University of Modena and Reggio Emilia
Department of Engineering "Enzo Ferrari"

XXXVII cycle of the Doctorate School in
Information and Communication Technologies

Integration of Vision and Language for Physical and Cognitive Human-Robot Interaction

Ph.D. Dissertation

Niyati Rawal

Advisors: Prof. Rita Cucchiara, Prof. Lorenzo Baraldi
Director of the School: Prof. Luigi Rovati

Modena, 2025

Advisors:

Prof. Rita Cucchiara	University of Modena and Reggio Emilia
Prof. Lorenzo Baraldi	University of Modena and Reggio Emilia

Director of the School:

Prof. Luigi Rovati	University of Modena and Reggio Emilia

Review Committee:

Prof. Abhinav Valada	University of Freiburg
Prof. Lamberto Ballan	University of Padova

The works described in this dissertation were supported by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie Action project, "Personalized Robotics as Service Oriented Applications" (PERSEO) Grant agreement No. 955778, and carried out in Information and Communication Technologies, at the AImageLab research laboratory of the University of Modena and Reggio Emilia.

This dissertation was typeset by the author using $\LaTeX\ 2_\varepsilon$, originally developed by Leslie Lamport and based on Donald Knuth's $\TeX$. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface.

In memory of my mentor, Daisaku Ikeda, and my grandfathers, Jagdish Rawal and Tilak Raj Ajmani.

Daisaku Ikeda shared the philosophy on living one's life positively with the world. My grandfathers always encouraged me in my studies since childhood and would have been proud of me.

Integration of Vision and Language for Physical and Cognitive Human-Robot Interaction

Abstract

While many researchers study computer vision, natural language processing, or robotics, the works proposed here lie at the intersection of these three domains. In this dissertation, two domains for applying Human-Robot Interaction (HRI) that combine vision and language are explored, namely physical HRI and cognitive HRI. For physical HRI, the task of Vision and Language Navigation (VLN) is considered. In VLN, there is an agent that can perceive the 360-degree view of the environment (vision) and has to follow the language instructions of the human such as "Go to the kitchen and clean the coffee table". For cognitive HRI, the task of multimodal empathetic dialogue generation is considered. In this task, input signals from facial expressions (vision) and the text of what the human says (language) are provided. The agent should respond to the human empathetically by considering these two multimodal input signals.

The first two works are related to physical HRI. The first work proposes a method to improve the navigation performance of an agent by augmenting already existing VLN datasets such as REVERIE. Specifically, a speaker model that generates language instructions for a sequence of images (for example, "Go to the sofa and bring me the remote control.") using an adversarial approach is proposed. In the second work, the speaker model is extended to generate dialogue whenever the navigation agent gets confused regarding where to go next.

Large Language Models (LLMs), such as ChatGPT, have become popular but these models are prone to giving long and neutral answers to assist humans in one way or another. The works proposed on cognitive HRI introduce ways to make artificial agents respond empathetically to humans. In the first work for cognitive HRI, an agent replies with parallel or reactive empathy to the human with a

certain facial expression and the text of what is said. Specifically, a Transformer encoder-decoder structure is used to respond to the human empathetically. The second work also consists of an agent that learns to respond to humans empathetically. However, this work makes use of only the Transformer decoder model to generate the dialogue response and the model is trained using Reinforcement Learning (RL) to respond in a manner that would make the human feel positive.

Finally, a generalized VLN agent is proposed as a work in progress. This agent can summarize a trajectory given a sequence of images, navigate and perform embodied questions and answering. To summarize, approaches based on Transformer models are proposed to enhance the performance of VLN agents for physical HRI tasks. Transformer models were also finetuned to learn to respond to humans empathetically for cognitive HRI tasks. While the two domains of physical HRI and cognitive HRI are kept segregated, ideally, a robot with general intelligence should be able to clean the house or bring a particular object (physical HRI) and be a social companion engaging in empathetic dialogue (cognitive HRI). In the future, a computational model that could perform both physical HRI and cognitive HRI could be developed to investigate how these two fields can interplay.

Keywords: Vision, Language, Robotics, Deep Learning, AI

Integrazione di visione e linguaggio per l'interazione fisica e cognitiva uomo-robot

Sommario

Mentre molti ricercatori studiano la computer vision, linguaggio o la robotica, i ricerche qui proposti si pongono nell'intersezione di questi tre domini. In questa dissertazione, verra' analizzata l'interazione tra umano e robot in due domini di applicazione differenti che combinano visione e linguaggio, ovvero l'interazione uomo-robot fisica e quella cognitiva. Per quanto riguarda l'interazione fisica uomo-robot, viene presa in considerazione la navigazione tramite visione e linguaggio. Nella navigazione tramite visione e linguaggio, un agente è in grado di percepire la vista a 360 gradi dell'ambiente e deve seguire le istruzioni linguistiche dell'uomo, come ad esempio "Vai in cucina e pulisci il tavolino". Per quanto riguarda l'interazione cognitiva uomo-robot, si considera la generazione di dialogo empatico multimodale. In questo setting, vengono forniti segnali di input dalle espressioni facciali (visione) e dal testo dettato dall'umano (linguaggio). L'agente deve rispondere all'uomo in modo empatico considerando i suddetti due segnali di input multimodali.

Verranno inizialmente presentati due ricerche riguardanti l'interazione fisica uomo-robot. Il primo propone un metodo per migliorare le prestazioni di navigazione di un agente, estendendo i dataset di navigazione tramite visione e linguaggio già esistenti, come REVERIE. In particolare, viene proposto un modello di speaker che genera istruzioni linguistiche per una sequenza di immagini (ad esempio, "Vai sul divano e portami il telecomando"). Nel secondo lavoro, il modello di speaker viene esteso per generare un dialogo ogni volta che l'agente di navigazione si confonde sulla direzione da prendere.

I Large Language Models, come ChatGPT, sono diventati popolari, ma questi modelli sono inclini a fornire risposte lunghe e neutre, per assistere gli esseri umani.

I lavori proposti sull'interazione cognitiva tra umano e robot introducono modi per assicurarsi che gli agenti artificiali rispondano in modo empatico agli esseri umani. Nel primo lavoro sull'interazione cognitiva uomo-robot, un agente risponde con empatia parallela o reattiva all'uomo con una certa espressione facciale e il testo di ciò che viene detto. In particolare, viene utilizzata una struttura Transformer encoder-decoder per rispondere all'umano in modo empatico. Anche il secondo studio consiste in un agente che impara a rispondere agli esseri umani in modo empatico. Tuttavia, questo lavoro utilizza solo il modello di Transformer decoder per generare la risposta al dialogo e il modello viene addestrato utilizzando il Reinforcement Learning per rispondere in modo da migliorare lo stato emotivo dell'umano.

Infine, viene proposto un agente di navigazione generalizzato con visione e linguaggio come work in progress. Questo agente può descrivere una traiettoria data una sequenza di immagini, navigare e generare risposte alle domande. In sintesi, sono stati proposti approcci basati su modelli di Transformer per migliorare le prestazioni degli agenti di navigazione tramite visione e linguaggio per compiti di interazione fisica uomo-robot. I modelli di Transformer sono stati messi a punto anche per imparare a rispondere agli esseri umani in modo empatico per i compiti di interazione cognitiva uomo-robot. Sebbene i due domini dell'interazione fisica uomo-robot e dell'interazione cognitiva uomo-robot siano tenuti separati, idealmente un robot con intelligenza generale dovrebbe essere in grado di pulire la casa o di portare un particolare oggetto e di essere un compagno sociale con capacità di dialogo empatico. In futuro, si potrebbe sviluppare un modello computazionale in grado di eseguire sia l'interazione fisica uomo-robot che l'interazione cognitiva uomo-robot, per studiare come questi due campi possano interagire.

Parole chiave: Visione, Linguaggio, Robotica, Deep Learning, IA

Contents

1
Introduction

No longer confined to the research community, advanced technologies like Artificial Intelligence (AI) and Deep Learning are now widely embraced by the general public. Several people who are not even from an AI background can now conveniently chat with artificial agents like ChatGPT that are capable of responding to most of the human prompts intelligently. ChatGPT has been trained on large amounts of textual data and the goal of the language model is to generate the next token. It is based on the decoder part of the Transformers model and can provide large amounts of text by generating the response token-by-token.

After the massive success of language models like ChatGPT, recently, there has been a growing focus on the development of multimodal Vision and Language models. Vision and Language models are now able to perform tasks such as image captioning, visual question and answering etc. Even roboticists are utilizing these language models (like ChatGPT) to teach robots to perform various tasks. Combining the modalities of vision, language and agency, this dissertation addresses the integration of Vision and Language modalities to perform Human-Robot Interaction (HRI).

While HRI can be classified into several domains, physical Human-Robot Interaction (pHRI) and cognitive Human-Robot Interaction (cHRI) are two of the most common and broad in terms of their definition. pHRI is when the robot has to make physical movements such as those of navigation or manipulation in a shared human-robot environment. On the other hand, in case of cHRI the robot has to understand human behavior, intentions and emotions and interact with the human in a natural and adaptive manner. It applies knowledge from cognitive science, psychology, and AI to develop robots that engage naturally and effectively with humans. By aligning robot behavior with human cognition, cHRI aims to create more intuitive, empathetic, and trustworthy human-robot partnerships.

The task of Vision-and-Language Navigation (VLN) is considered for pHRI. In Vision and Language Navigation, given a 360-degree view of the environment and language instructions such as "Go to the sofa and bring me the remote.", the robot has to navigate in the environment while following the instructions. It should be noted that the task is defined such that the robot has not previously explored the environment and does not know the location of the rooms or the kind of furniture that is there in the indoor environment.

For cHRI, the task of multimodal empathetic dialogue is considered. In this task, the robot perceives the facial expression of the person and listens to the language prompt and responds to the person in an empathetic manner. For example, if someone says "I am not feeling very well today" and looks sad, the robot would respond "I am sorry to hear that. Can I do something to make you feel better?".

Although in this dissertation two very different applications of pHRI and cHRI are considered, in terms of technicality, vision and language models are fine-tuned for most of the works presented here. The works here utilize the open source language models like Bert, GPT-2, Llama2 and Llama3 etc. Different approaches are proposed like adversarial training using BERT as a discriminator and GPT-2 as a generator for captioning image sequences, as well as reinforcement learning using facial expressions (as implicit human feedback) among others. Although the two domains of pHRI and cHRI are kept segregated here, in future, computational models can be developed to perform both kinds of tasks of VLN as well as multimodal empathetic dialogue generation.

1.1 Problem Statement

To come up with a generalized agent that can perform pHRI tasks such as VLN as well as cHRI tasks such as multimodal empathetic dialogue generation, years of research is required. In this dissertation, VLN and multimodal empathetic dialogue generation are treated as two separate tasks. Various methods to improve the performance of a VLN agent are proposed. This is followed by computational models to generate multimodal empathetic dialogue.

VLN is a complex task that involves an agent to perceive the 360-degree view of the environment and follow human language instructions such as “Go to the living room and clean the dining table”, while navigating in an indoor environment. As this task involves different modalities such as vision, language and robotics, the task is difficult and is hardly considered solved. The performance of language models have improved significantly in the recent years and this trend is currently being followed by the introduction of different Vision and Language models. Despite this, combining Vision and Language with robotics adds another layer of complication to the problem. The success rate of a robot to be able to navigate in an indoor environment accurately is still low.

It is still difficult for a VLN agent to reason about it’s environment effectively. One way to improve the accuracy of a VLN agent is to augment the VLN dataset and train the agent with a large dataset. Yet another way is to make the agent ask a question regarding where it should go next, whenever it gets confused. Another challenge is to build a generalized VLN agent that can summarize a trajectory and answer questions other than navigating.

With the advent of ChatGPT, humans are now able to talk to artificial agents. However, the responses given by ChatGPT are often neutral and aimed at assisting the human in one way or another. While simply plugging in ChatGPT into a robot would make the robot converse intelligently, this robot would not be able to make an emotional connection with its user. This inspires the need for an agent that can generate empathetic responses by considering a multimodal signal coming from a human's facial expressions and the text of what is being said.

This dissertation is a step towards solving the above problems of improving the performance of VLN agent and suggest ways to propose computational models for multimodal empathetic dialogue generation.

1.2 Organization

This dissertation presents ways to improve the performance of VLN agent and to interact with an artificial agent empathetically. Both these tasks integrate Vision and Language for HRI. In the following, a brief description of the organization of the work is given.

In Chapter 2, a detailed description of the current literature on Vision and Language Navigation and multimodal empathetic dialogue is presented. These works are going to be relevant throughout the thesis.

The following two chapters are on Vision and Language Navigation. In Chapter 3, a method to improve the performance of Vision and Language Navigation is presented. Specifically, a speaker model is introduced. This speaker model can describe the trajectory of the agent. By augmenting the data with the instructions provided by the speaker model for 128k unlabeled trajectories, it has been shown that the performance of navigation improves.

Chapter 4 contains an agent that can prompt for the dialogue whenever it gets confused regarding where to navigate. This way the agent does not always keep taking actions in the physical space unknowingly and the performance of navigation improves with the help the agent gets by the generated question and answer pair.

The next two Chapters are on multimodal empathetic dialogue generation. In Chapter 5, a robot that learns to respond to a human with parallel or reactive empathy is presented. This robot considers the facial expressions and the text of what the human says to respond empathetically.

Chaper 6 presents a robot that learns to generate a response to make the human feel positive. This computational model is trained using Reinforcement Learning (RL). It assigns a positive reward if the agent's response makes the human feel positive and a negative reward if the agent's response makes the human feel negative.

The following Chapter 7 introduces a generalized VLN agent in a continuous environment. This generalized agent can navigate, summarize a trajectory as well as perform embodied question and answering. By performing various kinds of tasks for navigation scenarios, the generalized agent for navigation can perform the individual tasks better as well.

Finally, the conclusions with some personal considerations and future work are presented in Chapter 8.

2

Literature Survey

This chapter presents a literature overview of the relevant work related to the tasks and settings proposed in this dissertation. First, the works related to VLN tasks are presented. This includes a Section on VLN (Section 2.1) and Vision and Dialogue Navigation (Section 2.2). These sections are followed by the section on text generation for visual navigation in Section 2.3. The section on multimodal Vision and Language models (Section 2.4) is relevant both for VLN and for multimodal empathetic dialogue generation studies. Section 2.8 contains a review of the literature on Embodied Question Answering (EQA) that is also linked to VLN. This is followed by the related works on empathy in social robots in Section 2.5 expressing the need to introduce multimodal empathetic dialogue in robots. Finally, Sections 2.6 and 2.7 present the overviews of Generative Adversarial Networks (GANs) and Reinforcement Learning (RL), respectively. While GAN is a specific approach used for a speaker model to improve the performance of VLN, RL is used for multimodal empathetic dialogue generation and is based on emotions of humans.

2.1 Vision and Language Navigation

In recent years, research aimed at the development of intelligent autonomous agents has acquired increasing interest with the release of simulation platforms like Gibson [145], Matterport3D [27], and Habitat [117], as well as datasets enabling object interaction [49, 99, 121]. Among the various embodied tasks that are the object of this research line, Vision-and-Language Navigation (VLN) aims to implement such agents with multimodal reasoning capabilities in both indoor and outdoor environments. In fact, VLN requires an agent to interpret human instructions, in the form of natural language text, while perceiving observations of the environment.

Among indoor VLN methods, [13] first tackled the task by adopting sequence-to-sequence long short-term memories for action inference. [46] started exploiting the panoramic observation space and introduced a module for the generation of synthetic instructions. [47] instead, used counterfactual thinking to perform data augmentation. More recently, [86, 87] proposed a model with a self-monitoring agent, and [71] used dynamic convolution filters. RCM [142] used a reinforcement learning training approach to improve cross-modal matching and [61] implemented graphs to model the relations between scenes, objects, and instructions. More recently, Transformer-based [135] models have become popular. Among these approaches, VLN↻BERT [63] models time dependencies using a recurrent BERT [39], while PTA [70] and HAMT [30] used Transformers to respectively perform multimodal fusion and exploit episode history. Topological maps and a dual-scale Transformer are proposed by [32] to consider both long-term action planning and fine-grained understanding. Some methods for VLN that tried to address this lack are proposed by [97] and [34]. However, [97] used preset language-assisted routes, and [34] limited the agent interaction to only one possible question, and the response given by the oracle is the next action on the shortest path to the goal, while our approach only exchanges textual information. Moving on to outdoor VLN approaches, the agent has to perform navigation in an urban environment where the visual appearance is more repetitive and clear landmarks are difficult to find. While StreetLearn [94] is the first dataset providing panoramic views of the streets of Manhattan and Pittsburg for navigation, it does not provide human-annotated instructions but only provides directions and street names toward the target location. Touchdown dataset [29, 93] introduces human instructions for a subset of the StreetLearn dataset. Another large-scale dialogue dataset is called "Talk The Walk" [38] and involves two agents (a "guide" and a "tourist") that communicate in natural language to achieve a common goal.

Current research on VLN methods is presented in Chapters 3 and 4.

2.2 Vision and Dialogue Navigation

Constraining the navigation in VLN to follow human instructions that are given only at the beginning of each episode could lead the agent to diverge from the correct trajectory when the match between instruction and visual cues is not clear. In this context, extending the task by allowing the agent to generate conversations with an oracle asking for new instructions could redirect the agent in the correct direction to the goal. However, this relaxation of the VLN task introduces new challenges defined by the generation of an appropriate question and by the decision of the most suitable moment for such interaction. The benchmark used to evaluate dialogue-based agents is defined by the contribution of [129], which introduced Cooperative Vision and Dialogue Navigation (CVDN), a dataset of over 2K embodied trajectories with human-human dialogues in the simulated indoor environments of Matterport3D [27], and Navigation from Dialog History (NDH), a task of 7K navigation episodes using CVDN dialogues as textual input. In particular, the CVDN dataset is annotated using two humans, a navigator and an oracle, where the first has to navigate toward a predefined target object while being able to ask the oracle for directions, and the oracle can access the shortest path trajectory from the current position of the navigator to the target. However, most of the existing studies tackling VDN use the dialogue only as an input for the navigation method [13, 30, 55, 104, 154, 157]. In these approaches, the agent does not generate dialogue. On the contrary, RMM [113] designed three agents, two of them are entitled of producing a dialogue aimed at a target object regularly, while the third is in charge of the navigation. [156] proposed a computational model that engages in dialogue only when the navigating agent is unsure of which action to take. However, the generated dialogue is based on textual templates and consists of questions that have affirmative or negative answers, with the navigation agent that is rewarded for producing questions that have "yes" as the answer. Yet another work introduces a model VISITRON that learns when to navigate and when to ask questions [122].

Research on Vision and Dialogue Navigation is written in Chapter 4.

2.3 Text Generation for Visual Navigation

To feed a sentence as an input in a language model, it is converted into discrete tokens. The goal of the language model is to predict the next token. Early language models consisted of RNNs [114] and LSTMs [60] could generate short texts. With the advent of LLMs, language models are now able to generate large amounts of text. RNNs have a hidden state vector that allows them to retain all the previous information about a sequence. As the sequence grows longer, RNNs are unable to make predictions based on the initial part of the sentence. This is due to the vanishing gradients. To solve this problem of vanishing gradients in RNNs, LSTMs were proposed. LSTMs have a forget gate, that controls the amount of information that needs to be remembered or forgotten. Later, the transformer models were introduced that have an attention mechanism to focus on particular words and use them to predict the predict the next word [135]. Transformer models directly take the whole sentence as an input, unlike RNNs and LSTMs that consider one word as an input at a time. The traditional transformer architecture comprises of an encoder-decoder structure. However, the LLMs contain only the decoder part of the transformer.

The idea of generating synthetic text for visual navigation has arisen naturally from the goal of improving the performance of a VLN agent. In fact, from the early work on VLN, a specific line of research focused on augmenting human-annotated datasets with well-formed synthetic instructions [46, 90]. For example, [155] converted the instructions provided by the Google Maps API in the

StreetLearn dataset to human-like instructions using a text-style transfer approach, showing improvements for outdoor VLN agents. Another line of research uses speaker models to generate textual instructions using sequences of images belonging to navigation trajectories. This framework can also be extended to unlabelled environments, as shown by [31]. Synthetically augmented datasets have been proven to improve the performance of navigation agents on several VLN datasets [31, 46, 53, 90, 141, 155].

Research on text Generation for navigation is written in Chapters 3,4 and 7.

2.4 Multimodal Vision and Language Models

Image captioning models, such as the Show and Tell model by Vinyals et al. [138], utilized encoder-decoder architectures where a Convolutional Neural Network (CNN) encodes the image and a Recurrent Neural Network (RNN) generates the caption. Subsequent models improved upon this by incorporating attention mechanisms, allowing the model to focus on specific regions of the image while generating each word [147].

In parallel, Visual Question Answering (VQA) emerged as a challenging task requiring models to reason about images and understand complex queries [14]. Models like the Stacked Attention Network applied hierarchical attention to relate question words with image regions [149]. The introduction of the VQA dataset spurred significant advancements by providing a large-scale benchmark.

The success of transformers in language modeling [135] inspired their application in multimodal contexts. Vision Transformers adapted the transformer architecture for image recognition tasks [11]. Combining these approaches, models like ViLT [68] processed images and text using a unified transformer architecture without relying on CNNs.

Vision and Language understanding models such as CLIP, GLIP etc. comprise of models that that can comprehend visual information in addition to language [75, 105]. Most of these models have zero-shot capabilities and excel at classification tasks. GLIP emphasizes object-level alignment through phrase grounding [75]. Some models deploy transformer architecture for performing diverse vision-language tasks, including VQA and image retrieval [17]. In the recent times, Vision and Language Understanding is not just limited to images but is also getting extended to videos [130, 146]. GPT-4V [9], Llava [83], Flamingo [10], PALM-E [43] are some of the models that use multimodal input and generate text. Recent models such as Video-Llama, Video-ChatGPT etc. use videos instead of images and text as an input to generate an output text [88, 152]. There are few works that generate multimodal output with multimodal input [127, 128, 144].

Research related to multimodal Vision and Language models is written in Chapters 3, 4, 5, 6 and 7.

