

DISSERTATION

EMBODIED MULTIMODAL REFERRING EXPRESSIONS GENERATION

Submitted by

Nada H. Alalyani

Department of Computer Science

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2024

Doctoral Committee:

Advisor: Nikhil Krishnaswamy

Francisco Ortega

Nathaniel Blanchard

Haonan Wang

Copyright by Nada H. Alalyani 2024

All Rights Reserved

ABSTRACT

EMBODIED MULTIMODAL REFERRING EXPRESSIONS GENERATION

Using both verbal and non-verbal modalities in generating definite descriptions of objects and locations is a critical human capability in collaborative interactions. Despite advancements in AI, embodied interactive virtual agents (IVAs) are not equipped to intelligently mix modalities to communicate their intents as humans do, which hampers naturalistic multimodal IVA. We introduce **SCMRE**, a situated corpus of multimodal referring expressions (MREs) intended for training generative AI systems in multimodal IVA, focusing on multimodal referring expressions. Our contributions include: 1) Developing an IVA platform that interprets human multimodal instructions and responds with language and gestures; 2) Providing 24 participants with 10 scenes, each involving ten equally-sized blocks randomly placed on a table. These interactions generated a dataset of 10,408 samples; 3) Analyzing SCMRE, revealing that the utilization of pointing significantly reduces the ambiguity of prompts and increases the efficiency of IVA's execution of humans' prompts; 4) Augmenting and synthesizing SCMRE, resulting in 22,159 samples to generate more data for model training; 5) Finetuning LLaMA 2-chat-13B for generating contextually-correct and situationally-fluent multimodal referring expressions; 6) Integrating the fine-tuned model into the IVA to evaluate the success of the generative model-enabled IVA in communication with humans; 7) Establishing the evaluation process which applies to both humans and IVAs and combines quantitative and qualitative metrics.

ACKNOWLEDGEMENTS

This thesis would not exist without the invaluable assistance and effort of many individuals. Firstly, I extend my sincere gratitude to my advisor, Prof. Nikhil Krishnaswamy, for his countless hours of advising. He imparted the importance of attention to detail, concept re-framing, and the ability to see the bigger picture, all with remarkable patience and a sense of enjoyment. In addition, I am profoundly thankful to my husband for providing me the opportunity to study abroad and for taking on numerous responsibilities to ensure my success. His unwavering support, both emotionally and practically, has been instrumental in this journey. By managing many of the day-to-day tasks and offering constant encouragement, he has allowed me to focus on my research and studies. His belief in my capabilities has been a driving force behind my perseverance and dedication. I also wish to express my deepest appreciation to my family, including my father, mother, sisters, and brothers, for their emotional support. Their encouragement and belief in me have been a constant source of strength. Lastly, I extend my gratitude to our participants for their contributions to the SCMRE data, despite their limited time and the lengthy process involved. Their participation was crucial to the success of this research. The commitment and cooperation of our participants have been essential in advancing our understanding of multimodal referring expressions, and their efforts are deeply appreciated.

DEDICATION

I would like to dedicate this thesis to my family. Their support has been invaluable throughout this journey.

TABLE OF CONTENTS

ABSTRACT	ii
ACKNOWLEDGEMENTS	iii
DEDICATION	iv
LIST OF TABLES	vii
LIST OF FIGURES	ix
Chapter 1 Introduction	1
1.1 Overview	1
1.2 Thesis Contributions	3
1.3 Related Work	4
Chapter 2 Interactive Virtual Agent Development	8
2.1 Overview	8
2.2 Evaluation of Google ASR Performance	9
2.2.1 Data Collection	10
2.2.2 Operationalizations and Calculations	11
2.2.3 Results and Analysis	12
2.3 Setting of the Experimental Platform	16
2.4 Interpretation of Multimodal REs	18
2.4.1 Language Preprocessing	19
2.4.2 Transformation to Predicate Logic Format	21
2.4.3 Attributive Referring Expressions Interpretation	26
2.4.4 Spatial Referring Expressions Interpretation	28
2.4.5 Historical Referring Expressions Interpretation	28
2.4.6 Non verbal Referring Expressions Interpretation	32
2.5 Deterministic Generation of MREs	32
2.6 Summary	34
Chapter 3 A Situated Corpus of MREs	35
3.1 Overview	35
3.2 Human-IVA Collaboration Data Collection	35
3.2.1 Participants	36
3.2.2 Procedures	37
3.2.3 Data Management	37
3.3 Dataset Statistics	38
3.3.1 Events statistics	39
3.3.2 References statistics	41
3.3.3 Prompts statistics	42
3.4 Data Analysis	43
3.5 Summary	45

Chapter 4	MRE Generation Model	46
4.1	Overview	46
4.2	Data Preparation	47
4.2.1	Data Annotation	48
4.2.2	Data Augmentation	48
4.2.3	Data Synthesizing	49
4.2.4	Prompt Engineering	49
4.2.5	Data Partitioning	50
4.3	MRE Generation Model	50
4.3.1	Model Architecture	50
4.3.2	Learning Experiments	51
4.3.3	Results	53
4.4	LLM-IVA Integration	57
4.5	Summary	59
Chapter 5	Evaluation of MREG Models	61
5.1	Overview	61
5.2	Evaluation Metrics	62
5.3	Automated Quantitative Evaluation	66
5.3.1	Contextual Awareness	66
5.3.2	Efficiency of Collaborative Task Completion	69
5.3.3	Contribution with Meaningful Content	71
5.4	Human Based Evaluation	72
5.4.1	Study Design	72
5.4.2	Results and Analysis	73
5.5	Correlation Analysis of Evaluation Metrics	76
5.6	Summary	79
Chapter 6	Conclusion	81
6.1	Thesis Contributions	81
6.2	Broader Implications	82
6.3	Future Research Directions	83
Bibliography	86

LIST OF TABLES

1.1	The existing corpora for generating and comprehending of referring expressions using language, gesture, or ensemble	6
2.1	The Concept of Word Error Rate: “S” substitution, “I” insertion, “D” deletion, “=” match	12
2.2	Statistical testing of the impact of accents’ variety and context restriction on the performance of Google Speech Recognizer using Welch Two Sample <i>t</i> -test.	13
2.3	Statistics (mean and SD) for open recognizer WER and accuracy and restricted recognizer WER and accuracy when recognizing speech from native and foreign speakers . .	13
2.4	The popular domain specific words were incorrectly recognized by Google speech recognizer	18
2.5	Mapping linguistic nouns, actions and relations to VoxSim entries	20
2.6	Transitive and non transitive verbs are used to distinguished between relational REs and transitive attributive REs	24
2.7	Predicate logic format (PLF) transformation for co-gestural verbal REs (Att_RE: Attributive RE, Trans_RE: Transitive RE, Rel_RE: Relational RE, Hist_RE: Historical RE, and Comp_RE: Compound RE). *Numerals in brackets denote variables that must be assigned from prior conversational or non-verbal context (e.g., “it,” “there,” etc.). . .	25
2.8	Transitive and non transitive verbs are used to distinguished between relational REs and transitive attributive REs	30
2.9	Sample of dialogue history, including previously mentioned actions and related objects after executing multimodal (co-gesure speech) or unimodal (speech only or pointing only) prompts.	30
3.1	Quantities of human generated events before removing duplicated referencing and non-referencing events	38
3.2	Quantities of human generated events after removing duplicated referencing and non-referencing events	39
3.3	Quantities of human generated events based on modalities used, including deictic gesture, speech only, or both	40
3.4	Diana generated behaviours using deictic gesture, speech only, or both to act on human generated events illustrated in Table 3.3	41
3.5	Quantities of human generated references based on modalities used including language, gesture and ensemble	41
3.6	Quantities of human generated references based on information used, including attributes, relations, and previous actions, to refer to targets	42
3.7	Quantities of human generated prompts based on information used, including attributes, relations, and previous actions	43
3.8	Quantities of human generated prompts based on modalities used to communicate intents	43
3.9	Contingency table of human RE ambiguity and modalities used: # ambiguous REs by modality type	44

4.1	Quantities of original, augmented, and synthesized datasets	49
4.2	Training, Validation, and Testing Sets	50
4.3	Hyper-parameters of Fine-tuning and Training Time for LLaMA Models	52
5.1	Contingency table of the IVA’s pointings to human targets and her contextual understanding: # of contextual understanding instances based on IVA’s pointings to the targets.	68
5.2	Quantities of testing videos and judgements for referring expressions with distinct modalities and speech types	73
5.3	Contingency table of IVA’s RE fluency and modalities used; Rel: Relational REs, Att: Attributive REs: fluency level by modality type	75
5.4	Contingency table of IVA’s RE clarity and modalities used; Rel: Relational REs, Att: Attributive REs: clarity level by modality type	75

LIST OF FIGURES

2.1	Selected scripts from EGGNOG dataset including highlighted examples of the most common referring expressions	11
2.2	WER and accuracy of OR and RR transcribing the voices recorded from different nationalities, N is the number of transcribed speech by each nationality	14
2.3	The influence of accents on Google ASR (a) Foreign Restricted Speech Recognizer WER vs Domestic Restricted Speech Recognizer WER (b) Foreign Open Speech Recognizer WER vs Domestic Open Speech Recognizer WER.	15
2.4	(a) Insertions, (b) Deletions, and (c) Substitutions occurred during GSR transcription over distinct nationalities. "N" is the total number of tokens in ground scripts of all transcribed audios	16
2.5	The influence of context restriction on Google ASR (a) Foreign Restricted Speech Recognizer WER vs Domestic Restricted Speech Recognizer WER (b) Foreign Open Speech Recognizer WER vs Domestic Open Speech Recognizer WER.	17
2.6	Experimental Diana System: the purple circle indicates where the user is pointing. Without disambiguation, any object within the pointing circle is a potential candidate for a deixis-only RE. Diana’s utterances appear on screen and are spoken aloud via TTS.	19
2.7	The process of speech interpretation (1) Speech is recognized through GSR (2) Tagging and dependency parsing are determined using StanfordCoreNLP (3) Computation of the extracted parse tree to transform the speech to its PLF (4) PLF interpreted and simulated using VoxML and VoxSim.	22
2.8	Transformation to predicates logic format requires 1) Connecting to StanfordCoreNLP, 2) Extracting the action, and 3) Finding the focus and destination objects from the provided linguistic prompts	23
2.9	Additional mandatory steps to compose the PLF 1) Finding determiners, 2) Finding Attributes, 3) Extracting distance relations to space, e.g., “there”, 4) Extracting relations to space, e.g., “to the right”	24
2.10	Transformation of 1) transitive attributive descriptions, 2) relational (to objects) descriptions, 3) historical descriptions to predicates form	26
2.11	Generating deictic gestures. Diana will respond to what she interprets the RE as referring to by pointing to it, which can be used to assess the correctness of her object grounding depending on which object the human actually intended to reference.	32
2.12	Sample dialogue: the interaction is unimodal (deictic gesture only).	33
2.13	Sample dialogue: the interaction from 1–12 is multimodal (co-gestural speech).	33
3.1	Data Collection procedures (a) Humans download the IVA and interact with her (b) The generated data is uploaded to the allocated OneDrive Space (c) The data is processed and merged into one CSV file	36
4.1	The main steps of data preparation, including annotation, augmentation, synthesizing, prompting, and splitting	47
4.2	The MREG architecture under exploration	52

4.3	The loss (a) and perplexity (b) of LLaMA-13B after one epoch of training. The loss (c) and perplexity (d) of LLaMA-7B after 1,107 steps of training	54
4.4	Quantities of Human and LLM Generated Pointing, Linguistic and Multimodal Referring Expressions.	55
4.5	The similarity between human-generated and LLM-generated MREs using BERT-Precision, BERT-Recall, BERT-F1, and Intersection of Union(IoU)	56
4.6	The distribution of BERT-F1 similarity results between human-generated and (13B-1 epoch) LLM-generated MREs	56
4.7	The performance of LLaMA models in predicting positions of target objects	57
4.8	Clarity level of (a) LLM-generated and (b) human-generated REs	57
4.9	Comparing human and IVA-generated REs for identical configurations.	58
4.10	The performance of LLaMA models in predicting positions of target objects	59
5.1	(a) Trial sample of the IVA’s ability to disambiguate the target; (b) Trial sample of human’s ability to correct misunderstanding.	67
5.2	Agent Pointing Success Rate (Metric #4: APSR) includes successful identifications of target objects both without requiring clarification and after requesting and receiving clarification.	68
5.3	Machine Interpretation of Human Communication (Metric #3: MIHC) categorized by (a) utilized modalities; and (b) speech type.	69
5.4	Agent Completion Time of Human Prompts (Metric #1: ACTHP) categorized by (a) Completion time of linguistic prompts; (b) Completion time of pointing-only prompts; and (c) Completion time of multimodal prompts; and (d) Human Interpretation Time of Machine Communication (Metric #2: HITMC)	70
5.5	Machine Multimodal Referencing Success (Metric #5: MMRS) categorized by (a) generation modalities only, and (b) speech type and generation modalities.	71
5.6	Crowdsourcing framework, including human judgments and related metrics, Metric # 6: MOISR, Metric # 7: HOISR, Metric # 8: MRFR, Metric # 9: HRFR, for evaluating MRE generation models.	73
5.7	Each HIT includes two tasks for quantitative and qualitative evaluation of (a) IVA REs and (b) human REs.	74
5.8	(a) Machine Object Identification Success Rate (MOISR) and Human Object Identification Success Rate (HOISR), and (b) Machine References Fluency Rate (MRFR) and Human References Fluency Rate (HRFR)	74

Chapter 1

Introduction

1.1 Overview

Natural language understanding (NLU) stands as a considerable challenge within the domain of artificial intelligence. It encompasses the ability to discern the semantic relationship, “meaning”, between linguistic expressions—be they individual words, phrases, or complete sentences—and underlying non-linguistic intents, as well as the capacity to articulate or communicate one’s own intentions [1].

Human language serves as a channel for communicating the physical world of speakers, with the concept of “symbol grounding” [2], facilitating the mapping of linguistic terms received from a speaker to their corresponding real-world referents. This process enables interlocutors to engage in reasoning and generate meaningful expressions. Nonetheless, recent advancements in language understanding models, exemplified by OpenAI’s ChatGPT, GPT-2 [3], and BERT [4], are constrained by accessing solely the textual representation (form) of language, thereby deviating from the actual use of the language. It may be tempting, therefore, to believe that language generation for conversational agents (CAs) is a solved problem. However, these recent large language models (LLMs) exhibit two significant limitations: firstly, they focus on learning only the syntactic aspects of words, such as Part of Speech [5], and their semantic features e.g., lexical similarity [6], which are insufficient for achieving a deep contextual understanding of language [1]. Secondly, their black-box nature leads to the argument that they are insufficient at demonstrating understanding of communicative intent and also cannot explain why or when these failures occur [1].

Given recent advancements in language modeling, the ability to seamlessly combine multiple modalities will be crucial for the next generation of conversational agents (CAs). As interactive agents become more sophisticated, interpreting both visual and linguistic contexts simultaneously, users will expect them to behave more like humans. Embodied agents, which represent human

interactions, provide the structures that make language "deeply understandable." A number of embodied conversation agents (ECAs) deployed in different task domains, including education [7], negotiations [8], and medical practitioner evaluation [9]. Indeed, various applications of embodied systems, such as robots aiding elderly individuals or virtual agents facilitating educational courses in subjects such as biology or geography, confirm the shared requirement for agents to engage in natural communication about their environments. Whether providing directions, describing spatial relations, or referencing physical objects, effective communication about the surrounding context is pivotal for these systems to interact seamlessly with humans.

As "Embodied worlds", either virtual, physical, or mixed-reality, are a uniquely useful tool to examine what NLP models in use meaningfully learn and "understand," a particular class of ECAs, namely the embodied Interactive Virtual Agent (IVA), is utilized as a measure of "deep understanding" as they provided new opportunities for grounding and interpreting visuals, gestures, locational information, or actions consequences. If one modality (e.g., language) is not communicative, another modality (e.g., gesture) can be used to explore the meaning. A communicative act, C_a , can be modeled as a tuple of expressions from the modalities available to the agent, which convey complementary or redundant information. For example, if the modalities involved are *Speech*, *Gesture*, *Facial expression*, *gaZe*, and *Action*, then $C_a = \langle S, G, F, Z, A \rangle$, of which any may be null or empty. Information in one channel may be supplemented by another channel, and they may disambiguate each other if properly aligned. For instance, if interpreted from the human's point of view, $C_a = \langle S = \text{"left"}, G = [Point_g \wedge Dir = \text{RIGHT}] \rangle$, this may signal a difference in the agents' relative frames of reference [10, 11]. This effect of embodiment is critical for deep understanding of situated language, vis-à-vis something as basic as directional terms.

Due to the fact that objects within a shared situated context as anchors for establishing mutual understanding between interlocutors, *Multimodal Referring Expressions* (MREs) have emerged as a valuable case study for understanding multimodal language use in context. A referring expression is a strategy, leveraging object characteristics and locations, to identify objects within a shared context. Incorporating multiple referring strategies, including speech and deictic gesture, is a

defined task to study the capability of IVAs to demonstrate: 1) the interpretation of the full meaning of intents in real-world human interactions, and 2) the generation of symmetrically-descriptive descriptions.

To enhance deep understanding for IVAs, we propose to explore human referential strategies via bidirectional human-IVA interactions to develop and evaluate an AI model, generating MREs that are fluent and clear, and symmetric and bidirectional in the context they exploit when compared to human-generated REs. This thesis was guided by following these research questions:

- **RQ.1:** How can an IVA effectively extract the communicative intents of others from their deixis and language synthesis?
- **RQ.2:** How can a large language model (LLM) effectively integrate multiple modalities, including speech and gesture, to convey its intentions similar to humans?
- **RQ.3:** What are the metrics that clearly establish whether both interlocutors are able to correctly and efficiently extract the communicative intents of the others from their behavior?

1.2 Thesis Contributions

Existing embodied multimodal language models, such as SayCan [12], PaLM-E [13], and RT-2 [14], have made significant strides in multimodal interaction and language comprehension, with a primary focus on integrating distinct modalities e.g., vision, and language to contextually interpret commands. Nonetheless, these models do not address the incorporation of gesture and speech despite its significance on the social fluency [15]. Therefore, our goal is to bridge this gap by developing an embodied large language model trained with SCMRE to generate fluent and clear referring expressions, with symmetrical and bidirectional use of language and deictic gesture. Our key contributions are:

- Developing an embodied IVA with the capability to interpret and respond using language and gestures to collect MREs from humans.
- Collecting the SCMRE corpus via bidirectional and symmetrical human-IVA interaction.

- Implementing a fine-tuned LLM for generating contextually correct and situationally fluent MREs.
- Applying quantitative and qualitative metrics to evaluate MRE generation for both humans and the IVA.

1.3 Related Work

Recent advancements in human-machine interaction, leveraging human cognition and behaviors, indicate potential for enabling human-like interactions with users [16, 17]. Nonetheless, it is argued that human-machine interaction systems lack of bidirectional and symmetrical recognition and generation of multimodal communication mechanisms [18]. This emphasizes the significance of machines in not solely recognizing but also generating multimodal communication within the domain of referring expressions.

Referring Expressions Generation The psycholinguistic literature shows the impact of deictic gesture on the successful communication of intent and reference for both speakers and hearers [19, 20]. Despite this significant contribution of deictic gesture to the successful communication, [19, 20], early RE generation research prioritized linguistic description, such as relative and absolute properties of objects (e.g., size and color) [21–23] and spatial references [24–27], and relational episodic descriptions [27]. Where non-verbal information, such as deictic gesture, is considered, much prior work focuses on RE comprehension rather than generation, e.g., [28–32]. Agent embodiment features were rarely integrated into generation [33, 34], with most studies treating generation and comprehension separately [27, 35] or it lacks the multimodal generation [36, 37]. The approaches have been used to generate referring expressions fall into two categories—rule based methods [24, 35, 38] and traditional learning methods [24–27].

Multimodal Generative LLMs AI demonstrated advances in the development of multimodal foundation models (MFMs) for multimodal generation [39]. Multimodal transformers, such as

[40], ViLBERT [41], VisualBERT [42], SimVLM [43], BLIP-2 [44] and Flamingo [45], process inputs from various modalities like text, images, and point clouds. Other models focus on processing video, audio, or 3D data understanding [46–48]. These models were pre-trained on large multimodal datasets containing images, audios and languages. Nevertheless, multimodal LLMs have not been pre-trained for understanding and generating a mix of deictic gestures and speech.

Datasets Several datasets and corpora contain human-generated descriptions of target objects in visual scenes, including Bishop [49], Drawer [50], GRE3D3 [51], TUNA [21], RS-VS [26], and recent corpora by Kunze et al. [24] and Doğan et al. [25]. Other RE corpora collected for the purpose of training comprehension models fall into three categories—verbal references only [52–59], gestures only [60–62], and embodied multimodal REs including language and gesture [60, 63–66]. The embodied multimodal referential behaviours were either generated by simulators, such as VoxSim [64], and CAESAR [65], or by humans while referring to static images [60], or referring to outdoor objects in front of the camera [66]. Therefore, we established humans-IVA interactions to elicit humans’ generated referring expressions while interacting with the embodied IVA considering humans expectations from IVAs in the real time as illustrated in Table 1.1.

Metrics Overlap in the properties of human and machine descriptions can be computed according to Dice Coefficient [67], MASI [68], Levenshtein Distance [69], BLEU [70], ROUGE [71], CIDER [72], or METEOR [73]. Alternatively, human judges can evaluate generated REs according to adequacy of reference or naturalness. While adequacy is evaluated by object identification tasks [24, 25, 27, 74], naturalness is evaluated by (1) metrics such as error rate, identification time, and reading time [75, 76] or (2) human ranking of generated references for objects in images or videos [24, 25, 64, 77]. Relatedly, Kozierok et al. [78] argue that evaluating multimodal interactions require a combination of quantitative and qualitative criteria, particularly in task-based situations. We therefore present a task-oriented setting designed to require MREs, and a proposal for qualitatively and quantitatively evaluating how non-verbal strategies complement verbal strategies for situated meaning [79].

Table 1.1: The existing corpora for generating and comprehending of referring expressions using language, gesture, or ensemble

Dataset	Lang	Gesture	Emb.	Type	Source	Inst.
Bishop Corpus [49]	✓	x	x	Cones im- ages	Hum- images	447
Drawer Corpus [50]	✓	x	x	Drawers Images	Hum- images	140
GRE3D3 Corpus [51]	✓	x	x	Sph-cube images	Hum- images	630
TUNA Corpus [21]	✓	x	x	People images	Hum- images	2,280
PointAt [60]	x	✓	✓	image	Lab	220
ReferAt [63]	✓	✓	✓	Video	Lab	242
RefIt [53]	✓	x	x	Image	Image CLEF	130,525
RefCOCO [54]	✓	x	x	Image	MSCOCO	142,209
IPO [61]	x	✓	✓	Image	Lab	287
Flickr30k entities [57]	✓	x	x	Image	Flickr30K	158,915
IMHF [62]	x	✓	✓	Image	Lab	1716
RefCOCO+ [55]	✓	x	x	Image	MSCOCO	141,564
RefCOCOg [56]	✓	x	x	Image	MSCOCO	104,560
Visual Genome Corpus [77]	✓	x	x	image	crowd- sourced	108,077
Kunze Corpus [24]	✓	x	x	image	crowd- sourced	200
GuessWhat? [58]	✓	x	x	Image	MSCOCO	155,280
CLEVR-Ref+ [52]	✓	x	x	Image	CLEVR	998,743
EMRE [64]	✓	✓	✓	Video	VoxSim	1,500
Cops-Ref [59]	✓	x	x	Image	COCO/Flickr	148,712
RS-VS Corpus [26]	✓	x	x	indoor im- ages	simulator	2,089
YouRefIt [66]	✓	x	x	Video	crowd- sourced	4,195
CAESAR-XL [65]	✓	✓	✓	Video	CAESAR	124,412
SCMRE	✓	✓	✓	Symbolic and Video	Human- IVA	22,159

Prior work on embodied agents argues for the role of embodiment in representing the salient content of objects in a scene [80], in contributing to mutual understanding [81], and in evaluating the outputs of interactive systems [82]. As such, we seek to design a task-oriented setting to solicit

MREs from humans, build and evaluate models for embodied IVA, with the goal of aiding the development of IVAs that interact with humans with symmetrical and bidirectional use of non-verbal and verbal behavior.

