” *Emotion*, vol. 21, no. 5, p. 1041, 2021.
- [65] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, “A survey of affect recognition methods: Audio, visual, and spontaneous expressions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.
- [66] D. Kwasny and D. Hemmerling, “Gender and Age Estimation Methods Based on Speech Using Deep Neural Networks,” *Sensors*, vol. 21, no. 14, p. 4785, Jan. 2021. [Online]. Available: <https://www.mdpi.com/1424-8220/21/14/4785>
- [67] M. Hao, W.-H. Cao, Z.-T. Liu, M. Wu, and P. Xiao, “Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features,” *Neurocomputing*, vol. 391, pp. 42–51, 2020.
- [68] D. Issa, M. F. Demirci, and A. Yazici, “Speech emotion recognition with deep convolutional neural networks,” *Biomedical Signal Processing and Control*, vol. 59, p. 101894, 2020.
- [69] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

- [70] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, “Conformer: Convolution-augmented transformer for speech recognition,” in *Interspeech*, 2020, pp. 5036–5040.
- [71] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2021.
- [72] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [73] J.-Y. Kim and S.-H. Lee, “Coordvit: A novel method of improve vision transformer-based speech emotion recognition using coordinate information concatenate,” in *2023 International Conference on Electronics, Information, and Communication (ICEIC)*, 2023, pp. 1–4.
- [74] S. Yang, P. Chi, Y. Chuang, C. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G. Lin, T. Huang, W. Tseng, K. Lee, D. Liu, Z. Huang, S. Dong, S. Li, S. Watanabe, A. Mohamed, and H. Lee, “Superb: Speech processing universal performance benchmark,” in *Interspeech*, 2021, pp. 3161–3165.
- [75] C.-Y. Ma, M.-H. Chen, Z. Kira, and G. AlRegib, “Ts-lstm and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition,” *Signal Processing: Image Communication*, vol. 71, pp. 76–87, 2019.
- [76] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, “Vivit: A video vision transformer,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6836–6846.
- [77] P. Xiang, K. Wu, C. Lin, and O. Bai, “Mtcaedfer: Multi-task cascaded autoencoder for dynamic facial expression recognition,” 2024. [Online]. Available: <https://arxiv.org/abs/2412.18988>

- [78] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, “A closer look at spatiotemporal convolutions for action recognition,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.
- [79] F. Zhao, C. Zhang, and B. Geng, “Deep multimodal data fusion,” *ACM Comput. Surv.*, vol. 56, no. 9, Apr. 2024. [Online]. Available: <https://doi.org/10.1145/3649447>
- [80] K. Gadzicki, R. Khamsehashari, and C. Zetsche, “Early vs late fusion in multimodal convolutional neural networks,” in *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, 2020, pp. 1–6.
- [81] E. Ghaleb, J. Niehues, and S. Asteriadis, “Joint modelling of audio-visual cues using attention mechanisms for emotion recognition,” *Multimedia Tools and Applications*, vol. 82, no. 8, pp. 11 239–11 264, 2023.
- [82] B. Mocanu, R. Tapu, and T. Zaharia, “Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning,” *Image and Vision Computing*, vol. 133, p. 104676, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0262885623000501>
- [83] L. Goncalves, S.-G. Leem, W.-C. Lin, B. Sisman, and C. Busso, “Versatile audio-visual learning for emotion recognition,” *IEEE Transactions on Affective Computing*, 2024.
- [84] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool, “Multi-task learning for dense prediction tasks: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1–1, 2021. [Online]. Available: <http://dx.doi.org/10.1109/TPAMI.2021.3054719>
- [85] J. Xie, J. Wang, Q. Wang, D. Yang, J. Gu, Y. Tang, and Y. I. Varatnitski, “A multimodal fusion emotion recognition method based on multitask learning and attention mechanism,” *Neurocomputing*, vol. 556, p. 126649, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0925231223007725>