2.5 Empathy in Social Robots

In recent years, robots have been developed beyond industrial applications where humans coexist in common spaces [151] such as in care homes [118, 137], at receptions [20, 54], as educational tutors [41, 100], supporting in household chores [150] etc. Unlike industrial robots, robots with social components are expected to be empathetic toward humans and other social agents with whom they share a common space. Understanding human affective cues and responding appropriately to humans is a requisite for any social robot. Various research explores the affective modeling of humans through the lens of social robots [26, 56, 59, 133]. Among various affective cues that provide information regarding human emotional states, the facial expression of emotion [76, 89] is salient. Correspondingly, affective behaviors [110, 116] generated by robots play significant roles when developing longitudinal relationships in social environments. Due to the prerequisites of understanding affective cues and behavior generation, social robots leverage insights from psychology and social sciences into engineering to ensure such robots can understand and respond appropriately to human expression of emotions [111]. It has been shown that social robots that are empathetic toward humans are considered friendly in interaction [73]. The facial expressions on social robots have been shown [40] to enhance recognition of expressions of emotion and improve the perception of warmth and attractiveness without creating uncanny valley effects. Apart from the ability to understand human affective cues and behavior generation, the physical aspect of social robots impacts the human's cognitive load [134].

In HHI, there is parallel empathy and a reactive empathy. Some researchers have tried to model parallel and reactive empathy in virtual agents and robots [15, 92]. For normal emotion intensity, the type of empathetic reaction, parallel or reactive is determined by the type of emotion [15]. For example, positive emotions i.e. happiness and surprise are responded with parallel empathy towards the user and negative emotions i.e. anger, fear, sadness and disgust are responded with reactive empathy towards the user.

Research related to empathy in social robots is written in Chapters 5 and 6.

2.6 Generative Adversarial Networks

Generative Adversarial Networks (GANs), introduced by Goodfellow et al. [51], have emerged as a transformative framework in generative modeling. GANs consist of two competing neural networks: a generator that generates synthetic data and a discriminator that evaluates its authenticity. This training paradigm has spurred significant advancements across various domains, including image synthesis and Natural Language Processing (NLP). It has proven remarkably effective in producing high-fidelity images and videos. Beyond the visual domain, recent work has applied GANs to speech synthesis, music generation, and even protein design, illustrating the versatility of GANs. While GANs have shown immense success in generating realistic data, especially in computer vision, their application to natural language processing (NLP), however, presents unique challenges due to the discrete nature of text, making gradient-based updates non-trivial. Over the years, researchers have proposed various approaches to adapt GANs for NLP tasks, including text generation, style transfer, etc. Previous works rely on REINFORCE [35] or using Gumbel-Softmax [120] to allow for backpropagation through discrete NLP tokens. Recently, Transformer networks were used in an adversarial manner, however, these approaches were used for the generation of images [64, 66].

Research related to GANs is written in Chapters 3.

2.7 Reinforcement Learning in NLP

In Reinforcement Learning (RL), an agent gets a positive or a negative reward when it takes certain actions in an environment. The goal is to maximize the reward by learning the kind of actions would give positive rewards to the agent. RL has become a prominent framework in natural language processing (NLP) for tasks that require sequential decision-making, optimization over long horizons, or the incorporation of human feedback.

Large Language Models (LLMs) that are trained on web-scale data use Reinforcement Learning with Human Feedback (RLHF) [125]. Humans provide explicit feedback on how they find the response of the agent. LLMs are then trained using Proximate Policy Optimization (PPO) and reward the agent positively if they find the response useful or negatively if the response is not useful. The training is done in a three step procedure. First, an LLM is trained or fine-tuned. This is followed by the fine-tuning of a reward model. Finally, the trained or fine-tuned model is trained using PPO where the rewards come from the fine-tuned reward model.

Research related to reinforcement learning is provided in Chapter 6.

2.8 Embodied Question Answering

Embodied Question Answering (EQA) extends the capabilities of Visual Question Answering (VQA) by introducing an embodied agent that actively navigates its environment to gather information and answer questions. In contrast to static VQA, EQA emphasizes active perception and decision-making within a dynamic context. Das *et al*. [36] formally introduced the EQA task, defining it as one where an agent starts in an unknown environment, receives a question (e.g., "What color is the car in the garage?"), and must physically explore its surroundings to provide the answer. This task inherently requires the integration of visual perception, navigation, and language understanding within a single system.

One key challenge in EQA is effective environment representation. Initial works relied on grid-based or pixel-based map representations, but recent advances have introduced semantic mapping techniques that allow agents to reason about objects and their relationships [28]. Semantic maps enable hierarchical planning, allowing agents to navigate more efficiently by focusing on task-relevant areas of the environment.

Another critical aspect of EQA is question grounding, where agents must interpret natural language queries into actionable goals. Early approaches utilized pre-defined templates for question understanding, but more sophisticated methods now leverage transformers and large language models to dynamically parse complex queries [121]. These systems can handle nuanced instructions that involve multi-step reasoning or conditional logic.

Interactive Question Answering (IQA), a related subfield, further expands on EQA by introducing physical interaction as part of the reasoning process. Gordon et al. [52] demonstrated IQA tasks where agents manipulate objects to uncover hidden information (e.g., opening a cabinet to identify its contents). This line of research highlights the importance of combining physical and cognitive skills for embodied agents.

Research related to EQA is presented in Chapter 7.

3

Vision and Language Model to Improve Navigation in an Indoor Environment

In the last decade, remarkable results in Natural Language Processing, Computer Vision, and Robotics have been witnessed [21]. More recently, increasing interest has been devoted to research at the intersection of these three domains [2, 22, 72]. In line with this trend, this chapter focuses on the Vision-and-Language Navigation (VLN) task. When performing VLN, an agent or a robot can perceive the 360° view of the environment and receive human instructions such as "*Go to the living room and bring me the remote on the table*". The agent has to follow the instructions and navigate in an unknown environment to reach the specified goal and stop there.

The development of VLN agents has significant implications for real-world applications, including assistive robots, autonomous vehicles, etc. These agents have the potential to operate in diverse environments, where they could assist in-

This Chapter is related to the publication "N. Rawal *et al.*, AIGeN: An Adversarial Approach for Instruction Generation in VLN, CVPRW 2024" [5]. See the list of publications on page 93 for more details.

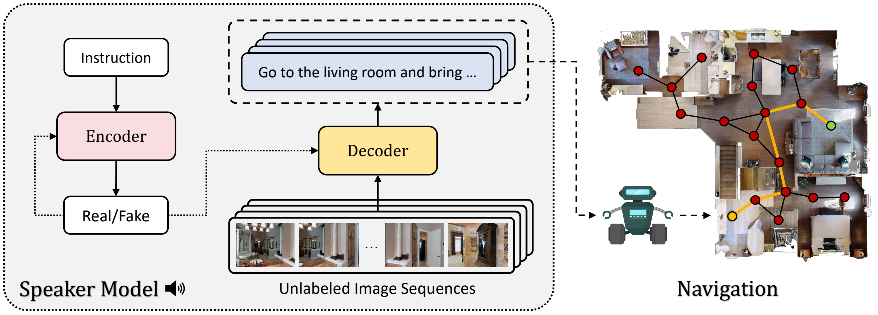


Figure 3.1: AIGeN is a novel GAN-like model for generating instructions given a sequence of images. Synthetic instructions can be used as training data for a VLN model to improve its navigation performance.

dividuals by guiding them through complex spaces, searching for specific objects, or helping with everyday tasks based on spoken or written instructions. The key challenges in VLN lie in integrating language comprehension with spatial awareness and decision making in real-time. Agents must recognize objects, understand spatial relationships, and interpret linguistic cues within the context of their environment.

It is not easy to find large datasets for VLN that are labeled. Human-generated instructions needed for training VLN architectures are difficult, costly, and time consuming to obtain. The resulting lack of annotated training data is one of the key factors in making VLN a challenging task. Recent work [31, 46, 53, 141] has focused on generating instructions at a lower cost by employing methods for the generation of synthetic instruction. For example, Guhur *et al.* [53] and Chen *et al.* [31] showed that generating synthetic instructions and augmenting the data, improves the navigation performance of the agent. Nevertheless, Guhur *et al.* [53] used image-caption pairs from the web on a prohibitive total number of 140K environments. Instead, Chen *et al.* [31] generated synthetic instructions using trajectories sampled on HM3D dataset [109] which is composed of 900 environments; the approach proposed in this chapter aims at improving the quality of the generation of such instructions.

In this chapter, AIGeN, a novel computational model that can generate synthetic instructions for unlabeled navigation paths, is proposed (see Figure 5.1

for an overview). The proposed model combines a multimodal Generative Pre-Trained Transformer (GPT) [23] and Bidirectional Encoder Representations from Transformers (BERT) [39] in an adversarial manner to generate high-quality instructions. In particular, the model consists of a Transformer decoder (GPT-2) that generates sentences describing the agent's path, using a sequence of images from the environment and the associated object detections extracted using Mask2Former [33]. The BERT-like encoder instead serves as a discriminator and is trained to distinguish between real and fake instructions for a given sequence of images. This is the first approach that combines Transformer networks and a GAN-like training procedure to generate synthetic navigation instructions.

Using this approach, the training data of REVERIE and R2R datasets are augmented using 217K Habitat-Matterport 3D Dataset (HM3D) trajectories and it is shown that AIGeN-generated instructions help to improve the results of a VLN model achieving state-of-the-art performance. Furthermore, the quality of the proposed method is validated by evaluating the generated instructions using image description metrics [124] and comparing the downstream navigation performance of different model configurations, showing that producing well-formed synthetic instructions is beneficial for the training of a navigation agent.

In summary, the contributions of this work are as follows. A novel computational model, AIGeN, was proposed. This model learns to generate synthetic instructions for a given sequence of images in an adversarial manner. By generating instructions for 217K unlabeled trajectories in the HM3D dataset, an improvement in the performance of an off-the-shelf VLN method is shown.

3.1 Proposed Method

AIGeN is used to describe an agent's trajectory in natural language and is composed of an instruction generator (visually depicted in Fig. 3.2) and an instruction discriminator. Both models are trained simultaneously using an adversarial approach, which aims to improve the performance of the generator and the quality of the generated instructions. The general model is shown in Fig. 5.2.

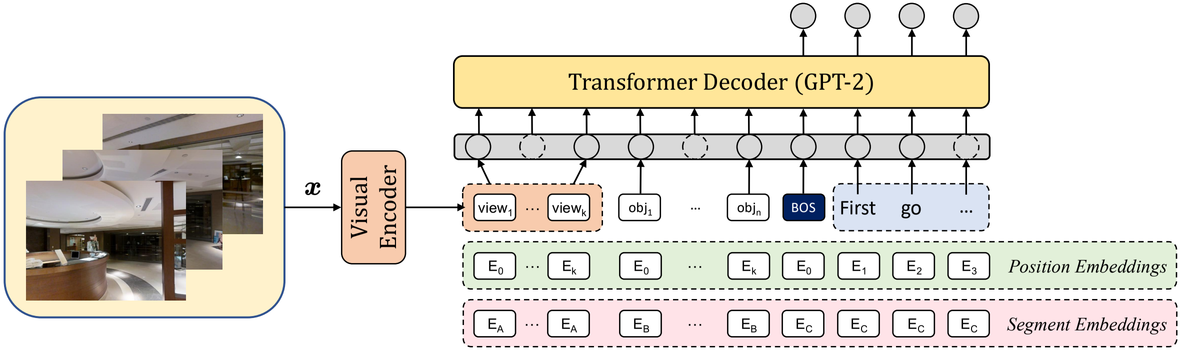


Figure 3.2: Architecture of the instruction generator model that generates VLN instructions. The model is fed with a sequence of images and object names, and prompted to generate instructions using a `BOS` token. The instructions are generated token by token until the `EOS` token.

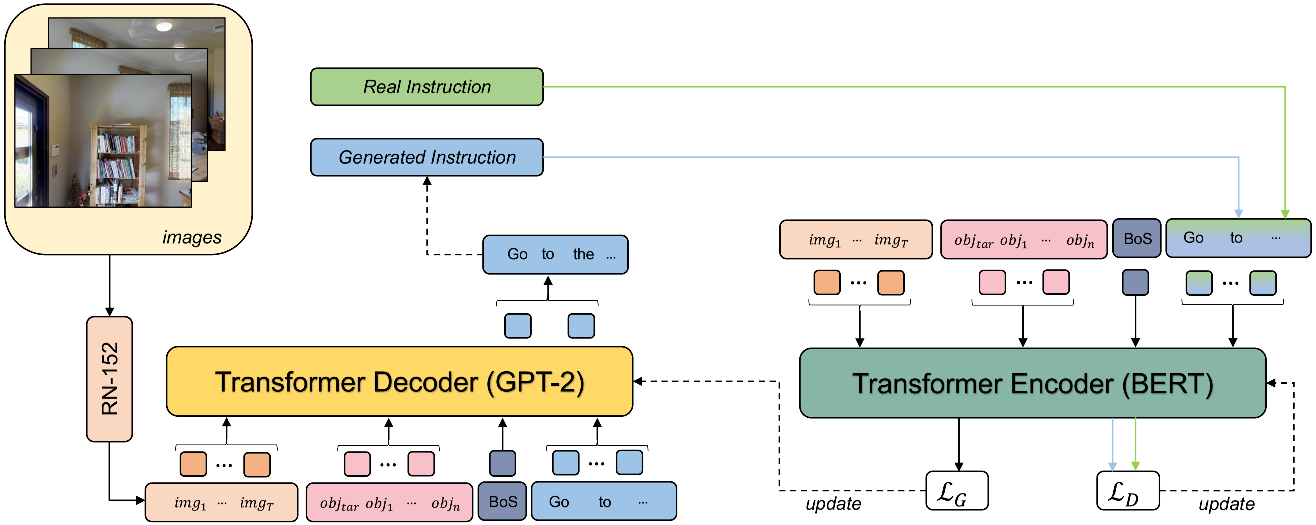


Figure 3.3: Schema of the proposed generative-adversarial framework for synthetic instruction generation. The GPT-2 decoder acts as a generator while the BERT encoder acts as a discriminator. Both models are trained simultaneously. The generator generates fake instructions token-by-token until it reaches the EOS token. The discriminator must detect whether the instructions corresponding to a given sequence of images are real (ground truth) or fake (generated).

3.1.1 Instruction Generator

The generator is a language model that is responsible for describing the actions needed to reach the target location. The proposed model uses a pre-trained language model that is fine-tuned conditioning on visual inputs to achieve multi-modal capabilities similar to Alayrac *et al.* [10]. However, instead of just using the GPT-2 decoder for instruction generation, it is used in combination with BERT in a GAN-like manner to generate better-quality instructions that are more diverse.

The overall approach takes the images from the environment the agent has to traverse and feeds them into a pre-trained ResNet-152 [57] encoder to obtain image features. Next, when generating instructions for Habitat-Matterport 3D (HM3D) environments, all objects in the given sequence of images are extracted using Mask2Former [33] trained on ADE20K to enrich the visual features with object names. This input is fed into the decoder along with the first `BOS` token using the concatenation of visual features and object names as a prompt for the language model. The GPT-2 decoder is trained to predict subsequent language tokens to generate a complete language instruction that describes the actions of the agent until the last image. Differently from [10], AIGeN flattens the tokens of both images and text before feeding them to the GPT-2 decoder. Furthermore, to effectively segregate visual information from textual information, position and segment embeddings are used in addition to tokens, as shown in Fig. 3.2. Formally,

$$y = \mathtt{GPT}\left(\left[\underbrace{v_0, .., v_t}_{\text{Images}}, \underbrace{o_{tgt}, o_0 .., o_n}_{\text{Objects}}, \mathtt{BOS}, \underbrace{i_1, .., i_m}_{\text{Instruction}}, \mathtt{EOS}\right]\right) \tag{3.1}$$

where $(v_0, ..., v_t)$ denotes the set of visual features for images of the trajectory, o_{tgt} indicates the target object label, $(o_0, ..., o_n)$ denote the names of the objects in the last image, `BOS` and `EOS` are begin of string and end of string tokens respectively. Consequently, $(i_1, ..., i_m)$ denotes the tokens that correspond to the instruction.

The result of the generation is defined starting from the output of the decoder following the `BOS`, autoregressively generating the instructions token by token and allowing backpropagation through the sampled tokens using the Gumbel-Softmax [120] trick.

3.1.2 Instruction Discriminator

The inputs of the BERT encoder are image features, tokens of the names of objects in the images, and an instruction. Similarly to the inputs for the GPT-2 decoder, the images are first fed into a pre-trained ResNet-152 [57] visual encoder to obtain image features. Then, a fully connected layer followed by a sigmoid function is used to process the results of the BERT output. The output is simply real or fake; real if the given instruction matches the given sequence of images, and fake otherwise. Finally, binary cross-entropy loss is used to minimize the error between the actual output as real or the generated output as fake. A well-trained discriminator should classify ground-truth instructions as real and generated instructions as fake. Like the decoder, the BERT input employs position embeddings and segment embeddings in addition to token embeddings.

3.1.3 Adversarial Training

The proposed method, shown in Fig. 5.2, follows an adversarial training approach, where the generator G is trained to fool the discriminator D, while the discriminator is taught to distinguish between real and fake instructions. The generator is trained to generate instructions as close to ground-truth instructions as possible by minimizing the cross-entropy loss between the generated instructions and the ground-truth instructions. The generator loss is defined as:

$$\mathcal{L}_G = -\log(D(I_G, \boldsymbol{x})), \tag{3.2}$$

where $I_G \in G(\boldsymbol{x})$ is the generated instruction and $\boldsymbol{x}$ is the sequence of images belonging to the trajectory.

The discriminator has to discriminate between ground-truth instructions as real instructions and generated instructions as fake instructions. Consequently, the discriminator loss is:

$$\mathcal{L}_D = -\log(1 - D(I_G, \boldsymbol{x})) - \log(D(I_R, \boldsymbol{x})), \tag{3.3}$$

where $I_R \in R(\boldsymbol{x})$ is the ground-truth instruction. Training is performed simultaneously on both the generator and the discriminator.

AIGeN-Llama [3] extends AIGeN by leveraging Llama's dual capabilities as both generator and discriminator, enabling adversarial training for high-quality instruction generation was also proposed.

3.2 Experimental Setup

3.2.1 Datasets

Experimental results are presented using the model proposed in Section 3.1 using two VLN datasets as a reference, REVERIE [103] and R2R [13]. Both datasets consist of navigation sequences composed of 360° images that are collected on the nodes of navigation graphs on Matterport3D environments [27] and each navigation sequence is associated with three ground-truth textual instructions. The main difference between the two datasets is that REVERIE instructions include interactions with specific target objects that the model is required to identify, whereas R2R instructions only specify the trajectory to be followed. In this study, only the frontal view of 360° images, with a field of view of 60° is considered.

3.2.2 Model Configuration

The training of AIGeN uses a learning rate of 0.0002, a batch size of 6, and Adam [69] as an optimizer. For the GPT-2 decoder, a medium-sized version is used with $L = 12$, $d = 768$, $H = 12$, where L is the number of layers, d is the model dimensionality, and H is the number of attention heads, unless otherwise stated. The visual features used by the model are extracted using ResNet-152. For the BERT encoder, a hidden size of 768, 12 layers, and an intermediate size

Table 3.1: VLN metrics for our approach and competitors on REVERIE Val Unseen. * denotes finetuning on our computing architecture.

	Val Unseen					
	TL	SPL↑	SR↑	OSR↑	RGS↑	RGSPL↑
Seq2Seq [13]	11.1	2.8	4.2	8.1	2.2	1.6
SMNA [86]	9.1	6.4	8.2	11.3	4.5	3.6
RCM [142]	12.0	7.0	9.3	14.2	4.9	3.9
FAST-MATTN [103]	45.3	7.2	14.4	28.2	7.8	4.7
SIA [82]	41.5	16.3	31.5	44.7	22.4	11.6
Airbert [53]	18.7	21.9	27.9	34.5	18.2	14.2
ProbES [80]	18.0	22.8	27.6	33.2	16.8	13.9
VLN↻BERT [63]	16.8	24.9	30.7	35.0	18.8	15.3
HAMT [30]	14.1	30.2	33.0	36.8	18.9	17.3
DUET [32]	22.1	33.7	47.0	51.1	32.2	23.0
KERM [79]	21.9	35.4	50.4	55.2	34.5	24.5
AZHP [48]	22.3	36.6	48.3	53.6	34.0	25.8
HM3D-AutoVLN* [31]	24.3	39.4	54.0	**60.0**	34.6	25.2
DUET + AIGeN	19.5	**41.9**	**54.4**	57.7	**35.1**	**26.9**

of 3072 are used. Overall, AIGeN has 289M model parameters and is trained for $\approx$ 36 hours on a single NVIDIA RTX6000 GPU.

3.2.3 Evaluation Criteria

The navigation performance was evaluated using the following navigation metrics: trajectory length in meters (TL); success rate (SR), *i.e.* the fraction of episodes in which the agent can reach the goal position within 3 meters; oracle success rate (OSR), that is, the success rate using an oracle stop policy; success rate weighted by path length (SPL); and navigation error (NE). The object grounding ability of the agents on the REVERIE dataset is evaluated using remote grounding success (RGS), which is the ratio of successfully followed instructions, and RGS weighted by path length (RGSPL).

Instead, to quantitatively evaluate the model on text generation, metrics commonly used for image description are used, namely BLEU [101], METEOR [16], ROUGE [81], CIDEr [136], and SPICE [12]*. All these metrics are obtained by comparing the predicted instruction with the ground truth instruction in terms of its n-grams (where an n-gram is a sequence of n consecutive words).

*Image description metrics are computed using the code provided at the following link: `https://github.com/tylin/coco-caption`

Table 3.2: VLN experiments on the Val Unseen split of R2R dataset.

	Val Unseen				Test Unseen			
	TL	SPL↑	SR↑	NE↓	TL	SPL↑	SR↑	NE↓
Seq2Seq [13]	8.39	-	22	7.81	8.13	18	20	7.85
PRESS [78]	10.36	45	49	5.28	10.77	45	49	5.49
SSM [139]	20.70	45	62	4.32	20.39	46	61	4.57
EnvDrop [126]	10.70	48	52	5.22	11.66	47	51	5.23
PREVALENT [55]	10.19	53	58	4.71	10.51	51	54	5.30
RelGraph [61]	9.99	53	57	4.73	10.29	52	55	4.75
ProbES [80]	11.58	55	61	4.03	12.43	56	62	4.20
Airbert [53]	11.78	56	62	4.01	12.41	57	62	4.13
VLN↻BERT [63]	12.01	57	63	3.93	12.35	57	63	4.09
MARVAL [67]	10.15	61	65	4.06	10.22	58	62	4.18
DUET [32]	13.94	60	72	3.31	14.73	59	69	3.65
KERM [79]	13.54	60	72	3.22	14.74	59	70	3.61
HAMT [30]	11.46	61	66	**2.29**	12.27	60	65	3.93
AZHP [48]	14.05	61	72	3.15	14.95	60	**71**	3.52
DUET + AIGeN	13.72	**63**	**73**	2.92	14.20	**61**	**71**	**3.33**

While all these metrics are commonly used for evaluating cross-modal description, only CIDEr and SPICE have been specifically designed for this task. BLEU, METEOR, and ROUGE have indeed been proposed to evaluate translation and summarization. According to recent literature, CIDEr showcases the best alignment with human judgment.