To achieve this goal, we begin by explaining the development stages of the embodied interactive virtual agent (Chapter 2). Next, we detail the establishment of human-IVA interaction and present the collected multimodal referring expressions (Chapter 3). Following this, we demonstrate the proposed intelligent model for generating multimodal referring expressions (Chapter 4). We then evaluate the IVA's generation and comprehension capabilities using qualitative and quantitative metrics applied to both humans and the IVA (Chapter 5). Lastly, we discussed the implications, broader impact, and limitations of the work, and provide future directions for MREG research (Chapter 6).

Chapter 2

Interactive Virtual Agent Development

2.1 Overview

In human communication, verbal referring expressions such as spatial descriptions (e.g., “Please bring the book beside the cup”), references to prior actions (e.g., “give me the cup you just put down”), or attributive references (e.g., “Pick up the red block”) are often supplemented by non-verbal cues to enhance comprehension and convey intentions efficiently. Understanding these expressions involves a complex interplay of cognitive processes, including parsing linguistic inputs to predicate logical forms (PLF), locating objects, resolving spatial relationships, recalling relevant historical actions, and interpreting pointing gestures. Inspired by human cognitive strategies, our research seeks to bridge the gap between embodied intelligent virtual agents (IVA) and cutting-edge natural language processing (NLP) technologies. By integrating VoxWorld [83], spatial relations recognition, speech recognition, we aim to enable IVAs to generate and interpret multimodal referring expressions in real-time interactions with humans. Our motivation for developing a bidirectional interaction scheme is to create methodologies for implementing a successful multimodal generative IVA. These methodologies are sufficient for training AI models to generate human-like multimodal referring expressions, and evaluating combined verbal and non-verbal behavior that apply equally well to human and IVA behaviors.

Prior to delving into generating and interpreting multimodal expressions, ensuring the quality of the Google Automatic Speech Recognizer (ASR) is imperative. Google ASR is a speaker-independent recognizer which means the translation results cannot vary relying on speakers. It decodes audios, recognize patterns, and transcribe speech by applying computations using deep neural networks with significant computing power that are made on Google’s servers. An anecdotal example highlights the importance of this step, where confusion between similar-sounding words led to system errors [84]. Therefore, we conducted an experiment to assess Google speech

recognition performance, focusing on addressing potential challenges in linguistic expression interpretation.

In this chapter, we present the IVA’s capabilities: 1) recognizing diverse verbal collaborative inputs, 2) transforming linguistic speech into PLF, 3) interpreting spatial referring expressions, 4) interpreting descriptions based on prior events, 5) interpreting non-verbal referring expressions, and 6) deterministically generating verbal and non-verbal behaviors. The code for PLF transformation, the relational referring expression interpretation and historical referring expression interpretation will be publicly available on Github¹.

The proposal of this work has been published in Proceedings of the First Workshop on Bridging Human Computer Interaction and Natural Language Processing [81]. Then, the details of this chapter has been published into GENE-2023, the generation and evaluation of non-verbal behavior for embodied agents workshop [85].

2.2 Evaluation of Google ASR Performance

This study evaluates Google ASR’s effectiveness through audio recordings in multimodal collaborative tasks involving humans and machines. High-quality speech input is essential to avoid computational interpretation issues due to recognition errors. The study objectives include examining Google ASR’s word error rate in recognizing vocabulary and phrases used in diverse collaborative tasks, investigating the impact of biasing the recognition algorithm towards domain-specific vocabulary, and identifying areas where recognition errors may cause downstream issues. The source code² and datasets³ used in the evaluations can be found on Github. This section primarily focuses on the evaluation of Google ASR. Initially, the evaluation methods will be outlined, followed by sections dedicated to analyzing the results of both domain-specific and open-domain tests.

¹https://github.com/nadahass/MREG_DIANA.git

²<https://rb.gy/gmxt2>

³https://github.com/nadahass/GSR_Evaluation.git

2.2.1 Data Collection

Participants. To evaluate the performance of Google ASR within the task domain, we collected voice recordings from 22 graduate students at Colorado State University via email. Of these participants, 15 were male and 7 were female, with ages ranging from 18 to 65 (mean = 20, SD = 9). The recruited participants involve diverse vocal profiles, including various accents and timbres, ensuring that speech models suitable for these voices are adequate for a crowd-sourced study. Participants originated from the United States (15), India (4), Mexico (1), China (1), and Iran (1). As compensation, each participant received a \$10 gift card from Amazon, Chipotle, or iTunes.

Test Data. participants were asked to record scripts containing domain-specific vocabulary sourced from the EGGNOG dataset [86]. This dataset comprises multimodal commands generated through human-human interaction in the blocks-World domain. We selected a total of 5 scripts, each containing 5-8 commands, resulting in approximately 30 commands of varying lengths (1-15 words). Our selection process revealed three primary referring strategies used by humans in the EGGNOG data: *Attributive REs* (describing object properties), *Relational REs* (describing objects in relation to each other), and *Historical REs* (describing previously mentioned or interacted-with objects). These selected scripts serve as the basis for our evaluation.

Procedure. Participants we recruited were requested to record their voice while reading all five scripts at their convenience, from their preferred location. We received five recordings from each participant, totaling 110 recordings. These audio files were then submitted to the Google Cloud Speech API for transcription, once using Google’s open vocabulary and once with our restricted vocabulary. This was done to evaluate the performance of Google Speech Recognition (GSR) under these two conditions. Upon receiving the transcripts, we calculated the Word Error Rate (WER) and accuracy detailed in Section 2.1, and also considered the domain tokens necessary to prevent any potential corruption of our experimental software.

Scripts
Put this blue block on the table (Attributive RE). Grab the block you just put on the table. Place it there. Take the yellow block and place it next to it. Take one more block, put it on the top of those two blocks. Stop, you are done.
Go there. Grab this purple block. Now put it in front of the cup. Pick up the block that you just put down to the left (Historical RE). Now, move that block and put it on top of these boxes. Wait! Move it more this way. Stop!
Go to that red block. Grab it and place it above these brown blocks. Nevermind! Turn right, Rotate the block that is beside the white block (Relational RE). Take that block and release it to the back of the blue block. That is Enough.
OK turn left, move the orange block to the right of that gray block. Wait! Grab the black block. Go forward put it next to that block. Go here, put this pink block between these ones. That is pretty good.
Go there! take that green plate and put it to the left of that mug. Go back! Push that glass now to the back of that bottles. Now you're going to bring this knife and drop it in the glass.

Figure 2.1: Selected scripts from EGGNOG dataset including highlighted examples of the most common referring expressions

2.2.2 Operationalizations and Calculations

We developed a Java program utilizing Google Cloud client libraries to conduct measurements of Word Error Rate (WER) and accuracy. Specifically, we implemented a function that iterates over audio files, sends each .wav file to the Google Cloud Speech API for transcription, and generates hypotheses. These hypotheses are then compared with the ground truth (references) to calculate statistics such as the number of deletions, insertions, substitutions, and correct words, enabling us to determine accuracy and WER. The source code for calculating WER is based on the ErrorRate class provided by the Google Cloud Platform ⁴. To evaluate the restricted recognizer, we sent specific domain phrases to the Google Speech API using the speechContexts parameter in each API call. In contrast, the open recognizer utilized Google’s open vocabulary. The reference, hypothesis, number of tokens, correct tokens, substitutions, insertions, deletions, reference length, differences between reference and hypothesis, calculated Word Error Rate (WER), calculated accuracy, and participant demographics (gender, accent, native language, and original country) were written to

⁴<https://github.com/GoogleCloudPlatform/java-docs-samples>

.csv files for each evaluation. Subsequently, these files were processed to determine the final accuracy, WER, and unrecognized words for each recognizer.

Word Error Rate (WER). This metric is a popular way of assessing the quality of a speech recognizer. WER [87] is calculated by finding the number of errors that resulted after the process of converting speech to text. To calculate the WER, the number of errors in the hypothesis is divided by the total number of words in the reference sentence (correct sentence). Equation 2.1 shows how the word error rate is calculated.

$$\text{WER} = \frac{\text{substitutions} + \text{deletions} + \text{insertions}}{N_{\text{Total}}} \quad (2.1)$$

Where:

- N_{total} be the total number of words.

As shown in Table 2.1, The errors that summarized by WER are incorrect substitutions, deletion or incorrect insertion of tokens that causes the clear difference between the hypothesis and the reference.

Table 2.1: The Concept of Word Error Rate: “S” substitution, “I” insertion, “D” deletion, “=” match

Reference	now	pick	up	this	red	block	
	=	D	S	=	=	=	I
Hypothesis	now		of	this	red	block	log

2.2.3 Results and Analysis

This section presents the evaluation results of both the open recognizer (OR) and restricted recognizer (RR) in terms of their Word Error Rate (WER) and accuracy across multiple accents. The objectives are to: 1) assess the ability of both open and restricted Google speech recognizers to transcribe domain-specific speech spoken in various accents, 2) investigate the impact of biasing the speech recognition model to domain tokens on WER and accuracy, and 3) explore common unrecognized tokens.

Table 2.2: Statistical testing of the impact of accents' variety and context restriction on the performance of Google Speech Recognizer using Welch Two Sample t -test.

	Accuracy Groups	P-value
Accent Impact	Non-Native RR vs Native RR	$8.39e - 05$
	Non-Native OR vs Native OR	$2.58e - 05$
Restriction Impact (RI)	OR vs RR	0.3433

Table 2.3: Statistics (mean and SD) for open recognizer WER and accuracy and restricted recognizer WER and accuracy when recognizing speech from native and foreign speakers

	Non-Native scripts N=35		Native scripts N=75		Mean
	Mean	SD	Mean	SD	AVG
OR_WER	0.23	0.272	0.11	0.107	0.17
OR_ACC	0.77	0.274	0.89	0.097	0.83
RR_WER	0.20	0.268	0.09	0.07	0.15
RR_ACC	0.80	0.270	0.91	0.07	0.85

As shown in Table 2.3, we received 110 recordings, including 35 non-native speech and 35 non-native speech, and sent them to Google ASR to be transcribed. As a result, 85% and 83% of the recorded speech were correctly transcribed by the restricted recognizer and open recognizer, respectively. Nonetheless, WER showed that 17% and 15% of the recordings were incorrectly transcribed by the restricted recognizer and open recognizer, respectively. Being closer to our data points, Figure 2.3 (a) and Figure 2.3 (b) illustrates how OR and RR worked with native and non-native speech. The WERs (OR: 11% and RR: 9%) are lower when transcribing native speech than those of non-native speech (OR: 23% and RR: 20%).

Figure 2.2 shows the accuracies and WERs for OR and RR over different nationalities including, Indian, Mexican, American, Chinese, and Persian. Total of recognition errors occurred on non-native speech, Indian (49.5% and 44.3%), Mexican (8% and 9%), Chinese (10% and 8%), and Persian (20% and 15%) were larger than those appeared on recognition of native speech (11% and 8%). There are differences between native speech and non-native speech recognition using the open recognizer and restricted recognizer (p -value: $8.39e-05 < 0.001$, p -value: $2.58e-05 < 0.001$, respectively, using Welch Two Sample t -test, as presented in Table 2.2). Variety of speech accents causes the low performance of OR and RR.

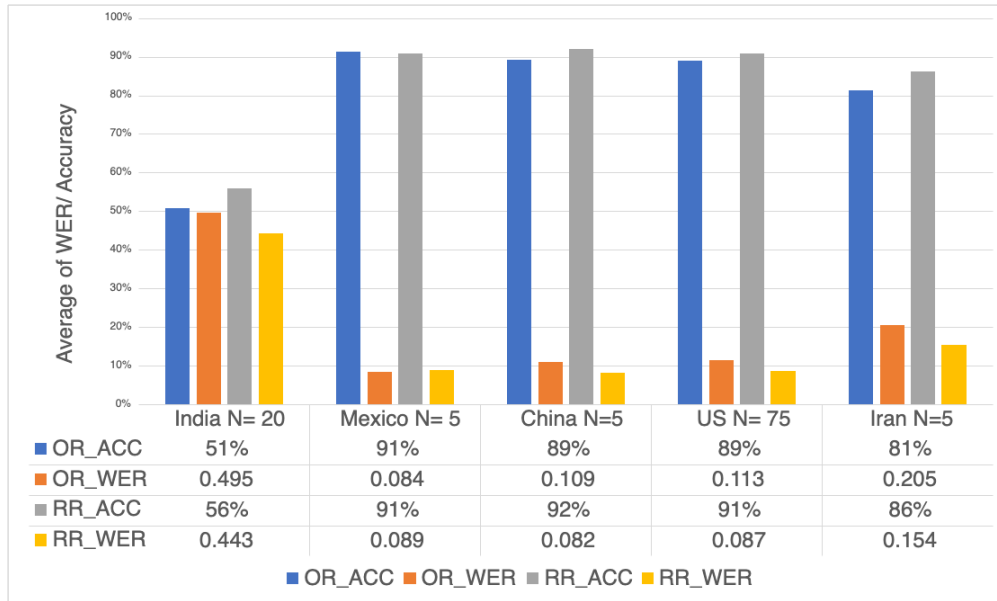
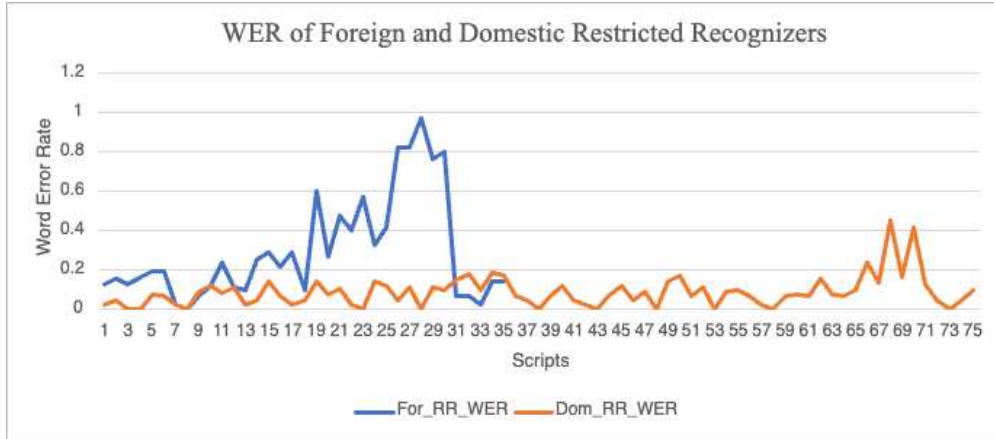


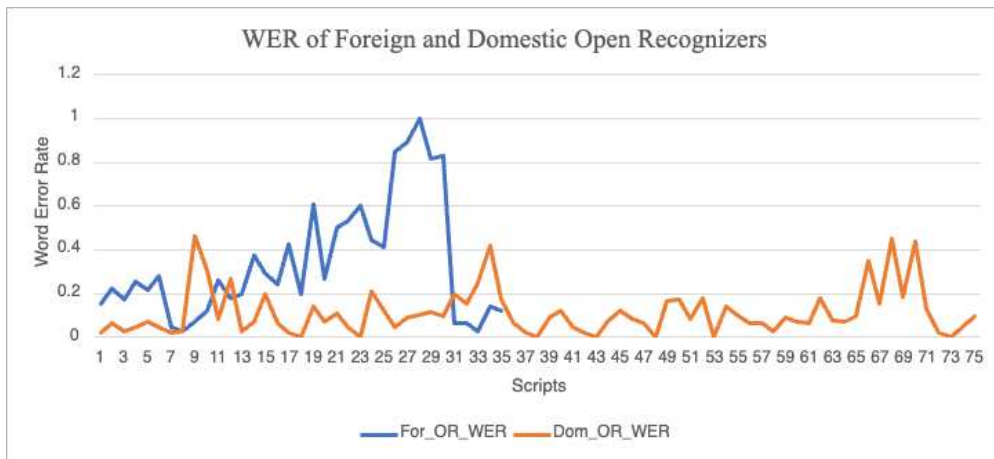
Figure 2.2: WER and accuracy of OR and RR transcribing the voices recorded from different nationalities, N is the number of transcribed speech by each nationality

Recognition errors, insertions (a), deletions (b), and substitutions(c) shown in Figure 2.4, varies among these nationalities. Indian accents affect the GSR performance by the highest rate of deletions (OR: 33% and RR: 32%) and substitutions (OR: 15% and RR: 12%). Similarly, considerable substitutions occurred when transcribing speech recorded by other nationalities; Mexican (OR: 7% and RR: 7%), American (OR: 7% and RR: 6%), Chinese (OR: 9% and RR: 7%), and Persian (OR: 14% and RR: 8%). Insertion error, however, rarely occurred (0% - 2%) by both recognizers for all solicited speech. Based on Spearman’s rank Correlation [88], substitutions (p -value: $2.20e-16 < 0.001$, $r = RR: 0.7$, OR: 0.9) and deletions (p -value: $2.20e-16 < 0.001$, $r = RR: 0.7$, OR: 0.8) raise the WER for OR and RR more than insertion (p -value: $0.00215 < 0.001$, $r = RR: 0.3$, OR: 0.3).

After adding domain words to Google ASR’s speech context, the difference is occurred in the performance of speech recognition. As presented in Figure 2.5, RR transcribed most of the audio files with high accuracy and low errors in compared to OR transcription for native and non-native speech. There is an impact of biasing the speech recognition model to domain words such that the accuracy became slightly higher, however, the difference (p -value: $0.3433 > 0.001$, using Welch Two Sample t -test, as presented in Table 2.2) is not statistically significant. Nonetheless, this slight



(a)



(b)

Figure 2.3: The influence of accents on Google ASR (a) Foreign Restricted Speech Recognizer WER vs Domestic Restricted Speech Recognizer WER (b) Foreign Open Speech Recognizer WER vs Domestic Open Speech Recognizer WER.

improvement in recognizing domain specific words is still useful to avoid any computational errors on those correctly transcribed tokens. Therefore, because of this improvements, we added domain specific phrases to model’s speech context.

Speech context restriction increased the GSR performance as shown in table 2.3, however, it still encounters errors (OR: 17%, RR: 15%). For native speech recognition, errors occur in homonyms (e.g., “there”, “their”). Nonetheless, non-native speech errors are caused by the nature of participants’ native languages, (e.g., “it”: “eat”, “there”: “deer”). Based on these results, two pre-processing steps were implemented before passing human’s speech to the agent’s internal pro-

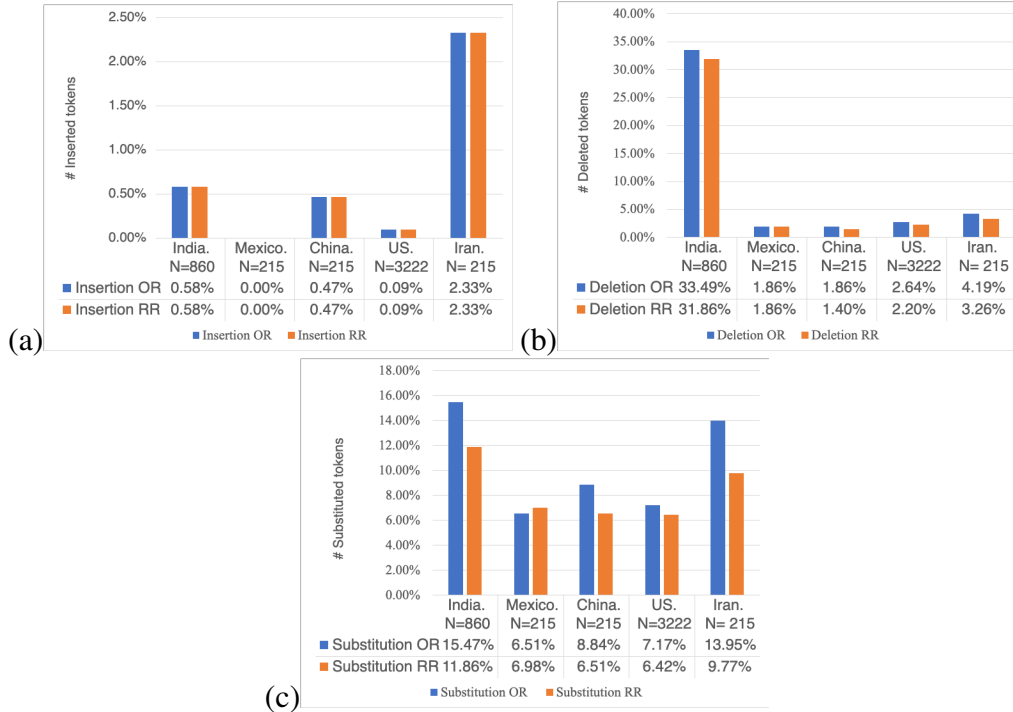
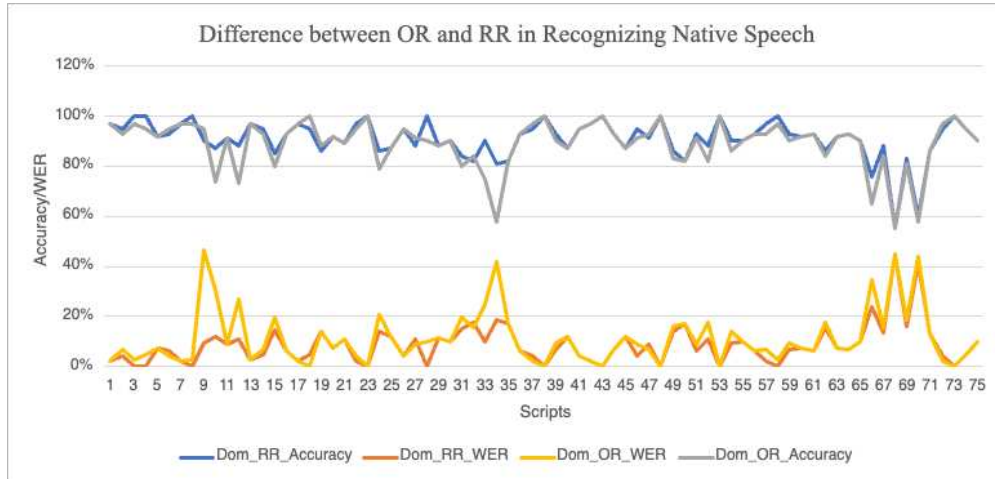


Figure 2.4: (a) Insertions, (b) Deletions, and (c) Substitutions occurred during GSR transcription over distinct nationalities. "N" is the total number of tokens in ground scripts of all transcribed audios

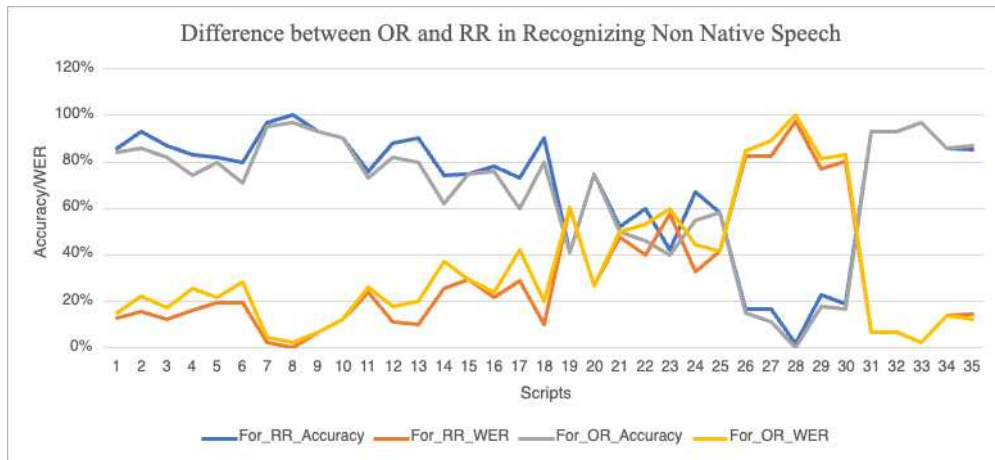
cessing. First, domain syntactic categories were incorporated into GSR’s speech context. Second, tokens listed in Table 2.4 were replaced with their correct counterparts. These pre-processing steps ensure linguistic inputs are refined for subsequent computations, such as language transformation to predicate logic format (Sec 2.7), VoxML modeling [89], and VoxSim simulation [83].