- [86] S. M. Anzalone, S. Boucenna, S. Ivaldi, and M. Chetouani, “Evaluating the Engagement with Social Robots,” *International Journal of Social Robotics*, vol. 7, no. 4, pp. 465–478, Aug. 2015. [Online]. Available: <https://doi.org/10.1007/s12369-015-0298-7>
- [87] E. REITER and R. DALE, “Building applied natural language generation systems,” *Natural Language Engineering*, vol. 3, no. 1, p. 57–87, 1997.
- [88] C. Dong, Y. Li, H. Gong, M. Chen, J. Li, Y. Shen, and M. Yang, “A survey of natural language generation,” *ACM Computing Surveys*, vol. 55, pp. 1 – 38, 2021.
- [89] A. Sherstinsky, “Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network,” *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020.
- [90] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, p. 1735–1780, Nov. 1997. [Online]. Available: <https://doi.org/10.1162/neco.1997.9.8.1735>
- [91] A. de Santana Correia and E. L. Colombini, “Attention, please! a survey of neural attention models in deep learning,” *Artificial Intelligence Review*, vol. 55, no. 8, pp. 6037–6124, 2022.
- [92] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, “An attentive survey of attention models,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 12, no. 5, pp. 1–32, 2021.
- [93] J. Lin, J. Ma, J. Zhu, and Y. Cui, “Short-term load forecasting based on lstm networks considering attention mechanism,” *International Journal of Electrical Power & Energy Systems*, vol. 137, p. 107818, 2022.
- [94] L. Feng, F. Tung, M. O. Ahmed, Y. Bengio, and H. Hajimirsadeghi, “Were rnns all we needed?” *arXiv preprint arXiv:2410.01201*, 2024.
- [95] M.-I. Stan and O. Rhodes, “Learning long sequences in spiking neural networks,” *Scientific Reports*, vol. 14, p.

- 21957, Sep. 2024. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11415486/>
- [96] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, “Improving language understanding by generative pre-training,” 2018.
- [97] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [98] B. C. Das, M. H. Amini, and Y. Wu, “Security and privacy challenges of large language models: A survey,” *ACM Computing Surveys*, vol. 57, no. 6, pp. 1–39, 2025.
- [99] A. Moroncelli, V. Soni, A. A. Shahid, M. Maccarini, M. Forgione, D. Piga, B. Spahiu, and L. Roveda, “Integrating reinforcement learning with foundation models for autonomous robotics: Methods and perspectives,” *arXiv preprint arXiv:2410.16411*, 2024.
- [100] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman *et al.*, “Do as i can, not as i say: Grounding language in robotic affordances,” *arXiv preprint arXiv:2204.01691*, 2022.
- [101] M. Maccarini, A. Moroncelli, E. Carpanzano, and L. Roveda, “Flare: Fine-tuned large language models for resource-efficient action generation in robotics,” *Procedia CIRP*, vol. 134, pp. 759–764, 2025.
- [102] P. Liu, D. F. Glas, T. Kanda, and H. Ishiguro, “Learning proactive behavior for interactive social robots,” *Autonomous Robots*, vol. 42, no. 5, pp. 1067–1085, Jun. 2018. [Online]. Available: <https://doi.org/10.1007/s10514-017-9671-8>
- [103] C. Sirithunge, A. G. B. P. Jayasekara, and D. P. Chandima, “Proactive Robots With the Perception of Nonverbal Human Behavior: A Review,” *IEEE Access*, vol. 7, pp. 77 308–77 327, 2019. [Online]. Available: <https://ieeexplore.ieee.org/document/8734064>

- [104] A. Garrell, M. Villamizar, F. Moreno-Noguer, and A. Sanfeliu, “Proactive behavior of an autonomous mobile robot for human-assisted learning,” Aug. 2013, pp. 107–113, iSSN: 1944-9437. [Online]. Available: <https://ieeexplore.ieee.org/document/6628463>
- [105] K. Hayashi, D. Sakamoto, T. Kanda, M. Shiomi, S. Koizumi, H. Ishiguro, T. Ogasawara, and N. Hagita, “Humanoid robots as a passive-social medium: a field experiment at a train station,” in *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, ser. HRI '07. New York, NY, USA: Association for Computing Machinery, Mar. 2007, pp. 137–144. [Online]. Available: <https://dl.acm.org/doi/10.1145/1228716.1228735>
- [106] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.
- [107] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, “Sigmoid loss for language image pre-training,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 11 975–11 986.
- [108] J. Li, D. Li, C. Xiong, and S. Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.
- [109] B. Zhang, K. Li, Z. Cheng, Z. Hu, Y. Yuan, G. Chen, S. Leng, Y. Jiang, H. Zhang, X. Li *et al.*, “Videollama 3: Frontier multimodal foundation models for image and video understanding,” *arXiv preprint arXiv:2501.13106*, 2025.
- [110] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu *et al.*, “Llava-onevision: Easy visual task transfer,” *arXiv preprint arXiv:2408.03326*, 2024.
- [111] Y. Zhang, B. Li, h. Liu, Y. j. Lee, L. Gui, D. Fu, J. Feng, Z. Liu, and C. Li, “Llava-next: A strong zero-shot