3.3 Experimental Results

3.3.1 Results on VLN

The navigation experiments on VLN are performed considering an off-the-shelf state-of-the-art VLN method fine-tuned with the instructions generated by the proposed approach. DUET [32] which is based on a dual-scale graph Transformer, is adopted to perform long-term action planning and fine-grained cross-modal understanding by exploiting topological maps built during the episode. The computational model DUET has a two-step training, as described by Chen *et al*. [32], that is composed of a pre-training phase on four auxiliary tasks and fine-tuning of the VLN task. The pre-training phase is performed following [31], and then fine-tune the pre-trained model by augmenting the REVERIE and R2R training splits with the synthetic instructions generated by the proposed approach.

Table 3.3: Image description experiments using our model with different size configurations and without using the detections on REVERIE.

	Val Seen				
Model	BLEU-1	METEOR	ROUGE	CIDEr	SPICE
AIGeN (Medium)	**0.487**	0.222	0.457	0.869	0.318
AIGeN w/o detect.	0.409	0.188	0.416	0.358	0.196
AIGeN	0.484	**0.228**	**0.465**	**0.890**	**0.329**
	Val Unseen				
Model	BLEU-1	METEOR	ROUGE	CIDEr	SPICE
AIGeN (Medium)	0.412	0.166	0.378	0.461	0.213
AIGeN w/o detect.	0.357	0.141	0.347	0.132	0.123
AIGeN	**0.421**	**0.179**	**0.393**	**0.486**	**0.228**

Table 3.4: Experimental comparison of VLN performance of different configurations of our model on REVERIE dataset.

	Val Unseen					
	TL	SPL↑	SR↑	OSR↑	RGS↑	RGSPL↑
AIGeN (Medium)	20.0	41.2	54.3	**60.1**	**35.1**	26.5
AIGeN w/o detect.	23.4	38.9	52.5	59.4	34.6	25.7
AIGeN	19.6	**41.9**	**54.4**	57.7	**35.1**	**26.9**

Table 3.5: Evaluation of the diversity of synthetic instructions using AIGeN before and after the GAN fine-tuning compared with the ground truth instructions on REVERIE training split.

	%Novel	Unigrams	Bigrams	Div-1	Div-2
Ground truth	-	3675	21551	0.019	0.113
AIGeN w/o adv. training	22.7%	2970	15928	0.016	0.086
AIGeN	100.0%	14783	43013	0.072	0.210

AIGeN-generated instructions are produced using 217K randomly sampled trajectories in the Habitat-Matterport 3D (HM3D) [109] dataset, a large-scale dataset of indoor photorealistic environments. The same paths released by Chen *et al*. [31] are used, and a synthetic instruction is generated for each sequence of observations. The navigation model is pre-trained and fine-tuned for a total of $\approx$ 32 hours, 12 and 20 respectively, on a single NVIDIA RTX6000 GPU, and is compared with current state-of-the-art VLN models on REVERIE and R2R.

The navigation results on REVERIE in Table 3.1 show that the proposed approach achieves state-of-the-art performance on SPL and both object grounding metrics while remaining competitive on the other metrics. In particular, the instruction set gives a boost in generating effective trajectories towards the goal, which is reflected by SPL and RGSPL. For example, the proposed approach shows an improvement of 8.2 and 3.9 on SPL and RGSPL with respect to the baseline DUET. For fairness of comparison, HM3D-AutoVLN was retrained on the computing infrastructure following [31] (denoted HM3D-AutoVLN*). The comparison between HM3D-AutoVLN* and the proposed approach highlights the effectiveness of AIGeN-generated instruction, as the difference between the two methods in VLN is defined by the quality of the synthetic instructions used to fine-tune DUET. When running this comparison, the proposed approach outperforms HM3D-AutoVLN* on the main navigation metrics, and in particular on the success rate weighted by path length (SPL) shows an improvement of 2.5.

Furthermore, analyzing the experiments on the R2R dataset in Tab. 3.2, the described method provides state-of-the-art results on SPL and SR for both validation and test unseen splits, proving that the propsed approach can generate synthetic instructions that are beneficial for multiple VLN datasets. When comparing the proposed method with the baseline DUET, the AIGeN-generated instructions allow an improvement in the SPL of 3.0 and 2.0, respectively, on the validation "unseen" and the test "unseen" splits. Overall, these results support the claim that well-formed synthetic instructions help the agent to learn better navigation and object localization in a VLN setting.

3.3.2 Ablation Study

The components of the proposed approach are validated by comparing different configurations of AIGeN on synthetic instructions generation in Table 3.3 and on navigation in Table 3.4.

Starting from Table 3.3, the first row shows the performance of the generator that is trained from scratch. As can be seen, training the instruction generation model from scratch provides a maximum CIDEr of 0.869 on the "seen" split and 0.378 on the "unseen" split, without reaching the results of the overall method. Following, the proposed model is ablated in terms of input modalities analyzing the contribution given by the object detections, and comparing it with a model that is trained using only visual features and textual instructions. All metrics related to AIGeN without detections, especially CIDEr and SPICE, are considerably lower than the metric values computed for AIGeN. This result confirms the importance of employing object detections as input features. The speculation is that when no object words are provided, the model is not able to identify which object in the scene it has to attend to. Therefore, the target objects in the generated sentences are often different from those in the ground-truth instructions even when the landmark is correctly recognized.

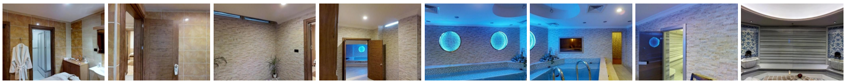
Moving on to the navigation experiments in Table 3.4, the performance of the VLN model is compared on the REVERIE dataset that augments the training data with the instructions generated by different configurations of the proposed approach. In this case, AIGeN surpasses its counterparts by an important margin on all the metrics. The improvement in the SPL over the Medium configuration that is trained from scratch is 0.7, and it becomes 3.0 when considering AIGeN w/o detections. The ablation study validates the effectiveness of the model components in both text generation and downstream vision and language navigation.

3.3.3 Diversity Analysis

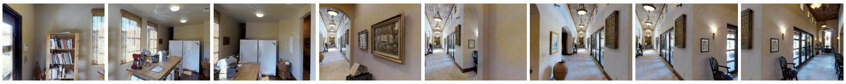
To assess the quality of synthetic instructions and measure the diversity of the generation, in Table 3.5 diversity metrics was computed commonly used for image captioning [124, 140] comparing the diversities of AIGeN-generated instructions with the ground-truth annotations of the trajectories of REVERIE training split. The comparison is performed by producing synthetic instruction using AIGeN before and after the adversarial fine-tuning phase.

The metrics used for this study are the ratio of the number of novel sentences, *i.e.* not contained in the dataset, to the number of ground-truth sentences, the number of unique words (Unigrams); the number of unique couples of consecutive words (Bigrams); Div-1; and Div-2. Div-1 and Div-2 are, respectively, the ratio of unique unigrams and bigrams to the total number of unigrams.

Looking at the results, it is evident that the fine-tuning phase using an adversarial approach helps to generate instructions that do not retrace the ground-

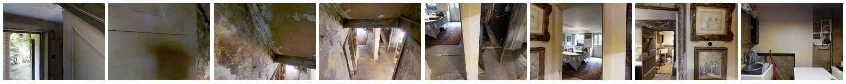


(a) **GT**: *Go to the white room with 4 basins opposite the lounge pool and turn on the faucet on the left closest to the entrance*
AIGeN: *Walk to the bathroom with the white sink and turn on the faucet*



(b) **GT**: *Go to the hallway with the black chairs on level 2 and sit in the second chair from your right*
AIGeN: *Go to the dining room and bring me the chair that is nearest to the door*

(c) **GT**: *Go to the lounge on level 1 and sit down on the armchair*
AIGeN: *Go to the dining room and with a lint roller clean the arm chair which is against the wall*



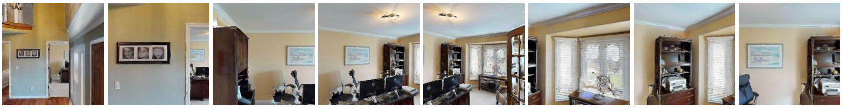
(d) **GT**: *Go to the laundry room on the first floor and sort out the cluttered cleaning products on the dryer*
AIGeN: *Go to the laundry room on level 1 and bring me the clothes on the rack opposite the dryer*

Figure 3.4: Sample image sequences from REVERIE Val Unseen split with corresponding ground-truth instruction and synthetic instructions generated using AIGeN. The images in each sequence have been reduced to 8 to facilitate the graphical presentation and only the frontal image of the panoramic observation at each timestep is shown.

truth instruction, with the consequence of improving the diversity of the dataset. In fact, while the generated instructions without using the GAN-like training present a small number of novel sentences, AIGeN returns a completely novel set of instructions. Furthermore, AIGeN with the adversarial fine-tuning can increase the number of unigrams and bigrams sampled from the word dictionary even with respect to the ground-truth annotations. This result is also reflected by the Div-1 and Div-2 metrics that present a significant increase with respect to the ground-truth instructions.

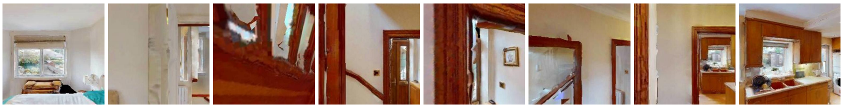
3.3.4 Qualitative Analysis

Finally, Fig. 3.4 and 3.5 show examples of sequences of input images with corresponding ground-truth instructions (if available) and generated instructions using AIGeN.



(e) **AIGeN**: *Go to the office on level 1 and clean the desk*

(f) **AIGeN**: *Go to the kitchen and turn off the light closest to the entrance*



(g) **AIGeN**: *Go to the laundry room and clean the mirror above the sink*

Figure 3.5: Sample image sequences from HM3D with corresponding synthetic instructions generated using AIGeN. The images in each sequence have been reduced to 8 to facilitate graphical presentation, and only the frontal image of the panoramic observation is shown at each timestep.

In Fig. 3.4 all four samples have been taken from the "unseen" validation split of REVERIE, so that AIGeN has never seen these environments during training. Two positive samples are provided in (a) and (b), as well as two negative ones in (c) and (d). For both (a) and (b), the generated instruction is similar to the ground-truth instruction and matches the given sequence of images. In fact, the landmarks, the bathroom in (a) and the dining room in the hallway in (b) are correctly recognized. Furthermore, both synthetic instructions refer to the correct target object, respectively, "faucet" and "chair". In the case of (c) instead, the generated instruction and the ground-truth instruction identify the wrong landmark, with the lounge that is recognized as a dining room by the proposed model. However, the target object "armchair" is still correctly recognized. Finally, in the case of (d), the generated instruction and the ground-truth one refer to different target objects. While the correct target objects are the "cluttered cleaning products" on the dryer, AIGeN refers to the "clothes on the rack" as the target objects. However, the model correctly identified the laundry room.

Moving on to Fig. 3.5, three qualitative samples using the HM3D dataset trajectories are presented. As HM3D is unlabeled, in this case, there is no ground truth annotation available, and only the instructions generated by AIGeN are provided. The first example (e) correctly identified the office and the target object "desk" in the final observation of the sequence. In the second trajectory shown in (f), the target object ("light") in the kitchen is correctly recognized. Finally, in the third trajectory (g), the kitchen is identified as a laundry room, and the correct target object "window" is misidentified as a "mirror".

These results demonstrate that even when the visual quality of the environment is low due to 3D reconstruction, AIGeN is capable of generating valuable instructions for vision-and-language navigation providing correct directions identifying objects and landmarks.

4

Dialogue Generation based on Vision for Cooperative Navigation

In recent years, the advances in Vision-and-Language research have substantially contributed towards the development of the smart embodied agents of the future. As also mentioned previously, VLN consists of an agent following human instructions while perceiving the environment. However, its standard definition forces the agent to follow textual instructions that are received once and only at the beginning of each episode. This formulation restricts the agent's freedom to interact with the surrounding environment during the duration of the navigation. Engaging in dialogue, instead, can help the agent navigate successfully in unknown environments by asking for help when the trajectory to the goal location is unclear. The capability to ask questions regarding its current location and where it should move next is a step towards building an intelligent, conversational agent that can communicate and interact with a human while per-

This Chapter is related to the publication "N. Rawal *et al.*, UNMuTe: Unifying Navigation and Multimodal Dialogue-like Text Generation, Submitted to omputer Vision and Image Understanding" [4]. See the list of publications on page 93 for more details.

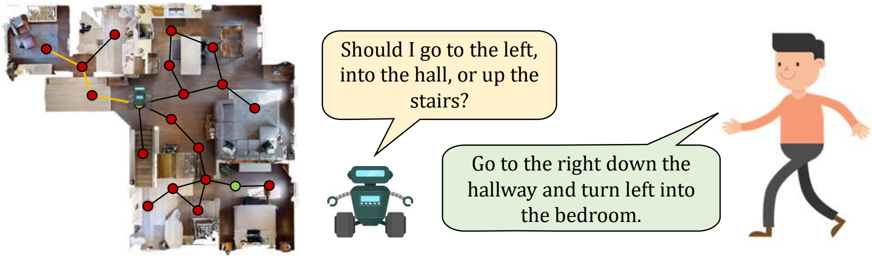


Figure 4.1: A novel computational model is proposed that learns to exchange dialogue during navigation when the agent is unsure of the action it should take in the environment. The proposed model allows the agent to (a) decide when to ask a question, (b) ask target-driven questions, (c) answer given questions, and more importantly, (d) navigate toward the goal.

forming intelligent navigation.

Vision and Dialogue Navigation (VDN) [129] consists of continuous communication and interaction between an agent and an oracle while performing navigation is the most appropriate candidate to achieve this goal. However, in addition to the navigation that is derived from VLN, in VDN some additional aspects need to be addressed: (a) selecting when is the appropriate time to ask a question, (b) deciding which question should be asked, and (c) determining how to answer a given query. In the task of VDN, no instructions are provided at first, but only the name of a target object; however, the agent can query and interact with another agent (the oracle) to gather information on how to navigate in an unseen environment. Nevertheless, most of the previous work in this field does not tackle the generation of dialogue, but performs the navigation task directly training the navigation agent with a human-annotated dialogue between a navigator and an oracle describing the path to a target object. This work differs from these approaches, as the model here is trained to equip a navigation agent with the ability to generate dialogue.

In this chapter, a novel method, called UNMuTe is proposed (see Fig. 4.1 for an overview). The model consists of two main modules: the first performs navigation or chooses whether to engage in dialogue, and the second generates navigation-based dialogue. The navigation part consists of a VLN method [32] that has been adapted to receive dialogue as input and has been equipped with a policy to decide when to generate dialogues. The dialogue part instead consists of a Generative Pre-trained Transformer (GPT-2) [108] model that is modified to generate pairs of questions and answers conditioned on the target object and the current position of the agent. The connection between these two components is given by a decision mechanism that regulates the generation of dialogue and must be based on the confidence of the navigator. When the navigator is uncertain of which direction it has to take, it should ask the oracle for help. Different dialogue activation policies are compared to study the effect of dialogue generation on navigation. In the experimental analysis, the effectiveness of the proposed model is proven using the main datasets on VDN [129], Cooperative Vision and Dialogue Navigation (CVDN) and Navigation from Dialog History (NDH), and proving that the proposed approach achieves state-of-the-art navigation results on this task.

4.1 Proposed Method

UNMuTe (visually depicted in Fig. 4.2 and 4.3) is composed of a navigation model that predicts the actions of the agent and a dialogue model that, when triggered, generates question-and-answer pairs that denote the trajectory to the goal. First, the dialogue model is individually trained so that the model can generate questions and answers. Next, the navigator model is trained with the help of the dialogue model. Specifically, the navigator model can consult the dialogue model when it is confused about which action to take. Given the current observation of the navigator and the target object, the dialogue model generates a question and an answer conditioned on the trajectory to the target. The navigator model uses the output of the dialogue model to select its next action, thereby improving the final navigation performance.

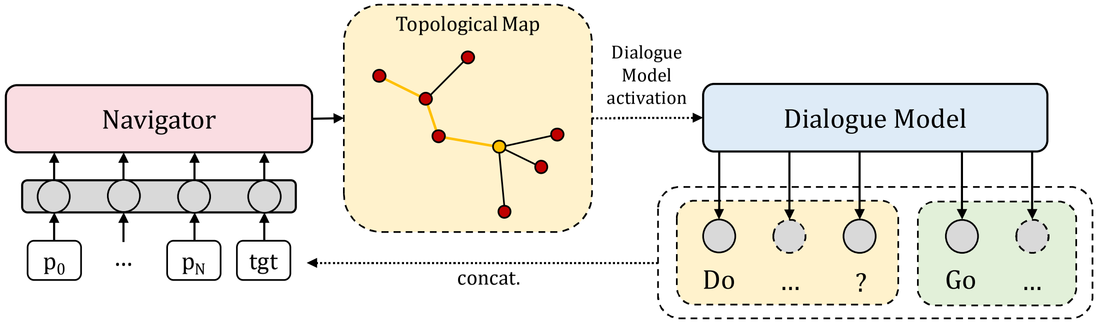


Figure 4.2: UNMuTe consists of a dialogue model that is based on a GPT-2 decoder and a navigation model that is based on a state-of-the-art navigator, *i.e.* DUET [32]. When DUET is uncertain of the action the agent should take, it outputs an action that prompts the dialogue model to generate a question and an answer regarding where the agent should move.

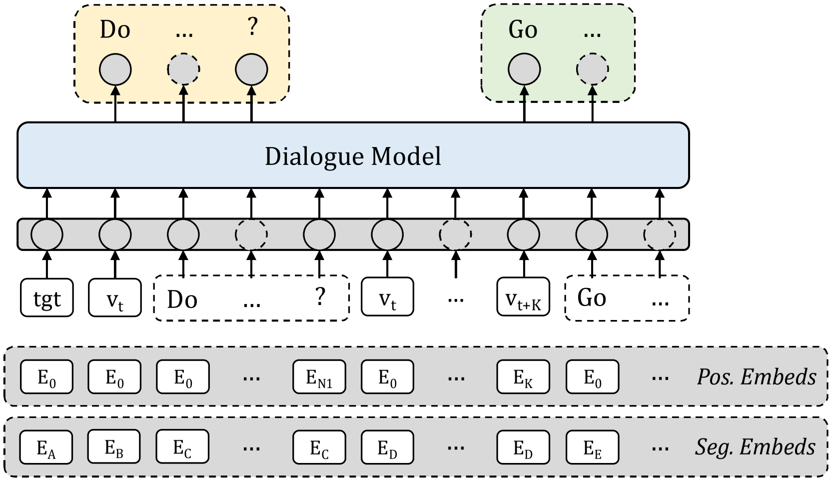


Figure 4.3: Dialogue model with corresponding inputs and outputs. The model is trained to predict the subsequent language token belonging to the sequence. To facilitate graphical presentation, special tokens such as `BOS` or `EOS` are omitted.

4

4

4.1.1 Dialogue Model

The dialogue model, shown in Fig. 4.3, is a single Generative Pre-trained Transformer (GPT-2) that generates question-and-answer pairs starting from the target object and the current observation of the agent. Inspired by [10], the dialogue model is fine-tuned by conditioning visual inputs to achieve multimodal capabilities using the trajectories and conversations contained in the CVDN dataset. The actual input of the dialogue model can be split into three components: the token of the target object label, the image features and textual tokens associated with the question, and the image features and textual tokens associated with the answer. Formally,

$$y = \mathtt{GPT}\Big(\Big[\mathtt{BOS}, \underbrace{o_{\text{tgt}}}_{\text{Target}}, \mathtt{EOS}, v_t, \mathtt{BOS}, \underbrace{q_1, .., q_n}_{\text{Question}}, \mathtt{EOS}, v_t, .., v_{t+k}, \mathtt{BOS}, \underbrace{a_1, .., a_m}_{\text{Answer}}, \mathtt{EOS}\Big]\Big) \tag{4.1}$$

where BOS and EOS are begin of string and end of string tokens respectively, o_{tgt} indicates the target object label, v_t the visual features related to the current observation of the agent, $(q_1, ..., q_n)$ the actual question tokens. Consequently, $(v_t, ..., v_{t+k})$ denotes the set of visual features for future images and $(a_1, ..., a_m)$ the tokens corresponding to the answer.

All the image features used for the dialogue model are extracted using a pre-trained visual encoder. During training, the dialogue model learns to predict the subsequent language token of both the question and the answer, starting from the BOS token. Instead, all the tokens following the image features are ignored. The generation of the question is influenced only by the current observation of the agent, while the answer is conditioned with k additional observations that are collected along the trajectory to the target. The trajectory to the goal is obtained using Dijkstra's algorithm on the navigation graph between the current node and the target node.

In addition to token embeddings, the proposed dialogue model uses position and segment embeddings to effectively segregate information regarding the different components and modalities of the input. This choice was inspired by Bidirectional Encoder Representations from Transformers (BERT) [39]. The segment embeddings in the proposed model are learnable. However, the position embeddings simply range from 0 to the last image, and for word tokens, from 0 to the last word token. During inference, the output of the dialogue model is generated token-by-token autoregressively, starting with the `BOS` token until the `EOS` token of the generated answer is produced.

4.1.2 Navigator Model

The navigator model consists of a modified variant of the Dual Scale Graph Transformer (DUET) [32]. DUET keeps track of the visited and observed nodes by producing a topological map of the environment. At each time step, the map is updated, storing the visual features associated with newly visited nodes and navigable nodes. Graph Transformers are used to combine a fine-scale encoding over the local observations and a coarse-scale encoding on the global map.

However, the original architecture of DUET prohibits backtracking by masking out visited nodes in the action space. Although this implementation holds when following the shortest trajectory from a certain position to the goal, it fails when the supervision is performed using human-generated trajectories like in CVDN, as they could contain backtracks. Therefore, it might be necessary to revisit the same node multiple times. DUET is modified accordingly to account for this behavior. Originally, DUET masks all the previously visited nodes to prevent the agent from revisiting these nodes. The masking of previously visited nodes is removed and only mask the current node to ensure that the agent does not remain on the same node.