2.3 Setting of the Experimental Platform

We developed a standalone version of the existing *Diana* system [18, 84], developed as a collaborative virtual agent who responds to instructions given via both live gesture and speech and collaborates with humans in situated task-based interactions. We adapted the system where human participants are presented with a sequence of 10 scenes, each involving (1) ten equally sized target blocks randomly placed on a table that (in simulated units in the Unity-based environment) is approximately 1.6m wide. There are two of each color of block: red, green, blue, pink, and yellow; and (2) two landmark objects (*plate* and *cup*) available for use when describing the target



(a)



(b)

Figure 2.5: The influence of context restriction on Google ASR (a) Foreign Restricted Speech Recognizer WER vs Domestic Restricted Speech Recognizer WER (b) Foreign Open Speech Recognizer WER vs Domestic Open Speech Recognizer WER.

blocks. This setting requires the IVA to ask for disambiguation based on factors like color and location if needed, and the human to provide complex descriptions including verbal (e.g., relational, historical) references, non-verbal (e.g., deictic pointing) references, or ensemble.

Diana initially asks a question, e.g., “Which object should we focus on?”, as shown in Figure 2.6, without providing any prior knowledge of what she understands, e.g., specific domain words or actions. Participants are informed that they are able to use multiple input channels, e.g., automatically recognized speech and mouse-based deixis, to clearly express their intent. To replicate the variability in pointing displayed in the Diana system with live gesture recognition, and

Table 2.4: The popular domain specific words were incorrectly recognized by Google speech recognizer

Correct form	Incorrect form	Correct form	Incorrect form
There	Their, they're, deer	Gray	Grave
This	Is, does, deep	Bottle	Water
Wait	Weight, rate	Mug	Month
Block	Rock, lock, blog	Pink	Paint, spring
	Blotch, coke	It	eat
	Black, blue, dog, clock	Knife	Nice
Scoot it	Scooter, screwed	Place it	Play sit, blessed
	Student	Push	Put
Now	Not	Grasp	Rasp, grass, graph, craft
Above	About	Glass	Class
Right	Light, site	Green	Cream
That	Dead, the, a, back	Put	What, but
Of	Out	Below	Pillow
Move it	Limit, modded	Brown	Problem
Your	You're	White	Bride, Y

the gesture-semantic notion of a *pointing cone* [90–92], the center of deixis fluctuates within a circle of radius $\pm 0.3\text{m}$ around the mouse location and the size of the deictic reticle (see Figure 2.6) randomly fluctuates in size within a range of 14–186% of the default radius (17.32cm). This variability prevents users from relying on fully accurate pointing with the mouse as a method of unambiguously indicating objects, and encourages the use of speech input for object specification.

2.4 Interpretation of Multimodal REs

Multimodal referring expressions can be considered special cases of *gesture utterances* as specified in [93], in that they contain a gestural component and a verbal component that must be unified for a complete interpretation by either human or machine. In addition, MREs may be mixed with unimodal REs in a discourse, but even unimodal REs may rely on meaning that was previously established in the discourse using multimodal communication.



Figure 2.6: Experimental Diana System: the purple circle indicates where the user is pointing. Without disambiguation, any object within the pointing circle is a potential candidate for a deixis-only RE. Diana’s utterances appear on screen and are spoken aloud via TTS.

2.4.1 Language Preprocessing

Given VoxWorld’s specific formatting needs for event visualization and simulation, it’s crucial to parse natural language input efficiently. This involves mapping tokens within entries to their synonyms, facilitating real-time visualization and simulation by VoxSim. Additionally, transforming the processed speech into a predicate logic format, as detailed in Section 2.7, enhances compatibility with VoxWorld’s requirements. In human-to-human interactions, participants are not always required to provide complete sentences with all components, including verbs and nouns; some elements may be omitted. A preprocessing step is employed to manage the various alterations that may arise in human prompts, ensuring effective communication. Due to the limited interpretation capabilities of VoxSim, actions, nouns, and relations are constrained. Therefore, synonyms of these entities, such as directive actions (e.g., “take”), prior actions (e.g., “took”), nouns (e.g., “box”), and relations (e.g., “above”), are gathered from interlocutors and mapped to corresponding VoxSim entities. For instance, “take” is mapped to “grasp,” “took” to “grasped,” “box” to “block,” and “above” to “on.” Refer to Table 2.5 for additional VoxWorld nouns, relations, and actions along with their synonyms.

Table 2.5: Mapping linguistic nouns, actions and relations to VoxSim entries

<i>VoxNouns</i>	Cup		Kinfe	Blocks	Block
Synonyms	Cylindrical block, cylindrical object, Rounded object, rounded block, Cylinder, mug, Glass, class		Nice	Boxes	Box
<i>VoxActions</i>	Ugrasp	Grasp	Grasped	Ungrasped	
Synonyms	Let go of, let go, drop, release, leave	Pick up, pick, grab, grasp, rasp, grass, hold, take	Picked up, grabed, took	Went of, dropped, unreleased	
<i>VoxActions</i>	Slid	Slide	Place	-	
Synonyms	Pushed, pulled	Move, push, pull	put	-	
<i>VoxRelations</i>	On	Behind	Under	Beside	
Synonyms	On top of, on the top of, Above	In back of, back of, In the back of, to back of, To the back of	Below	Next to	

Incomplete prompts often lead to significant corruption in subsequent semantic extraction, particularly when one or two speech parts are missing. Hence, it’s crucial to map incomplete speech to an acceptable form that the Stanford CoreNLP dependency parser can accurately extract semantics from. Two common scenarios are addressed: Firstly, when the action is missing and the prompt begins with a demonstrative (e.g., “the/this red block” instead of “pick up the/this red block”), the action “Grasp” is inserted at the start of the speech input. Secondly, if the prompt starts with a spatial relation entity (e.g., “behind the plate” rather than “take the red block behind the plate”) and clarification is asked by the agent (e.g., “Which red block?”), the human intent is extracted and placed between the action “Grasp” and the relational clause, resulting in a coherent prompt like “Grasp the red block behind the plate.” This properly mapped input is then processed by the parsers to generate the predicate form.

To rectify misunderstandings by the agent, humans utilize negative words like “Nevermind,” “No,” and “Stop.” These terms are not mapped or parsed to Predicate Logic Form, detailed in Section 2.7. Instead, undoing functionalities are triggered to revoke the last prompt and remove the last intended action object. In instances where these negative entities are coupled with another

event, such as “Take the green block - Nevermind take the red block,” the last event, “Take the green block” is omitted, and the new prompt is executed.

2.4.2 Transformation to Predicate Logic Format

To interpret the received speech, it recursively follows a set of rules relying on semantic relations (e.g., “Subj”, “Obj”, “Parataxis”, “Advmod”, “Amod”, etc.) between dependent entities, e.g., small, which modify governors, e.g., block, that are extracted by StanfordCoreNLP dependency parser [94] to compose linguistic constituents into a predicate logic form (PLF). As illustrated in Figure 2.7, the natural language input interpreted and simulated by the VoxWorld reasoning platform after being transformed into predicate logical representation. *Grasp(the(red(block)))* is an exemplar, encompassing an objects, action, and demonstrative based on discriminative syntactic and semantic features for each verbal expression, be it attributive, relational, or historical.

To explore the common features for each RE category, deterministic rules were implemented to connect to StanfordCoreNLP server and send linguistic prompts, 30 scripts that were selected from EGGNOG dataset as we mentioned in Sec 2.2.1 and shown in Figure 2.2, to the StanfordCoreNLP dependency parser in order to generate their dependency parsing in two forms; JSON format and graphical dependency tree to analyze the patterns of the resulted semantic relations and Part of Speech (POS) tags. We established common syntactic and semantic relations for each of the following verbal sets: 1) transitional attributive REs, describing properties for focus and destination objects (e.g., “Put the *red block*”, the focus object, “beside the *green block*”, the destination object), 2) non-transitive attributive REs, describing properties for focus object (e.g., “Grasp the *red block*”), 3) relational-to-objects REs, describing objects in relation to other objects, (e.g., “Move this block to *the left of the red block*”), 4) relational-to-space REs, describing objects in relation to the space (e.g., “Move the red block *to the left*”) and 5) historical REs, (e.g., “Grasp the block *you just put down*”).

We operationalized the transformation of natural language into related PLF by implementing a set of C# rules that recursively iterate over semantic relations between entities in the accessed

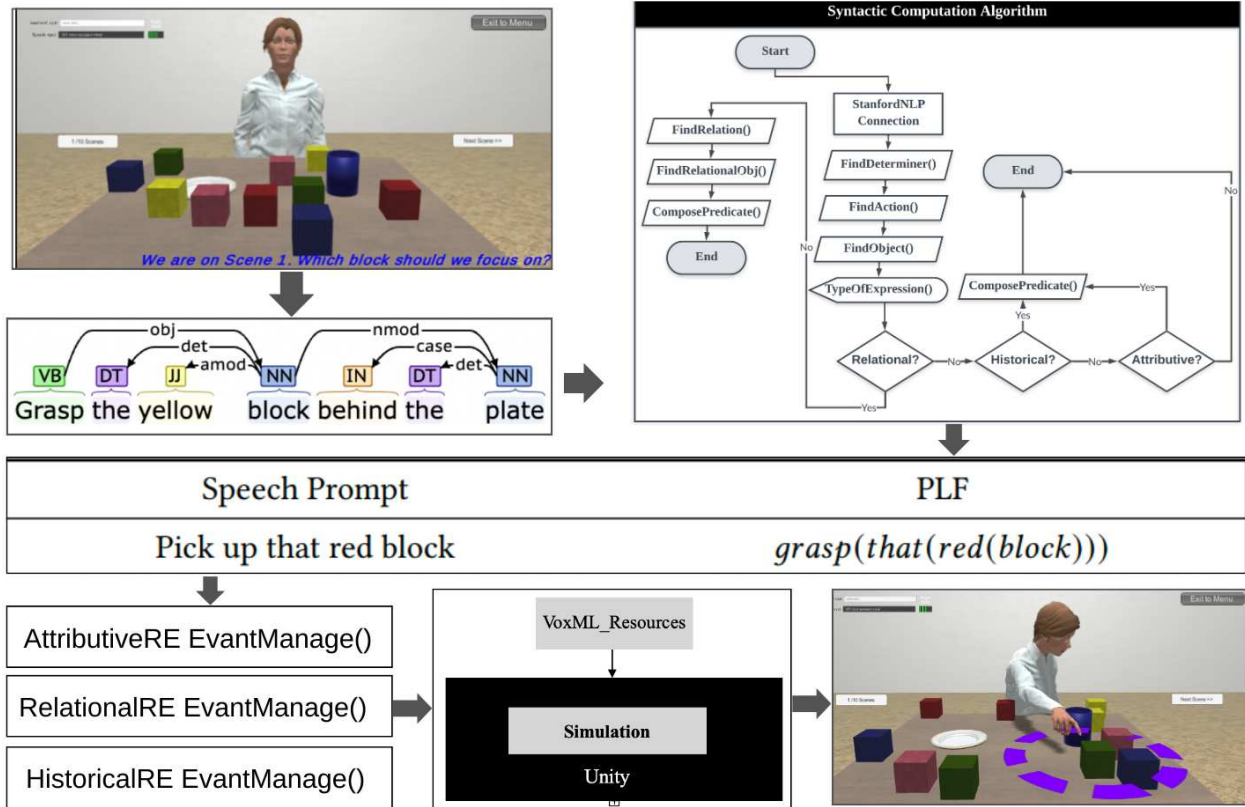


Figure 2.7: The process of speech interpretation (1) Speech is recognized through GSR (2) Tagging and dependency parsing are determined using StanfordCoreNLP (3) Computation of the extracted parse tree to transform the speech to its PLF (4) PLF interpreted and simulated using VoxML and VoxSim.

JSON output, as depicted in Figure 2.8. Initially, a webpage request is generated to the StanfordCoreNLP server to input the speech and retrieve the response as a JSON file containing the dependency tree and POS tags. Given that this algorithm is part of an IVA designed to collect data from students distributed throughout Colorado State University, remote access to the server must be accounted for. Consequently, a portion of a machine in the Computer Science building at Colorado State University was dedicated to hosting the StanfordCoreNLP server, along with its required libraries and operations for syntactically and semantically processing natural language inputs.

When providing a linguistic input along with its JSON dependency parsing, as illustrated in Figure 2.8(2), if the part-of-speech (POS) tag of the dependent entity of the “ROOT” dependency relation is a verb (VBZ), it is extracted as the action of the referring expression. This action can

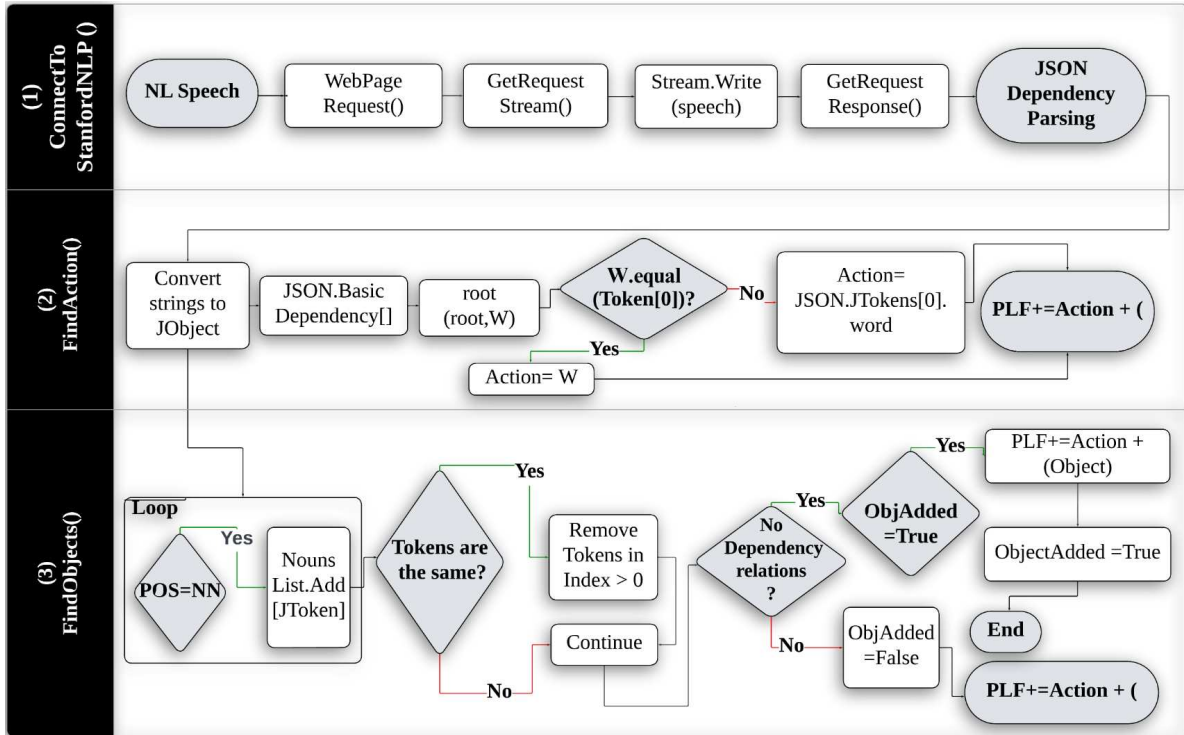


Figure 2.8: Transformation to predicates logic format requires 1) Connecting to StanforCoreNLP, 2) Extracting the action, and 3) Finding the focus and destination objects from the provided linguistic prompts

either be a transitional action (e.g., “push”, “move”, “put”) or a non-transitional action (e.g., “lift”, “take”, “grasp”). The extracted action is then placed at the beginning of the predicate form as *Action()*.

Based on our analysis of EGGNOG referring expressions, we propose that human prompts commonly entail two “Nominal Nouns (NNs)”: focus objects (the focus of discourse) and destination objects (objects to which other objects are moved). Given the JSON objects, as depicted in Figure 2.8(3), all tokens tagged with “NN” are collected. If the entire constituents, including adjectives, relations, and the destination object, are composed into the predicate form, the focus object is positioned at the end of the predicate form as *Action(FocusObject)*.

The algorithm iterates over the JSON entities to examine the dependency relations, which signify the semantics between dependent and governor constituents, including determiners, attributes, relations, etc. These are then utilized to compose the predicate form, as depicted in Figure 2.9 and Table 2.7. In the context of referring expressions, several semantic relations play important roles.

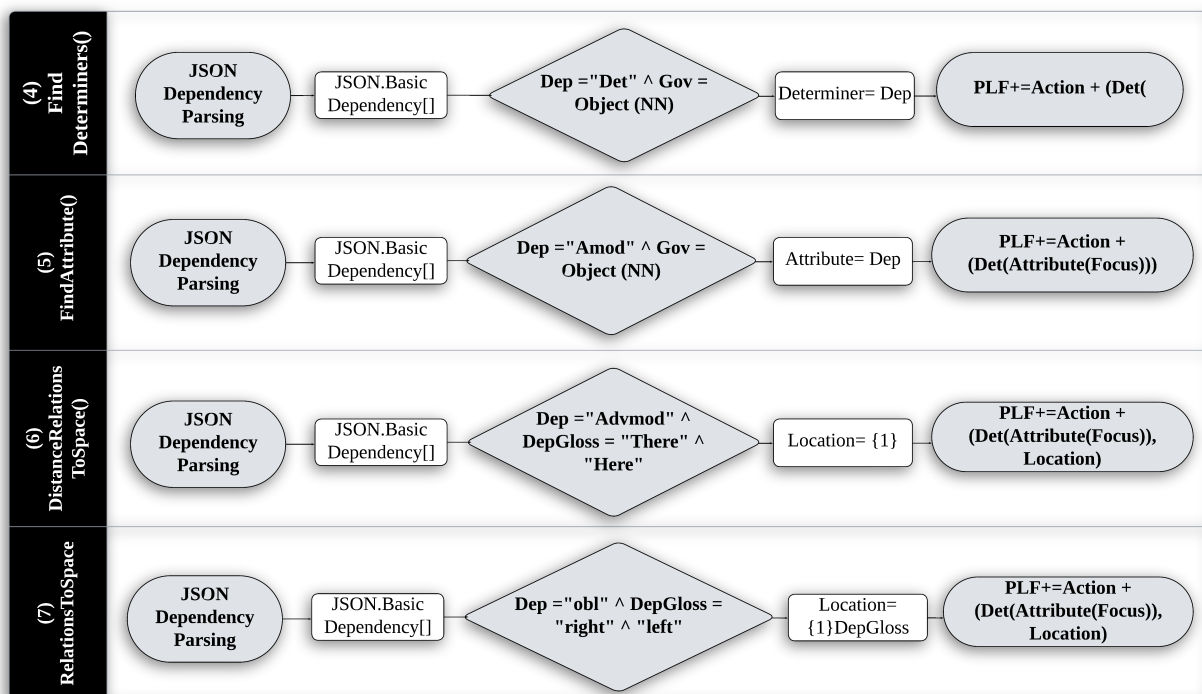


Figure 2.9: Additional mandatory steps to compose the PLF 1) Finding determiners, 2) Finding Attributes, 3) Extracting distance relations to space, e.g., “there”, 4) Extracting relations to space, e.g., “to the right”

“DT” denotes the determiner, specifying entities like “the,” “this,” or “that.” “Amod” signals attributes, properties of objects. Meanwhile, “Advmod” pertains to distance-related tokens such as “there,” crucial for spatial references. Additionally, “Obl” signifies prepositions like “right,” crucial for describing an object’s target location in space. The placeholder “{ 1 }” is integrated into the PLF to be replaced with the precise position in subsequent computations. Notably, while the first two relations modify objects, the latter two influence actions.

Table 2.6: Transitive and non transitive verbs are used to distinguished between relational REs and transitive attributive REs

Transitive verbs	Non-transitive verbs
Move, Put	Pick up, Lift, Go to
Push, Pull, Slide	Grab, Grasp, Take
Place, Shift	Let go of, Ungrasp
Scout, Servo, Bring	Drop, Release, Find

The extraction of semantics for complex referring expressions, including relational REs and historical REs, entails multiple iterations over JSON objects, as illustrated in Figure 2.10. Relational REs differentiate an object based on its relation to others, while historical REs utilize prior actions to identify an object. These prompts typically consist of two clauses, with the second clause elaborating on the previous one either with relative pronouns (e.g., “which” and “that”), identified by the “SBAR” entity, or without relative pronouns, identified by the “Parataxis” entity. Historical REs are distinguished by the presence of a pronoun constituent, signaled by the “PRP: you” part of speech in the second clause, and the “nsubj(you, put)” relation, where the governor is the prior action. This prior action is integrated before the focus object as an adjective in the PLF, either with “SBAR” (e.g., Table 2.7 #4) or with “Parataxis” (e.g., Table 2.7 #9). To identify relational REs, a

Table 2.7: Predicate logic format (PLF) transformation for co-gestural verbal REs (Att_RE: Attributive RE, Trans_RE: Transitive RE, Rel_RE: Relational RE, Hist_RE: Historical RE, and Comp_RE: Compound RE). *Numerals in brackets denote variables that must be assigned from prior conversational or non-verbal context (e.g., “it,” “there,” etc.).

No	Speech Prompt	PLF
1	Pick up that red block	<i>grasp(that(red(block)))</i>
2	Put this block to the right of the blue block	<i>put(this(block), right(the(blue(block))))</i>
3	Grasp the green block beside the plate	<i>grasp(the(green(beside_adj(plate(block))))))</i>
4	Lift the block that you just put down	<i>lift(the(put_adj(block)))</i>
5	Take this block and put it there	<i>take(this(block)) + Put({0}, {1})*</i>
6	Move the red block to the right	<i>move(the(red(block)), {1}right)</i>
7	Take the blue block to the right of the cup	<i>take(the(blue(right_adj(cup(block))))))</i>
8	Take the grey block in front of the plate	<i>take(the(grey(in_front_adj(plate(block))))))</i>
9	Lift the block you just put down	<i>lift(the(put_adj(block)))</i>
10	Take the blue block that is to the right of the cup	<i>take(the(blue(right_adj(cup(block))))))</i>

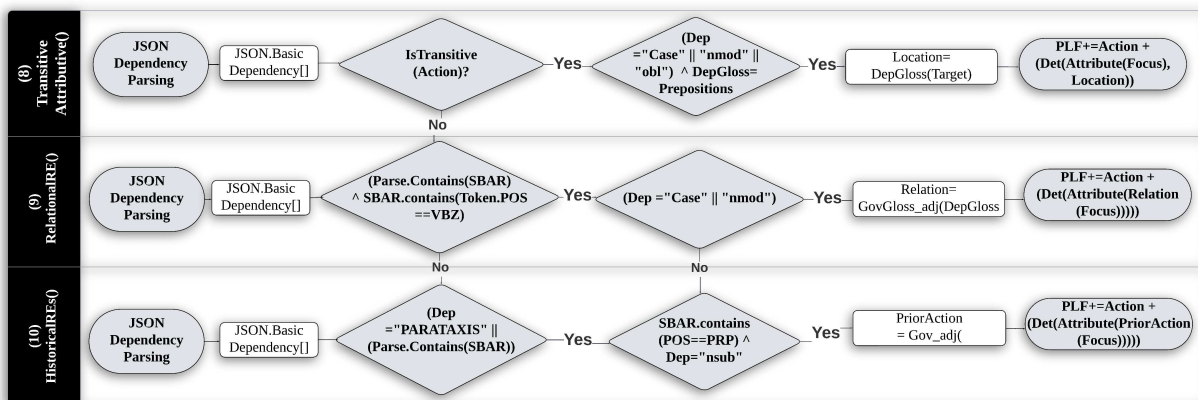


Figure 2.10: Transformation of 1) transitive attributive descriptions, 2) relational (to objects) descriptions, 3) historical descriptions to predicates form

token tagged with “VBZ” must exist with “SBAR” in the second clause (e.g., “that is” in Table 2.7 #10), or without verbs (e.g., Table 2.7 #3). The semantic relations between entities in relational prompts vary based on the number of prepositional tokens: 1) one token (e.g., “behind” in Table 2.7 #3), where the dependent entity of “Case(behind,plate)” is the preposition and its governor is the destination object, 2) two tokens (e.g., “right of” in Table 2.7 #7), or 3) three tokens (e.g., “in front of” in Table 2.7 #8). In the latter cases, the governor entity of “nmod(plate,right)” or “nmod(plate,front)” is the preposition, and its dependent is the destination object. This is distinct from transitive attributive REs (e.g., Table 2.7 #2), identified by the action category (transitive or non-transitive) as shown in Table 2.8. Additionally, compound conjunctions (CC) in two attributive and transitive prompts are recognized by inspecting a token tagged with CC, which is then replaced with “+”, as illustrated in Table 2.7 #5. These are executed by VoxWorld as two separate and consecutive prompts.