- video understanding model,” April 2024. [Online]. Available: <https://llava-vl.github.io/blog/2024-04-30-llava-next-video/>
- [112] K. A. Batterton and K. N. Hale, “The Likert Scale What It Is and How To Use It,” *Phalanx*, vol. 50, no. 2, pp. 32–39, 2017. [Online]. Available: <https://www.jstor.org/stable/26296382>
- [113] Y. Tang, J. Bi, S. Xu, L. Song, S. Liang, T. Wang, D. Zhang, J. An, J. Lin, R. Zhu *et al.*, “Video understanding with large language models: A survey,” *arXiv preprint arXiv:2312.17432*, 2023.
- [114] D. Chen and W. B. Dolan, “Collecting highly parallel data for paraphrase evaluation,” in *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 2011, pp. 190–200.
- [115] X. He, S. Chen, F. Ma, Z. Huang, X. Jin, Z. Liu, D. Fu, Y. Yang, J. Liu, and J. Feng, “Vlab: Enhancing video language pre-training by feature adapting and blending,” *IEEE Transactions on Multimedia*, 2024.
- [116] M. Sun, A. Farhadi, and S. Seitz, “Ranking domain-specific highlights by analyzing edited videos,” in *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I 13*. Springer, 2014, pp. 787–802.
- [117] M. Grunde-McLaughlin, R. Krishna, and M. Agrawala, “Agqa: A benchmark for compositional spatio-temporal reasoning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 287–11 297.
- [118] H. Zhang, X. Li, and L. Bing, “Video-LLaMA: An instruction-tuned audio-visual language model for video understanding,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Y. Feng and E. Lefever, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 543–553. [Online]. Available: <https://aclanthology.org/2023.emnlp-demo.49/>

- [119] G. A. Abbo and T. Belpaeme, “I Was Blind but Now I See: Implementing Vision-Enabled Dialogue in Social Robots,” Mar. 2025, pp. 1176–1180. [Online]. Available: <https://ieeexplore.ieee.org/document/10973830>
- [120] A. Favero, L. Zancato, M. Trager, S. Choudhary, P. Perera, A. Achille, A. Swaminathan, and S. Soatto, “Multi-modal hallucination control by visual information grounding,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 303–14 312.
- [121] J. Xu, H. Zhang, Q. Si, Y. Li, X. Lan, and T. Kong, “Towards unified interactive visual grounding in the wild,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 3288–3295.
- [122] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi, “Self-Instruct: Aligning Language Models with Self-Generated Instructions,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 13 484–13 508. [Online]. Available: <https://aclanthology.org/2023.acl-long.754/>
- [123] X. He, Z. Lin, Y. Gong, A.-L. Jin, H. Zhang, C. Lin, J. Jiao, S. M. Yiu, N. Duan, and W. Chen, “AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators,” in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, Y. Yang, A. Davani, A. Sil, and A. Kumar, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 165–190. [Online]. Available: <https://aclanthology.org/2024.naacl-industry.15/>
- [124] X. Xu, M. Li, C. Tao, T. Shen, R. Cheng, J. Li, C. Xu, D. Tao, and T. Zhou, “A survey on knowledge distillation of large language models,” *arXiv preprint arXiv:2402.13116*, 2024.
- [125] C.-Y. Hsieh, C.-L. Li, C.-K. Yeh, H. Nakhost, Y. Fujii, A. Ratner, R. Krishna, C.-Y. Lee, and T. Pfister, “Distill-

ing step-by-step! outperforming larger language models with less training data and smaller model sizes,” *arXiv preprint arXiv:2305.02301*, 2023.