The prediction of the next location, after this modification, considers an action space comprising all the possible navigable nodes in the graph instead of only the neighboring ones. In addition, the action space includes an additional possibility defined by the stop action. As in CVDN the only available textual input at the beginning of the episode is the target object, an instruction is mimicked including such object by prepending learnable prompt embeddings at the beginning of the input to the model. In the Experimental Results Section 4.3, experiments are conducted with different numbers of prompt embeddings and these results are compared with those without prompt embeddings.

4.1.3 Dialogue Exchange during Navigation

As represented in Fig. 4.2, UNMuTe consists of a dialogue model and a navigation model, where the navigation model can trigger the dialogue model to generate a pair of questions and answers when the trajectory to the target is not clear. In this respect, the confidence of the navigator can be quantified as the entropy $\mathcal{H}$ of its action probability distribution, which acquires higher values as the probability distribution approaches the uniform distribution. Therefore, the entropy $\mathcal{H}$ of the action probability distribution on the navigable nodes of the environment is computed at each time step. When the entropy $\mathcal{H}_t$ exceeds a threshold value α, the navigator triggers the dialogue model and dialogue generation is activated. The conversation returned by the dialogue pair is concatenated to the input of the navigator to recompute the probability distribution over the action space, and if $\mathcal{H}_{t+1} \leq \alpha$, the next viewpoint is selected for the navigation.

An empirical analysis of the choice of the entropy threshold is performed and the use of a learnable parameter $\hat{\alpha}$ is evaluated as the threshold. To this end, a binary cross-entropy loss is used to set a threshold value $\hat{\alpha}$ which is higher than the entropy in the nodes of the graph where the dataset contains dialogue annotations, and is lower otherwise:

$$\mathcal{L}_{QA} = \text{BCE}(q, \bar{q}), \quad \text{s.t.} \quad q = \frac{1}{1 + e^{(\hat{\alpha} - \mathcal{H}_t)}} \tag{4.2}$$

where $\bar{q}$ is 1 if a question is asked at time step t, and 0 otherwise. As DUET training is performed using both teacher forcing, *i.e.* following the ground truth trajectory, and by sampling the action probability distribution, $\mathcal{L}_{QA}$ is calculated only for the teacher forcing training stage. When actions are sampled, the value of q in Eq. 4.2 is only used to trigger the dialogue model.

Table 4.1: Hyperparameters related to the navigator model.

Navigator	
num text encoder layers:	9
num coarse-scale encoder layers:	4
num fine-scale encoder layers:	4
num pano layers:	2
max action length:	15
max instruction length:	512
training batch size:	2
learning rate:	10^{-5}
sample weight:	1.0
ml weight:	0.2

Table 4.2: Hyperparameters related to the dialogue model.

Dialogue Model	
num layers:	12
model dimensionality:	768
num attention heads:	12
training batch size:	12
learning rate:	10^{-4}
max instruction length:	1024
num imgs used to generate question:	1
num future imgs to generate answer:	20
optimizer:	adam

4.2 Experimental Setup

4.2.1 Datasets

The effectiveness of UNMuTe is evaluated on VDN using both CVDN and NDH datasets. CVDN contains 2050 navigation trajectories performed on a total of 83 environments of Matterport3D [27], while NDH is composed of 7K navigation episodes obtained by splitting CVDN trajectories in multiple instances. The navigation episodes are performed on navigation graphs where each node is defined by a 360° RGB observation. Even if the navigation module exploits the complete panoramic image to compute its output, the dialogue model uses only frontal crops of 60° to generate the conversation pairs forcing the generated text to refer to the scene in the direction of the agent.

4.2.2 Model Configuration

4

In Tab. 4.1 and Tab. 4.2, the most relevant hyperparameter values used to implement the models composing UNMuTe are shown. For the GPT-2 decoder, a medium-sized, pre-trained version with L = 12, d = 768, H = 12 is used, where L is the number of layers, d is the model dimensionality, and H is the number of attention heads respectively. The resulting dialogue model contains 124M parameters and was trained for approximately 6 hours on a single NVIDIA RTX6000 GPU. The navigation model (164M parameters) was finetuned for 48 hours each on a single NVIDIA RTX6000 GPU. The visual features used by UNMuTe are extracted using ResNet-152 model. However, navigation results are also compared with different image feature extractors such as BLIP, ViT, CLIP etc. in the following section on experimental results.

4.2.3 Evaluation Criteria

The experimental results contained in this section are compared with the current state-of-the-art methods on both CVDN and NDH datasets. While the evaluation using NDH dataset is more popular, interactive experiments on CVDN are only performed by RMM [113] and SCoA [156]. RMM uses two speaker models that regularly generate questions and answers, while SCoA uses a model to predict when to generate dialogue and selects the most appropriate question among a set of question templates. The main competitor on NDH are instead, HAMT [30] and VISITRON [122]. HAMT encodes episode history and uses it as an additional modality with text and images to predict its actions, while VISITRON trains a multimodal Transformer encoder and an LSTM decoder to predict navigation actions and when to exchange dialogue.

The metrics employed for the navigation experiments are goal progress (GP), *i.e.* the mean reduction in Euclidean distance between the starting position and to final position with respect to the target; success rate (SR), *i.e.* the fraction of episodes where the agent can reach the goal position within 3 meters; success rate weighted by path length (SPL); and normalized Dynamic Time Warping (nDTW) as defined by [65].

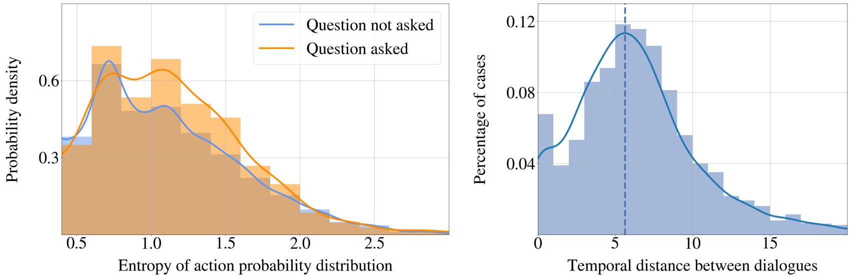


Figure 4.4: Probability distributions of the entropy of the action probability and the temporal distances between dialogues on the training split of CVDN.

4.3 Experimental Results

4.3.1 CVDN Experiments

The experiments performed on the CVDN dataset are presented in Tab. 4.3 and showcase the quality of the overall approach in an interactive setting. In fact, during the navigation using the episodes of CVDN, the model has to autonomously trigger the dialogue model to generate question-and-answer pairs to guide its movement toward the target.

Different configurations of UNMuTe are compared, using the learnable threshold presented in Sec. 4.1.3, and a policy that activates at regular time intervals. The latter is obtained on the basis of the distribution of the training split of CVDN (shown in Fig. 4.4), by considering the mode of the temporal distance between ground-truth dialogues. As the mode of the temporal distances distribution is 5.63, question-and-answer pairs are generated every 4, 5, and 6 time steps during training and evaluation on the CVDN task. Triggering the dialogue model every 5 time steps achieves a state-of-the-art success rate of 7.73 and SPL of 9.62. State-of-the-art goal progress of 13.35 meters is obtained by the model with a learnable entropy threshold, thus confirming the effectiveness of this strategy. UNMuTe is also compared with the current state-of-the-art methods, which, however, do not evaluate in terms of SPL, SR, and nDTW, but only present GP results. All configurations of UNMuTe present better results than the competitors, with the best configuration that overcomes SCoA by 2.16 meters in terms

Table 4.3: Navigation results for our approach and recent methods on the "val unseen" split of CVDN.

	Val Unseen			
	GP	SPL	SR	nDTW
$RMM_{n=3}$ + Oracle Stopping [113]	8.9	-	-	-
SCoA [156]	11.19	-	-	-
UNMuTe (threshold)	**13.35**	5.39	7.31	24.81
UNMuTe (4 time steps)	12.68	3.62	5.00	24.44
UNMuTe (5 time steps)	13.13	**7.73**	**9.62**	**25.76**
UNMuTe (6 time steps)	12.31	4.81	5.77	23.65

Table 4.4: Comparison of navigation results with different image feature extractors on CVDN val unseen.

	Val Unseen			
	GP	SPL	SR	nDTW
UNMuTe (BLIP)	12.05	4.97	6.54	21.67
UNMuTe (ViT-L/16)	12.29	5.99	**8.46**	24.78
UNMuTe (CLIP ViT-L/14)	12.21	4.78	6.15	23.78
UNMuTe (CLIP RN50)	11.83	4.94	6.15	**25.11**
UNMuTe (ResNet50)	12.34	**6.68**	8.08	23.80
UNMuTe (ResNet152)	**13.35**	5.39	7.31	24.81

Table 4.5: Comparison of navigation results with different numbers of prompt embeddings on CVDN val unseen.

	Val Unseen			
	GP	SPL	SR	nDTW
UNMuTe (w/o prompts)	11.97	**8.12**	**10.77**	**25.48**
UNMuTe (4 prompts)	**13.35**	5.39	7.31	24.81
UNMuTe (8 prompts)	11.96	8.49	11.92	23.30

of goal progress.

Experiments using Different Extracted Image Features. In Tab. 4.4, the most appropriate pre-trained visual encoder is selected for the extraction of the image features for the dialogue model assessing the results of different models: ResNet152 [57], ResNet50 [57], CLIP [105], BLIP [74] and ViT-L/16 [42]. In the case of CLIP, the variants exploiting ViT-L/14 and RN50 as backbones are considered. Following previous work on Vision-and-Dialog Navigation, models with better goal progress are prioritized. It was found out that the navigation results of the agent using image features extracted with ResNet-152 achieved the best performance. The goal progress for UNMuTe using ResNet152 features is

Table 4.6: Comparison of navigation results with different constant thresholds on CVDN val unseen.

	Val Unseen			
	GP	SPL	SR	nDTW
UNMuTe (learnable thr.)	**13.35**	5.39	7.31	**24.81**
UNMuTe (thresh=0.9)	12.99	**6.99**	**9.62**	24.21
UNMuTe (thresh=1.0)	11.27	5.28	6.92	22.14
UNMuTe (thresh=1.1)	12.03	5.62	7.31	23.45

better than the other configurations by at least 1.01 meters.

Experiments using Different Prompt Embedding Sizes. An ablation study on the navigation performance of UNMuTe was performed using different numbers of learnable prompt embeddings at the beginning of the instruction used by the navigator. A model not using learnable prompt embeddings was compared with models using respectively 4 and 8 learnable prompt embeddings. For all the navigators considered in this experiment, the questions were asked using the learnable entropy threshold. As it can be seen in Tab. 4.5, UNMuTe with 4 learnable prompts has the best performance in terms of goal progress (GP) with an increase of 1.39 meters over UNMuTe with 8 prompt embeddings and 1.38 meters over the model that does not use prompt embeddings.

Experiment using Different Constant Thresholds. Experiments were also performed considering different constant threshold values in comparison to the model using the learnable threshold. Considering the action probability distribution of the navigator when the questions are and are not asked in Fig. 4.4 of the main paper, the threshold is set to 0.9, 1.0, and 1.1 choosing values that separate the two distributions. However, looking at the results in Tab 4.6, UNMuTe with a learnable threshold value performs better than all the baselines using fixed threshold values with a minimum improvement in terms of goal progress of 0.36 meters.

4.3.2 NDH Task

The navigation experiments of UNMuTe are complemented with experiments on the NDH task. NDH consists of navigation episodes using dialogue instances as textual input. To this end, the dialogue annotations and the trajectories of

Table 4.7: Navigation metrics for our approach and competitors on the "val unseen" and "test unseen" splits of the NDH dataset.

	Val Unseen			Test Unseen		
	GP	SPL	SR	GP	SPL	SR
Seq2Seq [13]	2.10	-	-	2.35	16	-
PREVALENT [55]	3.15	-	-	2.44	24	-
CMN [157]	2.97	-	-	2.95	1	-
HOP [104]	4.41	-	-	3.24	-	-
HAMT [30]	5.13	-	-	5.58	7	-
ScoA [156]	2.91	-	-	3.37	15	-
VISITRON [122]	3.25	11	27	3.11	12	-
VISITRON (Best SPL) [122]	2.71	25	33	2.40	25	-
UNMuTe (Planner)	4.98	**49**	**60**	4.03	**47**	**56**
UNMuTe (Player)	**5.88**	22	36	**5.75**	22	35

CVDN are split to form a total of 7K navigation episodes. Before training the navigation model for the task, question-and-answer pairs are generated using the dialogue model for each trajectory in the training split of NDH. Consequently, DUET is trained on the resulting double-sized dataset, augmented with synthetically generated dialogues.

As it can be seen from Tab. 4.7, state-of-the-art results are achieved on both "val unseen" and "test unseen" splits of NDH. In particular, UNMuTe trained on the trajectory performed by the human annotator (Player) achieves goal progress of 5.88 and 5.75 for the "val unseen" and "test unseen" respectively. UNMuTe trained on the shortest path trajectory (Planner), instead, achieves a SPL and SR of 49 and 60 on "val unseen" and of 47 and 56 on "test unseen". The high difference in the SPL and SR of the agents trained on the planner and player trajectories is due to the fact that the agent uses the shortest path annotation in the case of the planner trajectory. Instead, the player trajectory often includes mistakes and reconsiderations, thus requiring the agent to backtrack to a previously visited node and lowering the values of SPL. In the table, the first section comprises studies that employ ground-truth dialogue annotations as instruction. These do not generate their own dialogues but simply use the dialogue provided in the NDH task for navigation. The second section, instead, reports methods that generate additional synthetic dialogues. Overall, UNMuTe achieves top-1 performance on all metrics of the NDH task.

Table 4.8: Evaluation in terms of text generation quality.

	Val Unseen				
	BLEU-1	METEOR	ROUGE	CIDEr	SPICE
Questioner	0.201	0.092	0.179	0.181	0.089
Oracle w/o future images	0.214	0.091	0.177	0.111	0.088
Oracle w/o target object	0.228	0.098	0.192	0.145	0.094
Oracle	0.237	0.098	0.200	0.179	0.109

4.3.3 Dialogue Generation

In this section, the capability of the dialogue model to generate proper question-and-answer pairs is discussed. To this aim, the generated questions and answers are compared with human annotations using NLP and reference-based description metrics like BLEU [101], ROUGE [81], METEOR [16], CIDEr [136], and SPICE [12]. Results are reported in Tab. 4.8. Here, the question is asked by the "navigator" (upper portion of the table) and the answer is given by the "oracle" (lower part of the table). For calculating different metric scores, the predicted sentences are compared with the ground-truth ones in terms of their n-grams (*i.e.* a sequence of n consecutive words). BLEU, METEOR, and ROUGE are commonly used for the task of evaluating translation and summarization, while CIDEr and SPICE have been specifically designed for the task of image description and are also employed in VLN works in which synthetic instructions are generated [124]. As can be seen, most of the metric values are above 0.20 for generating an answer close to the ground-truth answer, which outlines the linguistic capabilities of the proposed model. It was further noticed that the metric values for the "navigator" role are lower than those of the "oracle", *i.e.* the model is better at generating correct answers rather than asking proper questions. This is because there can be greater diversity in the generation of a question than that of the answer, which is instead more objective and should match the actions in the given trajectory.

Future Images for Answer Generation. The contribution given by the incorporation of images extracted from the future trajectory (*i.e.* $(v_{t+1}, \ldots, v_{t+k})$ in Eq. 4.1) is then validated during the generation of answers in the dialogue model. This is done by comparing UNMuTe with the answers of a dialogue model trained without using future images. The results are provided in the lower part of Tab. 4.8. Comparing the two oracles it was observed that, the oracle that does not employ future images undergoes a drastic reduction in performance on the "val unseen" split. In fact, the CIDEr score in "val unseen" decreases from 0.179 to 0.111. Overall, this underlines the effectiveness of employing future frames as a conditioning signal for the dialogue model.

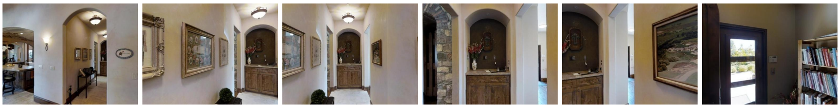
Target object for Answer Generation. The contribution given by the target object (*i.e.* o_{tgt} was validated in Eq. 4.1) during the generation of answers in the dialogue model. For this, UNMuTe was compared with the answers of a dialogue model trained without using the target object. The results are provided in the lower part of Tab. 4.8. It can be observed that, the oracle without the target object undergoes a reduction in performance on the "val unseen" split. The CIDEr score in "val unseen" decreases from 0.179 to 0.145. Overall, this shows that employing the target object as a conditioning signal for the dialogue model is beneficial for the generation of the answers.

4.3.4 Qualitative Generation Samples

To showcase the quality of the proposed approach, three examples of generated dialogues are reported in Fig. 4.5. For all three examples, the question and answer generated by UNMuTe appropriately describe the path that the agent should take. Noticeably, even if the ground-truth answer annotation of the first sample contains a mistake (the instruction is asking the agent to turn left rather than turning *right*), UNMuTe generates a correct answer, by asking the agent to turn right towards the living room. The second example consists of a yes-or-no interaction where the agent answers affirmatively to go down the long hall. In the third example, the agent asks a reasonable question on whether it should go right or left and the answer is clear and concise: go right, head into the hallway, and take a right to the stairs. As can be observed, these examples outline the effectiveness of the dialogue model and its ability to generate appropriate questions and answers for a given sequence of images.

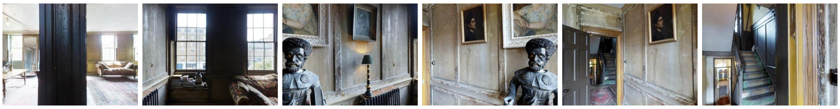
GT: *Should I go back down this hall?*
It wants you to make a left turn and go in that family room.

UNMuTe: *Which way from here?*
Make a right and go towards the living room.



GT: *Lt straight or rt?*
Turn right, then all the way down the hallway, there will be a room at the end of the hallway on the right.

UNMuTe: *Do I go down the long hallway here?*
Yes, go down the long hall to the living room.



GT: *Okay. Left, right, center left, or center right?*
Take the right narrow doorway and look for more stairs that continue down. Take them all way to the bottom.

UNMuTe: *Should I go to the left or right side of the room?*
Make a right and head into the hallway and then make a right into the stairs.

Figure 4.5: Sample paths taken from the CVDN "val unseen" split, together with the corresponding ground-truth interactions and generated ones. The number of depicted steps has been artificially reduced to 6 to facilitate the graphical presentation. Only the frontal image of the panoramic observation at each timestep is shown.

5

Robots with Parallel and Reactive Empathy for Dialogue Generation

Transformer models have changed the way humans interact with artificial agents. Recently, the decoder part of the transformer has been used independently to generate large amounts of text. Large Language Models (LLMs) such as ChatGPT [107] and Llama [131] have gained popularity and allowed the general public to interact with artificial agents in terms of natural language. Despite this, these models do not understand human emotions, and it seems that the goal of the models is to assist humans in some way or another. For example, when ChatGPT was given the following prompt: "I am scared of talking in front of so many people", it responded that it was normal to feel scared when talking in front of a large group. But with a few strategies, one can build confidence and manage that fear. This is followed by a list of points on how one can overcome stage fright. Humans do not necessarily always give each other tips

This Chapter is related to the planned publication "N. Rawal *et al.*, MATE: Multimodal Agents that Talk and Empathesize, Submitted to Proceedings of the International Conference on Image Analysis and Processing" [8]. See the list of publications on page 93 for more details.

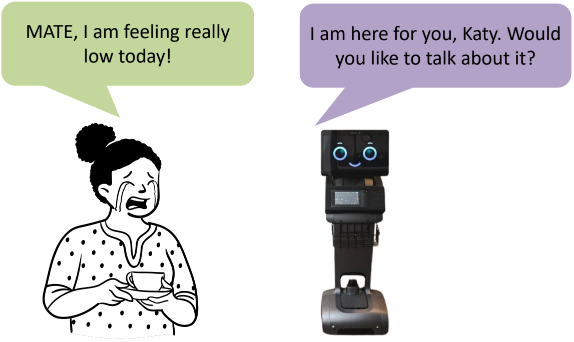


Figure 5.1: MATE, a multimodal agent that talks and empathesize is proposed. MATE takes the human's facial expressions and text as an input and responds back to the human in an empathetic manner. The empathetic output given by the robot consists of the textual response as well as a facial expression.

to improve something. Just a small encouragement saying "You can do this. I believe in you" might be sufficient in this case.

As social robots are getting increasingly popular in a plethora of domains, it is crucial for these robots to understand human emotions and respond to humans in an empathetic manner. Deep learning algorithms are already quite good at recognizing human facial expressions and even being able to classify human emotions coming from a multi-modal input signal that combines modalities such as facial image, body gestures, text, audio, and physiological signals [1]. When these deep learning algorithms are plugged into robots, the robots are capable of recognizing human emotions. Some of these robots can even express emotions through facial expressions or gestures [110, 148].

In Human-Human Interaction (HHI), there is parallel empathy and reactive empathy [37]. Parallel empathy is when a human responds to another human with the same facial expression, such as a human smiles when smiling at [18, 58, 123]. Reactive empathy is when a human generates an emotion in response to the emotion of another human being. Robots should also be capable of responding to the human with both parallel and reactive empathy depending on the situation. For example, a robot could be happy when the human shares a happy news or it could be sad when the human is angry at it.

In this study, a computational model is proposed that responds to human empathy by considering a multimodal input of the person's facial expressions and the text content of what the person says (see Figure 5.1 for an overview). The proposed computational model is based on a transformer encoder-decoder structure. The encoder part of the transformer takes the person's facial expressions, the textual prompt, and the target emotion as input. The embeddings that are the output from the transformer encoder are then fed into the decoder part as input. The decoder part of the transformer is fine-tuned on Llama3 [44] model that returns the textual response to a given input prompt. To evaluate the fine-tuned model, responses are recorded for 17 prompts with varying emotions. A survey is conducted in which the responses generated by the model are rated as appropriate and empathetic on a five-point likert scale. The responses are also compared with those generated by ChatGPT. Finally, the results are validated using a statistical test. The responses generated by MATE are rated more appropriate and empathetic compared to those generated by ChatGPT. The results are also statistically significant.

The contributions of this work are as follows. A new empathetic dialogue dataset is created that contains dialogues between two people, where the second person responds to the first person using parallel or reactive empathy. A novel computational model MATE is proposed. MATE is based on a transformer architecture for generating empathetic textual responses. The responses generated by MATE are shown to be both appropriate and empathetic by conducting a survey where participants see videos of the robot responding to various prompts.