2.4.3 Attributive Referring Expressions Interpretation

The attributive RE algorithm interprets attributes of objects, such as color or demonstratives (e.g., “the,” “this,” etc.). There are two subcategories of attributive REs: Transitive REs, which involve an attributive RE embedded within a transitive clause (see Table 2.7 #6), and Compound REs, which entail multiple events, as shown in Table 2.7 #5. A simple, non-transitive, non-compound

```

Data: List of local variables
Result: event ← Action(Focus)
RelList ← SpatialTokensList;
relation ← Prompt.Intersect(RelList);
multiAppendDelimiter ← " + "
if "adj" ∈ Prompt then
  | if relation ≠ null then
  | | Call RelationalREsInterpretation Algorithm 2
  | else
  | | Call HistoricalREsInterpretation Algorithm 3
  | end
end
else
  | if multiAppendDelimiter ∈ Prompt then
  | | EventsList ← Prompt.Split(" + ")
  | | foreach event ∈ EventsList do
  | | | EventManager.InsertEvent(event)
  | | end
  | else
  | | EventManager.InsertEvent(event)
  | end
end
if matches > 1 then
  | AskForDisambiguation()
end

```

Algorithm 1: Attributive RE Interpretation (*Prompt*)

attributive RE is exemplified in Table 2.7 #1. This utterance also illustrates how demonstratives may serve to deictically disambiguate the target.

Algorithm 1 outlines the process of discerning the type of a referring entry and subsequently interpreting and acting upon it. Relational REs and historical REs are differentiated from attributive ones by the presence of the “adj” token. If a spatial relation exists in the expression, it indicates a relational RE; otherwise, it signifies a historical RE. Attributive REs, whether transitive or non-transitive, are added to the EventManager, the gateway of the VoxWorld platform, if the prompt comprises only one event. For prompts with multiple events, they are split by the delimiter “+”, and then inserted into the EventManager over multiple iterations, determined by the event count. If the generated description is unambiguous and the focus object can be discriminated successfully, the agent follows the EventManager rules to simulate the requested event. Nonetheless, if the focus

object is preceded by demonstratives and the shared space encompasses more than one match to the uttered intended object, the agent cannot discriminate the focus object solely through speech. In such cases, humans will be prompted for additional information to clarify their intent.

2.4.4 Spatial Referring Expressions Interpretation

Algorithm 2 outlines the computations of relational REs interpretation. It utilizes the relational PLF to resolve spatial relations between objects and identify the focus object. VoxWorld employs calculi such as RCC-3D [95] to track spatial relations, aiding in the extraction of the focus object, human intent, and destination object relative to the focus object, along with their relation. This facilitates the localization of the focus object and subsequent action upon it.

The algorithm begins by extracting the spatial relation from the uttered speech and then proceeds to extract attributes and objects based on predefined patterns. The spatial relation is extracted by intersecting the prompt’s tokens with known spatial relations that are enumerated in Table 2.8. Then, the attributes and objects, including the focus and destination objects, are extracted by string indexing of this pattern as “Take (*action*) + red (*attribute*)/ none + block (*object*) + beside (*relation*) + plate (*object*)”. The prompt’s focus object, destination object and their spatial relation are recursively compared to all tracked two relative objects by VoxWorld’s relations tracker. If objects with their relation are tracked by VoxWorld, the agent efficiently act on the prompt, otherwise, if the objects are shown on the scene, but the relation is not tracked, the agent suggests to act on an object that is similar to the uttered object. For example, if a human asks to act on object with an untracked relation “Grasp the red block to the right of the plate”, the agent suggests to grasp one of the red blocks in the shared space. Moreover, in case the human asks to act on objects are not placed in the common space, the agent responds negatively but offers helpful suggestions.

2.4.5 Historical Referring Expressions Interpretation

In addition to language and deictic gesture, prior actions contribute to building speakers’ knowledge of descriptions of objects as defined by Grice’s maxim of quantity [96]. Therefore, we integrated a dialogue history to the IVA. Two stacks stores all requested actions along with

```

Data: List of local variables
Result:  $event \leftarrow Action(Focus)$ 
 $RelList \leftarrow SpatialTokensList;$ 
 $relation \leftarrow Prompt.Intersect(RelList);$ 
 $Action \leftarrow Prompt[0];$ 
 $FocusType \leftarrow Prompt[Prompt.Length - 1];$ 
if  $relationIndex$  is 2 then
  |  $targetObj = Prompt[3];$ 
  |  $FocusColor \leftarrow Prompt[1];$ 
  |  $Focus \leftarrow FocusColor.Concat(FocusType);$ 
else
  | if  $relationIndex$  is 1 then
  | |  $relation \leftarrow Prompt[1];$ 
  | |  $targetObj = Prompt[2];$ 
  | |  $Focus \leftarrow FocusType;$ 
  | end
end
if  $relation == left$  then
  |  $relation \leftarrow right;$ 
else
  | if  $relation == right$  then
  | |  $relation \leftarrow left;$ 
  | end
end
if  $(targetObj \parallel Focus) \notin SceneObjects$  then
  |  $Agent.Speech : "There is no object here!"$ 
end
foreach  $DictEntry \in relationTracker.relations$  do
  | if  $DictEntry.Keys[0] = targetObj \wedge DictEntry.Keys[1] = Focus$  then
  | |  $RelValue \leftarrow DictEntry.Values[0];$ 
  | | if  $RelValue == relation$  then
  | | |  $event \leftarrow Action + (+Focus+);$ 
  | | |  $EventManager.InsertEvent(event)$  VoxWorld Executes the event
  | | |  $Found \leftarrow True;$ 
  | | | break
  | | end
  | end
end
if  $Found \neq True$  then
  |  $Agent.Speech: "Do you mean this?"$ 
  |  $Agent.Point: "Object"$ 
end

```

Algorithm 2: Relational RE Interpretation ($List < String > Prompt$)

Table 2.8: Transitive and non transitive verbs are used to distinguished between relational REs and transitive attributive REs

Known Spatial Relations			
In	On	Atop	Port
Starboard	Afore	Astern	At
Behind	In front of	Beside	Near
Left of	Right of	Center of	Edge of
Under	Underneath	Below	Against
Here	There	Right	left

destination objects, and accommodates interpretations of verbal, gestural, and multimodal inputs. These stored actions are available for use by both humans and the IVA to refer to objects that may have previously been interacted with.

Table 2.9: Sample of dialogue history, including previously mentioned actions and related objects after executing multimodal (co-gesture speech) or unimodal (speech only or pointing only) prompts.

No.	Modality	PLF	Actions Stack	Referents Stack
4	Speech Only	<i>grasp(the(adj_slid((block)))</i>	grasp slide put put	Green2 Green2 Red1 Green1
3	Multimodal	<i>slide(Green2; left(the(plate)))</i>	slide put put	Green2 Red1 Green1
2	Pointing Only	<i>put(Red1; left(the(plate)))</i>	put put	Red1 Green1
1	Pointing Only	<i>put(Green1; < X', Y', Z' >)</i>	put	Green1

As shown in Algorithm 3, after recognizing the human’s intent and executing the parsed-out prompt, the IVA pushes the action and referent (extracted from the PLF of the prompt) to two separate stacks (an actions stack and a referents stack) as shown by Table 2.9, #1–3. If the human uses a mention of a previously executed action to indicate an object as in Table 2.9, #4, the IVA visits the dialogue history to 1) retrieve the most recently referenced object that is relevant to the provided action (in this case, *Green2* block, as it satisfies the *adj_slid(·)* predicate), 2) push the new most recent action and referent onto the stack for future retrieval if necessary.

```

Data: List of local variables
Result:  $event \leftarrow Action(Focus)$ 
 $DHE \leftarrow DialogueHistoryEventStack;$ 
 $DHO \leftarrow DialogueHistoryObjectStack;$ 
if  $DHE.Count == 0$  then
  |  $Agent.Speech: "No Events received yet!"$ 
else
  |  $priorActionIndex \leftarrow FindIndex(Prompt, x \Rightarrow x.Contains("adj"));$ 
  |  $Action \leftarrow Prompt[0];$ 
  |  $FocusType \leftarrow Prompt[Prompt.Length - 1];$ 
  | if  $priorActionIndex$  is 2 then
  | |  $priorAction \leftarrow Prompt[2];$ 
  | |  $FocusColor \leftarrow Prompt[1];$ 
  | |  $Focus \leftarrow FocusColor.Concat(FocusType);$ 
  | else
  | | if  $priorActionIndex$  is 1 then
  | | |  $priorAction \leftarrow Prompt[1];$ 
  | | |  $Focus \leftarrow FocusType;$ 
  | | end
  | end
  | while  $DHE.Count > 0$  do
  | |  $LastEvent \leftarrow DHE.Pop();$ 
  | |  $LastObject \leftarrow DHO.Pop();$ 
  | | if  $(LastEvent = priorAction) \wedge (LastObject = Focus)$  then
  | | |  $TempEventStack.Push(LastEvent)$ 
  | | |  $TempObjStack.Push(LastObject)$ 
  | | |  $event \leftarrow Action + (+Focus+);$ 
  | | |  $EventManager.InserEvent(event)$  VoxWorld Executes the event
  | | |  $Found \leftarrow True;$ 
  | | | break
  | | else
  | | |  $DHE.Push(LastEvent)$ 
  | | |  $DHO.Push(LastObject)$ 
  | | end
  | end
  | if  $Found \neq True$  then
  | |  $Agent.Speech: "Sorry I cannot remember this!"$ 
  | end
end

```

Algorithm 3: Historical RE Interpretation ($List < String > Prompt$)

2.4.6 Non verbal Referring Expressions Interpretation

Deictic inputs can be generated by humans along with clear speech, attributive REs, relational REs, or historical REs to ensure clarity in descriptions. They can resolve the meaning of demonstratives in ambiguous speech within attributive REs or after the agent requests disambiguation, using either pointing-only REs or co-gestural REs. Pointing gesture can also be sequential pointing prompts with no speech, e.g., Figure 2.12, in which the human points to the focus object, “BlueBlock”, then to the destination object/location to instruct Diana to put the focus object on the pointed at object/location, “GreenBlock”.

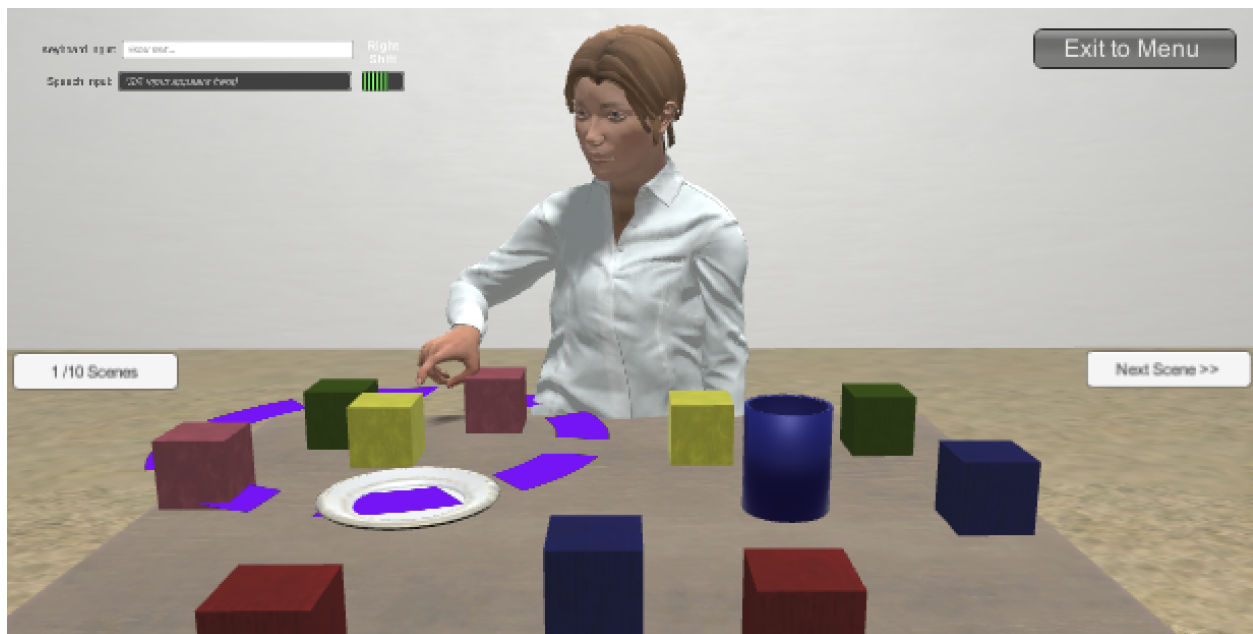


Figure 2.11: Generating deictic gestures. Diana will respond to what she interprets the RE as referring to by pointing to it, which can be used to assess the correctness of her object grounding depending on which object the human actually intended to reference.

2.5 Deterministic Generation of MREs

In addition to interpreting multimodal inputs, being able to generate non-verbal behavior is essential for interactive agents to add social fluency to the interaction [15]. Diana is able to generate speech via text-to-speech, deictic gesture via animation and inverse kinematics executed on her body rig, and action by manipulating virtual objects in the scene. At this stage, the generation

capability can be summarized as follows: (1) When the human indicates a block without supplying an action to execute, Diana points to it, confirming understanding of the RE with her own deictic RE, as shown in Figure 2.11. (2) She directly acts on all aforementioned verbal prompts (e.g., multimodal commands in Figure 2.13, #1–12) by either disambiguating candidate destination objects or carrying out the requested action in the virtual space. (3) She also acts on non-verbal prompts (e.g., unimodal commands in Figure 2.12 from 1-4) by performing the denoted actions after the human specifies the focus and target locations. (4) As shown in Figure ??, she expresses emotions (e.g., confusion and joy), in response to human inputs, such as being confused when there is an ambiguity in RE or action interpretation, or joy at having interpreted an input successfully. Appropriate generation, then, becomes a question of correctly generating the content of an utterance, movement through space of a gesture, or specific facial expression at the right time, to serve a communicative purpose.

[1] HUMAN: [*Points to a blue block*]
 [2] DIANA: [*Points to the same blue block*]
 [3] HUMAN: [*Points over the green block*]
 [4] DIANA: [*Puts the blue block on the green block*]

Figure 2.12: Sample dialogue: the interaction is unimodal (deictic gesture only).

[1] HUMAN: Take the red block and put it to the left of the plate.
 [2] DIANA: Which red block?
 [3] HUMAN: This one. [*Points to a red block*]
 [4] DIANA: OK! [*Acts on the red block*]
 [5] HUMAN: Grasp this pink block and put it on the red block. [*points to a pink block*]
 [6] DIANA: Do you mean the red block I just put down? [*Points to a red block*]
 [7] HUMAN: No, the red block behind the blue block.
 [8] DIANA: OK! [*Acts on the red block*]
 [9] HUMAN: Pick up the block that you just grasped.
 [10] DIANA: I've just grasped the pink block (that is) on the red block. Is this what you mean? [*points to a pink block*]
 [11] HUMAN: Yes.
 [12] DIANA: OK! [*Acts on the pink block*]

Figure 2.13: Sample dialogue: the interaction from 1–12 is multimodal (co-gestural speech).

2.6 Summary

In this chapter, we anticipated the complex dynamics of human communication [97], wherein verbal referring expressions are often complemented by non-verbal cues to effectively convey intentions and enhance comprehension. We have identified the cognitive processes involved in understanding such expressions, spanning from resolving spatial relationships to interpreting historical actions and gestures. Motivated by these cognitive strategies, our research endeavors to bridge the visual and kinematic capabilities of embodied intelligent virtual agents (IVA) and state-of-the-art natural language processing (NLP) technologies. By integrating VoxWorld, spatial relations recognition, and speech recognition, our aim is to empower IVAs to generate and interpret multimodal referring expressions in real-time interactions with humans. This bidirectional interaction scheme not only seeks to facilitate the implementation of successful multimodal generative IVAs but also lays the groundwork for training AI models to mimic human-like multimodal behaviors. Through the developed IVA, encompassing the transformation of linguistic speech to predicate logic format, interpretation of spatial referring expressions, prior-events based descriptions, and non-verbal cues, and deterministic generation of verbal and non-verbal behaviors, methodologies can be created to pave the path towards more sophisticated and intuitive human-IVA interactions.

Chapter 3

A Situated Corpus of MREs

3.1 Overview

Numerous corpora were collected for multimodal referencing research, but referential behaviors were either generated by simulators (e.g., VoxSim [64], CLEVER [52], CAESAR [65]), by humans referring to static images [23], or indoor objects in front of a camera [66]. Our goal is to explore how humans use multiple modalities when collaborating with IVAs and answering the key research question: *How can an IVA effectively extract the communicative intents of others from their deixis and language synthesis?* Therefore, we collected **SCMRE**, a situated corpus of multimodal reffering expressions to investigate humans-IVA understanding and referential capabilities, and to train generative AI systems for MREs. In this chapter, we elaborate on our data elicitation process and present statistics and analysis of participants' referring behavior during task-oriented dialogues. This explores whether humans and IVAs are able to extract intents using MREs. The collected dataset is publicly available on GitHub⁵ for use by other researchers in analyzing multimodal behaviors or training generative models. This work has been submitted to the 26th International Conference on Human-Computer Interaction (HCI International 2024).

3.2 Human-IVA Collaboration Data Collection

As our goal is to explore human referential strategies in collaborative interaction and develop a multimodal generative model of referring expressions, we organized several sessions of human-IVA interactions for an object identification task. This section details the process of collecting SCMRE, highlighting the recruitment of human participants, the procedures followed as shown in Figure 3.1, the collaboration between humans and the IVA, and the management of the collected data.

⁵https://github.com/nadahass/SCMRE_Dataset.git

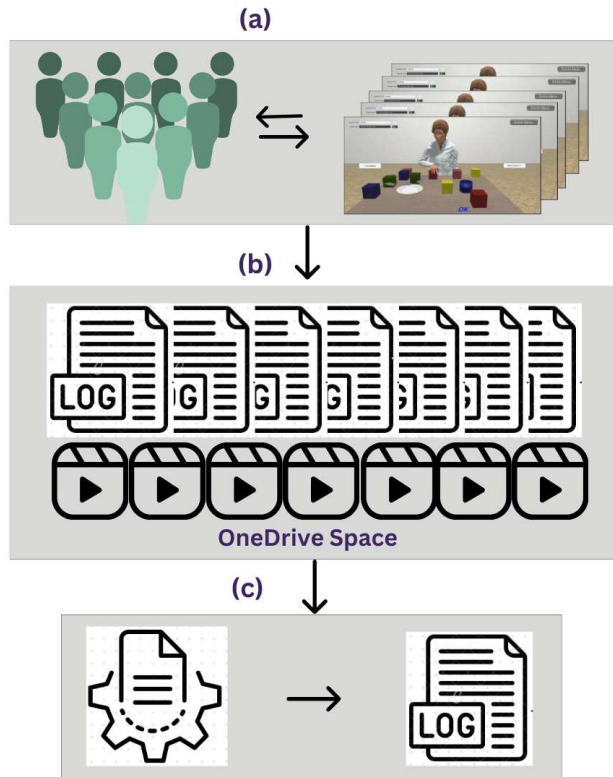


Figure 3.1: Data Collection procedures (a) Humans download the IVA and interact with her (b) The generated data is uploaded to the allocated OneDrive Space (c) The data is processed and merged into one CSV file

3.2.1 Participants

After obtaining approval from the IRB, we recruited 24 participants from Computer Science department in Colorado State University to use REs of different types and strategies while collaborating with the IVA to perform the required task detailed in the following part 3.2.2. People are eligible for participation if they fit the criteria of being over 18 years old, speaking English fluently, having access to a computer with a good-quality microphone to make a clear recording. Twelve participants were male and twelve of them were female. Their ages ranged from 18 to 35 (mean = 27, SD = 4.21). Their native languages are English, Hindi, Spanish, Chinese, Persian, Arabic, Assamese, and Nepalese. For participating in the study, participants receive a \$10 gift card from an Amazon merchant if they are recruited from outside a course, or an extra credit of 2% of the total class grade if recruited from a course.

3.2.2 Procedures

Conducting interactions between humans and an embodied IVA is the main aim of this experiment to understand human referring behaviors while collaborating with the agent. Therefore, participants were asked to download an executable application from an accessible OneDrive space. This application provides participants with the developed IVA for an object identification task (See Chapter 2). Specifically, participants are asked to generate different referring strategies that indicate 10 randomly placed target blocks in each of 10 scenes by language, gesture or ensemble when the IVA asked to refer to a block. During human-agent interaction sessions, referencing is considered successful when the IVA is able to ground the human’s MRE to the same object as the human intends to describe. We anticipated that participants need one minute or less to describe each block, which leads them to spend roughly 1 hour and 40 minutes, as maximum, to finish the task. The IVA’s and participant’s movements parameters including target objects, distance from the agent, object coordinates, relations in the scene, previously referred-to objects; utterance, modality(ies), and demonstratives, are logged for analysis, training and evaluation of multimodal referring expression generation models. In order to review the logged data about how participants respond to the questions, and about the scene, they were asked to record their interactions while referring to objects including their mouse movements and their voice without including their identities. They were required to make these recordings at their own locations to achieve more diversity in our samples and quicker data collection. Participants also completed a survey to provide demographic information, such as gender, nationality, native language, and to note any errors observed during the interaction.

3.2.3 Data Management

The data received from each participant included CSV files for each scene, resulting in a total of 10 scenes and 10 corresponding videos per participant. Participants were instructed to upload their data to a designated OneDrive space. To ensure the validity and reliability of the data, we reviewed each scene’s data, including CSV files and videos to check for any logging errors or

duplication. This careful verification process was essential to maintain the integrity of the dataset. After confirming the accuracy of the data, we merged all the generated referring expressions into a single comprehensive CSV file. This file serves as the primary dataset for analysis, preparation, and training purposes, facilitating the development and evaluation of multimodal referring expression generation models.