- [126] S. Wang, Y. Liu, Y. Xu, C. Zhu, and M. Zeng, “Want to reduce labeling cost? GPT-3 can help,” in *Findings of the Association for Computational Linguistics: EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 4195–4205. [Online]. Available: <https://aclanthology.org/2021.findings-emnlp.354/>
- [127] N. Savage, “Robots rise to meet the challenge of caring for old people,” *Nature*, vol. 601, no. 7893, pp. S8–S10, Jan. 2022. [Online]. Available: <https://www.nature.com/articles/d41586-022-00072-z>
- [128] H. Nap, d. Kort, Y.A.W., and W. IJsselsteijn, “Senior gamers : preferences, motivations and needs,” *Gerontechnology*, vol. 8, no. 4, pp. 247–262, 2009.
- [129] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, “Simple online and realtime tracking,” in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 3464–3468.
- [130] T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi, “Toward robust and unconstrained full range of rotation head pose estimation,” *IEEE Transactions on Image Processing*, vol. 33, pp. 2377–2387, 2024.
- [131] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, “Repvvg: Making vgg-style convnets great again,” *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13 728–13 737, 2021.
- [132] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.
- [133] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee *et al.*, “Mediapipe: A framework for perceiving and processing reality,”

- in *Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR)*, vol. 2019, 2019.
- [134] C. Feichtenhofer, H. Fan, J. Malik, and K. He, “Slowfast networks for video recognition,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.
- [135] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.
- [136] M. Colledanchise and P. Ogren, *Behavior Trees in Robotics and AI: An Introduction*, 1st ed. USA: CRC Press, Inc., Jun. 2018.
- [137] S. Cooper and S. Lemaignan, “Towards using behaviour trees for long-term social robot behaviour,” in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2022, pp. 737–741.
- [138] C. Paxton, A. Hundt, F. Jonathan, K. Guerin, and G. D. Hager, “Costar: Instructing collaborative robots with behavior trees and vision,” in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 564–571.
- [139] Z. Wang, B. Xu, and H.-J. Zhou, “Social cycling and conditional responses in the Rock-Paper-Scissors game,” *Scientific Reports*, vol. 4, no. 1, p. 5830, Jul. 2014.
- [140] F. Alonso-Martin, S. Carrasco-Martínez, J. J. Gamboa-Montero, E. Fernández-Rodicio, and M. Á. Salichs, “Introducing psychology strategies to increase engagement on social robots,” in *Social Robotics*, F. Cavallo, J.-J. Cabibihan, L. Fiorini, A. Sorrentino, H. He, X. Liu, Y. Matsumoto, and S. S. Ge, Eds. Cham: Springer Nature Switzerland, 2022, pp. 378–387.
- [141] T. Bocklisch, J. Faulkner, N. Pawlowski, and A. Nichol, “Rasa: Open source language understanding and dialogue management,” *ArXiv*, vol. abs/1712.05181, 2017.

- [142] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, “A comprehensive overview of large language models,” *arXiv preprint arXiv:2307.06435*, 2023.
- [143] OpenAI, “Chatgpt: Language model developed by openai,” <https://openai.com>, 2024.
- [144] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “Vqa: Visual question answering,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.
- [145] A. Fan, M. Lewis, and Y. Dauphin, “Hierarchical Neural Story Generation,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, I. Gurevych and Y. Miyao, Eds. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 889–898. [Online]. Available: <https://aclanthology.org/P18-1082/>
- [146] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra, “Diverse beam search: Decoding diverse solutions from neural sequence models,” *arXiv preprint arXiv:1610.02424*, 2016.
- [147] B. Li, K. Zhang, H. Zhang, D. Guo, R. Zhang, F. Li, Y. Zhang, Z. Liu, and C. Li, “Llava-next: Stronger llms supercharge multimodal capabilities in the wild,” 2024.
- [148] A. Suhr, M. Lewis, J. Yeh, and Y. Artzi, “A corpus of natural language for visual reasoning,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017, pp. 217–223.
- [149] J. Li, K. Pan, Z. Ge, M. Gao, H. Zhang, W. Ji, W. Zhang, T.-S. Chua, S. Tang, and Y. Zhuang, “Empowering vision-language models to follow interleaved vision-language instructions,” *CoRR*, 2023.
- [150] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan *et al.*, “Deepseek-v3 technical report,” *arXiv preprint arXiv:2412.19437*, 2024.