5.1 Proposed Method

MATE is a multimodal agent that talks and empathizes. It is based on a transformer encoder-decoder structure. The facial image, the input text that contains what the person says, and a target emotion are used as input and converted to embeddings using a transformer encoder. These embeddings are used as prompts for the transformer decoder part, which is a pre-trained Llama3 [44] model. Llama3 is fine-tuned to output the dialogue response for the given input prompt.

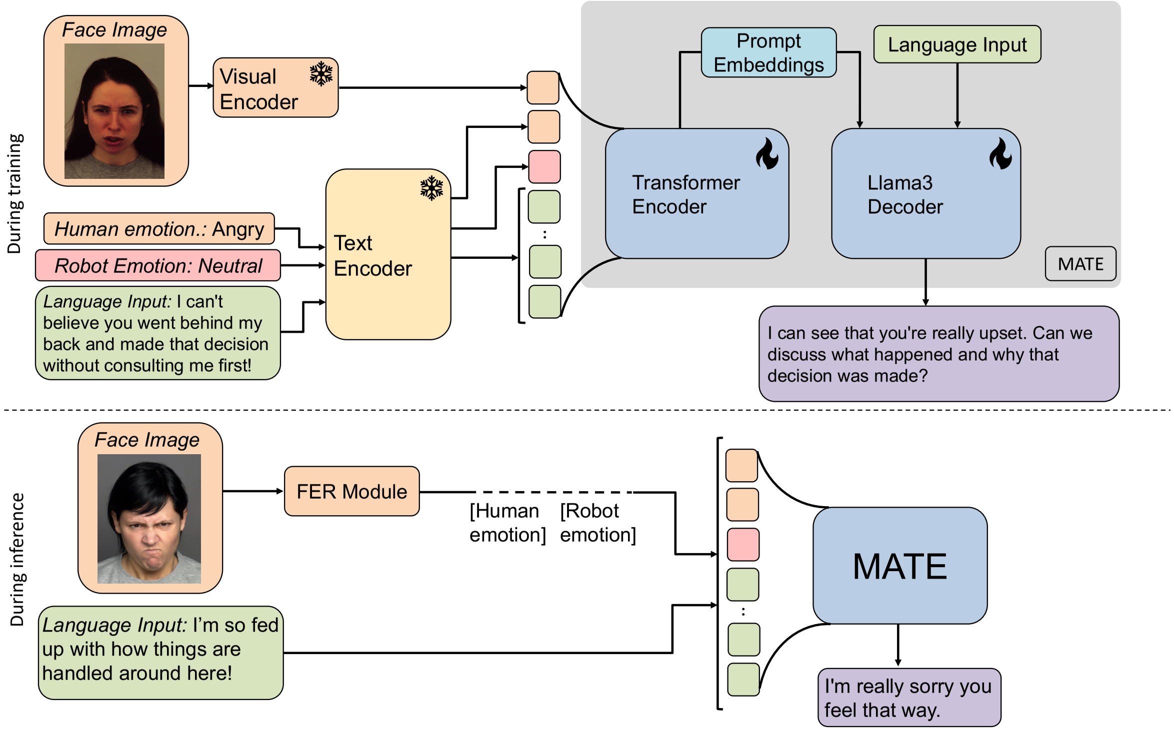


Figure 5.2: The training process of the proposed computational model is as follows. The facial features, ground truth facial expression label of the human, the emotion of the agent and what the human says are fed as input to the transformer encoder model which gives the embeddings as an output. These prompt embeddings are fed into the Llama3 decoder and the response is generated token by token. During inference, the `Facial Expression` is given by the FER Module. Based on the recognized facial of the human, the emotion of the agent is selected. After using the human prompt, human facial features, the recognized facial expression of the human and the emotion of the agent as an input to the proposed model, the response is generated from the fine-tuned model.

5.1.1 Creating the Empathetic-HRI Dataset

To create the Empathetic-HRI dataset, ChatGPT is prompted as follows: "Create a two line dialogue between two people `[SITUATION]` such that person 1 feels `[EMOTION 1]` and person 2 gives a `[EMOTION 2]` response". For the list of situations considered, see Tab. 6.2. For the combinations of `[EMOTION 1]` and `[EMOTION 2]`, the positive emotions are responded with parallel empathy and the negative emotions are responded with reactive empathy. Tab. 5.2 shows the possible combinations of [EMOTION 1] and [EMOTION 2]. The language data is paired with images from the KDEF dataset [85], showing various facial expressions to make the dataset multimodal. In the end, the Empathetic-HRI dataset has 2800 dialogues, showing various emotions.

5.1.2 Transformer Model Architecture for Empathetic Response Generation

The proposed model comprises of a transformer encoder and decoder structure. First, the image v is fed to ResNet-152 [57] pre-trained on VGG-Face Dataset [25] to extract the facial features. These facial features are then passed through a linear layer and converted to embeddings. The current emotion $\mathcal{E}_h$ of the person (eg. "happy") is coverted into a global vector using Glove42B tokens [102]. This global vector is also passed through a linear layer and converted into embeddings. Similarly, the emotion with which the agent should reply $\mathcal{E}_a$ (the emotion corresponding to EMOTION 2 in the dataset) is also coverted into a global vector using Glove42B tokens which is then passed through a linear layer and converted into embeddings. For tokenizing the words of the human $\mathcal{T}$, Llama3 tokenizer is used which are also converted into embeddings. The four kinds of embeddings for faces, emotion of the person, emotion of the agent, as well as the words of the person are concatenated and sent as input to the transformer encoder model. Formally,

$$y = \texttt{Encoder}\left(\left[\underbrace{v}_{\text{Image}}, \underbrace{\mathcal{E}_h}_{\text{Human Emotion}}, \underbrace{\mathcal{E}_r}_{\text{Agent Emotion}}, \underbrace{\mathcal{T}}_{\text{Human Text}}\right]\right) \tag{5.1}$$

Table 5.1: The situations and the number of the corresponding dialogues that were considered for building the Empathetic-HRI dataset.

SITUATION	No. of dialogues
None	700
At work	200
In school	200
While travelling	200
In a hospital	200
At a ski resort	100
In a coffee shop	100
At the gym	100
At a party	100
In a restaurant	100
At a park	100
At the beach	100
At a concert	100
At a wedding	100
At a supermarket	100
At a museum	100
At a sports event	100
At a movie theater	100

where v denotes the set of visual features for the facial expression image, $\mathcal{E}_h$ indicates the ground truth emotion label of the human, $\mathcal{E}_a$ indicates the emotion of the agent, $\mathcal{T}$ denote the human text input tokens.

The embeddings that are an output of the transformer encoder are then fed into the transformer decoder model (i.e. Llama3) as prompt embeddings. The prompt embeddings that are fed into the Llama3 decoder are concatenated with the embeddings of the language input. Formally,

$$y = \texttt{Llama3}\left(\left[\underbrace{P_v}_{\text{Image}}, \underbrace{P_h}_{\text{Human Emotion}}, \underbrace{P_a}_{\text{Agent Emotion}}, \underbrace{\mathcal{T}}_{\text{Human Text}}, \texttt{BOS}, \underbrace{i_1, .., i_m}_{\text{Instruction}}, \texttt{EOS}\right]\right) \tag{5.2}$$

where P_v denotes the prompt embeddings of visual features, P_h indicates the prompt embeddings of the human emotion label, P_a indicates the prompt embeddings of the agent emotion label, $\mathcal{T}$ denote the human text input tokens, `BOS` and `EOS` are begin of string and end of string tokens respectively. Consequently,

Table 5.2: To build the Empathetic-HRI dataset, ChatGPT is prompted to "Create a two line dialogue between two people `[SITUATION]` such that person 1 feels `[EMOTION 1]` and person 2 gives a `[EMOTION 2]` response". The combinations of `EMOTION 1` and `EMOTION 2` are listed below.

EMOTION 1	Empathy Type	EMOTION 2
happy	parallel	happy
surprise	parallel	surprise
sad	parallel	sad
sad	reactive	neutral
angry	reactive	sad
angry	reactive	neutral
neutral	parallel	neutral
neutral	reactive	happy
fearful	reactive	neutral
disgusted	reactive	neutral

$(i_1, ..., i_m)$ denotes the tokens that correspond to the response.

The Llama3 decoder is fine-tuned to respond to a given query and facial expression in a supervised manner. The goal is to reduce the cross-entropy loss between the predicted responses and the ground truth responses. Specifically,

$$\begin{aligned} l_n &= -\sum_{c=1}^{C} y_{n,c} \log \frac{\exp(x_{n,c})}{\sum_{i=1}^{C} \exp(x_{n,i})} \\ L &= \frac{\sum_{n=1}^{N} l_n}{N}, \end{aligned} \tag{5.3}$$

where x is the input, y is the target, C is the number of tokens in the dictionary, and N spans the minibatch dimension.

During the testing phase, first the facial expression of the person is recognized using FIXR [89] that is trained on three different datasets[95, 50, 77]. Depending upon the recognized facial expression, the emotion of the agent is decided using Table 5.2. If there is more than one possibility of reacting with an emotion for a given human emotion, eg. when the human emotion is detected as sad, the agent's emotion is chosen between sad and neutral, where sad is a parallel emotion type while neutral is a reactive emotion type. Similarly to the method

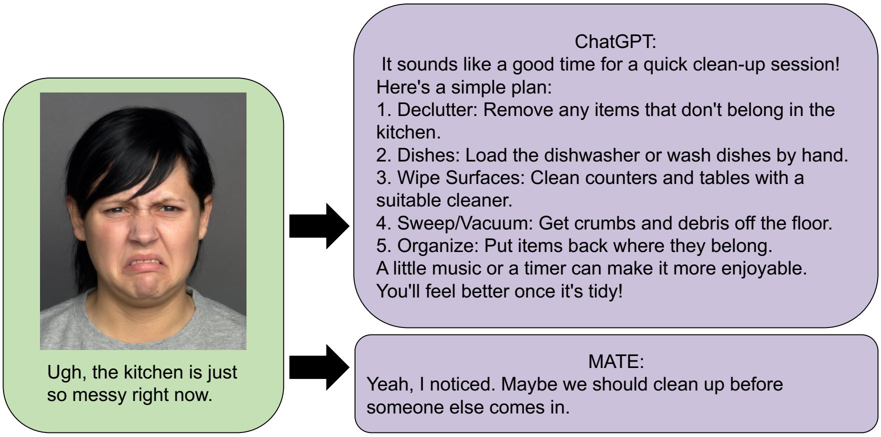


Figure 5.3: A qualitative sample where the prompt consist of a disgusted emotion. The sample compares the response of MATE with that of ChatGPT-4o for a given human facial expression and input prompt.

5

during fine-tuning of the model, first the embeddings for the four inputs, i.e. facial features of the human, recognized facial expression, agent's emotion, and the words that the human says are extracted. These are then concatenated and sent as an input to the transformer encoder. After getting the embeddings from the transformer encoder model, these become an input for the transformer decoder model i.e. Llama3, followed by the embeddings of the language input tokens. Llama3 predicts the output response token-by-token until it reaches the `EoS` token. Fig. 5.2 shows the proposed model architecture for generating empathetic responses.

5.2 Experimental Evaluation

AdamW optimizer with a learning rate $0.2e - 5$ is used to fine-tune the model.

5.2.1 Qualitative Results

In Fig. 5.3, an example of a face image from the FACES dataset [45] is presented along with its corresponding input text that is fed into the proposed model and compared to the response given by ChatGPT-4o. The example is of disgusted

Table 5.3: Sample inputs of the proposed model and their corresponding outputs are provided below.

Human	Input	Agent	MATE
afraid	I heard some strange, unusual noise in the streets last night. It was a bit scary.	neutral	It's probably just the sound of the city. We live in a busy area. Nothing to worry about.
angry	I can't believe my boss ignored all my suggestions in the meeting today, like my ideas don't matter.	neutral	I understand how that feels. I've been there before, and it hurts to feel that your hard work isn't valued.
happy	I finally booked my dream vacation to New Zealand for next month. I've been wanting to go there for years.	happy	That's amazing! I've heard it's absolutely stunning. You're going to have the time of your life.
neutral	I finished the report earlier than I thought. Can you help me check it?	neutral	Sure, let's take a look together.
sad	I didn't get the job I was hoping for even though I really thought I had a good chance this time	sad	The news hit me like a punch in the gut. I know how much you wanted this, and I'm so sorry you didn't get it.

emotion, where the human exclaims that the kitchen is so messy. The algorithm correctly recognizes the human emotion as disgust and responded with a neutral emotion that it noticed and that they should clean up before someone else came in. On the other hand, ChatGPT gave a long list of tips on how to clean up a space.

In addition to this qualitative sample, more qualitative samples of afraid, angry, happy, neutral, and sad expressions are presented in Tab. 5.3. All these samples are also accompanied by their corresponding facial images, but in the table, the images are omitted to save space. As can be seen in the table, all the responses given by MATE are appropriate and empathetic.

5.2.2 Results of the survey

For this study, the responses to certain prompts and face images in the FACES dataset [45] are obtained using MATE. To compare the responses of the model with those by ChatGPT-4o, the agent's responses are recorded for 17 different prompts with varying emotions, and an anonymized survey is conducted asking human participants to rate whether the responses given by the agent are appropriate and empathetic. If the responses are not empathetic, human participants could write in free text how they would respond to the given query.

The survey that was created consisted of the 17 qualitative examples. Each qualitative example consisted of a human image, the language text that the human said and responses given my MATE as well as by ChatGPT-4o. The survey was answered by 20 human participants. Of these, 14 were men and six were women. The survey data was collected anonymously, and each participant had to answer mainly two questions for every qualitative example. The first question was whether they agreed that the response given by the robot was appropriate. This question had to be answered on five-point Likert scale where one meant strongly disagree and five meant strongly agree that the response given was appropriate. The second question if the response given by the robot was empathetic. This question also had to be answered on five-point Likert scale where one meant strongly disagree and five meant strongly agree that the response given was empathetic. However, if the human participant said that the responses given by both the models were not empathetic, they had free space to write in text how they would have responded to the given input prompt.

To analyze the survey results of the participants, the paired-samples T-test is conducted. For the first questions where the participants had to select to what extent they agreed the responses given by ChatGPT and MATE were appropriate, the t-value of MATE − ChatGPT was 2.683 and the p-value was 0.015. As the p-value was less than 0.05, the results are statistically significant. Figure 5.4(a) show the comparison plot of the two models.

For the second question of the survey where the participants had to select to what degree they agreed that the responses given by ChatGPT and MATE were

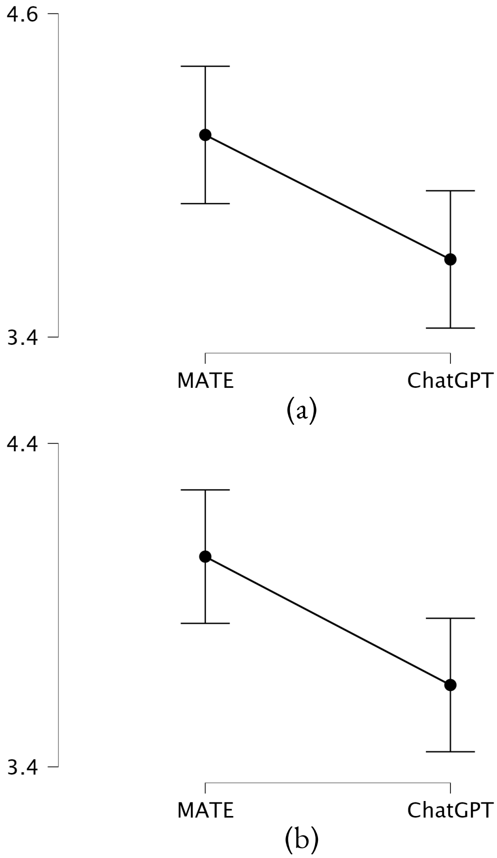


Figure 5.4: Comparison of responses generated by MATE and those generated by ChatGPT using the paired-samples T-test. The appropriateness of the responses is measured in (a), and the empathetic nature of the responses is measured in (b).

empathetic, the t-value of MATE − ChatGPT was 2.847 and the p-value was 0.010. As the p-value was less than 0.05, the results are statistically significant. Figure 5.4(b) show the comparison plot of the two models.

5.3 Limitations of the Current Study

When the dataset is created using ChatGPT, it is always assumed that there is no mismatch between the facial expression and the generated dialogue. However, in real-life scenarios, it could be that the human's facial expression does not match the content of what is being said by them. In this case where there is a mismatch between facial expression and the content of what is being said, the LLM agent should be able to say something like "Congratulations on your promotion! That's a fantastic accomplishment. You seem a bit down, though. Is everything okay?". The current textual data that is collected using ChatGPT-4o is generated using the prompt "Create a two line dialogue between two people `[SITUATION]` such that person 1 feels `[EMOTION 1]` and person 2 gives a `[EMOTION 2]` response". To account for the contradiction in the facial image and the textual prompt, ChatGPT would have to be prompted differently or use the multimodal features of ChatGPT and also send the facial image as input. Additional data would have to be collected and the current dataset would have to be expanded to account for this behavior.

Currently, the emotions of humans are recognized only by considering their facial expressions. Previous research has shown that considering a multimodal signal for human emotion recognition is more accurate than relying on a single signal. It would be better to recognize the emotions through a multimodal signal consisting of facial expressions and speech and adapt the robot's response according to the recognized emotion.

6

Using Implicit Rewards with Human Emotions for Empathetic Dialogue

The previous chapter introduced that Large Language Models (LLMs) like ChaptGPT are often prone to giving long and neutral answers. Although simply plugging in ChatGPT into a robot might help it converse intelligently, it would fail to make an emotional connection with the human. As social robots such as Nao, Pepper, Furhat, etc. are increasingly being used in different domains such as education, healthcare centers, etc., these robots must be empathetic[6]. Using this as motivation, this chapter introduces yet another computational model that can engage in an empathetic dialogue with a human.

The computational model proposed in the previous chapter was based on a transformer encoder-decoder structure. In this chapter, the computational model is based solely on the decoder part of the transformer. The model is called Emma, an Empathetic Multimodal Agent. Figure 5.1 shows an overview of the proposed

---

This Chapter is related to the publication "N. Rawal *et al.*, Multimodal Dialogue for Empathetic Human-Robot Interaction, Submitted to International Conference on Social Robotics" [7]. See the list of publications on page 93 for more details.

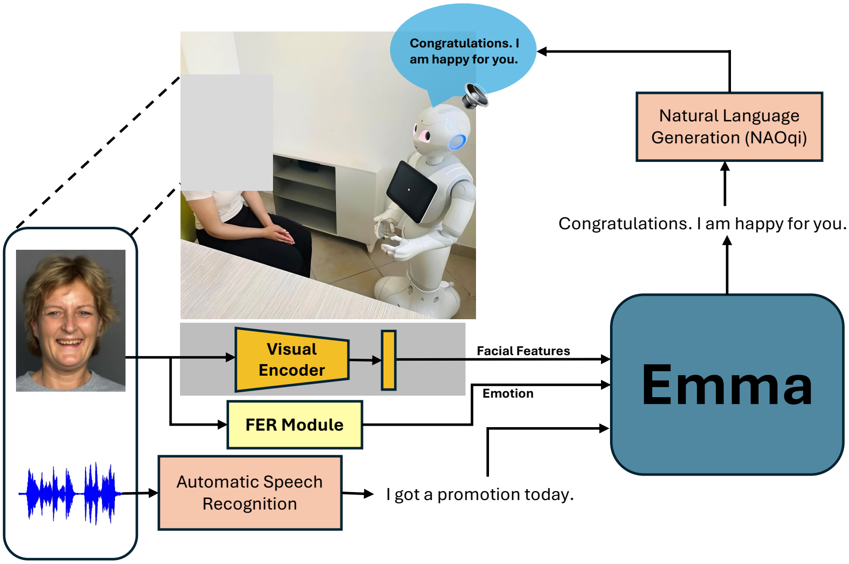


Figure 6.1: A computational model is proposed that responds empathetically to humans, taking their emotions into account. The model is fine-tuned on Llama2 using multimodal data consisting of varying facial expressions corresponding to their dialogues. The fine-tuned model is trained using reinforcement learning to respond in a manner to make the human feel positive.

computational model. To compensate for the need of having large-scale training data, an approach that is trending in the literature is followed where the language dataset is created using one LLM and another LLM is fine-tuned to make it multimodal [83]. The model is trained in a three-step procedure. First, Llama 2[132] is fine-tuned on the empathetic dialogue dataset that is created to generate emotional responses using ChatGPT. Llama2[132] is made multimodal so that it can take into account both human facial expressions and what they are saying. This is achieved by concatenating tokens from facial features with those of the text. Next, a reward model is trained based on the change in the person's emotion after receiving a response from the agent. If the person's emotion becomes positive, the reward is positive, and the reward is negative otherwise. This is a novel component where the reward comes implicitly from the person's emotional response what is being said by the agent. Finally, the fine-tuned model is trained using Proximal Policy Optimization (PPO) [119] so that the agent responds to humans in an empathetic manner to make the human feel positive. For example, when the human says, "I am tired of this. I can't take it anymore", the agent responds "Don't worry, you'll get through it". The study is validated by conducting a survey and experiments in a Human-Robot Interaction (HRI) scenario using the robot Pepper. The responses given by the model are shown to be more empathetic and human-like compared to the multimodal model ChatGPT-4o [98]. The results of the HRI experiment also show that the responses given by the model are both appropriate and empathetic. This proves the effectiveness of the proposed method. In summary, the contributions of this study are as follows. First, a dataset is built consisting of 5600 emotional dialogues with their corresponding facial expressions, by combining the textual dialogue generated by ChatGPT-3.5 Turbo with appropriate images of facial expressions. Second, Llama-2 is fine-tuned in a multimodal manner using the dataset that was created. The multimodality is achieved by concatenating the facial feature tokens with that of the text of what is said by the human. The fine-tuned model that is trained using RL can engage in an empathetic multimodal dialogue with the human. Finally, a survey is conducted in which the responses generated by the model are compared with those of ChatGPT. In addition, an HRI is performed experiment with a Pepper robot and evaluated the model by surveying the participants.

6.1 Proposed Method

In this work, a novel, empathetic, multimodal agent (Emma) is proposed that can respond appropriately to a person considering their emotions. The contributions are threefold. First, the dialogue dataset is built using ChatGPT and these dialogues are paired with appropriate facial expressions from the KDEF dataset [85]. Next, Llama2 is trained using the dataset that was created in a multimodal manner by concatenating the image features with the textual dialogue as input. This model is then trained using Reinforcement Learning (PPO) where the reward comes from the human's emotional response to the agent. Finally, a survey is conducted to compare the proposed model with ChatGPT and performed an HRI study with the Pepper robot.