3.3 Dataset Statistics

We divided our presentation of dataset quantities into three parts: namely, events statistics, prompts statistics and references statistics. A prompt might involve more than one event that includes actions and referring expressions. In events statistics, we count the number of events to execute each prompt. In references statistics, we count all referents in each generated prompt. In prompts statistics, we count the number of human commands that are provided to the agent to act on objects. For example, if a human asks the IVA to, “Put the red block on the green block”, the statistics are: 1) one event: *Put(the(RedBlock, on(the(GreenBlock))))*, 2) two references, one for the focus object (*RedBlock*) and another for the target object (*GreenBlock*), and 3) one prompt which is the generated request. For pointing only prompts, if a human pointed to a block (e.g., *RedBlock*) then pointed to another block (e.g., *GreenBlock*), the statistics are: 1) two events: *Grasp(the(RedBlock))* and *Put(the(RedBlock, on(the(GreenBlock))))*, 2) two references, one for the focus object (*RedBlock*) and another for the target object (*GreenBlock*), and 3) one prompt which is the generated request.

Table 3.1: Quantities of human generated events before removing duplicated referencing and non-referencing events

Total of Events (with duplication)	11,976
Total of executed and not executed referencing events	10,670
Total of non-referencing events	1,306
Total of duplicated referencing events	948
Total of duplicated non-referencing events	620

3.3.1 Events statistics

The construction process of SCMRE corpus relies on the number of provided events such that each event is inserted as a sample. Each verbal-only or multimodal prompt is added as a sample as the agent receives only one event including action and referring expressions to execute the prompt. However, a pointing-only prompt, e.g, a deictic prompt of moving a block to another is inserted as two samples. Theses two events are: 1) a pointing event to the focus object: *Grasp(focusObj)*, and 2) a pointing event to the target object: *Put(focusObj, on(targetObj))* to execute the prompt. The elicitation process resulted in a total of 11,976 events as shown in Table 3.1, including 10,670 executed and not executed referencing events and 1,306 non-referencing events. The events generated by participants varied in the range of 258 to 801 (mean = 444.6, SD = 171.9). Moreover, the data includes 194 recorded videos spanned over around 36 hours, ranging from 24 minutes to 4 hours (mean of 01:27:52, SD = 0.945 hours). The number of samples that are generated by 24 participants is significantly larger than what we anticipated (2400 samples). Some participants generated multiple descriptions for each object, while others provided data for more than the 10 required scenes.

Table 3.2: Quantities of human generated events after removing duplicated referencing and non-referencing events

Total of Events	10,408
Total of executed pointing-only events	7,681
Total of executed transitive-attributive events	551
Total of executed attributive events	641
Total of executed relational events	369
Total of executed historical events	27
Total of executed non-referencing events	686
Total of not executed verbal events	453

As humans are able to use text inputs in addition to speech to communicate their intents with the IVA, their switching from keyboard entries to voice entries without removing the previous textual prompt from the text field led to duplicate 948 referencing events and 620 non-referencing events. After removing these duplications, the final dataset comprised 10,408 events. This in-

cluded 7,681 pointing-only references, 551 transitive attributive events, 641 attributive events, 369 relational events, 27 historical events, 453 not executed verbal and and multimodal events, and 686 non-referencing events. The non-referencing events encompassed undoing actions, such as “Nevermind,” 428 events, and “Stop,” 23 events, refusal prompts like “No,” 118 events, and affirmative prompts, e.g., “Yes,” 117 events. In terms of modalities used by humans, 575 events were generated by mixing deixis and language, 7,681 events were generated using pointing-only, and 2,152 events were generated using speech-only. This distribution is presented in Table 3.2.

Table 3.3: Quantities of human generated events based on modalities used, including deictic gesture, speech only, or both

Humans’ Used Modalities			
Events	Modalities	Quantity	Total
Attributive Multimodal Events	Multimodal	186	575
Transitive Attributive Multimodal Events		302	
Relational Multimodal Events		48	
Historical Multimodal Events		3	
Not executed multimodal events		36	
Focus and target pointing	Pointing-Only	7,681	7,681
Attributive Speech Only Events	Speech-Only	455	2,152
Transitive attributive speech only events		249	
Relational speech-only events		321	
Historical speech-only events		24	
Non-referencing speech-only		686	
Not executed speech-only events		417	
Total			10,408

The IVA acted to each human-generated event, which resulting in 10,408 IVA-based reactions, as shown in Table 3.4. While interacting with humans, the IVA generated 5,271 multimodal actions in response to 539 multimodal events, 3,628 pointing-only events, 686 non-referencing events, and 418 speech-only events. Moreover, the agent react unimodally using either deictic gesture or language. She acted to human pointing to 4,053 focus objects to confirm her understanding of interlocutors’ pointing using her deictic reference. Using language, she asked for more information to disambiguate 1,084 events.

Table 3.4: Diana generated behaviours using deictic gesture, speech only, or both to act on human generated events illustrated in Table 3.3

Diana Used Modalities			
Diana’s behaviours	Modalities	Quantity	Total
Acting to multimodal events	Multimodal	539	5,271
Acting to pointing-only events		3,628	
Acting to non-referencing events		686	
Acting to speech-only events		418	
Pointing to the focus objects	Pointing-Only	4,053	4,053
Asking for disambiguation	Speech-Only	1,084	1,084
Total		10,408	

Table 3.5: Quantities of human generated references based on modalities used including language, gesture and ensemble

Total of References	10,273
Total of executed pointing-only REs	7,681
Total of executed speech-only REs	1,298
Total of executed multimodal REs	841

3.3.2 References statistics

As we explore human use of referring expressions— including types (e.g., block, cup, plate), attributes (e.g., red, large, wide), and relations (e.g., to the left, on, beside)—we focus in this section on presenting the quantities and categories of references utilized in each generated command. In human-IVA interactions, humans asked the agent to act on shared objects using 10,273 references by exploiting information of objects characteristics and locations using language and gesture. These references include 7,681 pointing-only references, 1,298 speech-only references, or 841 multimodal references, as demonstrated in Table 3.5. If language modality is used with or without gesture, different contextual information is leveraged to communicate intents. This information includes relations between objects to generate relational REs, color of objects to generated attributive REs, and actions performed on previous referred-to objects to generate historical REs. As shown in Table 3.6, SCMRE corpus includes 186 multimodal attributive REs, 455 speech-only attributive REs, 48 multimodal relational REs, 321 speech-only relational REs, 3 multimodal historical REs, 24 speech-only historical REs, 7,681 pointing-only transitive attributive REs, 604

multimodal transitive attributive REs, and 498 speech only transitive attributive REs. Every *one* transitive attributive event includes *two* referred-to objects, a focus object and a target object, to move the focus to the target object.

Table 3.6: Quantities of human generated references based on information used, including attributes, relations, and previous actions, to refer to targets

Total of References	10,273
Pointing only transitive attributive REs	7,681
Multimodal transitive attributive REs	604.00
Speech only transitive attributive REs	498.00
Multimodal attributive REs	186.00
Speech-only attributive REs	455.00
Multimodal relational REs	48.00
Speech-only relational REs	321.00
Multimodal historical REs	3.00
Speech-only historical REs	24.00

3.3.3 Prompts statistics

In human-to-human collaborations, typically the multimodal referring expression is a part of a provided prompt. Mainly, the referential prompts are composed of actions, e.g., “grasp”, “put”, “move”, and referring expressions. This section presents the quantities and categories of human generated prompts in human-IVA interactions. To perform the required task, humans prompted Diana to act on objects using syntactic composition of actions and references that describe focus objects, target objects, or both using gesture-only, speech-only or both. Prompts is exemplified by “move this red block to the blue block”, which demonstrates a one prompt and two references. In terms of modalities used, human-IVA interactions resulted with 6,780 prompts, including 4,035 pointing-only prompts, 1,735 speech only prompts counting non-referencing and referencing prompts, and 539 multimodal prompts, as presented in Table 3.8. In terms of the contextual information used for discriminating objects, the quantity of prompts in each of attributive, relational and historical categories are identical to the quantity of references in each of those categories as shown in Table 3.7 and 3.6 since each command has one reference only. Nevertheless, as the number of

referents in each command is not counted, the quantity of transitive-attributive prompts is significantly lower than the quantity of transitive-attributive references. Specifically, 4,053 Pointing only, 302 multimodal, and 249 speech only transitive attributive prompts.

Table 3.7: Quantities of human generated prompts based on information used, including attributes, relations, and previous actions

Total of Prompts	6,780
Pointing only transitive attributive prompts	4,053
Multimodal transitive attributive prompts	302.00
Speech only transitive attributive prompts	249.00
Multimodal attributive prompts	186.00
Speech-only attributive prompts	455.00
Multimodal relational prompts	48.00
Speech-only relational prompts	321.00
Multimodal historical prompts	3.00
Speech-only historical prompts	24.00

Table 3.8: Quantities of human generated prompts based on modalities used to communicate intents

Total of Prompts	6,780
Total of executed pointing-only prompts	4,053
Total of executed speech-only prompts	1,735
Total of executed multimodal prompts	539

3.4 Data Analysis

After presenting the descriptive statistics of the collected data, we proceed to a detailed analysis to answer the above question and uncover insights. This section outlines the key findings from our data analysis, exploring the relationships between modalities utilization and extraction of intent meaning.

In terms of pointing, it reduces cognitive load [19] in human-human interactions. Similarly, this is observed with the IVA, such that when she misunderstands or asks for more information, the human recognizes her verbal questions or the failed multimodal actions and provide clarification using verbal language, pointing gestures, or a combination of both. The IVA points to the human

targets to confirm her successful comprehension or ask again for disambiguation. As shown in the contingency table, Table 3.9, the agent shows her understanding of the human’s intended meaning when providing a sequence of pointing REs or co-gestural speech (Multimodal REs) without asking for more clarification by pointing to the referents. Nonetheless, relying solely on speech for communication often necessitates the agent to seek additional information, such as gestures, to precisely identify and point to the target. When the speech-only REs are relational REs or historical REs, leveraging recognized relations or historical dialogue, respectively, disambiguation is less frequent. However, if the agent fails to recognize the relation or previous actions, clarification is requested by the IVA.

Table 3.9: Contingency table of human RE ambiguity and modalities used: # ambiguous REs by modality type

Modality	Did Diana Disambiguate?	
	No	Yes
Multimodal RE	539	0
Pointing Only RE	7,681	0
Speech Only RE	1,104	1,084
P-value	< 2.2e-16	

Table 3.9 shows that the IVA asked to clarify 1084 events; namely the attributive REs (e.g., “the red block”), unrecognized spatial REs, and not-executed REs. On the other hand, 1104 verbal events, including relational RE, pronoun (e.g., “it”), historical REs, attributive REs with no demonstratives for the target (e.g., “take red block”), and non-referencing speech, were clear to the IVA. As Pearson’s χ^2 -test [98] is used for categorical data [99], it indicates that a relationship exists between the modalities used and the level of ambiguity, such that use of pointing and complex linguistics significantly reduces the ambiguity level of the prompt (p -value < 0.001). The relationship indicates that incorporating pointing and complex linguistic REs significantly reduces the ambiguity of prompts and enables the IVA extracting the meaning of human intents. This finding implies that these modalities enhance the clarity and effectiveness of communication between humans and the IVA. A more extensive analysis and evaluation, covering detailed methodologies,

comprehensive data interpretation, and an in-depth discussion of findings, are presented in Chapter 5. This chapter delves into the collection and exploration of data, investigating whether the agent can extract the contextual meaning of intents.

3.5 Summary

This chapter has detailed the comprehensive process of collecting SCMRE, aimed at training generative AI systems for multimodal human-IVA interactions involving both language and gestures. The recruitment of participants was designed to ensure diversity and relevance, resulting in a well-rounded sample of individuals from various nationalities and age groups. The procedures followed ensured systematic data collection while allowing participants to interact naturally with the IVA in a shared virtual space.

The collaboration between humans and the IVA was organized to elicit a wide range of referring expressions, encompassing verbal and non-verbal cues. The management of collected data was rigorous, involving careful logging and validation to ensure the reliability and validity of the dataset. We presented detailed statistics of the collected data, which included a variety of referring expressions and interaction modalities. This dataset not only supports the training of AI systems but also provides insights into human referential strategies in collaborative settings.

The quantities of SCMRE trials demonstrated the IVA's ability to retrieve and disambiguate objects referred to by humans. Additionally, the trials highlighted humans' capability to communicate intents using language, deictic gestures, or a combination of both, and emphasized the ease of use during interactions. The findings from this chapter indicate the potential of multimodal interactions in enhancing human-IVA communication and provide a solid foundation for further research and development in this field.

Chapter 4

MRE Generation Model

4.1 Overview

Multimodal generative AI models represent a significant advancement in the field of artificial intelligence, integrating multiple forms of data such as text, images, audio, and gestures to generate coherent and contextually appropriate outputs [39]. These models are designed to emulate human-like understanding and generation of information across different modalities, facilitating more natural and effective human-computer interactions. Gesture and speech integration is another critical area of multimodal AI. Since there are no existing models that combine these modalities, we aim to improve human-computer interaction clarity and naturalness by understanding and generating non-verbal cues alongside spoken language. The research question we addressed in this chapter is: *How can a large language model (LLM) effectively integrate multiple modalities, including speech and gesture, to convey its intentions similar to humans?*

The *neurosymbolic* approach to AI has long argued for *structured representation* [100, 101], and similar arguments have also been made by deep learning luminaries e.g., Bengio [102]. Introducing preexisting structure into NLP facilitates higher-order reasoning, but a hybrid approach also does so at larger scales than purely symbolic systems, by using flexible deep-learning representations to process the structured input. Therefore, SCMRE, our collected human-IVA interactions with symbolic referential structures, was used to train a generative AI model for integration into embodied IVAs to generate MREs. LLM outputs serve as inputs to the IVA's simulators, such as VoxSim, enabling the generation of coordinated speech and gestures to simulate more human-like interaction patterns. This integration is particularly beneficial in applications that utilize MREs, such as instruction assistants, where the ability to understand and generate gestures can significantly enhance the user experience.

In this chapter, we delve into: 1) data preparation, discussing the processes of data augmentation, synthesizing, prompting and splitting, 2) model architecture, describing the process of producing MREs that closely approximate human responses within a given environment, 3) learning experiments, outlining the experimental setup, training procedures, and variations in experiments, 4) learning results, presenting the quantitative results, comparing them with human-generated MREs, and discussing the implications.

The work described in this chapter has been published to the *26th International Conference on Human Computer Interaction* (HCI International 2024). The code for LLM fine-tuning, training, and testing is available on GitHub⁶, and the trained generative LLM is publicly accessible on Huggingface⁷.

4.2 Data Preparation

To create a robust and diverse dataset that ensures that an LLM trained over it can contextually generate MREs, four key preparation steps were applied to the SCMRE dataset: annotation, augmentation, synthesizing, prompting and splitting, as illustrated in Figure 4.1.

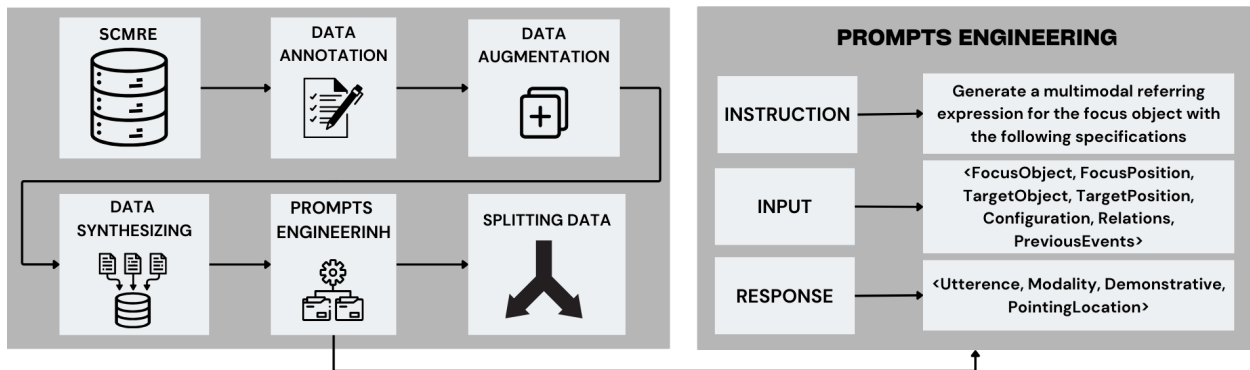


Figure 4.1: The main steps of data preparation, including annotation, augmentation, synthesizing, prompting, and splitting

⁶https://github.com/nadahass/MREG_LLM.git

⁷<https://huggingface.co/Nadahass/MREG-LORA8-13b-4batches-1epochs-q8-lr1e-5>

4.2.1 Data Annotation

This step addressed 453 prompts that were not executed because they could not be parsed by the IVA’s parser component. One such example is “move blue block in corner to pink block,” where failure to correctly parse “in corner” resulted in an invalid PLF form. This failure prevented the identification of target objects and associated parameters. Thus, we annotated these prompts to include all necessary parameters. The parameters that were logged for these prompts include timestamps, utterances, relations, configurations, and previous events. To contribute effectively to the exploration of human referential behaviors and the learning of our proposed model, we needed to add the remaining parameters. These missing parameters are focus objects, destination objects, focus positions, target positions, and demonstratives. We systematically review these prompts and their corresponding videos to predict the focus and target objects, extract their positions from the generated configurations, and identify the demonstratives within the linguistic prompts.

4.2.2 Data Augmentation

We utilized a data augmentation method to increase both the size and diversity of our SCMRE dataset. Specifically, we employed synonym augmentation from the NLPAug library [103] to expand the range of multimodal and speech-only referring expressions. Each original expression was augmented to produce three similar expressions. To maintain semantic similarity to the ground truth MRE, we systemically reviewed and adjusted the augmented expressions by replacing less popular or informative words to align with our specific requirements. We then used BERT Score [104] to assess semantic similarity between augmented REs to human REs using the cosine similarity of their respective embedding vectors. We achieve an average BERT-Precision of 97.1%, BERT-Recall of 97.6%, and BERT-F1 97.3%. The dataset was expanded to include 16,527 events, comprising 2,296 multimodal REs, 6,550 speech-only REs, and 7,681 pointing-only REs. Both multimodal REs and speech-only REs obtained significant increases compared to their original counts, as detailed in Table 4.1.

4.2.3 Data Synthesizing

Despite the expansions resulting from augmentation, the dataset remained imbalanced, particularly in multimodal REs, potentially affecting the robustness of MRE generative model training. To augment the dataset with diverse multimodal samples, we synthesized individual pointing-only and speech-only samples to create new multimodal RE samples. This process involved identifying instances where both deictic gestures and speech were used to refer to the same object at the same spatial location. By aligning these expressions based on their shared focus object and position, we created composite samples that incorporate both modalities. Following this synthesis approach, the dataset includes 7,928 multimodal samples. Specifically, 5,632 new multimodal RE samples were generated, increasing the dataset size from 16,527 to 22,159, as shown in Table 4.1. This enhancement contributes to a more balanced dataset, improving the results of MRE generative models learning as shown in Chapter 3. The augmentation and synthesizing code is publicly available on GitHub⁸.

Table 4.1: Quantities of original, augmented, and synthesized datasets

Dataset	Speech-Only REs	Multimodal REs	Pointing-Only REs	Total
Original Dataset	2,152	575	7,681	10,408
Augmented Dataset	6,550	2,296	7,681	16,527
Synthesized Dataset	6,550	7,928	7,681	22,159

4.2.4 Prompt Engineering

Prompting and instruction tuning are established methods used to elicit a response from a Large Language Model (LLM). In prompting process, an instruction p , manually composed, is concatenated to the input x . Then, the response is sampled from the probability distribution, denoted as $\hat{y} \sim p(\mathbf{y} \mid \mathbf{p}, \mathbf{x}; \theta_{LLM})$ [105].

We used Alpaca [106] as the basis for our MRE-generating model. Alpaca’s Instruction-following models [106] require structuring the data in a way that aligns with the model’s archi-

⁸https://github.com/nadahass/SCMRE_Dataset.git

Table 4.2: Training, Validation, and Testing Sets

Dataset	Train. Set	Valid. Set	Test. Set	Total
Original Dataset	8,325	1,665	417	10,407
Enhanced Dataset	17,727	3,545	887	22,159

ture, incorporating instructions, inputs, and outputs consistently throughout the dataset. This involved concatenating a set of columns for both the input and output components as shown in Figure 4.1. The input tuple includes configuration, relations, focus object, target object, and previous events, while the output tuple comprises the utterance, modality, demonstrative, and pointing location.

4.2.5 Data Partitioning

We divided the original and enhanced dataset into three subsets: a training set, validation set and a testing set. The training set, comprising 80% of the total data, was used to train models. The validation and testing sets, consisting of 80% and 20% of the remaining 20% of the total data, was reserved for evaluating the model’s performance. The testing data, comprising 20% of the validation and testing data, was used to assess the model’s generalization ability on unseen data. Table 4.2 illustrates the resulting number of samples in each set for both original and enhanced datasets. To ensure an unbiased representation of the data, the datasets was shuffled and the division was performed randomly.

4.3 MRE Generation Model

4.3.1 Model Architecture

We used open-weight LLaMA2-chat models [107] to conduct parameter-efficient fine-tuning for generating contextually-correct and situationally-fluent referring expressions, including language and gesture. As illustrated in Figure 4.2, the model takes a query, representing the target object O , its position P , relations R , configurations C , and previous events H ; and outputs a descriptor tuple, $\langle \text{Modality}, \text{Utterance}, \text{Location}, \text{Demonstratives} \rangle$. $M \in \{\text{Gesture}, \text{Language},$

Ensemble}, U is a decoded sentence embedding, L is the location the gesture grounds to, and $D \in \{the, this, that\}$. Depending on the value of M , some of the other parameters may be empty by default. The query constitutes a description of the environment in which the agent is situated, along with an utterance prompting for a referring expression, and the model is optimized to generate output that approximates what a human would say in response to the prompt, while remaining situationally-grounded, fluent, natural, and referring to the correct object. The query $\langle O = \text{RedBlock}, P = \langle X, Y, Z \rangle, \text{None}, \text{None}, R = [\text{Right}(\text{RedBlock}, \text{GreenBlock}), \dots], C = [\langle X', Y', Z' \rangle, \dots], H = [\text{Put}(\text{YellowBlock}), \dots] \rangle$, represents the target object (the red block), the current spatial arrangement, associated relations and previous events. The corresponding output, $\langle \text{multimodal}, \text{pick the red block}, \langle X'', Y'', Z'' \rangle, \text{the} \rangle$, embodies the elements of the generated multimodal referring expression. Here, this output prompts the agent to utter “pick the red block” while pointing to the location denoted by Cartesian coordinates $\langle X'', Y'', Z'' \rangle$.

Given the instructional prompts as detailed in Section 4.2.4, Llama2-chat’s tokenizer generates embedding vectors for inputs, converting tokens into numerical indices used as model inputs. An end-of-sequence token $\langle \text{EOS} \rangle$ is appended to the input sequence, with the label matching the input and an attention mask set to 1. The decoding-only transformer, which includes multihead masked self-attention and neural networks, selects the next token from the probability distribution over tokens. This process continues iteratively until the $\langle \text{EOS} \rangle$ token is reached or the maximum length is achieved.