- [151] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, “Qwen3 technical report,” *arXiv preprint arXiv:2505.09388*, 2025.
- [152] F. Jiang, “Identifying and mitigating vulnerabilities in llm-integrated applications,” Master’s thesis, University of Washington, 2024.
- [153] O. Canévet, W. He, P. Motlicek, and J.-M. Odobez, “The MuMMER Data Set for Robot Perception in Multi-party HRI Scenarios,” Aug. 2020, pp. 1294–1300, iSSN: 1944-9437. [Online]. Available: <https://ieeexplore.ieee.org/document/9223340>
- [154] I. D. Gebru, S. Ba, X. Li, and R. Horaud, “Audio-visual speaker diarization based on spatiotemporal bayesian fusion,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 5, pp. 1086–1099, 2017.
- [155] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh *et al.*, “Mixed precision training,” *arXiv preprint arXiv:1710.03740*, 2017.
- [156] A. Ben-Youssef, C. Clavel, S. Essid, M. Bilac, M. Chamoux, and A. Lim, “UE-HRI: a new dataset for the study of user engagement in spontaneous human-robot interactions,” in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ser. ICMI ’17. New York, NY, USA: Association for Computing Machinery, Nov. 2017, pp. 464–472. [Online]. Available: <https://doi.org/10.1145/3136755.3136814>
- [157] A. B. Youssef, G. Varni, S. Essid, and C. Clavel, “On-the-fly detection of user engagement decrease in spontaneous human-robot interaction using recurrent and deep neural networks,” *International Journal of Social Robotics*, vol. 11, pp. 815 – 828, 2019.
- [158] G. Hoffman and X. Zhao, “A primer for conducting experiments in human-robot interaction,” *J. Hum.-Robot Interact.*, vol. 10, no. 1, oct 2020. [Online]. Available: <https://doi.org/10.1145/3412374>

- [159] C. Bartneck, “Godspeed Questionnaire Series: Translations and Usage,” C. U. Krägeloh, M. Alyami, and O. N. Medvedev, Eds. Cham: Springer International Publishing, 2023, pp. 1–35. [Online]. Available: https://doi.org/10.1007/978-3-030-89738-3_24-1
- [160] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, “G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences,” *Behavior Research Methods*, vol. 39, no. 2, pp. 175–191, May 2007. [Online]. Available: <https://doi.org/10.3758/BF03193146>
- [161] M. Schrum, M. Ghuy, E. Hedlund-botti, M. Natarajan, M. Johnson, and M. Gombolay, “Concerning Trends in Likert Scale Usage in Human-robot Interaction: Towards Improving Best Practices,” *J. Hum.-Robot Interact.*, vol. 12, no. 3, pp. 33:1–33:32, Apr. 2023. [Online]. Available: <https://dl.acm.org/doi/10.1145/3572784>
- [162] L. J. Cronbach, “Coefficient alpha and the internal structure of tests,” *Psychometrika*, vol. 16, no. 3, pp. 297–334, Sep. 1951. [Online]. Available: <https://doi.org/10.1007/BF02310555>
- [163] M. Tavakol and R. Dennick, “Making sense of Cronbach’s alpha,” *International Journal of Medical Education*, vol. 2, pp. 53–55, Jun. 2011. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4205511/>
- [164] H. L. O’Brien, P. Cairns, and M. Hall, “A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form,” *International Journal of Human-Computer Studies*, vol. 112, pp. 28–39, Apr. 2018. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1071581918300041>
- [165] F. Wilcoxon, “Individual comparisons by ranking methods,” *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945. [Online]. Available: <http://www.jstor.org/stable/3001968>
- [166] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, “Crema-d: Crowd-sourced emotional multimodal actors dataset,” *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

- [167] S. R. Livingstone and F. A. Russo, “The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english,” *PloS one*, vol. 13, no. 5, p. e0196391, 2018.
- [168] A. Quattoni and A. Torralba, “Recognizing indoor scenes,” in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 413–420.
- [169] S. Bubeck, V. Chadrsekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg *et al.*, “Sparks of artificial general intelligence: Early experiments with gpt-4,” 2023.
- [170] D. G. Bonett and T. A. Wright, “Cronbach’s alpha reliability: Interval estimation, hypothesis testing, and sample size planning,” *Journal of Organizational Behavior*, vol. 36, no. 1, pp. 3–15, 2015. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/job.1960>
- [171] G. De Simone, L. Greco, A. Saggese, and M. Vento, “Robot to human interaction with multi-modal conversational engagement,” in *2025 IEEE International Conference on Simulation, Modeling, and Programming for Autonomous Robots (SIMPAN)*. IEEE, 2025, pp. 1–8.
- [172] —, “Multimodal audio-visual emotion recognition for social robotics,” in *Computer Analysis of Images and Patterns*, M. Castrillón-Santana, C. M. Travieso-González, O. Deniz Suarez, D. Freire-Obregón, D. Hernández-Sosa, J. Lorenzo-Navarro, and O. J. Santana, Eds. Cham: Springer Nature Switzerland, 2026, pp. 245–255.
- [173] G. De Simone, L. Greco, A. Saggese, and M. Vento, “Integrating visual and audio cues for emotion and gender recognition: A multi modal and multi task approach,” *Information Fusion*, vol. 130, p. 104071, 2026. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1566253525011339>
- [174] G. De Simone, L. Roveda, A. Saggese, and M. Vento, “Towards proactive social robots: Distilling visual knowledge from large vision-language models,” in *Proceedings of the*

- IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2025, pp. 2492–2502.
- [175] G. De Simone, A. Saggese, P. Foggia, and M. Vento, “Deep learning based empty shelf detection based on autonomous mobile robot,” *Computer Vision and Image Understanding*, vol. 265, p. 104697, 2026. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1077314226000640>
- [176] G. De Simone, A. Greco, F. Rosa, A. Saggese, and M. Vento, “Context-aware data augmentation for enhanced speech command recognition in industrial environments,” *Scientific Reports*, vol. 15, no. 1, p. 17445, 2025.
- [177] V. Solfrizzi, C. Chiapparino, G. Castellano, B. De Carolis, D. Lofrese, G. Palestra, A. Toma, L. Perla, S. Massaro, M. T. Santacroce *et al.*, “Sister: Social robots to support biopsychosocial frailty,” in *Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI*, 2025, pp. 1–6.
- [178] G. De Simone, P. Foggia, F. Rosa, A. Saggese, M. Vento *et al.*, “Evolving roles of intelligent robots in social and industrial domains,” in *Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI*, 2025, pp. 1–6.

List of Figures

1.1	The illustration provides a clear depiction of the essential components necessary for the operational framework of a social robot. The presence of the video camera icon and the audio icon signifies that the tasks can be addressed using either video data, audio data, or both, as required.	6
2.1	The JAMES socially aware robot bartender [37] . . .	14
2.2	The figure below illustrates the system architecture proposed in [37], representing the core components of the "James Project". This architectural framework highlights the primary modules and their interactions within the system. The robot's interaction with the external environment commences with data acquisition through its various sensors. These sensory inputs are then processed by modules responsible for speech recognition and video analysis. The output from these modules is subsequently directed to the Social State Recognizer , a critical component that discerns which users from the environment wish to engage with the robot. Following this initial analysis, the Social Skills Executor comes into play, responsible for the implementation of the engagement policy. It processes the information from the "Social State Recognizer" and selects the specific user with whom the robot should engage. Finally, the robot's response and actions are executed through its actuators, translating the chosen interaction into tangible and perceptible behaviors. . . .	15
2.3	General architecture of the gaze estimation system proposed by [39]	17

2.4	Architecture of the system proposed by [40]	18
2.5	The overarching structure of our multi-modal engagement prediction model encompasses three distinct phases: feature extraction, feature processing, and engagement classification.	23
2.6	The architecture of the method proposed in the study by Zhang et al. [28] is structured around a neural network divided into two distinct subnetworks, each specialized in video and audio analysis, respectively. The video analysis subnetwork is designed to process input data in the form of a 4-dimensional tensor. This tensor encapsulates temporal information, height, width, and the channels of each frame. These components are integral to the analysis and understanding of video content. Conversely, the audio analysis subnetwork operates on a different type of input, specifically a 13-dimensional MFCC.	24
2.7	Architecture of TalkNet [48]	25
2.8	The figures illustrate the steps involved in computing both the cross-attention and the self-attention , respectively.	26
2.9	Architecture of the proposed method	28
2.10	Taxonomy of methods presented in the literature for addressing the problem of multi-engagement.	30
3.1	Complete Behavioral Tree of the proposed system, used to model the robot's behavior during interaction with users.	54