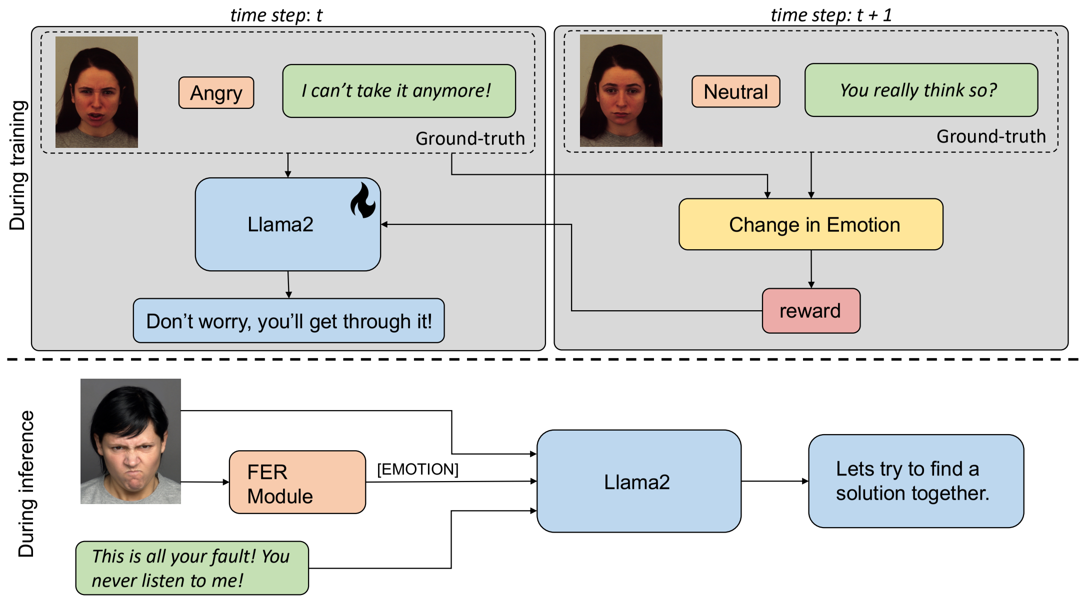


Figure 6.2: The training process of the proposed computational model is as follows. The facial features, ground-truth facial expression label of the person, and what they say are fed as input to the Llama2 model, which generates an appropriate response, token by token. If the emotion of the person, after listening to the agent's response, becomes positive, the agent gets a positive reward; otherwise, it gets a negative reward. The reward is then backpropagated through the Llama2 model. This helps the model learn that it should only say things that would make the human feel positive. During inference, the `Facial Expression` is given by the FER Module. This is followed by the response of the fine-tuned Llama2 model.

6.1.1 Creating the dataset

A dataset is created with 5600 dialogues using ChatGPT-3.5 turbo and these dialogues are paired with appropriate facial expression images from the KDEF dataset [85]. The prompt that is used for ChatGPT-3.5 Turbo is as follows: "Create a three-line dialogue between two people `[SITUATION]` such that the expression of the first person changes from `[EMOTION 1]` to `[EMOTION 2]` and what the second person says makes the first person feel `[EMOTION 2]`". Here, `SITUATION` stands for the categories as shown in Table 6.2. For `[EMOTION 1]` and `[EMOTION 2]`, the 20 combinations that are considered are shown in Table 6.1. The emotions in Table 6.1 consist of the seven basic emotions i.e. sad, happy, angry, afraid, neutral, surprise and disgust. For extremely negative emotions like anger and disgust, the change in the first person's emotion is considered to be positive if it can change from anger or disgust to neutral. The KDEF dataset consists of facial expressions of 70 people, of which 35 were female and 35 were male. The facial expression images of these 70 people were repeated in different situations presented in Table 6.2, to make the dataset diverse and to allow the agent to appropriately respond to a wide range of prompts. The dataset will be released upon the final publication.

6.1.2 Finetuning Llama2

The model follows a three-step training procedure. First, Llama2 is fine-tuned on the empathetic dialogue dataset $\mathcal{D}$. The dataset $\mathcal{D}$ consists of a facial expression image $\mathcal{I}$, input of human text $\mathcal{T}$ and ground truth response y. The Llama2 model takes the image of facial expression $\mathcal{I}$ and human text as input $\mathcal{T}$ to generate empathetic responses. The image $\mathcal{I}$ is fed to ResNet-152 [57] pre-trained on VGG-Face Dataset [25] to extract the facial features. These facial features are then passed through a linear layer and converted to embeddings. The image embeddings are concatenated with the token embeddings of the language part to make Llama2 multimodal. The ground truth facial expression label $\mathcal{E}$ is also prepended as a word to the language part. This concatenation of the embeddings

Table 6.1: ChatGPT-3.5 turbo was prompted to "Create a three-line dialogue between two people such that the expression of the first person changes from `[EMOTION 1]` to `[EMOTION 2]` and what the second person says makes the first person feel `[EMOTION 2]`" to build the empathetic dialogue dataset. The following combinations of `EMOTION 1` and `EMOTION 2` were considered.

Emotion 1	Emotion 2	Change in Emotions
sad	happy	positive
happy	sad	negative
sad	angry	negative
angry	neutral	positive
sad	neutral	positive
neutral	sad	negative
neutral	angry	negative
neutral	happy	positive
sad	surprise	positive
surprise	sad	negative
neutral	afraid	negative
afraid	neutral	positive
neutral	disgust	negative
disgust	neutral	positive
happy	neutral	negative
surprise	happy	positive
happy	angry	negative
angry	happy	positive
happy	happy	positive
sad	sad	negative

is sent to the Llama2 model as input. Formally,

$$y = \texttt{Llama2}\left(\left[\underbrace{v}_{\text{Image}}, \underbrace{\mathcal{E}}_{\text{Emotion}}, \underbrace{\mathcal{T}}_{\text{Human Text}}, \texttt{BOS}, \underbrace{i_1, .., i_m}_{\text{Instruction}}, \texttt{EOS}\right]\right) \tag{6.1}$$

where v denotes the set of visual features for the facial expression image, $\mathcal{E}$ indicates the ground truth emotion label, $\mathcal{T}$ denote the human text input tokens, `BOS` and `EOS` are begin of string and end of string tokens respectively. Consequently, $(i_1, ..., i_m)$ denotes the tokens that correspond to the response.

Llama2 is then fine-tuned to generate the response given in the dataset for a given query and facial expression in a supervised manner. For fine-tuning the model, the goal is to reduce the cross-entropy loss between the predicted responses and the ground-truth responses. The loss is backpropagated, and the weights of

Table 6.2: The list of situations and their corresponding number of dialogues that were considered for building the empathetic dialogue dataset.

Situation	No. of Dialogues
None	1400
At work	400
In school	400
At a ski resort	200
While travelling	400
In a hospital	400
In a coffee shop	200
At the gym	200
At a party	200
In a restaurant	200
At a park	200
At the beach	200
At a concert	200
At a wedding	200
At a supermarket	200
At a museum	200
At a sports event	200
At a movie theater	200

the decoder model are updated during fine-tuning. For testing, the weights of the decoder remain fixed, and the responses are predicted in forward passes for a given facial expression and a textual prompt until it reaches the `EoS` token.

The reward model is designed to evaluate the quality of responses generated by the LLM by assessing their emotional impact on the human user. It predicts a reward score based on the human emotional response to the LLM's output. The reward is based on the change in the user's emotion before and after the response. If the human feels positive compared to before, it gets a reward r of $+1$. Otherwise, it gets a reward r of -1. As a second step, the reward model is fine-tuned to be able to output the reward based on the emotion the human would feel as a result of what the agent said. The column that contains the "change in emotion" in Tab. 6.1 is used to indicate the ground-truth positive reward or the negative reward during the finetuning of the reward model which is a model for sequence classification. Mean squared error is used to minimize the difference

6

between the predicted reward and the ground-truth reward. Formally,

$$y = \texttt{OpenLlama}\left(\left[\underbrace{v}_{\text{Image}}, \underbrace{\mathcal{E}}_{\text{Emotion}}, \underbrace{\mathcal{T}}_{\text{Human Text}}, \underbrace{\mathcal{R}}_{\text{Response}}, \underbrace{r}_{\text{Reward}}\right]\right) \tag{6.2}$$

where v denotes the set of visual features for the facial expression image, $\mathcal{E}$ indicates the ground truth emotion label, $\mathcal{T}$ denote the human text input tokens, $\mathcal{R}$ is the agent response. Consequently, r is the reward to the agent's response.

As the final step, the fine-tuned Llama2 model and the trained reward model are used to output the reward to the generated response and use Proximal Policy Optimization (PPO) to update the Llama2 model. The PPO objective is given as follows:

$$\mathcal{L}(\theta) = \hat{\mathbb{E}}_t \left[min \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{old}(a_t|s_t)} \hat{A}_t, clip \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{old}(a_t|s_t)}, 1 - \varepsilon, 1 + \varepsilon \right) \hat{A}_t \right) \right] \tag{6.3}$$

where $\frac{\pi_\theta(a_t|s_t)}{\pi_{old}(a_t|s_t)}$ is the ratio of the new policy's probability over the old policy's probability, ε is a hyperparameter that determines how much the new policy can deviate from the old policy and $\hat{A}_t$ is the advantage based on the value function.

PPO maximizes the reward while ensuring that the updated policy does not deviate too far from the current policy. This means that the proposed model implicitly learns to generate responses that make the human feel positive. Figure 5.2 shows the overall training procedure of the model.

6.1.3 HRI experiment

The robot and a camera with inbuilt microphone were connected to a computer. The camera is placed diagonally on the table between the robot and the human. When a human speaks, the microphone captures the audio input. This speech input is then processed using Whisper [106], a robust speech-to-text model, to accurately convert spoken language into textual data. At the same time, the human's facial expression is captured by the camera and the facial features are extracted using ResNet-152 [57] pre-trained on VGG-Face Dataset [25]. The human's facial expression is recognized using the FER Module. FIXR [89] is employed as the FER module. FIXR is trained on a diverse range of in-the-wild datasets, including AffectNet [95], the CelebA dataset [50], and the Reliable Facial Expression Recognition Dataset [77]. To enhance its robustness, the dark experience replay approach [24] is integrated into the model.

The facial features, resulting text and recognized emotion is fed into the proposed model as input. The model outputs the text which is spoken by the Pepper robot.

6.2 Experimental Results

For the first step of training, i.e., to fine-tune the Llama2 model in a supervised manner, the AdamW [84] optimizer with a learning rate of $0.2e - 4$ is used. The training batch size is three. The model is fine-tuned using LoRA. For finetuning of the reward model, the AdamW optimizer is used with a learning rate of $0.2e - 5$. Finally, for PPO, Adam Optimizer is used with a learning rate of $0.2e - 7$. The total time taken to fine-tune the model is 44 hours on a single NVIDIA RTX6000 GPU.

6.2.1 Qualitative Samples

In this section, some responses are provided as qualitative samples evaluated on the FACES dataset [45]. The figure 6.3 shows a sample prompt and corresponding response to fear emotion. The lady in the picture says she has stage-fright and cannot speak in front of all the people. In response to this prompt and facial expression, while ChatGPT gave a long response covering several tips on how to overcome nervousness before speaking in public, Emma replied that it would be there to help her.

Five other qualitative examples are provided covering various emotions such as angry, disgust, happy, neutral and sad in Tables 6.3 and 6.4. If the lengths of responses for all these emotional input prompts are considered, it can be clearly seen that the lengths of responses given by ChatGPT are much longer than those

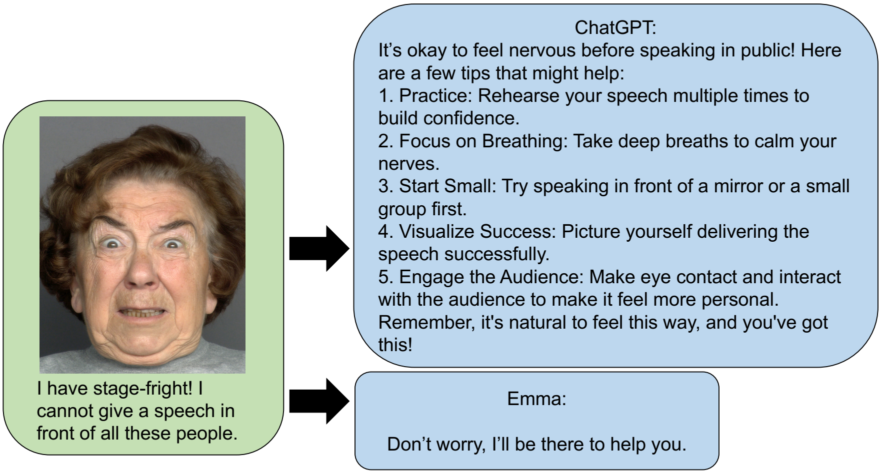


Figure 6.3: A qualitative sample where the prompt consist of a fearful emotion is provided.

given by Emma. Therefore, the responses given by ChatGPT are hardly conversational and not the kind that humans usually use.

6.2.2 Survey Results

To evaluate the performance of the model, a survey is conducted with 17 prompts and their corresponding responses, with varying emotions. For each prompt, two corresponding responses are gathered; the first response is generated by the proposed model, and the other response is generated by ChatGPT-4o. For each prompt-response pair, participants are asked to rate on a five-point Likert scale [96] to what degree they agree that the responses are appropriate, empathetic, and human-like. The responses ranged from strongly disagree (1) to strongly agree (5). There was also an option to respond in free text if both responses were not empathetic, according to the participants.

The survey results containing 17 prompts and their corresponding responses generated by Emma and those generated by ChatGPT were compared using the paired-samples T-test. 20 participants took the survey of which 15 were men and five were women. For the first question, participants were asked to rate if they agreed that the responses generated by Emma and those generated by ChatGPT

Table 6.3: Sample inputs prompts and their corresponding outputs for emotions happy and neutral using ChatGPT and Emma are provided below.

Human	Prompt	ChatGPT	Emma
happy	I got a gold medal for getting the first place in mathematics olympiad.	Congratulations on winning the gold medal in the mathematics olympiad! That's an incredible achievement. You must be very proud of your hard work and dedication. Keep up the great work!	Oh, that's great! Congratulations on your victory.
neutral	I am not good at cooking.	That's okay! Cooking is a skill that takes practice. Here are some tips that might help you improve: 1. Start Simple: Begin with easy recipes that have a few ingredients. 2. Follow Recipes: Pay close attention to instructions and measurements. 3. Learn Basic Skills: Practice chopping, sautéing, and boiling. 4. Use Fresh Ingredients: Fresh ingredients can improve the flavor of your dishes. 5. Taste as You Go: Adjust seasonings as needed. 6. Ask for Help: Don't hesitate to ask friends or family for advice. Would you like some simple recipe ideas to get you started?	Well, maybe you can try making something simple like grilled cheese.

were appropriate. Figure 6.4(a) shows the comparison plot of the two models, Emma and ChatGPT, for this question. The t-value of Emma − ChatGPT was 1.992, and the p-value was 0.061. The results given by the proposed model were found to be appropriate compared to ChatGPT. As the p-value was $0.061(> 0.05)$, the results are close to significance.

For the second question, participants were asked if they agreed that the responses generated by Emma and ChatGPT were empathetic. Figure 6.4(b) shows the comparison plot of the two models, Emma and ChatGPT, for this question. The t-value of Emma − ChatGPT was 2.099, and the p-value was 0.049. The results provided by the proposed model were found to be empathetic compared to ChatGPT, and the results were statistically significant as the p-value was less than 0.05.

In the third question, participants were asked if they agreed that a human or friend was likely to give the same response as the robot. Figure 6.4(c) shows the comparison plot of the two models, Emma and ChatGPT, for this question. The t-value of Emma − ChatGPT was 8.048, and the p-value was < 0.001. The results given by the proposed model were found to be more human-like compared to ChatGPT, and they were statistically significant.

6.2.3 HRI Experiment Evaluation

Pepper robot was used to carry out an HRI experiment. The experiment procedure was as follows.

- First, the participants sat on the chair that was diagonally in line with the robot and the camera that was on top of the table. The camera also had an inbuilt microphone and could listen to the participants.
- In the HRI experiment, the robot starts with a greeting “Hi” and asks the human to share the happiest moment of their life first.
- When a human speaks, the speech is converted to text using Whisper [106]. Next, the converted text is fed to the proposed model as input. After listening to the human response, the robot speaks the output of the model.

Table 6.4: Sample inputs prompts and their corresponding outputs for emotions angry, disgust and sad using ChatGPT and Emma are provided below.

Human	Prompt	ChatGPT	Emma
angry	I am sick and tired of this job. I can't take it anymore.	I'm sorry to hear that you're feeling this way. It's important to prioritize your mental health and well-being. Consider reaching out to someone you trust or a mental health professional for support. Reflecting on what changes you'd like to see or exploring new opportunities might also help. You deserve to feel good about your work.	I understand how you feel, but remember why you started this journey for a reason.
disgust	There is so much filth in this room. I can't stand it.	It looks like the person in the image is showing a strong reaction, possibly disgust. If you're in a situation with a messy room, maybe consider tidying up bit by bit to make it more comfortable. Would you like some organization tips?	I heard they found a way to clean it all up.
sad	I can't believe I worked so hard for the physics exam, and yet I failed.	I'm sorry to hear that. It can be really tough when you put in so much effort and don't get the result you hoped for. Remember that setbacks are part of the learning process, and there are always opportunities to improve. If you need someone to talk to or strategies to help in the future, I'm here to help.	Oh, well, don't worry, I can help you study for the next one.

6

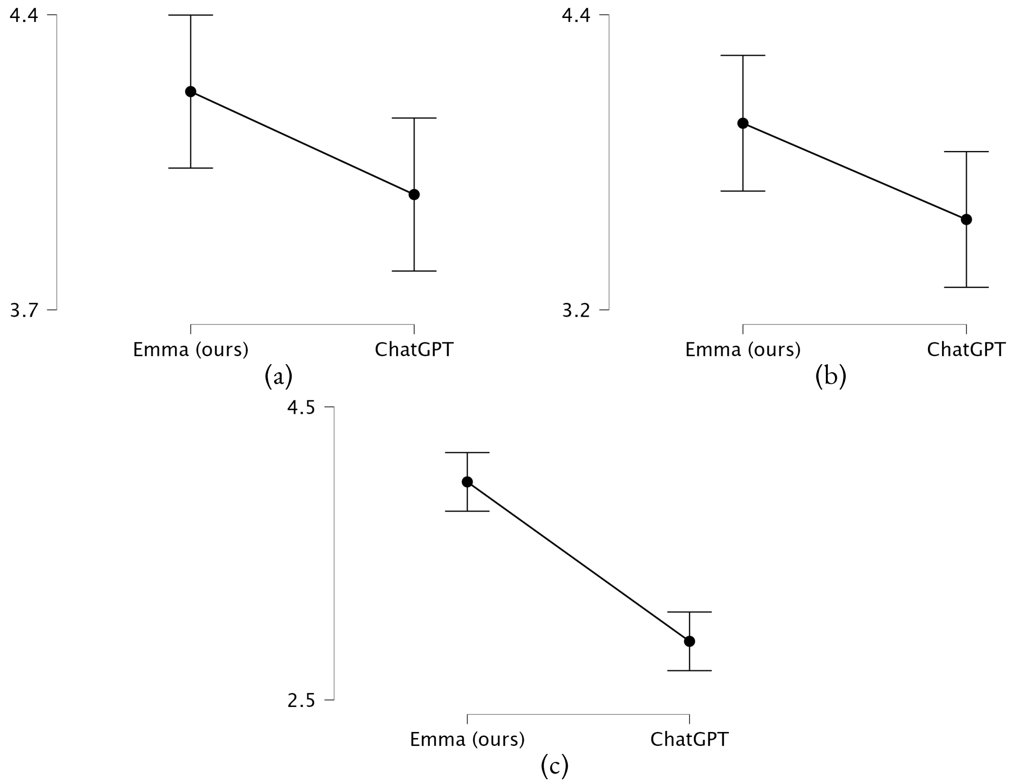


Figure 6.4: Comparison of responses generated by Emma and those generated by ChatGPT using the paired-samples T-test. The appropriateness of the responses is measured in (a), and the empathetic nature of the responses is measured in (b), and (c) measures if the responses are human-like.

Table 6.5: The appropriateness of the answers given by the model is measured. Here, the Likert scale (1) indicates that the participants strongly disagree that the responses are appropriate, and the Likert scale (5) indicates that the participants strongly agree that the responses are appropriate.

Likert Scale Values	1	2	3	4	5
H_0 (same as chance)	6	6	6	6	6
Observations	0	2	5	16	7

Table 6.6: Whether the responses given by the model are empathetic or not is measured. Here, the Likert scale (1) indicates that the participants strongly disagree that the responses are empathetic and the Likert scale (5) indicates that the participants strongly agree that the responses are empathetic.

Likert Scale Values	1	2	3	4	5
H_0 (same as chance)	6	6	6	6	6
Observations	0	2	8	13	7

- Next, the robot asks the human to share something that makes them feel sad.
- After listening to what the human has to say, the robot responds in the same manner as before.
- Finally, the HRI experiment is concluded.

Thereafter, a short survey was conducted, and the participants were asked if the robot communicated appropriately with them and if the robot response was empathetic. These questions were measured on a Likert scale of one to five, where one is ”strongly disagree” and five is ”strongly agree”. In the end, the participants were asked to answer in detail what they thought of the experiment overall.

The results of the HRI experiment survey are as follows. The survey was conducted with a total of 30 participants, of whom 23% were female and 77% were male. The majority of the participants were students aged between 26 and 35 years, with limited or no prior experience with robots. In the first question, the participants were asked whether the robot responded appropriately to them. Then a Chi-squared test was performed using the contingency tables provided in Tab.6.5 and Tab.6.6 as the basis for the analysis.

The null hypothesis (H_0) assumed that the evaluations were distributed by chance. To measure the appropriateness of the robot's responses, the Chi-squared statistic was calculated as 12.713, with a p-value of less than 0.013. Since the p-value was below the threshold of 0.05, the results were statistically significant. Similarly, to evaluate whether the participants perceived the robot's responses as empathetic, the Chi-squared statistic was found to be 10.941, with a p-value of less than 0.027. Again, the results were statistically significant as the p-value was below 0.05.

These findings indicate that the responses generated by the proposed approach were both appropriate and empathetic. Overall, the results demonstrate that participants were able to interact with the Pepper robot effectively and receive appropriate responses using the Emma model.