4.3.2 Learning Experiments

We fine-tuned multiple LLMs using Low-Rank Adapters (LoRA [108]) to enhance parameter and memory efficiency. LLaMA 2-chat [107], developed by Meta AI, includes large-scale generative language models available in distinct parameter sizes: 7B, 13B, 70B, and empirical studies indicate that even the LLaMA 2-chat-13B model, with just $\frac{1}{10}$ of the parameters, surpasses GPT-3 (175B) [109] in most benchmark evaluations. For this study, we selected LLaMA 2-chat-7B⁹ and

⁹<https://huggingface.co/meta-llama/Llama-2-7b-chat-hf>

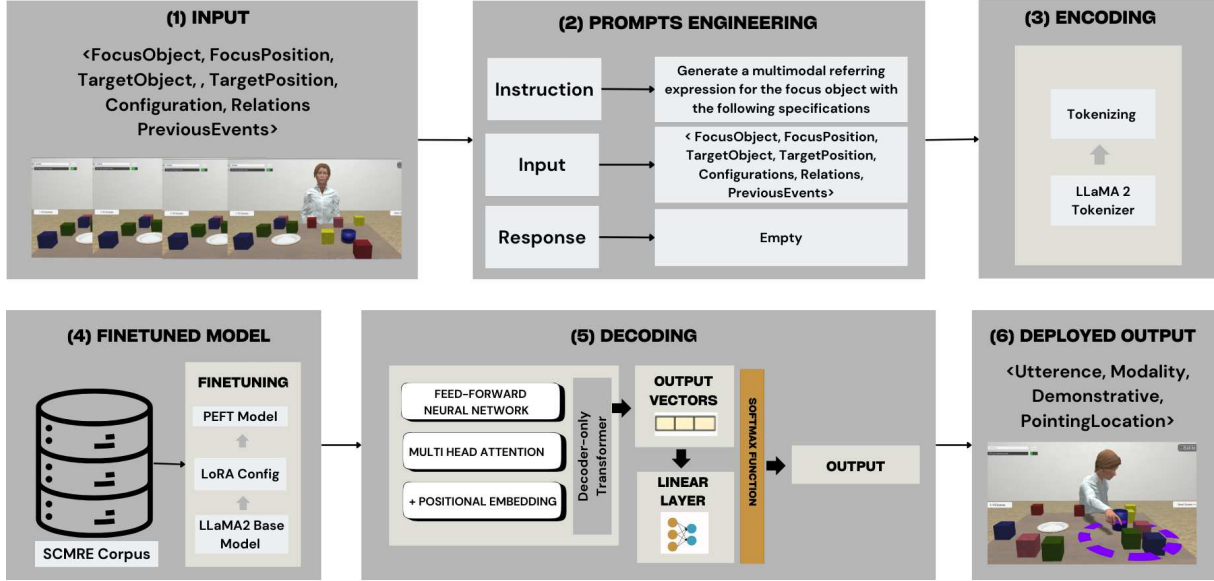


Figure 4.2: The MREG architecture under exploration

Table 4.3: Hyper-parameters of Fine-tuning and Training Time for LLaMA Models

Models	Dataset	Learning rate	Epochs	Steps	Training Time (hh:mm:ss)
LLaMA-13B	8,325	3×10^{-4}	0.14	300	3 : 54 : 29
LLaMA-13B	17,727	3×10^{-4}	0.07	300	6 : 49 : 15
LLaMA-7B	17,727	2×10^{-5}	0.25	1,107	9 : 00 : 35
LLaMA-13B	17,727	3×10^{-4}	1	4,432	48 : 30 : 10

LLaMA 2-chat-13B¹⁰ as our foundational experimental models. To enable loading these models, fitting them into memory, and speeding up inference, we employed 8-bits quantization to represent weights with lower-precision data types. We use LLaMA2-chat in this study, which for convenience is hereafter simply referred to as “LLaMA.” We applied patches to the LoRA modules for the query, and value matrices, set their rank to 8, a scaled factor to 16, a dropout rate of 0.05, and task type to CAUSAL_LM. This setting reduced the trainable parameters from 13,022,417,920 parameters to 6,553,600 parameters, allowing models to be processed on 2 NVIDIA RTX A6000-49GB GPUs.

According to Alpaca-LORA implementation, we utilized a learning rate of 2×10^{-5} for LLaMA-7B and 3×10^{-4} for LLaMA-13B. The fine-tuning process included one LLaMA-7B model that

¹⁰<https://huggingface.co/meta-llama/Llama-2-13b-chat-hf>

was fine-tuned for 1, 107 steps, and three LLaMA-13B models were fine-tuned, two for 300 steps each, and one for 4, 432 steps. We applied *AdamW* as a stochastic optimization method with a global batch size of 4 and precision of *fp16*. We incorporated warm-up steps of 100 and validation steps of 100 for all models. The checkpoint with the best cross-entropy on development set was retained. Table 4.3 lists the hyper-parameters, training sets and training time that are related to each model.

4.3.3 Results

Loss Entropy. The loss curve for LLaMA-13B in Figure 4.3a, trained for 4,430 steps (1 epoch), shows slower convergence; however, it achieves lower loss values compared to LLaMA-7B in Figure 4.3c, which was trained for 1,107 steps (0.25 epochs). The fine-tuned LLaMA-13B reached training and evaluation losses of 0.517 and 0.515, respectively, while the LLaMA-7B obtained 0.576 and 0.575.

Perplexity. As depicted in Figure 4.3b,d, the perplexity of both models decreases steadily as training progresses, indicating that both fine-tuned models are learning and improving their predictions over time. Nonetheless, the LLaMA-13B model demonstrates a more rapid decrease in perplexity compared to the LLaMA-7B model. The fine-tuned LLaMA-7B achieved training and evaluation perplexity values of 1.777 and 1.779, respectively, whereas the LLaMA-13B recorded values of 1.676 and 1.674. This suggests that LLaMA-13B converges faster and achieves better performance more quickly.

Comparisons between Human and LLM Utilization of Referring Strategies. We tested the performance of fine-tuned Alpaca LoRA-based models—namely LLaMA-7B and LLaMA-13B—in integrating gesture and speech for referential behaviors across various parameterizations. Using datasets of 10K and 22K samples and varying training epochs and step counts (see Table 4.3), it was observed that the performance improved with larger datasets, models, and more training steps. The LLaMA-13B model, trained for one epoch on a test set of 887 samples, demonstrated the best

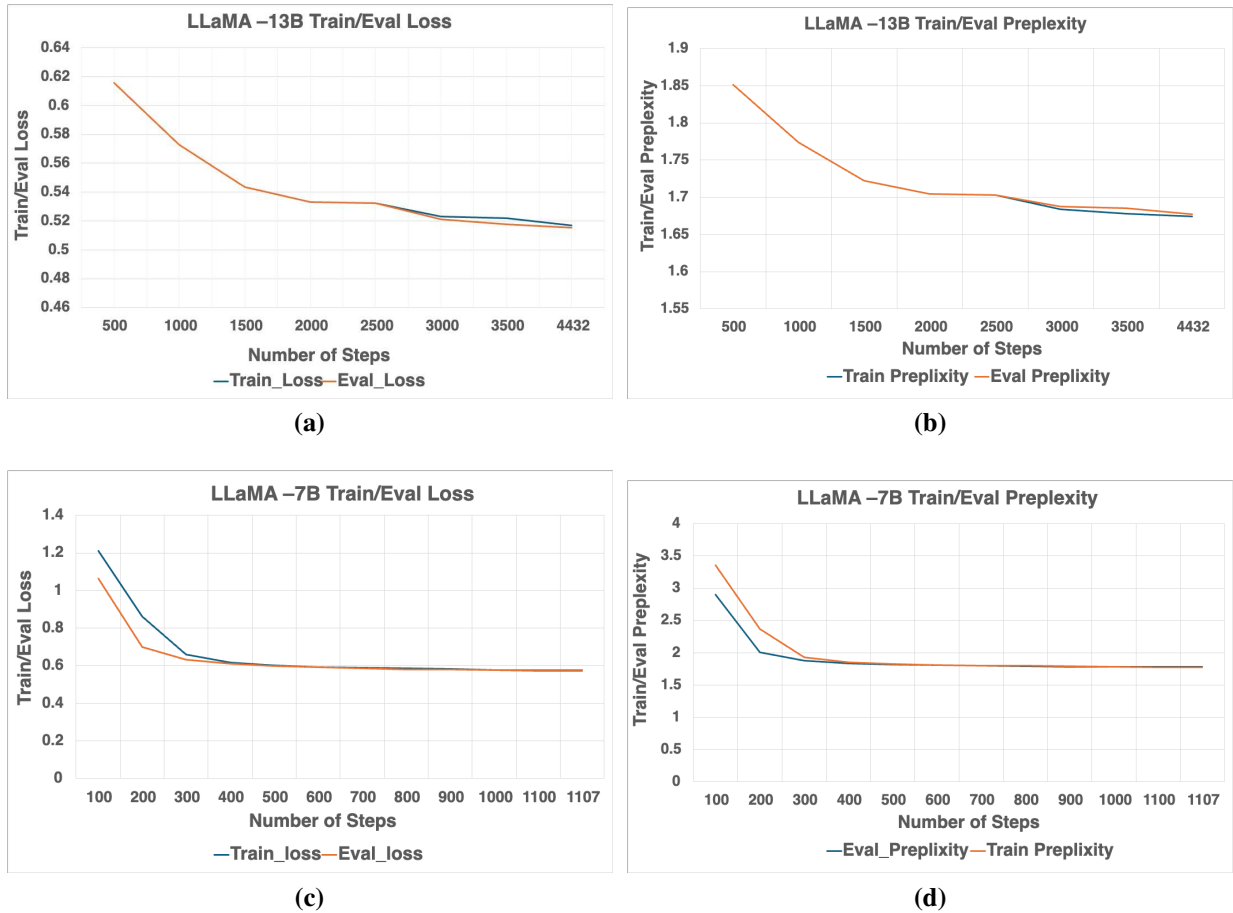


Figure 4.3: The loss (a) and perplexity (b) of LLaMA-13B after one epoch of training. The loss (c) and perplexity (d) of LLaMA-7B after 1,107 steps of training

performance in mixing modalities for generating referring expressions as depicted in Figure 4.4d. It generates 40.61% of multimodal REs, 13.91% speech-only REs, and 45.48% of pointing-only REs, closely resembling human utilization of modalities when generating REs as in Figure 4.4e: 43.55%, 22.29%, 34.16%, respectively. Nevertheless, pointing-only REs dominate with the tuned LLaMA-13B model trained on the original dataset. In Figure 4.4a, they account for 96% of outputs. On the LLaMA-7B model (Figure 4.4b), they account for 55.13%, and on the enhanced dataset with fewer steps (Figure 4.4c), they account for at 54.31%.

Similarity between human-generated and LLM-generated MREs. Successful generation results in a descriptor tuple that includes speech, demonstrative, gesture, and the target location for the specified target object and scene configuration. The multimodal generated description should

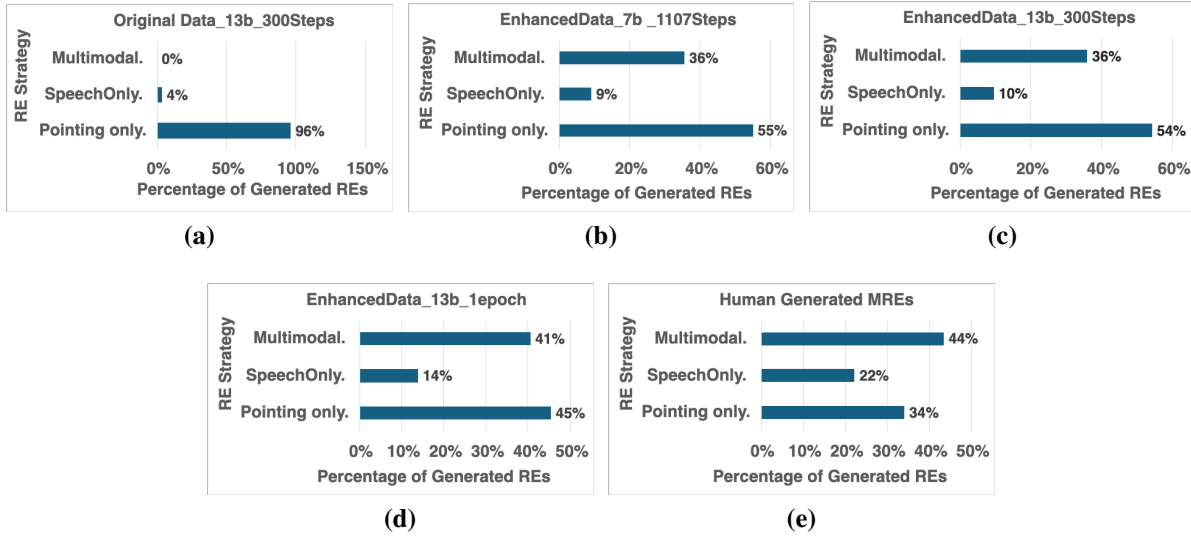


Figure 4.4: Quantities of Human and LLM Generated Pointing, Linguistic and Multimodal Referring Expressions.

maintain semantic similarity to the ground truth MRE. Semantic similarity must be attained at both the speech and position levels. Similarity results are illustrated in Figure 4.5 for the four fine-tuned models, utilizing automatic evaluation metrics that compare model outputs to human REs in the same situation. The tuned LLaMA-13B model for one epoch surpasses all models in achieving similarity to human outputs on both the tuple and speech levels. It achieves an average BERT-Precision of 93%, BERT-Recall of 93%, BERT-F1 of 93%, and IoU of 72% on the tuple level, and an average BERT-Precision of 91%, BERT-Recall of 92%, and BERT-F1 of 91% on the speech level. Figure 4.6 depicts the distribution of similarity results of BERT-F1 between human-generated tuples and the dominant LLaMA-13B model-generated tuples. Approximately 350 samples exhibit similarity results ranging from 98% to 100%. The remaining low-similarity results occur due to the divergence in generated modalities compared to human samples.

Correctness of Generated Positions. Given our goal of ensuring the LLM produces contextually appropriate referring expressions, it is essential to assess their positional accuracy. The fine-tuned LLaMA-13B, trained for one epoch, achieved remarkable performance, with an average accuracy, precision, recall, and F1-score of 99% for correctly generated positions. Performance

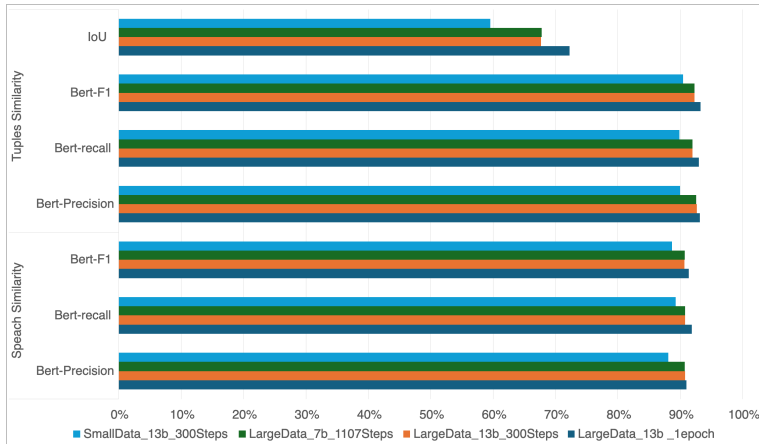


Figure 4.5: The similarity between human-generated and LLM-generated MREs using BERT-Precision, BERT-Recall, BERT-F1, and Intersection of Union(IoU)

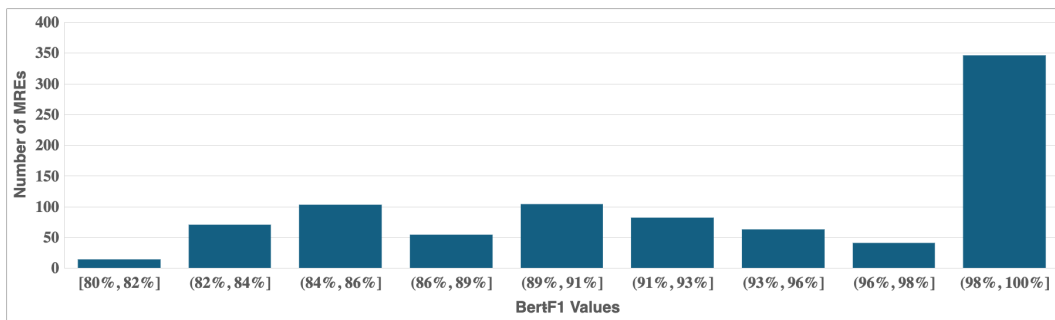


Figure 4.6: The distribution of BERT-F1 similarity results between human-generated and (13B-1 epoch) LLM-generated MREs

for LLaMA-13B with fewer steps and LLaMA-7B was notably lower, reaching 86% and 97% respective for the LLaMA-13B models, and 87% for LLaMA-7B, as shown by Figure 4.7.

Clarity of Generated MREs. As the goal is to maintain human-level clarity for the fine-tuned LLM, we compared human and LLM-generated REs based on the same provided contextual information. Figure 4.8 categorizes the combined strategies used by human and LLM to convey information about the target object. Humans utilized multimodal relational REs, multimodal attributive REs, speech-only relational REs, and speech-only attributive REs. The fine-tuned model utilized all the above strategies except speech-only attributive REs. This is a feature of the fine-tuned model, as using only object attributes without additional clarification often requires interlocutors to seek disambiguation, leading to inefficient communication of intent. Figure 4.9 presents exam-

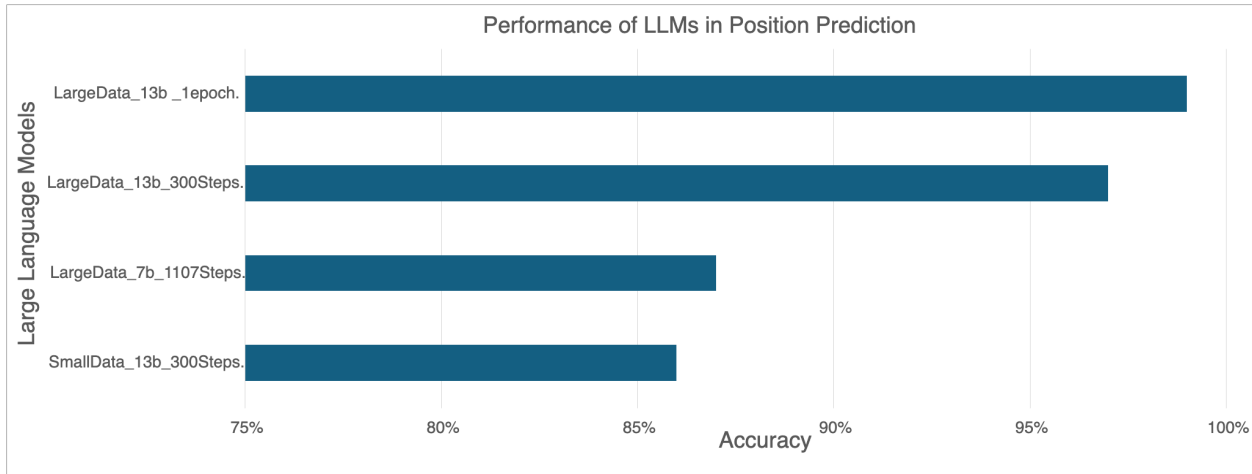


Figure 4.7: The performance of LLaMA models in predicting positions of target objects

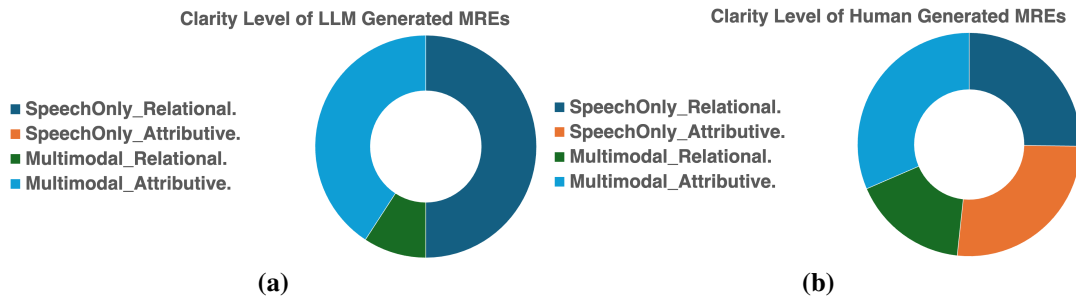


Figure 4.8: Clarity level of (a) LLM-generated and (b) human-generated REs

ples of all combinations of RE strategies for both humans and the fine-tuned model when referring to the same target objects in identical situations.

4.4 LLM-IVA Integration

We used server-based approach, llama.cpp¹¹ [107], to integrate the LLM with the IVA. This integration leverages the flexibility, efficiency, and scalability of the llama.cpp server to deploy, manage, and scale LLMs in real-time applications, facilitating more efficient and effective IVA-human interactions.

To set up the llama.cpp server, we first installed the necessary dependencies, cloned the repository, built the server, and configured the settings to specify the model path, server details, and

¹¹<https://github.com/ggerganov/llama.cpp.git>

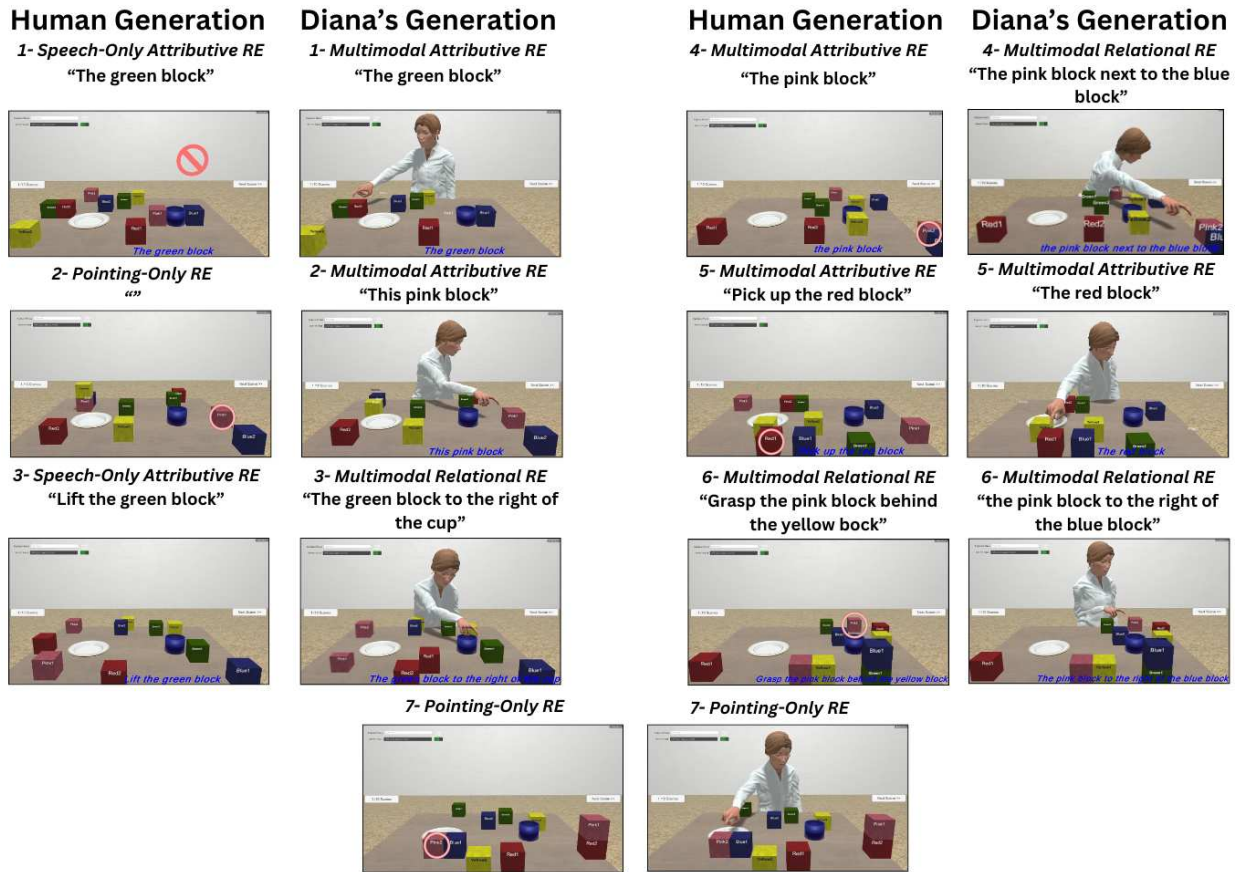


Figure 4.9: Comparing human and IVA-generated REs for identical configurations.

usage parameters. For optimal performance and compatibility with the Unity-developed IVA, we convert the LLM to GGUF format, ensuring efficient and streamlined inference. Once configured, the server can be started, allowing the IVA to communicate with it using a REST API. This setup ensures efficient query submission and response retrieval, as illustrated in Figure 4.10, with the server handling the computational load of generating language outputs.