- 3.2 High-level architecture of the proposed system. Video and audio data gathered by sensors are processed by the *perception* layer, which includes modules for detecting and tracking users' faces in the scene, estimating head pose for each user, detecting speakers, recognizing hand gestures, estimating the direction of arrival of sound (DOA), and detecting voice activity (VAD). The processed data is then passed to the *reasoning* layer, which comprises a Finite State Machine (FSM) and a Behavior Trees (BTs), together with a dialogue management module. This layer models the system's behavior by considering both perceptual data and interaction history. Finally, the output is transmitted to the robot's actuators to perform actions such as arm movements, head rotations and robot speech. 55
- 3.3 Proposed architecture for solving the multi-engagement problem. The system takes both video and audio data as input. Face detection and tracking are applied to all users within the scene, followed by two deep neural networks that perform detection of *active speakers* and head pose estimation. This information is then given as input to a rule-based classifier to determine which user to engage with, effectively managing the conversation between multiple users and robot. 56

- 3.4 Overall architecture of the proposed MAGNET. The audio encoder processes MFCC features using Conv(1+1)D operations for feature extraction. In parallel, the video encoder receives a video clip as input, applies uniform temporal subsampling along the temporal axis, and extracts features using 3D convolutions and bottleneck layers. For each convolutional block, the number of input and output channels for the video and audio encoders are denoted as C_{v-in} , C_{a-in} and C_{v-out} , C_{a-out} , respectively. The embeddings generated by both encoders are temporally aggregated using pooling operations and then normalized. The resulting feature vectors are concatenated along the embedding dimension and progressively reduced using a series of 1D Residual Blocks. The final representation is passed to the classification head. 60
- 3.5 Finite State Machine (FSM): the high-level controller designed to manage the system's modalities. Each state within the FSM is associated with a specific Behavior Tree (BT). The conversation modality serves as the initial state of the FSM. Transitions between modalities can be triggered by user requests, which can be performed either through voice commands or by pressing the physical robot buttons (bumpers). 67
- 3.6 Behavior Tree for conversational modality. This BT models the robot's behavior during interactions with the user. It specifically implements multi-engagement strategies and formulates responses to user requests. 69
- 3.7 Multi-Engagement Tree. This BT manages the system's ability to handle multiple engagements. Specifically, it uses data obtained from the perceptive layer to select the appropriate user to engage in the scene. Additionally, it determines whether to disengage the currently engaged user, taking into account the context and the state of the conversation. 70
- 3.8 Speech To Text Tree. This Behavior Tree is designed to await a speech request from the user and subsequently invokes the necessary services to transcribe the input audio into text. 71

- 3.9 Natural Language Processing Tree: This BT gathers contextual data from the scene, including the image of the environment, soft biometrics of users, and identified objects, along with the transcribed user request. It then calls the dialogue server, passing this information, and subsequently manages the server's response. 72
- 3.10 Text to Speech Tree: This Behavior Tree receives textual input and invokes the TTS service to convert the text into audio, which is then played through the robotic platform. 73
- 3.11 **Game Modality Tree:** This Behavior Tree (BT) models the system's behavior during the "rock, paper, scissors" game with the user. During the initialization step, the robot explains the game rules to the user. The Speech to Text, Text to Speech, and multi-engagement tasks are handled as previously discussed in the reported section. The "Game Logic" subtree manages the gameplay by detecting the user's hand gesture, determining the winner, and communicating the result. 74
- 3.12 Behavior Tree (BT) managing the gameplay. Initially, it verifies whether the rules have been explained, then starts the *rock, paper, scissors* animation by moving the robot's arms. Subsequently, it recognizes the user's hand gesture and randomly selects the robot's move. Finally, the winner is determined and communicated to the user using both visual cues (icons displayed on the tablet) and auditory feedback (speech and specific sounds). 75
- 3.13 List of the allowed game moves: *rock, paper, scissor*. 76
- 3.14 Tablet view during the game session. The right side displays the video feed captured by the camera, along with the detected game move chosen by the user (in this specific case it is paper). On the left side, the option selected by the proposed system is shown, which is scissors in this case. 77