7

Towards a Generalized Navigation Agent in a Continuous Environment

Embodied tasks span a wide variety of activities, each requiring agents to perceive, reason, and act within an environment. Navigation tasks, such as Object Goal navigation (ObjectNav) [19], involve exploring spaces to reach specific locations or find objects, emphasizing spatial awareness and path planning. Instruction following tasks require agents to interpret and execute natural language commands, often involving multi-step reasoning, such as "Turn left, go to the table, and pick up the red book". Embodied Question Answering (EQA) tasks push agents to gather information from their environment to respond to queries like, "What is the color of the chair in the living room?". In this chapter, a generalized navigation agent is proposed that can navigate, generate instructions for trajectories, and perform embodied question and answering. This chapter is a work in progress. The results of the experiments are not presented. However, the proposed methodology is presented here.

Previous works have tried to model a generalized embodied agent in a discrete environment [55, 104, 153]. Pre-training approaches for vision-and-language navigation (VLN) that aims to create a generic agent have been proposed. The ar-

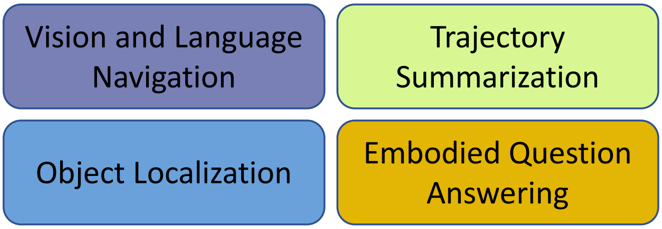


Figure 7.1: A generalized agent is proposed that can navigate, perform object localization, trjectory summarization and embodied question answering

chitectures proposed previously are based on transformer models and involve pretraining and fine-tuning [55, 104]. In one of these works, the transformer model is pre-trained on instruction-trajectory pair and then the model is fine-tuned on R2R, RxR, REVERIE and NDH datasets [104]. Finally, one of the most recent works fine-tune LLMs to build a generalized embodied agent [153]. Inspired by this, in this chapter, a generalized embodied agent is proposed that fine-tunes the Vision and Language model. A difference between prior works and the proposed method in this chapter is that all previous works have tried to model a generalized agent in a discrete environment. This is the first work to attempt to model a generalized navigation agent in a continuous environment. While agents operate on predefined nodes in graphical spaces in the discrete environment, real-world scenarios often involve a continuous environment with dynamic and unstructured layouts. The transition from discrete to continuous environments introduces significant challenges as the action space is larger in continuous environments compared to discrete environments.

One of the previous works discretizes a continuous action space for the Vision and Language Navigation (VLN) task in a continuous environment by iteratively predicting waypoints until a target location is reached [62]. The work presented in this chapter builds upon this work and extends the VLN task in continuous environment to a generalized agent that can, in addition to navigation, summarize trajectories and perform EQA.

7.1 Proposed Method

In this work, a Vision and Language model is fine-tuned to build a generalized agent that can navigate (both using low-level and high-level instructions), summarize trajectories, and perform EQA, in a continuous environment. The following subsections provide the proposed methodology in detail.

7.1.1 Visual Encoder

For any arbitrary point in the open space of an environment, its RGB and depth panoramas are captured. These images are processed by visual encoders to generate a sequence of RGB and depth features denoted v^{RGB} and v^{D}, respectively. Two ResNet-50 [57] models are utilized for encoding RGB and depth images: one pre-trained on ImageNet [115] and the other pre-trained for point-goal navigation [143]. A non-linear layer combines each pair of RGB and depth features to produce v^{RGBD}.

7.1.2 Waypoint Prediction for Navigation Tasks

Each panorama is broken into 12 single-view images spaced 30 degrees apart. After extracting the features of each of the 12 single view images v_i^{RGBD}, the Vision and Language model processes all the 12 representations to model relationships and infer adjacent waypoints. Since each single-view image is square with a 90-degree field of view, covering three 30-degree sectors, the self-attention for each is restricted to include only one neighboring representation on either side. The feature tokens output by the Vision and Language model, each encapsulating information from a sector centered on the corresponding image are passed to a classifier to predict a heatmap with dimensions of 120 angles by 12 distances. Each angle corresponds to a 3-degree increment, while distances range from 0.25 meters to 3.00 meters, with 0.25-meter intervals. Applying non-maximum suppression (NMS) to the resulting heatmap yields K neighboring waypoints.

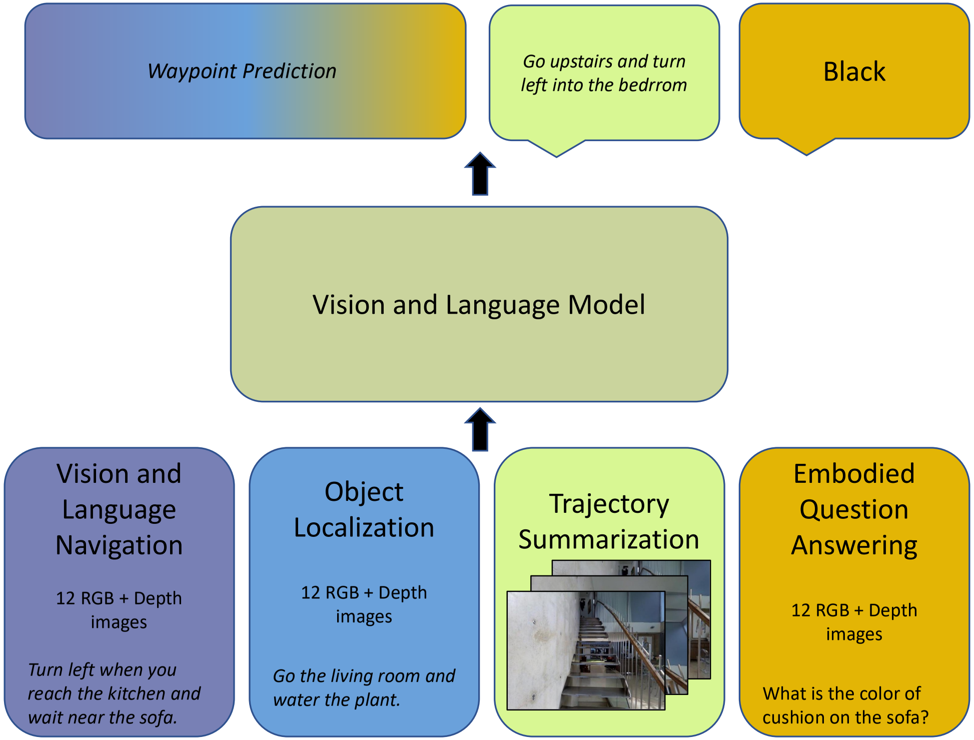


Figure 7.2: Vision and Language model is fine-tuned to perform Vision and Language Navigation, object localization, trajectory summarization and embodied question answering. For navigation tasks 12 RGB and depth images are fed into ResNet-50 models and the features are extracted. These image features along with the corresponding language instruction tokens are fed into the Vision and Language model which predicts the waypoints. For the task of trajectory summarization, the model is fed with frontal images and the language instruction is generated by the Vision and Language model.

7.1.3 Integrated Multi-Task Approach

The tasks, Vision and Language Navigation, Object Localization, Trajectory Summarization and Embodied Question Answering are used to build a generalized embodied agent. In the following, the methodology of each task is described in detail.

Vision and Language Navigation. This task involves guiding the agent to navigate within a 3D environment following language-based instructions. The panoramic view is segmented into 12 RGB and depth images, which serve as input to the Vision and Language model, enabling the prediction of waypoints as described in Section 7.1.2.

Object Localization. It also requires the agent to navigate and localize a target object according to high-level instructions. The panoramic image is again split into 12 RGB and depth views, which are provided as input to the Vision and Language model for waypoint prediction. The waypoints are predicted iteratively until the agent reaches the target object and predicts the stop action.

Trajectory Summarization. This task takes a sequence of frontal images as input and generates the language instruction for the given sequence of images.

Embodied Question Answering. In this task, the agent has to navigate and answer the question in text. The input of the Vision and Language model is panoramic image divided into 12 RGB and depth images. The model predicts the next waypoints iteratively until the agent reaches its target location. Once the 'stop' action is predicted, the question is answered in text by the same Vision and Language model.

8
Conclusions and Future Works

This dissertation is positioned at the confluence of three dynamic fields: Computer Vision, Natural Language Processing (NLP), and Robotics. By exploring and innovating within this interdisciplinary space, it aims to bridge the gap between how robots perceive, understand, and interact with humans in both physical and cognitive dimensions. The work delves into two core topics—Vision and Language Navigation (VLN) and multimodal empathetic dialogue—which together encompass the essence of modern HRI. Below is an expanded discussion of the contributions, followed by reflections on future directions and concluding remarks.

8.1 Contributions of the Thesis

8.1.1 Vision and Language Model to Improve Navigation

A computational model was proposed for the generation of synthetic instruction, based on the state-of-the-art NLP models GPT-2 and BERT. The model leverages a GAN-like structure and takes sequences of images as input to generate instructions for the path traversed by the agent to reach its target location. The model has been trained and validated on REVERIE instructions and achieved high image description metric results when comparing the generated instructions with the ground-truth instructions. Subsequently, using synthetic instructions to augment a VLN dataset, such as REVERIE or R2R, improves the performance of a VLN method achieving state-of-the-art performance on both navigation and object grounding metrics.

8.1.2 Dialogue Generation for Cooperative Navigation

A novel computational model that engages in dialogue while navigating is presented. The proposed architecture consists of a dialogue model and a navigator model: a fine-tuned GPT-2 decoder produces synthetic dialogues, and the navigation is predicted using a modified DUET model. The GPT-2 decoder is a multimodal text generator trained to generate questions using as input the target object and the current observation of the agent, while answers include future images along the trajectory to the goal. The modified DUET model is then trained to navigate using both ground truth annotation and generated dialogues.

Further, an entropy-based "whether-to-ask" policy is learned by minimizing a binary cross-entropy loss that predicts when it is beneficial to generate new dialogues. As a result, UNMuTe learns to navigate more efficiently. We validated the effectiveness of our approach by performing extensive experiments triggering the dialogue model under different policies and settings. The final model achieves state-of-the-art performance on the most common VDN datasets.

8.1.3 Robots with Parallel and Reactive Empathy

An Empathetic-HRI dataset was created by prompting ChatGPT to generate two-line dialogues between two individuals, where Person 1 feels `[EMOTION 1]` and Person 2 provides a `[EMOTION 2]` response. The dialogues were generated for various types of situations, such as at work, at school, while traveling, etc. After the dataset was created, the proposed multimodal model was fine-tuned. The proposed model is based on a transformer encoder-decoder structure designed to generate responses for given prompts by considering both facial expressions and the spoken words of the human.

After fine-tuning, the proposed model generated responses for 17 prompts with different emotions, which were then compared to responses generated by ChatGPT-4o. A survey was conducted to gather human participants' opinions on whether the responses from the two models were appropriate and empathetic. A paired-samples t-test was conducted, revealing significant results. These results indicated that the responses generated by our model were both more appropriate and empathetic compared to those from ChatGPT-4o.

8.1.4 Using Implicit Rewards with Human Emotions

A computational model is proposed that sees the person's facial expression and listens to what they are saying to respond to them in an appropriate and empathetic manner. The model is fine-tuned using reinforcement learning (RL) to respond positively. Qualitative examples based on various emotions were provided to show the kind of responses generated by the model. To evaluate the model quantitatively, a survey was conducted with 17 prompts of varying emotions, where we compared the responses generated by our model with those generated by ChatGPT-4o. The responses generated by our model were found to be both empathetic and human-like, compared to ChatGPT-4o, using the paired-samples T-test. Moreover, a Human-Robot Interaction (HRI) experiment was conducted in which the robot asked participants to share the happiest moments of their lives and something that made them feel sad. After the participants shared these with the robot, the robot responded to them. After the experiment, a survey was conducted in which the participants were if the responses given by the robot were appropriate and empathetic on a five-point Likert scale. A chi-square test was performed to show that the results obtained were statistically significant and that our computational model responded appropriately and empathetically.

8.1.5 Towards a Generalized Navigation Agent

A generalized navigation agent in a continuous environment is proposed. Besides performing Vision and Language Navigation, this generalized agent can also perform object localization, trajectory summarization and Embodied Question Answering. The proposed model fine-tunes a Vision and Language model to perform these tasks. As this chapter is a work in progress, it only presented the proposed methodology without the experimental results.

8.2 Future Works

The research presented in this dissertation opens several promising avenues for future exploration in Vision and Language Navigation (VLN) and Multimodal Empathetic Dialogue. By extending the capabilities and addressing the limitations of current models, future work can push the boundaries of human-robot interaction (HRI) in both physical and cognitive domains. Below are the key directions for future research:

Real World Deployment. While the experiments detailed in Chapters 3, 4, and 5 demonstrated promising results in controlled settings, their effectiveness in real-world environments remains unvalidated. Future research could address this gap by deploying these models in dynamic, unstructured environments, tackling challenges such as handling noisy and incomplete inputs, adapting to the variability and unpredictability of human behavior, and scaling to diverse, large-scale applications. Real-world deployment would necessitate robust mechanisms for real-time adaptation, continuous learning from user interactions, and efficient resource management to ensure responsiveness and reliability. Bridging the gap between theoretical performance and practical utility will be crucial for these systems to function effectively in real-world scenarios, such as assistive robotics, public spaces, and personalized support settings.

Human-in-the-loop Systems. This represent a promising approach to enhance human-robot interaction (HRI) by actively involving humans in the operational loop, enabling dynamic learning and adaptation. Interactive feedback mechanisms allow systems to incorporate real-time user feedback to refine their behavior and improve performance. By continuously learning from human inputs during interactions, robots can adjust to individual preferences, resolve ambiguities, and deliver more natural and intuitive responses, fostering a sense of collaboration and trust. Especially the work presented in Chapter 4 can be extended to a human-in-the-loop system, where the agent can ask human a question whenever it gets confused regarding where to navigate.

In this manner, collaborative AI frameworks enable robots and humans to work as partners in solving tasks, with robots leveraging human expertise and guidance while providing assistance in areas requiring precision, automation, or efficiency. Such systems not only boost task effectiveness but also allow robots to learn iteratively from human interactions, making them more adaptable and capable over time.

Enhancing Trust through Explainability. Future work on building systems capable of explaining their decisions and actions to users can focus on developing robust Explainable AI (XAI) frameworks tailored for HRI. Such systems should provide transparent, user-friendly justifications for their behavior, bridging the gap between complex machine reasoning and human understanding. This involves creating models that not only make accurate decisions but also generate clear, interpretable explanations in real-time, suitable for diverse users with varying levels of technical expertise. By integrating natural language explanations with visual or interactive demonstrations, these systems can help users comprehend the underlying logic of decisions, especially in critical applications like healthcare, autonomous navigation, or emotional support. For example, Poppi *et al*. introduce methods for autonomous agents to provide interpretable explanations and recount their navigation decisions, enhancing transparency and user trust in navigation systems [2]. Research can explore how explanations can adapt dynamically to user preferences and contexts, further enhancing trust and engagement. Ultimately, developing explainable systems will not only improve user confidence but also foster collaboration, as users will be more likely to guide, correct, and work alongside systems they can understand.

Cross-Task Integration for a Unified HRI. In this dissertation the two domains of pHRI and cHRI are kept segregated. Future systems could seamlessly integrate physical tasks (e.g., navigation) with cognitive tasks (e.g., empathetic dialogue) to provide holistic interaction experiences. For instance, a robot could navigate to a user while simultaneously providing emotional support or engaging in conversation. A computational model to integrate both pHRI and cHRI tasks could be proposed. For example, this could probably be achieved by using continual learning [91, 112], where the robot is first taught a cHRI task like multimodal empathetic response generation and then a pHRI task like vision and language navigation, without making it forget about the task that was learned previously.

8.3 Final Remarks

This dissertation was completed as part of the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie Action "Personalized Robotics as Service Oriented Applications" (PERSEO), Grant agreement no. 955778, and carried out in Information and Communication Technologies, at the AImageLab research laboratory of the University of Modena and Reggio Emilia. The works described in this dissertation also contributed to the research of the project "Fit for Medical Robotics" ("Fit4MedRob"), funded by the Italian Ministry of University and Research.

This dissertation serves as a stepping stone toward the development of human-centric AI systems. By addressing the dual dimensions of physical and cognitive interaction, it opens new avenues for research and practical applications in HRI. The work undertaken during this Ph.D. program, including publishing research papers, presenting findings at conferences, and collaborating with interdisciplinary teams, has significantly contributed to advancing the state-of-the-art in computer vision, NLP, and robotics.

8.4 Ph.D. Activities

The final section presents a list of the main activities carried out during the Ph.D. program.

8.4.1 Exchange Periods

15 January - 16 February 2024: PERSEO Secondment at Cognitive Robotics Lab (CoRoLab), University of Manchester, United Kingdom ;

15 July - 31 October 2024: Research Exchange at Robot Learning Lab, University of Freiburg, Germany;

9 September - 13 September 2024: PERSEO Secondment at Omitech SRL, Italy.

8.4.2 Conferences and Workshops Attended

25 November - 27 November 2021: PERSEO First Residential Workshop, Pisa, Italy;

21 June - 23 June 2022: PERSEO Second Residential Workshop, Barcelona, Spain;

29 August - 2 September 2022: 31st IEEE International Conference on Robot & Human Interactive Communication (RO-MAN), 2022, Naples, Italy;

21 February - 23 February 2023: PERSEO Third Residential Workshop, Modena, Italy;

13 March - 16 March 2023: ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2023, Stockholm, Sweden;

18 September - 20 September 2023: PERSEO Fourth Residential Workshop, Vienna, Austria;

4 March - 6 March 2024: PERSEO Fifth Residential Workshop, Munich, Germany;

30 September - 1 October 2024: 17th International Workshop on Human-Friendly Robotics (HFR), 2024, Lugano, Switzerland.

8.4.3 Schools

23 March - 25 March 2022: PERSEO Winter School on Ethics and Interdisciplinary Research Methods for Personalized Robotics, Amsterdam, Netherlands (Remote);

22 August - 26 August 2022: 5th Advanced Online and Onsite Course on Data Science and Machine Learning, Siena, Italy;
20 September - 22 September 2022: PERSEO Summer School on Service-Based and Cloud Robotics, Palma de Majorca, Spain.

List of Publications

[1] Maharjan, R. S., Rawal, N., Romeo, M., Baraldi, L., Cucchiara, R., and Cangelosi, A. (2025). Multimodal emotion recognition in conversation via possible speaker's audio and visual sequence selection. In *Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing*.

[2] Poppi, S., Bigazzi, R., Rawal, N., Cornia, M., Cascianelli, S., Baraldi, L., and Cucchiara, R. (2023). Towards Explainable Navigation and Recounting. In *Proceedings of the International Conference on Image Analysis and Processing*.

[3] Rawal, N., Baraldi, L., Cucchiara, R., et al. (2025). AIGeN-Llama: An Adversarial Approach for Instruction Generation in VLN using Llama2 Model. In *Proceedings of the 21st Conference on Information and Research Science Connecting to Digital and Library Science*.

[4] Rawal, N., Bigazzi, R., Baraldi, L., and Cucchiara, R. Unmute: Unifying Navigation and Multimodal Dialogue-Like Text Generation. *Submitted to Computer Vision and Image Understanding*.

[5] Rawal, N., Bigazzi, R., Baraldi, L., and Cucchiara, R. (2024a). AIGeN: An Adversarial Approach for Instruction Generation in VLN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 2070–2080.

[6] Rawal, N., Maharjan, R. S., Romeo, M., Bigazzi, R., Baraldi, L., Cucchiara, R., and Cangelosi, A. (2024b). Intelligent multimodal artificial agents that talk and express emotions. In *International Workshop on Human-Friendly Robotics*, pages 240–254. Springer.

[7] Rawal, N., Maharjan, R. S., Salici, G., Riccardo, C., Romeo, M., Bigazzi, R., Baraldi, L., Roberto, V., Cucchiara, R., and Cangelosi, A. Multimodal Dialogue for Empathetic Human-Robot Interaction. *Submitted to International Conference on Social Robotics*.

[8] Rawal, N., Xia, M., Tessaro, D., Baraldi, L., and Cucchiara, R. MATE: Multimodal Agents that Talk and Empathesize. *Submitted to Proceedings of the International Conference on Image Analysis and Processing*.

Bibliography

[9] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

[10] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. (2022). Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

[11] Alexey, D. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*.

[12] Anderson, P., Fernando, B., Johnson, M., and Gould, S. (2016). SPICE: Semantic Propositional Image Caption Evaluation. In *Proceedings of the European Conference on Computer Vision*.

[13] Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., and Van Den Hengel, A. (2018). Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[14] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

[15] Bagheri, E. (2022). Improving human-robot interaction through empathy and transparency.

[16] Banerjee, S. and Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics Workshops*.

[17] Bao, H., Wang, W., Dong, L., Liu, Q., Mohammed, O. K., Aggarwal, K., Som, S., Piao, S., and Wei, F. (2022). Vlmo: Unified vision-language pre-

training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912.

[18] Barger, P. B. and Grandey, A. A. (2006). Service with a smile and encounter satisfaction: Emotional contagion and appraisal mechanisms. *Academy of management journal*, 49(6):1229–1238.

[19] Batra, D., Gokaslan, A., Kembhavi, A., Maksymets, O., Mottaghi, R., Savva, M., Toshev, A., and Wijmans, E. (2020). Objectnav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171*.

[20] Bazzano, F. and Lamberti, F. (2018). Human-robot interfaces for interactive receptionist systems and wayfinding applications. *Robotics*, 7(3):56.

[21] Bigazzi, R., Baraldi, L., Kousik, S., Cucchiara, R., and Pavone, M. (2024). Mapping high-level semantic regions in indoor environments without object recognition. *arXiv preprint arXiv:2403.07076*.

[22] Bigazzi, R., Cornia, M., Cascianelli, S., Baraldi, L., and Cucchiara, R. (2023). Embodied Agents for Efficient Exploration and Smart Scene Description. In *Proceedings of the IEEE International Conference on Robotics and Automation*.

[23] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

[24] Buzzega, P., Boschini, M., Porrello, A., Abati, D., and Calderara, S. (2020). Dark experience for general continual learning: a strong, simple baseline. *Advances in Neural Information Processing Systems*, 33:15920–15930.

[25] Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE.

[26] Castillo, J. C., Castro-González, Á., Alonso-Martín, F., Fernández-Caballero, A., and Salichs, M. Á. (2018). Emotion detection and regulation from personal assistant robot in smart environment. *Personal assistants: Emerging computational technologies*, pages 179–195.