The communication protocol between the IVA and the llama.cpp server involves C# based functions that send requests with specific inference parameters using, such as max length, temperature, top-k, and top-p sampling. These parameters control the behavior of the LLM, affecting the randomness, diversity, and length of the generated responses. For instance, a lower temperature value makes the model's predictions more deterministic, while a higher value increases creativity and variability in the responses.

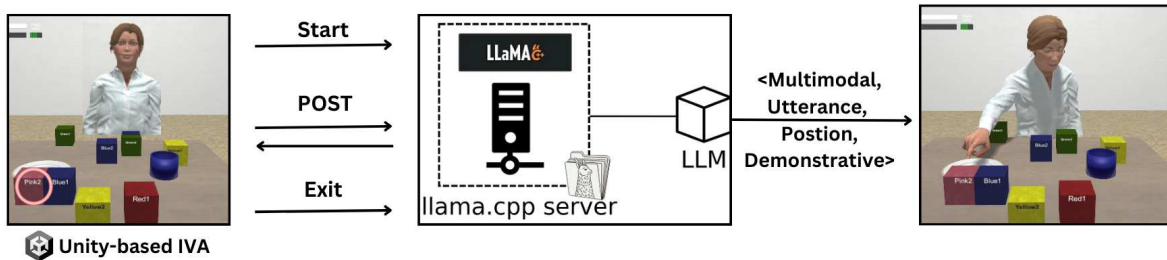


Figure 4.10: The performance of LLaMA models in predicting positions of target objects

Practically, the IVA can send a query to the server to generate a descriptive response. For example, when asked to describe an object, e.g., “RedBlock1”, on a table with specific configurations, the server respond with (Multimodal, “Grasp the red block”, (X'' , Y'' , Z'' >), “the”). The IVA, then, is able to use speech (generated via text-to-speech) to utter the generated speech and deictic gesture (generated via animated gesture) to refer to the “RedBlock1”, facilitating clear and effective communication.

Ensuring low latency is crucial for real-time applications, particularly in LLM-IVA integration. We applied several strategies, including optimizing server configurations and leveraging two GPU acceleration. GPUs handle parallel processing tasks required by LLMs more efficiently than CPUs, enabling the system to process more queries simultaneously. In addition, we used caching mechanisms reducing repetitive processing. Moreover, model optimizations, such as pruning and quantization, reducing computational demands, were considered. Nonetheless, the scarcity of GPU resources lead to delays in processing requests. Continuous improvement through regular updates and fine-tuning of the model based on user interactions can further enhance the IVA’s performance in real-time interaction.

4.5 Summary

In this chapter, we detailed the fine-tuning process of a LLM to generate contextually appropriate and situationally fluent referring expressions. The chapter covered various aspects, including data preparation techniques such as augmentation and synthesizing, the model architecture, and

the outcomes of the fine-tuning process. Additionally, we discussed the integration of the LLM with IVAs, highlighting the importance of optimizing advanced inference parameters and leveraging GPU acceleration to ensure low latency. The findings indicate a significant similarity between human-generated and LLM-generated MREs, as well as the correctness of generated positions, and a favorable perplexity score. These results suggest that deploying a generative LLM within interactive systems to integrate gestures and language is highly promising. This approach enhances task-based interactions, enabling more natural and effective human-computer collaboration.

The implication of fine-tuning an LLM, as opposed to using a deterministic approach for generating MREs, is significant. Fine-tuning allows for the incorporation of contextual distinctions, enabling the model to generate more natural and human-like expressions. This leads to more effective and engaging interactions between IVAs and users. In contrast, deterministic methods often lack the flexibility to handle the complexity and variability of real-world interactions, resulting in less fluid and more rigid communication. By leveraging the strengths of fine-tuned LLMs, we aimed to improve the overall clarity and naturalness of IVA-generated communication.

Chapter 5

Evaluation of MREG Models

5.1 Overview

With the aim of enabling bidirectional communication between machines and humans using multimodal referring expressions as a testbed use case, it is essential to enumerate specific evaluative properties using a set of metrics. These metrics help demonstrate the effectiveness of a fully-symmetrical system compared to one that maintains communicative asymmetry between the two interlocutors. The key research question with evaluation is: *What are the metrics that clearly establish whether both interlocutors are able to correctly and efficiently extract the communicative intents of the others from their behavior?* Therefore, good metrics will answer if the non-verbal behavior generation methods is effectively contributing to both interlocutor’s understanding, as defined as the ability to extract communicative intent from utterances and actions.

We consider key properties that are related to deictic and linguistic referencing, as used in the evaluation of human-machine collaboration [78], and propose quantitative and qualitative metrics that assess the following properties of multimodal RE usage in a task-based environment: 1) efficient collaborative task completion, 2) Effective understanding of contexts, 3) contribution of meaningful content and 4) human satisfaction of communication clarity and fluency. We used SCMRE corpus to evaluate multimodal communication strategies. Moreover, we utilized the LLM’s generated MREs to evaluate the success of the fine-tuned model’s generation. The results utilizing a subset of the proposed metrics have been published in GENE-2023, the Generation and Evaluation of Non-Verbal Behavior for Embodied Agents workshop [85], as well as in the 26th International Conference on Human-Computer Interaction (HCI International 2024).

5.2 Evaluation Metrics

To evaluate the success of the IVA w.r.t. the key characteristics of human-machine collaboration from Sec 5.1, we define 9 metrics as follows:

(1) Agent Completion Time of Human Prompts (ACTHP): Time from human generation of a referring expression, using multimodal REs, verbal only REs, or pointing only REs to task completion by the agent (See equation 5.1). ACTHP serves as the dependent variable, while C_{IVA} and G_{HUM} are regarded as independent variables. This metric can be utilized based on the categories of modalities used in the research. Therefore, we excluded these previously proposed metrics: Multimodal Prompt Completion Efficiency (MPCE) and Linguistic Prompt Completion Efficiency (LPCE). These metrics were intended for measuring the completion times when using multimodal REs vs. verbal only REs.

$$ACTHP = C_{IVA} - G_{HUM} \quad (5.1)$$

Where:

- C_{IVA} be the task completion time by the agent.
- G_{HUM} be the target generation time by the human.

(2) Human Interpretation Time of Machine Communication (HITMC): Time from generation of machine's communication (e.g, asking for disambiguation) to target identification by human (See equation 5.2). The HITMC is a dependent variable, whereas I_{HUM} and G_{IVA} are independent variables.

$$HITMC = I_{HUM} - G_{IVA} \quad (5.2)$$

Where:

- I_{HUM} be the target identification time by human.
- G_{IVA} be the communication generation time by the IVA.

The metrics #1 and #2 mentioned above are derived directly from data logged during human-agent interactions to assess the human-IVA's *efficiency* in interpreting, responding and completing

collaborative tasks. The IVA and humans work together in a way that makes the best use of time, effort, and resources to achieve the task effectively. The previously proposed metric, Human-Machine Completion Efficiency (HMCE), is omitted because these metrics #1 and #2 measure the efficiency for each prompt instead of the entire interaction. This enables researchers to investigate specific areas for improvement based on each prompt.

(3) Machine Interpretation of Human Communication (MIHC): Rate of correctly executed prompts (See equation 5.3). $N_{\text{correct_prompts}}$ and $N_{\text{total_prompts}}$ are independent variables, while MIHC is a dependent variable. This metric can be applied based on categories of modalities and speech types in the research. Thus, we excluded the proposed metrics: Machine Interpretation of Relational REs (MIRRE) and Machine Interpretation of Historical REs (MIHRE). The MIHC metric can then be generalized to different multimodal communication contexts, whether using relational and historical REs or not.

$$\text{MIHC} = \frac{N_{\text{correct_prompts}}}{N_{\text{total_prompts}}} \quad (5.3)$$

Where:

- $N_{\text{correct_prompts}}$ be the number of correctly executed prompts.
- $N_{\text{total_prompts}}$ be the total number of prompts given.

(4) Agent Pointing Success Rate (APSR): Rate of agent successfully pointing out the target object (See equation 5.4). The APSR is a dependent variable, while $P_{\text{correct_pointings}}$ and $N_{\text{total_prompts}}$ are independent variables.

$$\text{APSR} = \frac{P_{\text{correct_pointings}}}{N_{\text{total_prompts}}} \quad (5.4)$$

Where:

- $P_{\text{correct_pointings}}$ be the number of successful pointings used to identify the target object.
- $N_{\text{total_prompts}}$ be the total number of prompts given.

The metrics #3 and #4 mentioned above are derived directly from data logged during human-agent interactions to assess the IVA's *contextual awareness*. The IVA's capability of referencing and understanding contexts, including the linguistic and deictic context.

(5) Machine Multimodal Referencing Success (MMRS): Rate of machine-generated multimodal referring expressions compared to total number of generated REs (See equation 5.5). MMRS is a dependent variable whereas $N_{\text{mm_REs}}$ and $N_{\text{total_REs}}$ are independent variables. This metric is derived directly from data generated by the LLM using the testing data to assess the IVA’s *contribution with meaningful contents*, specifically the machine’s ability to communicate its intents to humans in a situation-appropriate manner to provide meaningful content. This metric can be used based on categories of speech types used in the research. This can be generalized to different multimodal communication contexts either using relational and historical or not. Therefore, we omitted these proposed metrics: Machine Relational Referencing Success (MRRS) and Machine Historical Referencing Success (MHRS). These metrics were intended for measuring the rates of generated spatial references and previous-action based references.

$$\text{MMRS} = \frac{N_{\text{mm_REs}}}{N_{\text{total_REs}}} \quad (5.5)$$

Where:

- $N_{\text{mm_REs}}$ be the number of machine-generated MREs.
- $N_{\text{total_REs}}$ be the total number of generated referring expressions.

The following metrics 6–9 are collected *post facto* from the judgments of 3rd-party evaluators (see Section 5.4) to measure the *clarity and fluency* of IVA-generated REs. While the clarity refers to how easily understandable the generated REs, fluency, pertains to the naturalness and smoothness of the generated REs. By involving human evaluators, we ensure that the IVA’s generation is not only automatically accurate but also align with human expectations. This is invaluable for refining the model to produce more human-like and contextually appropriate REs.

(6) Machine Object Identification Success Rate (MOISR): Rate of correctly identified objects (by machine), as shown in equation 5.6. While $M_{\text{correct_identifications}}$ and $M_{\text{total_referrings}}$ are independent variables, the MOISR is a dependent variable.

$$\text{MOISR} = \frac{M_{\text{correct_identifications}}}{M_{\text{total_referrings}}} \quad (5.6)$$

Where:

- $M_{\text{correct_identifications}}$ be the number of correctly identified objects.
- $M_{\text{total_referrings}}$ be the total number of machine REs given.

(7) Human Object Identification Success Rate (HOISR): Rate of correctly identified objects (by humans), as shown in equation 5.7. $H_{\text{correct_identifications}}$ and $H_{\text{total_referrings}}$ are independent variables, while the HOISR is a dependent variable.

$$\text{HOISR} = \frac{H_{\text{correct_identifications}}}{H_{\text{total_referrings}}} \quad (5.7)$$

Where:

- $H_{\text{correct_identifications}}$ be the number of correctly identified objects.
- $H_{\text{total_referrings}}$ be the total number of human REs given.

(8) Machine References Fluency Rate (MRFR): Rate of top-rated machine references according to 3rd-party human judgments, as shown in equation 5.8. MRFR is the dependent variable, while $M_{\text{top_rated_references}}$ and $M_{\text{total_referrings}}$ are independent variables.

$$\text{MRFR} = \frac{M_{\text{top_rated_references}}}{M_{\text{total_referrings}}} \quad (5.8)$$

Where:

- $M_{\text{top_rated_references}}$ be the number of top-rated machine REs.
- $M_{\text{total_referrings}}$ be the total number of machine REs given.

(9) Human References Fluency Rate (HRFR): Rate of the top-rated human references according to 3rd-party human judgments, as shown in equation 5.9. While the HRFR is a dependent variable, $H_{\text{top_rated_references}}$ and $H_{\text{total_referrings}}$ are independent variables.

$$\text{HRFR} = \frac{H_{\text{top_rated_references}}}{H_{\text{total_referrings}}} \quad (5.9)$$

Where:

- $H_{\text{top_rated_references}}$ be the number of top-rated human REs.
- $H_{\text{total_referrings}}$ be the total number of human REs given.

Some other metrics, such as Proceed Without Reset (PWR), were proposed to evaluate reliability, but we excluded it. Although PWR is important, it is not related directly to multimodal interaction and understanding; it is intended for systems usability and reliability. Additionally, we could not use this metric as some participants had memory issues affecting the app’s execution, which were not related to the app itself. Moreover, we omitted Machine Appropriate Response Success Rate (MARSR), measuring the rate of IVA responses to human prompts that are not followed by a negative response. However, we observed that it is not necessarily indicative of an inappropriate response if a human uses negative words like “nevermind.” They sometimes simply be changing their target, which means the IVA correctly understood the intent. The proposed metric, Mutual Contribution Success Rate (MCSR), is also omitted. Due to the lack of available GPUs, we were unable to evaluate mutual interaction, and thus, this metric was not tested or included in our analysis.

These 9 remaining evaluation metrics are rigorously utilized to assess and analyze the IVA’s interpretation and generation capabilities during human interaction, both quantitatively and qualitatively, as presented in subsequent sections.

5.3 Automated Quantitative Evaluation

Using the quantitative metrics from 1-5 with Human-IVA logged data, presented in Section 3.3, and LLM-generated data in Section 4.3.3, we obtained results regarding the ease of IVA disambiguation, human recognition efficiency of IVA intent from verbal and non-verbal behavior, the IVA’s abilities in completing different human referential prompts, and IVA’s ability to generate multimodal communication.

5.3.1 Contextual Awareness

In this section, we evaluate the IVA’s ability to reference and understand contexts, involving both linguistic and deictic contexts, to extract the meaning of human intents using Agent Pointing Success Rate (APSR) and Machine Interpretation of Human Communication (MIHC). In Fig-

```

[2023-06-07-11-51-05] -----Pointing to the FOCUS without Speech-----
[2023-06-07-11-51-05] user:intent: object _ focus | BlueBlock1
[2023-06-07-11-51-05] Focus object position | (0.2, 1.1, -0.3)
[2023-06-07-11-51-05] Distance from agent to focus obj:1.188692
[2023-06-07-11-51-05] Diana pointed | BlueBlock1
[2023-06-07-11-51-06] -----Pointing to the TARGET without Speech -----
[2023-06-07-11-51-06] target object | YellowBlock1
[2023-06-07-11-51-06] target object position | (0.0, 1.1, -0.2)
[2023-06-07-11-51-06] Distance from agent to focus obj:1.188692
[2023-06-07-11-51-06] user:intent:event | put(BlueBlock1,behind(YellowBlock1))
[2023-06-07-11-51-06] agent executed | put(BlueBlock1,behind(YellowBlock1))
[2023-06-07-11-51-06] -----
[2023-06-07-11-51-06] Relations: under | Table and PinkBlock1 + under | Table and BlueBlock2 ...
[2023-06-07-11-51-06] Configurations: RedBlock1: (0.583337500: 1.124870000: 0.429246700) ...
[2023-06-07-11-51-06] -----
[2023-06-07-11-51-07] user speech | pick up the Yellow Block
[2023-06-07-11-51-07] Parsed speech | grasp(the(yellow(block)))
[2023-06-07-11-51-07] Diana | Which Yellow Block?
[2023-06-07-11-51-08] -----Pointing to the FOCUS After Disambiguation-----
[2023-06-07-11-51-08] user:intent: object _ focus | YellowBlock2
[2023-06-07-11-51-08] Focus object position | (0.5, 1.1, 0.1)
[2023-06-07-11-51-08] Distance from agent to focus obj:1.447748
[2023-06-07-11-51-09] Diana executes | grasp(the(yellow(block)))

```

(a)

```

[2023-06-07-13-38-16] User speech | Grasp this block
[2023-06-07-13-38-16] Parsed speech | Grasp(this(block))
[2023-06-07-13-38-17] Diana | OK!
[2023-06-07-13-38-17] User speech | No, this block
[2023-06-07-13-38-18] Diana | OK!
[2023-06-07-13-38-39] User speech | Move the green block to the left of the plate
[2023-06-07-13-38-39] Parsed speech | slide(the(green(block),left(the(plate))))
[2023-06-07-13-38-39] User speech | nevermind
[2023-06-07-13-38-40] Diana | OK! Nevermind.

```

(b)

Figure 5.1: (a) Trial sample of the IVA’s ability to disambiguate the target; (b) Trial sample of human’s ability to correct misunderstanding.

Figure 5.1a, interlocutors’ moves, including actions, speech, and gestures, are logged with their timestamps. We see that the human started pointing to the focus object (*BlueBlock1*) and moving it behind *YellowBlock1*. Logs also include the positions of each, distance from agent to each, and the agent’s action after pointing to each of the two blocks. The human then used language only (“Pick up the yellow block”) to instruct Diana to pick up *YellowBlock2*. This instruction required Diana ask for disambiguation: “Which yellow block?”, as there are two in the scene. To disambiguate, the human uses pointing, and the object, its position, and distance are logged along with Diana’s action. Additionally, when Diana has a misunderstanding, the human can correct it using language, deictic gesture, or both (Figure 5.1b). Diana confirms that disambiguation was successful using deictic gesture to the correct object, resulting in bidirectional communication where both human and agent use verbal and non-verbal behavior.

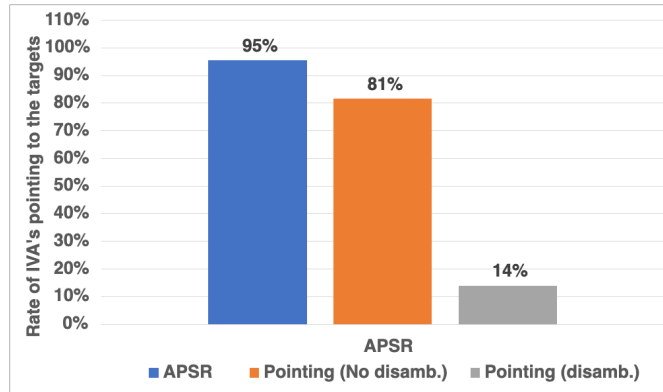


Figure 5.2: Agent Pointing Success Rate (Metric #4: APSR) includes successful identifications of target objects both without requiring clarification and after requesting and receiving clarification.

This demonstrates Diana’s ability to accurately disambiguate the object referred to by the human and to reference the human’s target, showcasing her comprehension of the prompted object as shown in 5.2a. Using APSR, we found that Diana successfully identified and pointed to 95% of the target objects when she extracted the correct intent. Of these, 81% of the prompts were clear without requiring additional clarification, while 14% were ambiguous, prompting her to ask for more information. These findings confirms the significant relationship between multimodal responding, including pointing and speech, and contextual understanding (p -value < 0.001 using Pearson’s χ^2 -test) as illustrated in Table 5.1. To quantitatively measure the IVA’s situational un-

Table 5.1: Contingency table of the IVA’s pointings to human targets and her contextual understanding: # of contextual understanding instances based on IVA’s pointings to the targets.

IVA’s Pointings	Did Diana Understand?	
	No	Yes
No	453	0
Yes	0	9,268
P-value	< 2.2e-16	

derstanding capabilities, we used MIHC. As illustrated in Figure 5.3a, while the IVA successfully interpreted 5.18% of multimodal REs, 73.80% pointing-only expressions, and 16.67% of speech-only expressions, 4.36% of multimodal and speech-only expressions were not correctly interpreted by the IVA. In terms of speech type, she correctly interpreted 11.45% of attributive REs, 3.55% of

relation REs, 0.26% of historical REs and 6.95% of non-referencing expressions as shown in Figure 5.3b. Overall, the IVA successfully interpreted 95% of human communications, while 5% were not correctly interpreted. By maintaining a high success rate in both APSR and MIHC metrics, the IVA demonstrates robust contextual understanding and responsiveness, crucial for efficient and effective human-agent interactions.

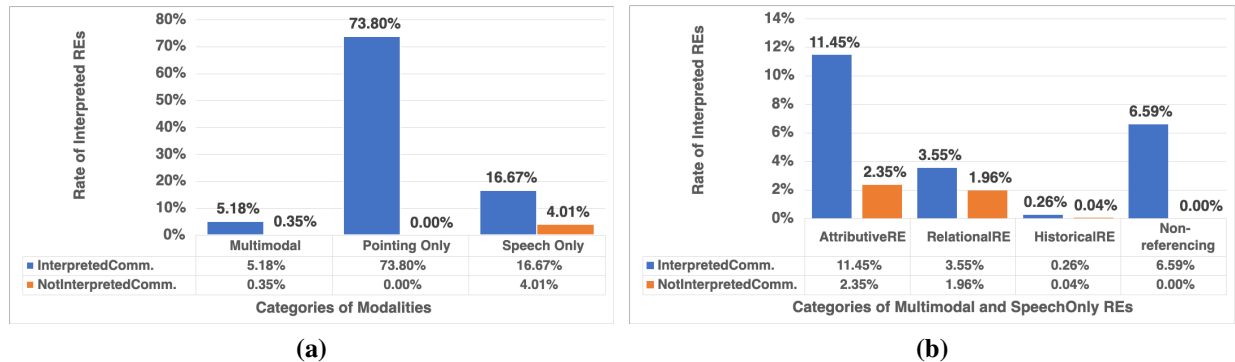


Figure 5.3: Machine Interpretation of Human Communication (Metric #3: MIHC) categorized by (a) utilized modalities; and (b) speech type.