3.15	Label generation pipeline. A video sample from the dataset is provided as input to the VLM teacher to generate contextual captions. These captions are then passed to the LLM teacher, which produces an augmented set of context-aware sentences. The resulting dataset consists of video samples paired with multiple associated sentences.	81
4.1	Image from the MIVIA-ME dataset illustrating two users interacting with the robot. The user on the left is gazing at the robot but not speaking, as highlighted by a yellow rectangle, while the user on the right is both gazing at and speaking to the robot, indicated by a green rectangle.	90
4.2	Questions from the Godspeed Questionnaire assessing user perceptions of <i>Likeability</i> and <i>Perceived Intelligence</i>	94
4.3	Pepper robot in action during the <i>Maker Faire</i> and <i>Futuro Remoto</i> exhibition.	99
4.4	Accuracy of the proposed MAGNET solution, by varying the number of frames (in the range 4-32), and compared with the different backbones architectures and multimodal baselines.	115
4.5	Mean token accuracy on the validation set.	119
4.6	Web interface of the evaluation framework. Evaluation instructions are displayed in the blue box at the top. Below, a video player presents the video sample to be assessed. The sentence to be evaluated appears in the red box, followed by a 5-point Likert scale for rating. At the top of the web interface, the following labeling instructions are displayed: "Evaluate the sentence generated by a social robot to greet user(s). The system uses the video below to generate a greeting sentence. Press play and watch the video. Then, read the sentence shown in the red box titled <i>Greeting Sentence</i> , and provide your evaluation using the Likert scale below."	121

List of Tables

2.1	Results of the method proposed in [28]	25
3.1	MAGNET-A architecture. The model takes as input an MFCC representation of the audio, where B denotes the batch size, C_a the number of input channels (initialized to 1), $MFCC$ the number of MFC Coefficients, and T_a the temporal dimension. For each layer, the number of input channels C_{a-in} and output channels C_{a-out} are reported. For pooling operations, kernel size, stride, and padding parameters are specified.	62
3.3	Design of the task-specific branches. For each layer, the transformation of features is reported, specifically illustrating how the input embedding size E_{av} is progressively reduced and how the task-specific representation E'_{av} is subsequently expanded.	64
4.1	Results obtained for Precision (P), Recall (R), and F1 score (F1), using different model settings: (1) only the <i>head pose</i> module, (2) only the <i>active speaker</i> module, and (3) both the <i>head pose</i> and the <i>active speaker</i> modules.	91
4.2	System performance on MIVIA-ME dataset in terms of <i>Precision</i> , <i>Recall</i> , and <i>F1-score</i>	92
4.3	Summary of the results obtained for the constructs <i>Engagement (ENG)</i> , <i>Likeability (LI)</i> , and <i>Perceived Intelligence (PI)</i> . The table reports the mean and standard deviation (SD) for Group 1 (Baseline system) and Group 2 (Proposed system), along with the p-values, z-scores obtained from the statistical tests, and the corresponding effect size and statistical power.	97

4.4	Summary of the survey responses for each question, including statistics such as mean, standard deviation (std), minimum (min), maximum (max), and quartiles (25%, 50%, and 75%).	100
4.5	Results for the emotion recognition and gender recognition tasks in a single-task framework, both single modality and multi modality. For each metric, the values corresponding to the best-performing models are highlighted in bold. Note that M stays for Million, K stays for Kilo.	106
4.6	Results for the emotion recognition and gender recognition tasks in a multitask soft-sharing framework, both single modality and multi modality. For each metric, the values corresponding to the best performing models are highlighted in bold. Note that M stays for Million.	108
4.7	Results for the emotion recognition and gender recognition tasks in a multitask hard-sharing framework, across single modality and multi modality. For each metric, the values corresponding to the best performing models are highlighted in bold. Note that M stays for Million.	111
4.8	Comparison of the proposed MAGNET and its single task variant with state of the art approaches on CREMA-D dataset. Several metrics are reported, depending on the ones available for that specific method.	112
4.9	Comparison of the proposed MAGNET with state of the art approaches on RAVDESS dataset. Several metrics are reported, depending on the ones available for that specific method.	114
4.10	Emotion and Gender accuracy by varying the α value in GradNorm.	115
4.11	Inference time comparison among different models, ordered by average time (expressed in second). . . .	117
4.12	Mean Token Accuracy (MTA) achieved by the evaluated models on test set at varying levels of temporal video context T	120
4.13	Summary of the survey responses for each question, including statistics such as mean, standard deviation (std), and quartiles (25%, 50%, and 75%).	122

- 4.14 Memory usage and inference time (mean and standard deviation) for different configurations of the LLaVA-OneVision models on the NVIDIA Jetson Orin AGX device. Results are reported for varying numbers of input frames T and generation lengths (#Toks).124