[27] Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., and Zhang, Y. (2017). Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*.

[28] Chaplot, D. S., Salakhutdinov, R., Gupta, A., and Gupta, S. (2020). Neural topological slam for visual navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12875–12884.

[29] Chen, H., Suhr, A., Misra, D., Snavely, N., and Artzi, Y. (2019). Touchdown: Natural Language Navigation and Spatial Reasoning in Visual Street Environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[30] Chen, S., Guhur, P.-L., Schmid, C., and Laptev, I. (2021). History Aware Multimodal Transformer for Vision-and-Language Navigation. *Advances in Neural Information Processing Systems*.

[31] Chen, S., Guhur, P.-L., Tapaswi, M., Schmid, C., and Laptev, I. (2022a). Learning from Unlabeled 3D Environments for Vision-and-Language Navigation. In *Proceedings of the European Conference on Computer Vision*.

[32] Chen, S., Guhur, P.-L., Tapaswi, M., Schmid, C., and Laptev, I. (2022b). Think Global, Act Local: Dual-scale Graph Transformer for Vision-and-Language Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[33] Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., and Girdhar, R. (2022). Masked-attention Mask Transformer for Universal Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[34] Chi, T.-C., Shen, M., Eric, M., Kim, S., and Hakkani-tur, D. (2020). Just ask: An interactive learning framework for vision and language navigation. In *Proceedings of the Conference on Artificial Intelligence*.

[35] Dai, B., Fidler, S., Urtasun, R., and Lin, D. (2017). Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE international conference on computer vision*, pages 2970–2979.

[36] Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., and Batra, D. (2018). Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10.

[37] Davis, M. (2018). *Empathy: A Social Psychological Approach*.

[38] De Vries, H., Shuster, K., Batra, D., Parikh, D., Weston, J., and Kiela, D. (2018). Talk the Walk: Navigating New York City through Grounded Dialogue. *arXiv preprint arXiv:1807.03367*.

[39] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

[40] Dong, J., Santiago-Anaya, A., and Jeon, M. (2023). Facial expressions increase emotion recognition clarity and improve warmth and attractiveness on a humanoid robot without adding the uncanny valley. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 67, pages 933–939. SAGE Publications Sage CA: Los Angeles, CA.

[41] Donnermann, M., Schaper, P., and Lugrin, B. (2022). Social robots in applied settings: A long-term study on adaptive robotic tutors in higher education. *Frontiers in Robotics and AI*, 9:831633.

[42] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations*.

[43] Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al. (2023). Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

[44] Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

[45] Ebner, N. C., Riediger, M., and Lindenberger, U. (2010). Faces—a database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior research methods*, 42:351–362.

[46] Fried, D., Hu, R., Cirik, V., Rohrbach, A., Andreas, J., Morency, L.-P., Berg-Kirkpatrick, T., Saenko, K., Klein, D., and Darrell, T. (2018). Speaker-

follower models for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 31.

[47] Fu, T.-J., Wang, X. E., Peterson, M. F., Grafton, S. T., Eckstein, M. P., and Wang, W. Y. (2020). Counterfactual Vision-and-Language Navigation via Adversarial Path Sampler. In *Proceedings of the European Conference on Computer Vision*, pages 71–86. Springer.

[48] Gao, C., Peng, X., Yan, M., Wang, H., Yang, L., Ren, H., Li, H., and Liu, S. (2023). Adaptive Zone-aware Hierarchical Planner for Vision-Language Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[49] Gao, X., Gao, Q., Gong, R., Lin, K., Thattai, G., and Sukhatme, G. S. (2022). DialFRED: Dialogue-Enabled Agents for Embodied Instruction Following. *IEEE Robotics and Automation Letters*, 7.

[50] Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., et al. (2013). Challenges in representation learning: A report on three machine learning contests. In *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*, pages 117–124. Springer.

[51] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*.

[52] Gordon, D., Kembhavi, A., Rastegari, M., Redmon, J., Fox, D., and Farhadi, A. (2018). Iqa: Visual question answering in interactive environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4089–4098.

[53] Guhur, P.-L., Tapaswi, M., Chen, S., Laptev, I., and Schmid, C. (2021). Airbert: In-Domain Pretraining for Vision-and-Language Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[54] Gunson, N., Garcia, D. H., Sieińska, W., Addlesee, A., Dondrup, C., Lemon, O., Part, J. L., and Yu, Y. (2022). A visually-aware conversational robot receptionist. In *Association for Computational Linguistics (ACL)*.

[55] Hao, W., Li, C., Li, X., Carin, L., and Gao, J. (2020). Towards learning a generic agent for vision-and-language navigation via pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13137–13146.

[56] Häring, M., Bee, N., and André, E. (2011). Creation and evaluation of emotion expression with body movement, sound and eye color for humanoid robots. In *2011 RO-MAN*, pages 204–209. IEEE.

[57] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

[58] Hennig-Thurau, T., Groth, M., Paul, M., and Gremler, D. D. (2006). Are all smiles created equal? how emotional contagion and emotional labor affect service relationships. *Journal of marketing*, 70(3):58–73.

[59] Heredia, J., Lopes-Silva, E., Cardinale, Y., Diaz-Amado, J., Dongo, I., Graterol, W., and Aguilera, A. (2022). Adaptive multimodal emotion detection architecture for social robots. *IEEE Access*, 10:20727–20744.

[60] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[61] Hong, Y., Rodriguez, C., Qi, Y., Wu, Q., and Gould, S. (2020). Language and visual entity relationship graph for agent navigation. *Advances in Neural Information Processing Systems*, 33:7685–7696.

[62] Hong, Y., Wang, Z., Wu, Q., and Gould, S. (2022). Bridging the gap between learning in discrete and continuous environments for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15439–15449.

[63] Hong, Y., Wu, Q., Qi, Y., Rodriguez-Opazo, C., and Gould, S. (2021). VLN BERT: A Recurrent Vision-and-Language BERT for Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[64] Hudson, D. A. and Zitnick, L. (2021). Generative Adversarial Transformers. In *Proceedings of the International Conference on Machine Learning*.

[65] Ilharco, G., Jain, V., Ku, A., Ie, E., and Baldridge, J. (2019). General evaluation for instruction conditioned navigation using dynamic time warping. *Advances in Neural Information Processing Systems*.

[66] Jiang, Y., Chang, S., and Wang, Z. (2021). TransGAN: Two Pure Transformers Can Make One Strong GAN, and That Can Scale Up. *Advances in Neural Information Processing Systems*.

[67] Kamath, A., Anderson, P., Wang, S., Koh, J. Y., Ku, A., Waters, A., Yang, Y., Baldridge, J., and Parekh, Z. (2023). A New Path: Scaling Vision-and-Language Navigation with Synthetic Instructions and Imitation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[68] Kim, W., Son, B., and Kim, I. (2021). Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR.

[69] Kingma, D. P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. *Proceedings of the International Conference on Learning Representations*.

[70] Landi, F., Baraldi, L., Cornia, M., Corsini, M., and Cucchiara, R. (2021). Multimodal Attention Networks for Low-Level Vision-and-Language Navigation. *Computer Vision and Image Understanding*.

[71] Landi, F., Baraldi, L., Corsini, M., and Cucchiara, R. (2019). Embodied Vision-and-Language Navigation with Dynamic Convolutional Filters. In *Proceedings of the British Machine Vision Conference*.

[72] Landi, F., Bigazzi, R., Cornia, M., Cascianelli, S., Baraldi, L., and Cucchiara, R. (2022). Spot the Difference: A Novel Task for Embodied Agents in Changing Environments. In *Proceedings of the International Conference on Pattern Recognition*.

[73] Leite, I., Pereira, A., Mascarenhas, S., Martinho, C., Prada, R., and Paiva, A. (2013). The influence of empathy in human–robot relations. *International journal of human-computer studies*, 71(3):250–260.

[74] Li, J., Li, D., Xiong, C., and Hoi, S. (2022a). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the International Conference on Machine Learning*.

[75] Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., et al. (2022b). Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975.

[76] Li, S. and Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(3):1195–1215.

[77] Li, S., Deng, W., and Du, J. (2017). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2852–2861.

[78] Li, X., Li, C., Xia, Q., Bisk, Y., Celikyilmaz, A., Gao, J., Smith, N., and Choi, Y. (2019). Robust Navigation with Language Pretraining and Stochastic Sampling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[79] Li, X., Wang, Z., Yang, J., Wang, Y., and Jiang, S. (2023). KERM: Knowledge Enhanced Reasoning for Vision-and-Language Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[80] Liang, X., Zhu, F., Li, L., Xu, H., and Liang, X. (2022). Visual-Language Navigation Pretraining via Prompt-based Environmental Self-exploration. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[81] Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics Workshops*.

[82] Lin, X., Li, G., and Yu, Y. (2021). Scene-Intuitive Agent for Remote Embodied Visual Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[83] Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2024). Visual instruction tuning. *Advances in neural information processing systems*, 36.

[84] Loshchilov, I., Hutter, F., et al. (2017). Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5.

[85] Lundqvist, D., Flykt, A., and Öhman, A. (1998). Karolinska directed emotional faces. *PsycTESTS Dataset*, 91:630.

[86] Ma, C.-Y., Lu, J., Wu, Z., AlRegib, G., Kira, Z., Socher, R., and Xiong, C. (2019a). Self-Monitoring Navigation Agent via Auxiliary Progress Estimation. In *Proceedings of the International Conference on Learning Representations*.

[87] Ma, C.-Y., Wu, Z., AlRegib, G., Xiong, C., and Kira, Z. (2019b). The Regretful Agent: Heuristic-Aided Navigation through Progress Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[88] Maaz, M., Rasheed, H., Khan, S., and Khan, F. S. (2023). Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

[89] Maharjan, R. S., Romeo, M., and Cangelosi, A. (2023). Faces are domains: Domain incremental learning for expression recognition. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

[90] Majumdar, A., Shrivastava, A., Lee, S., Anderson, P., Parikh, D., and Batra, D. (2020). Improving vision-and-language navigation with image-text pairs from the web. In *Proceedings of the European Conference on Computer Vision*.

[91] McCloskey, M. and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

[92] McQuiggan, S. W., Robison, J. L., Phillips, R., and Lester, J. C. (2008). Modeling parallel and reactive empathy in virtual agents: An inductive approach. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*, pages 167–174. Citeseer.

[93] Mehta, H., Artzi, Y., Baldridge, J., Ie, E., and Mirowski, P. (2020). Retouchdown: Releasing Touchdown on StreetLearn as a Public Resource for Language Grounding Tasks in Street View. In *Proceedings of the Third International Workshop on Spatial Language Understanding*.

[94] Mirowski, P., Grimes, M., Malinowski, M., Hermann, K. M., Anderson, K., Teplyashin, D., Simonyan, K., Zisserman, A., Hadsell, R., et al. (2018).

Learning to navigate in cities without a map. *Advances in neural information processing systems*, 31.

[95] Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31.

[96] Nemoto, T. and Beglar, D. (2014). Likert-scale questionnaires. In *JALT 2013 conference proceedings*, volume 108, pages 1–6.

[97] Nguyen, K. and Daumé III, H. (2019). Help, anna! visual navigation with natural multimodal assistance via retrospective curiosity-encouraging imitation learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[98] OpenAI (2024). Chatgpt.

[99] Padmakumar, A., Thomason, J., Shrivastava, A., Lange, P., Narayan-Chen, A., Gella, S., Piramuthu, R., Tur, G., and Hakkani-Tur, D. (2022). TEACh: Task-Driven Embodied Agents That Chat. In *Proceedings of the Conference on Artificial Intelligence*, volume 36, pages 2017–2025.

[100] Pai, R. Y., Shetty, A., Dinesh, T. K., Shetty, A. D., and Pillai, N. (2024). Effectiveness of social robots as a tutoring and learning companion: a bibliometric analysis. *Cogent Business & Management*, 11(1):2299075.

[101] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[102] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

[103] Qi, Y., Wu, Q., Anderson, P., Wang, X., Wang, W. Y., Shen, C., and Hengel, A. v. d. (2020). REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[104] Qiao, Y., Qi, Y., Hong, Y., Yu, Z., Wang, P., and Wu, Q. (2022). Hop: history-and-order aware pre-training for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15418–15427.

[105] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

[106] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

[107] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training. *OpenAI*.

[108] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*.

[109] Ramakrishnan, S. K., Gokaslan, A., Wijmans, E., Maksymets, O., Clegg, A., Turner, J., Undersander, E., Galuba, W., Westbury, A., Chang, A. X., et al. (2021). Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D Environments for Embodied AI. *Advances in Neural Information Processing Systems*.

[110] Rawal, N., Koert, D., Turan, C., Kersting, K., Peters, J., and Stock-Homburg, R. (2022). ExGenNet: Learning to Generate Robotic Facial Expression using Facial Expression Recognition. *Frontiers in Robotics and AI*, 8:730317.

[111] Rawal, N. and Stock-Homburg, R. M. (2022). Facial emotion expressions in human–robot interaction: A survey. *International Journal of Social Robotics*, 14(7):1583–1604.

[112] Robins, A. (1995). Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146.

[113] Roman, H. R., Bisk, Y., Thomason, J., Celikyilmaz, A., and Gao, J. (2020). Rmm: A recursive mental model for dialog navigation. *arXiv preprint arXiv:2005.00728*.

[114] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

[115] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252.

[116] Salem, M., Kopp, S., Wachsmuth, I., Rohlfing, K., and Joublin, F. (2012). Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics*, 4:201–217.

[117] Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., et al. (2019). Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347.

[118] Sawik, B., Tobis, S., Baum, E., Suwalska, A., Kropińska, S., Stachnik, K., Pérez-Bernabeu, E., Cildoz, M., Agustin, A., and Wieczorowska-Tobis, K. (2023). Robots for elderly care: Review, multi-criteria optimization model and qualitative case study. In *Healthcare*, volume 11, page 1286. MDPI.

[119] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

[120] Shetty, R., Rohrbach, M., Anne Hendricks, L., Fritz, M., and Schiele, B. (2017). Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4135–4144.

[121] Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L., and Fox, D. (2020). Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749.

[122] Shrivastava, A., Gopalakrishnan, K., Liu, Y., Piramuthu, R., Tür, G., Parikh, D., and Hakkani-Tür, D. (2021). Visitron: Visual semantics-aligned interactively trained object-navigator. *arXiv preprint arXiv:2105.11589*.

[123] Smith, L. W. and Rose, R. L. (2020). Service with a smiley face: Emojional contagion in digitally mediated relationships. *International Journal of Research in Marketing*, 37(2):301–319.

[124] Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., and Cucchiara, R. (2022). From Show to Tell: A Survey on Deep Learning-based

Image Captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[125] Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. (2020). Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

[126] Tan, H., Yu, L., and Bansal, M. (2019). Learning to Navigate Unseen Environments: Back Translation with Environmental Dropout. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

[127] Tang, Z., Yang, Z., Khademi, M., Liu, Y., Zhu, C., and Bansal, M. (2024a). Codi-2: In-context interleaved and interactive any-to-any generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27425–27434.

[128] Tang, Z., Yang, Z., Zhu, C., Zeng, M., and Bansal, M. (2024b). Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36.

[129] Thomason, J., Murray, M., Cakmak, M., and Zettlemoyer, L. (2020). Vision-and-dialog navigation. In *Proceedings of the Conference on Robot Learning*.

[130] Tong, Z., Song, Y., Wang, J., and Wang, L. (2022). Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093.

[131] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023a). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

[132] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023b). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

[133] Tsiourti, C., Weiss, A., Wac, K., and Vincze, M. (2019). Multimodal integration of emotional signals from voice, body, and context: Effects of (in)

congruence on emotion recognition and attitudes towards robots. *International Journal of Social Robotics*, 11:555–573.

[134] van Otterdijk, M., Laeng, B., Lindblom, D. S., and Torresen, J. (2024). The effect of expressive robot behavior on users' mental effort: A pupillometry study. *IEEE Transactions on Cognitive and Developmental Systems*.

[135] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

[136] Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). CIDEr: Consensus-based Image Description Evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[137] Vercelli, A., Rainero, I., Ciferri, L., Boido, M., and Pirri, F. (2018). Robots in elderly care. *DigitCult-Scientific Journal on Digital Cultures*, 2(2):37–50.

[138] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

[139] Wang, H., Wang, W., Liang, W., Xiong, C., and Shen, J. (2021a). Structured Scene Memory for Vision-Language Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[140] Wang, Q., Wan, J., and Chan, A. B. (2020). On diversity in image captioning: Metrics and methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):1035–1049.

[141] Wang, S., Montgomery, C., Orbay, J., Birodkar, V., Faust, A., Gur, I., Jaques, N., Waters, A., Baldridge, J., and Anderson, P. (2021b). Less is More: Generating Grounded Navigation Instructions from Landmarks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[142] Wang, X., Huang, Q., Celikyilmaz, A., Gao, J., Shen, D., Wang, Y.-F., Wang, W. Y., and Zhang, L. (2019). Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[143] Wijmans, E., Kadian, A., Morcos, A., Lee, S., Essa, I., Parikh, D., Savva, M., and Batra, D. (2019). Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. *arXiv preprint arXiv:1911.00357*.

[144] Wu, S., Fei, H., Qu, L., Ji, W., and Chua, T.-S. (2023). Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*.

[145] Xia, F., Zamir, A. R., He, Z., Sax, A., Malik, J., and Savarese, S. (2018). Gibson Env: Real-World Perception for Embodied Agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[146] Xu, H., Ghosh, G., Huang, P.-Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., and Feichtenhofer, C. (2021). Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*.

[147] Xu, K. (2015). Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*.

[148] Yang, D., Sato, W., Liu, Q., Minato, T., Namba, S., and Nishida, S. (2022). Optimizing facial expressions of an android robot effectively: a bayesian optimization approach. In *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*, pages 542–549. IEEE.

[149] Yang, Z., He, X., Gao, J., Deng, L., and Smola, A. (2016). Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29.

[150] Yenamandra, S., Ramachandran, A., Yadav, K., Wang, A., Khanna, M., Gervet, T., Yang, T.-Y., Jain, V., Clegg, A. W., Turner, J., Kira, Z., Savva, M., Chang, A., Chaplot, D. S., Batra, D., Mottaghi, R., Bisk, Y., and Paxton, C. (2023). Homerobot: Open vocab mobile manipulation.

[151] Youssef, K., Said, S., Alkork, S., and Beyrouthy, T. (2022). A survey on recent advances in social robotics. *Robotics*, 11(4):75.

[152] Zhang, H., Li, X., and Bing, L. (2023). Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

[153] Zheng, D., Huang, S., Zhao, L., Zhong, Y., and Wang, L. (2024). Towards learning a generalist model for embodied navigation. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13624–13634.

[154] Zheng, Q., Liu, D., Wang, C., Zhang, J., Wang, D., and Tao, D. (2023). Esceme: Vision-and-language navigation with episodic scene memory. *arXiv preprint arXiv:2303.01032*.

[155] Zhu, W., Wang, X. E., Fu, T.-J., Yan, A., Narayana, P., Sone, K., Basu, S., and Wang, W. Y. (2020a). Multimodal text style transfer for outdoor vision-and-language navigation. *arXiv preprint arXiv:2007.00229*.

[156] Zhu, Y., Weng, Y., Zhu, F., Liang, X., Ye, Q., Lu, Y., and Jiao, J. (2021). Self-motivated communication agent for real-world vision-dialog navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1594–1603.

[157] Zhu, Y., Zhu, F., Zhan, Z., Lin, B., Jiao, J., Chang, X., and Liang, X. (2020b). Vision-dialog navigation by exploring cross-modal memory. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10730–10739.

Glossary

AI branch of computer science that develops machines and software with human-like intelligence. 1

BLEU bilingual evaluation understudy. 25, 26, 47

cHRI interact with the humans in a natural and adaptive manner while understanding their behavior, intentions and emotions. 2, 3, 89

CIDEr consensus-based image description evaluation. 25, 26, 47, 48

Deep Learning a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks. 1

EQA Embodied Question Answering. 7, 79

GANs Generative Adversarial Networks. 7

HRI Human-Robot Interaction. 1, 2, 65, 85, 89

LLMs Large Language Models. 63

METEOR metric for evaluation of translation with explicit ordering. 25, 26, 47

ObjectNav Object Goal navigation. 79

pHRI make physical movements such as those of navigation or manipulation in a shared Human-Robot environment. 2, 3, 89

RL area of machine learning concerned with how intelligent agents ought to take actions in an environment in order to maximize the notion of cumulative reward. Reinforcement learning is one of three basic machine learning paradigms, alongside supervised learning and unsupervised learning. 4, 7, 65

ROUGE recall oriented understudy of gisting evaluation. 25, 26, 47

SPICE semantic propositional image caption evaluation. 25, 47

SPL success weighted by path length. 25

SR success rate. 25

VDN Vision and Dialogue Navigation. 34

VLN Vision-and-Language Navigation. 2–4, 7

XAI Explainable AI. 89

Acknowledgments

I would like to take this opportunity to express my heartfelt gratitude to everyone who has supported and guided me throughout the journey of completing this dissertation.

First, I am deeply indebted to my supervisors, Prof. Rita Cucchiara and Prof. Lorenzo Baraldi, for their invaluable guidance, encouragement, unwavering support and giving me the opportunity to work on such an interesting topic. Their expertise, insightful advice, and constructive feedback have been a constant source of inspiration and have played a crucial role in shaping the direction and quality of this research.

I am also grateful to Committee Members Prof. Abhinav Valada and Prof. Lamberto Ballan for their thoughtful suggestions, which have enhanced my work. Your suggestions and advice have been immensely helpful.

I also extend my sincere thanks to my colleagues and friends at the AImage Lab, University of Modena and Reggio Emilia. Your collaboration and encouragement have made this journey both productive and enjoyable. Special thanks to Roberto Bigazzi for his technical assistance. I would also like to thank all the collaborators in the Marie Sklodowska-Curie Action project, "Personalized Robotics as Service Oriented Applications" (PERSEO), especially Rahul Singh Maharjan. I would also like to thank Prof. Angelo Cangelosi and Prof. Abhinav Valada for their collaboration.

I am deeply grateful to my family for their unconditional love and encouragement. To my parents and my sister, your belief in me and your constant support have been the pillars that kept me grounded and motivated throughout this process. Special thanks to my sister, Kriti Rawal for designing the cover page of this dissertation.

I would also like to acknowledge the European Union's Horizon 2020 research and innovation program for providing financial support and resources that have been instrumental in the successful completion of this research.

Finally, I wish to thank all those who contributed in some way or another to the completion of this dissertation. Your support and kindness have meant more to me than words can express.

Thank you to everyone for making this milestone possible.