5.3.2 Efficiency of Collaborative Task Completion

In this section, we evaluate the IVA-humans efficiency in completing tasks using Agent Completion Time of Human Prompts (ACTHP) and Human Interpretation Time of Machine Communication (HIEMC). Based on ACTHP results illustrated in Figure 5.4a, b and c, we observed that the IVA’s efficiency of completing prompts increases when humans use pointing or multimodal references to communicate their intents. Specifically, Diana spent approximately 3 seconds or more to understand the meaning of humans’ intents and respond to 46.14% of speech-only prompts, as shown in Figure 5.4b. Nonetheless, giving deictic and multimodal communication, Diana demonstrates increased efficiency by spending 3 seconds or more on 20.03% of multimodal REs and on 24.13% of pointing-only REs as illustrated in Figure 5.4a and c. We conclude that the use of pointing significantly increases the efficiency of IVA’s understanding of the meaning of humans’ intents and completing the task (p -value $< 2.2e - 16$ using ANOVA test [110]). On the other

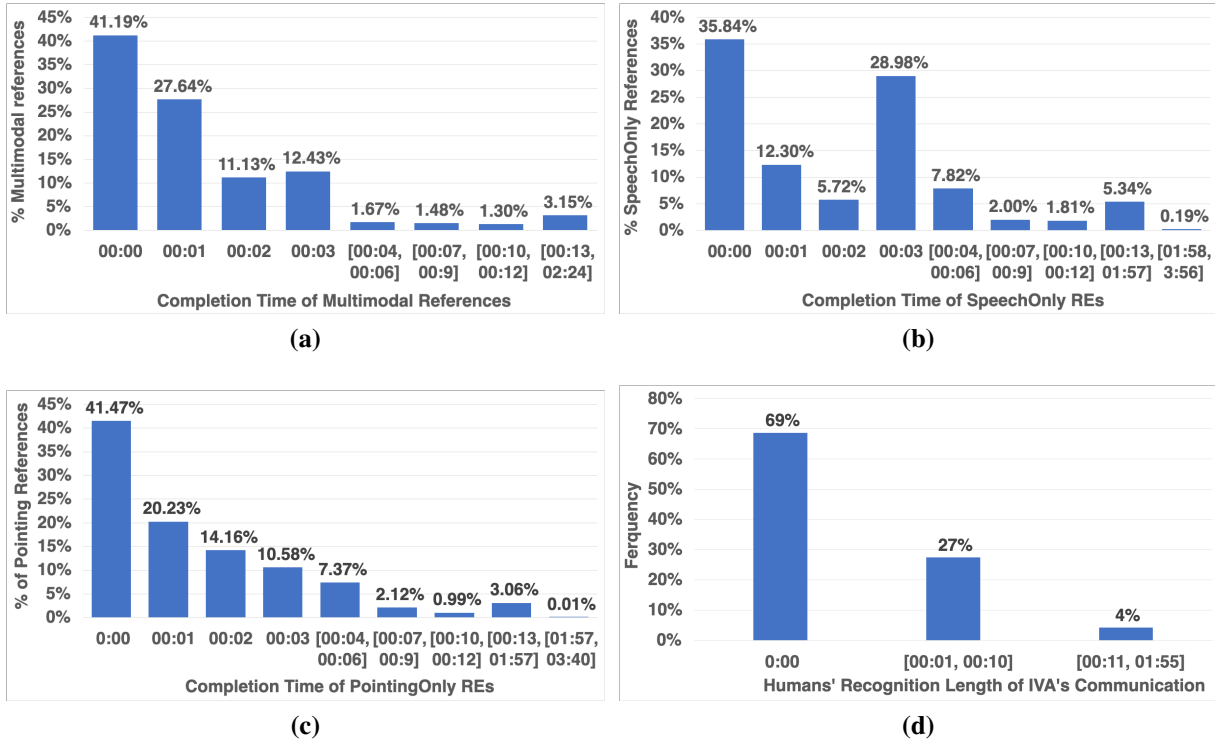


Figure 5.4: Agent Completion Time of Human Prompts (Metric #1: ACTHP) categorized by (a) Completion time of linguistic prompts; (b) Completion time of pointing-only prompts; and (c) Completion time of multimodal prompts; and (d) Human Interpretation Time of Machine Communication (Metric #2: HITMC)

hand, using HITMC, we noticed distinct response times by humans to the IVA's communication (see Figure 5.4d). While 69% of the human responses were immediate, taking 0 seconds or less, 31% of the human communications required between 1 second and approximately 2 minutes to recognize and respond to Diana's communication.

For a set of prompts, the humans' recognition time, as shown in Figure 5.4d, and Diana's completion time, as depicted in Figure 5.4a, b, and c, exceed 3 seconds and can reach up to approximately 2 minutes. This delay is due to slow human responses to Diana's disambiguation questions, speech recognition issues, and failures in parsing speech to a PLF, as observed in human recordings of interactions. Addressing these challenges and maintaining lower ACEHP and HIEMC result in efficient collaborative interactions between IVAs and humans.

5.3.3 Contribution with Meaningful Content

Based on the significant reduction in ambiguity levels observed when humans used co-gestural REs while interacting with the IVA (see Section 3.4), we evaluate whether LLM-generated references maintain this level of clarity using Machine Multimodal Referencing Success (MMRS). As illustrated by Figure 5.5, the LLM generated 40% of multimodal REs including 33.1% attributive REs and 7.5% relational REs, 45.5% of pointing-only REs, and 13.9% speech-only REs, including only relational REs. Nonetheless, the LLM generated none of the historical REs because of

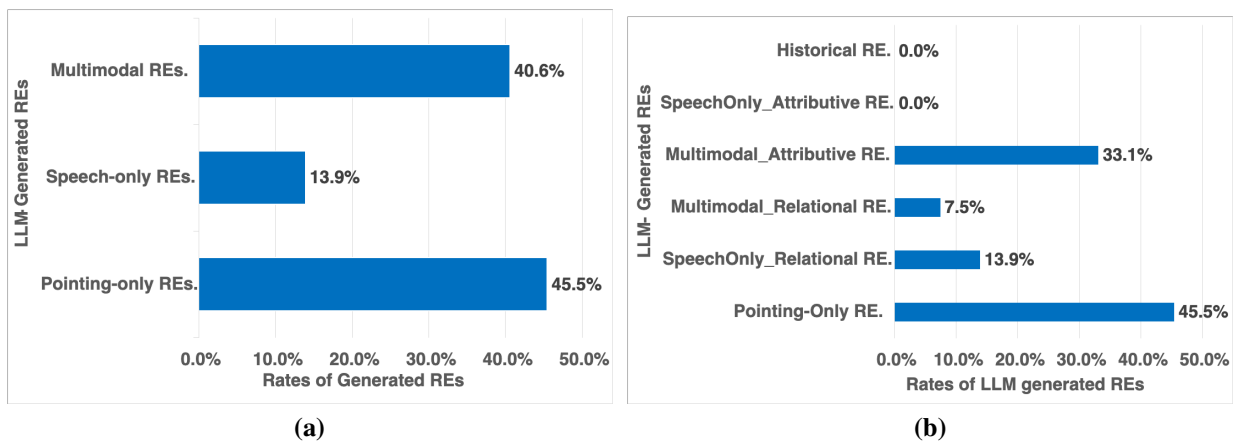


Figure 5.5: Machine Multimodal Referencing Success (Metric #5: MMRS) categorized by (a) generation modalities only, and (b) speech type and generation modalities.

the lack of prior-actions based data to train the LLM on. Similarly, the LLM generated 0% of the speech-only attributive REs. However, this is considered an advantage of the LLM, as this category often requires interlocutors to disambiguate as illustrated in the contingency table in Section 3.4, which leads to reduced efficiency of task completion. These results show the capacity of a generative model to contextually integrate gestures and language. Higher MMRS thereby enhances task-based interactions and facilitates more natural and bidirectional human-IVA communication.

5.4 Human Based Evaluation

Alongside the automatic quantitative evaluations, we conducted two human-based experiments on Amazon Mechanical Turk (AMT) to assess the fluency and clarity of IVA and human-generated MREs. We proposed two criteria for evaluating the generated MREs: 1) A qualitative comparison of IVA with human-generated MREs, using MRFR, the rate of top-rated machine references based on third-party human judgments, and HRFR, the rate of top-rated human references based on third-party human judgments, through preference ordering, 2) quantitative comparison of IVA with human-generated MREs, using MOISR, the rate of correctly identified objects (by machine), and HOISR, the rate of correctly identified objects (by humans). Evaluation data and results are publicly available on GitHub¹².

5.4.1 Study Design

As illustrated in Figure 5.6, we selected 50 human MREs from the SCMRE dataset. These were compared with 50 REs generated by the IVA in the same situation when driven by a generative model trained over the human data. A total of 100 videos were collected. The referencing strategies examined for each of human and IVA generation are pointing only REs, relational speech-only REs, attributive speech-only REs, relational multimodal REs and attributive multimodal REs as shown in Table 5.2.

Videos were used in a set of AMT human intelligence tasks (HITs), each involving workers rating 1 video for *both* fluency and clarity, including 1 machine generated RE or 1 human REs, for a total of 100 HITs. As shown in Figure 5.7, workers first identified the target object mentioned, then, they rate the fluency of each video description on a Likert scale (5 = best, 1 = worst). Each video was completed by 10 workers, for a total of 2,000 individual judgments for fluency and clarity. We recruited workers fluent in English between 18 and 60 years old. They were given 1 hour per task and were compensated \$0.75 for each HIT.

¹²<https://github.com/nadahass/Human-based-Evaluation-of-EMREG>

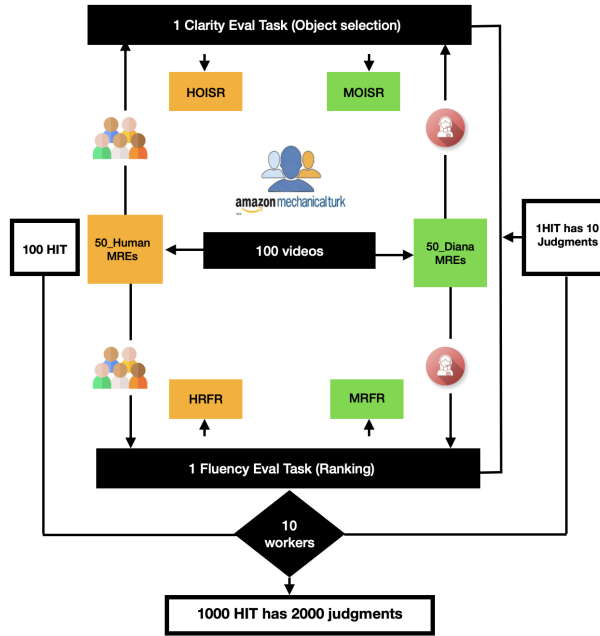


Figure 5.6: Crowdsourcing framework, including human judgments and related metrics, Metric # 6: MOISR, Metric # 7: HOISR, Metric # 8: MRFR, Metric # 9: HRFR, for evaluating MRE generation models.

Table 5.2: Quantities of testing videos and judgements for referring expressions with distinct modalities and speech types

Human-Generated REs	# Testing REs	# Judgements	IVA-Generated REs	# Testing REs	# Judgements
Multimodal_Attributive	21	210	Multimodal_Attributive	33	330
Multimodal_Relational	14	140	Multimodal_Relational	6	60
Speech_Attributive	3	30	Speech_Attributive	0	0
Speech_Relational	9	90	Speech_Relational	7	70
Pointing	3	30	Pointing	4	40
Total	50	500	Total	50	500

5.4.2 Results and Analysis

Upon analyzing 1,000 assignments, it was found that 300 were rejected for not following instructions or attempting to game the system, and were subsequently republished. Workers, with an average lifetime approval rate of 100%, invested approximately 30 minutes on average to complete the tasks, indicating thorough engagement.

The rates of identifying objects referred to by humans and IVA were compared to the intended objects in the dataset. As shown in Figure 5.8a, the overall HOISR and MOISR were 82.6% and 89.4%, respectively, demonstrating that the clarity level of IVA-generated MREs strongly competes

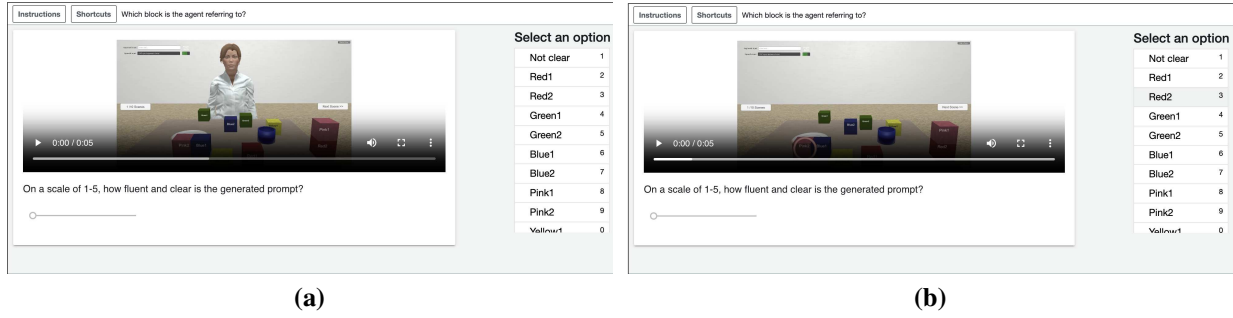


Figure 5.7: Each HIT includes two tasks for quantitative and qualitative evaluation of (a) IVA REs and (b) human REs.

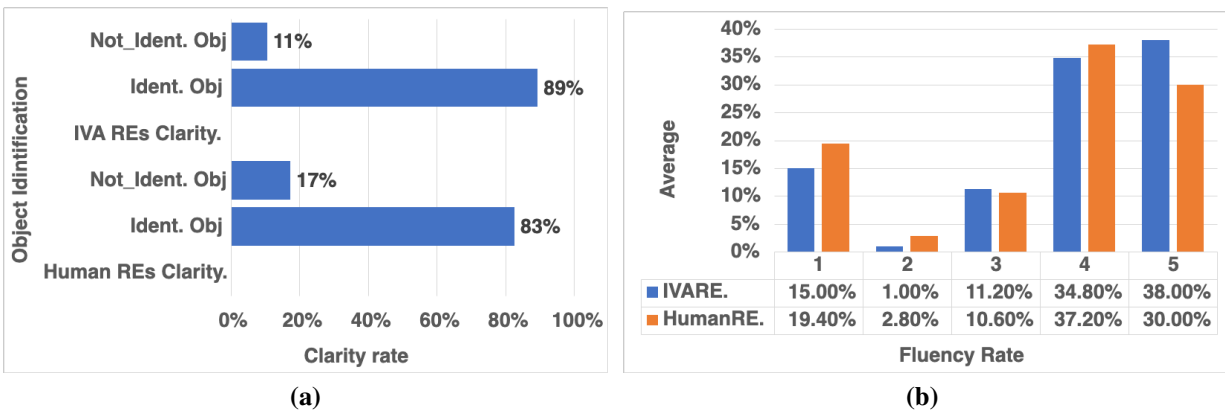


Figure 5.8: (a) Machine Object Identification Success Rate (MOISR) and Human Object Identification Success Rate (HOISR), and (b) Machine References Fluency Rate (MRFR) and Human References Fluency Rate (HRFR)

with human-generated MREs (p -value $< 2.2e - 16$ using Pearson’s χ^2 -test). For the fluency task, Figure 5.8b shows MRFR of 73% and HRFR of 67% , with ratings at “4” and “5”. The fluency between human and LLM-generated REs demonstrates no significant difference, as indicated by a p -value of 0.5529 (p -value = 0.5529 using Pearson’s χ^2 -test). This indicates that the generative LLM is capable of producing REs of comparable quality to those generated by humans.

In terms of utilized modalities, the findings in Table 5.3 confirm that integrating multiple modalities by the LLM significantly improves the fluency of REs. Pointing-only REs and speech-only REs received 55% and 31.43% of “1”, “2” and “3” ratings, respectively, whereas multimodal REs received fewer low ratings. Higher ratings of “4” and “5” were received for multimodal relational REs and multimodal attributive REs, with fluency rates of 78.33% and 79.58%, respectively.

The p -value of 0.00405, using Pearson’s χ^2 -test, indicates the importance of employing mixed modalities to achieve more natural, fluent and effective interactions.

Table 5.3: Contingency table of IVA’s RE fluency and modalities used; Rel: Relational REs, Att: Attributive REs: fluency level by modality type

Fluency	Multimodal-Rel	Multimodal-Att	Pointing	SpeechOnly-Rel
1	13.33%	13.64%	22.50%	18.57%
2	0.00%	0.61%	2.50%	2.86%
3	8.33%	9.70%	30.00%	10.00%
4	33.33%	38.67%	27.50%	38.57%
5	45.00%	40.91%	17.50%	30.00%
P-value		0.00405		

Table 5.4: Contingency table of IVA’s RE clarity and modalities used; Rel: Relational REs, Att: Attributive REs: clarity level by modality type

Clear?	Multimodal-Rel	Multimodal-Att	Pointing	SpeechOnly-Rel
No	15%	7%	0	30%
Yes	85%	93%	100%	70%
P-value		2.373e-08		

Furthermore, the findings (see Table 5.4) indicate that incorporating pointing gestures significantly enhances the clarity of referring expressions compared to using speech alone (p -value: $2.373 \times 10^{-8} < 0.001$ using Pearson’s χ^2 -test). Pointing provides a direct visual cue that guides the listener’s attention to the specific object being referred to. While speech conveys detailed descriptive information, pointing gestures provide spatial context. This reduces ambiguity, especially in environments where multiple objects could match the spoken description. These are obvious with 85%, 93% and 100% correctly located objects described by multimodal relational REs, multimodal attributive REs and pointing-only REs, respectively. This has significant implications for the design of IVAs, suggesting they should leverage multimodal communication to enhance overall user satisfaction and task performance.

5.5 Correlation Analysis of Evaluation Metrics

This section analyzes the correlation between key properties of collaborative interaction and their associated metrics. We explore how these metrics relate to their properties, providing comprehensive insights into the performance and interaction dynamics of IVAs.

Efficiency of Collaborative Task Completion. Human and agents collaborate in the way that consumes less amount of time to complete the collaborative tasks. Extracting the intents is the major aspect in this research that could consumes time to continue the collaboration and successfully accomplish the tasks. Given the significance of completion time in measuring system efficiency [111], we proposed the *Agent Completion Time of Human Prompts* (ACTHP) and *Human Interpretation Time of Machine Communication* (HITMC) to assess how efficiently the IVA interpret and executes prompts, as well as how humans interpret and respond to the IVA’s instructions. Low ACTHP values indicate effective understanding and execution of human commands, which is essential for collaborative task completion. This metric assesses the IVA’s versatility in quickly interpreting various forms of human communication, whether linguistic, pointing-only, or multimodal as shown in Figure 5.3. Additionally, ACTHP indicate the existance of execution bottlenecks and areas for improvement, leading to faster and more efficient task completion.

Similarly, HITMC metric, which measures the time spent to interpret and respond to the agent, indicates how efficiently the IVA’s messages and prompts are understood by humans. In human interactions, it it argued that the low response latency associates with the high quality of perceived communication [112]. Therefore, the low HITMC suggests that the IVA’s communication is clear and easily interpreted by humans, which is essential for seamless collaboration. Efficient human responses are critical for maintaining the flow of collaborative tasks and ensuring timely task completion. Furthermore, the difference in task completion time using pointing-only, speech-only REs, and multimodal REs, detailed in Section 5.3.2, highlights the significant relationship between understanding efficiency and ACTHP and HITMC metrics (p -value < 0.001 , using ANOVA test).

Low values in both ACTHP and HITMC metrics indicate well-balanced and efficient extraction of intents meaning between the IVA and humans, resulting in more effective task completion.

Contextual Awareness. Being contextually aware is the ability to reference and understand contexts. “Understanding” is defined as the capability of machines to extract the “meaning” of intent, communicate it or reproduce it [1]. Thus, we proposed *Machine Interpretation of Human Communication* (MIHC) and *Agent Pointing Success Rate* (APSR) metrics to evaluate the IVA’s contextual understanding.

MIHC measures the IVA’s ability to accurately interpret various forms of human communication, including linguistic inputs, gestures, and multimodal prompts. This ensures that the agent comprehends instructions correctly, leading to fewer errors and misunderstandings. As discussed in Section 3.4, there exists a notable difference in the IVA’s interpretation of human REs across various modalities and speech types. This variance underscores the correlation between the depth of understanding and the contextual information conveyed to the agent (p -value < 0.001 , using Pearson’s χ^2 -test). The IVA demonstrates proficiency in interpreting contextual REs, such as multimodal and some of relational REs, without requiring clarification. In contrast, the interpretation of linguistic REs alone sometimes leads to misunderstandings for the IVA. APSR, on the other hand, evaluates the accuracy with which the IVA points to the target objects specified by humans. A strong relation between agent pointings to the targets and her understanding (p -value < 0.001 , using Pearson’s χ^2 -test) as shown in Section 5.3.1. A high success rate in this metric demonstrates the agent’s ability to respond appropriately to deictic gestures and other references, ensuring correct understanding and effective communication.

Both MIHC and APSR metrics together reflect the IVA’s understanding level of the situational context. This includes recognizing the relevance of objects and actions within the given environment, which is critical for performing tasks that require contextual sensitivity. Analyzing the MIHC and APSR metrics can help identify specific areas where the IVA may struggle with context interpretation or response accuracy. Maintaining high success rates in both APSR and MIHC metrics demonstrates the IVA’s robust contextual understanding and responsiveness, which are crucial

for efficient and effective human-agent interactions. This leads to improved task performance, user satisfaction, and overall system reliability.

Contribution with Meaningful Content. According to our results in Section 3.4, multimodal REs demonstrated a strong impact in providing the necessary information to understand the meaning of intents and reduce the ambiguity of references (p -value < 0.001 , using Pearson’s χ^2 -test). Therefore, we proposed the *Machine Multimodal Referencing Success* (MMRS) metric to assess how effectively the LLM integrates multiple modalities to convey meaningful information. By assessing the IVA’s success in multimodal referencing, MMRS ensures that the content generated by the LLM is not only accurate, but also meaningful and helpful in the given context. High MMRS, as illustrated in Section 5.3.3, indicates that the LLM can effectively combine different communication forms (verbal and non-verbal) to deliver coherent and comprehensive references. This ability enable human interlocutors to effectively extract the intended meaning, leading to enhance the overall quality of interaction. Additionally, as MMRS quantifies the generated communication types, it aids continuous improvement of the LLM’s performance in conveying meaningful content. It also allows for the comparison of different versions of LLMs, ensuring that advancements in multimodal referencing are effectively captured and evaluated.

Clarity and Fluency. As human satisfaction is essential to evaluate the success of human-IVA collaboration [78], we suggested the *Machine Object Identification Success Rate* (MOISR), and *Human Object Identification Success Rate* (HOISR) to evaluate the effectiveness of object identification by both machines and humans. Moreover, we proposed the *Machine References Fluency Rate* (MRFR) and *Human References Fluency Rate* (HRFR) to assess the fluency and naturalness of the interactions between the IVA and humans.

MOISR measures how successfully humans identifies targets based on machine-generated REs, reflecting the system’s ability to correctly generate REs. HOISR, however, evaluates the human’s ability to correctly identify objects described by humans when collaborating with the IVA. These metrics indicate the clarity and effectiveness of the IVA’s communication and whether it competes human communication. As indicated in Section 5.4.2, there exists a notable distinction between

REs generated by humans and those generated by the IVA (p -value $< 2.2e - 16$ using χ^2 -test). Humans demonstrate a higher accuracy in identifying objects referenced by IVA-generated descriptions compared to human-generated ones. Similarly, MRFR evaluate the human’s satisfaction on the LLM’s proficiency in generating smooth, coherent, and contextually appropriate references, while HRFR measures the human’s satisfaction on human proficiency in generating natural REs. The fluency between human-generated and LLM-generated referring expressions shows no significant difference, as indicated by a p -value of 0.5529 using the χ^2 -test as outlined in Section 5.4.2. Together, these metrics provide comprehensive insights into the fluency and naturalness of IVA communication, highlighting areas for improvement in machine performance to meet human expectations. Moreover, these metrics are utilized to analyze human-IVA collaboration in real-time, advancing both machine performance and human interaction REs.

5.6 Summary

As interactive agents become more widespread in everyday use, developers need principled ways of evaluating their behavior. Modern generative large language models require new evaluation methods beyond traditional metrics like accuracy, precision, and recall on benchmark datasets. Factors such as fluency, reliability, correctability, and ease of use must also be considered, especially when non-linguistic modalities are involved, as with embodied IVAs. In this chapter, we proposed a quantitative and qualitative evaluation framework to assess the quality of generated multimodal referring expressions, including language, gesture, and actions grounded in a shared virtual environment.

Our approach involved developing a set of metrics for evaluating the quality of REs, applying to both human and IVA-generated expressions using combined verbal and non-verbal information. These metrics focus on the fluency and effectiveness of the communication in achieving its intended purpose. By examining the detailed results obtained from users interacting with the experimental platform, we were able to analyze the IVA’s references based on our proposed metrics. This analysis demonstrated the utility of these metrics in evaluating the contribution of non-verbal

information towards the effectiveness and efficiency of bidirectional interpretation and generation of REs.

The evaluation framework targets the understanding, generation, efficiency and fluency of referring expressions, proposing a set of qualitative and quantitative metrics that we believe will generalize to other situated tasks requiring the use of MREs. Such tasks include robot instruction, situated collaborative problem-solving, and other contexts where precise and efficient communication is crucial (Bradford et al., 2023).

Findings of these metrics highlight the importance of multimodal communication in enhancing the effectiveness of interactive agents. By integrating gestures, language, and context-aware actions, IVAs can achieve more natural and effective interactions with human users. This research provides a foundation for future work in the field of IVA and multimodal interaction, offering a robust framework for the assessment of dialogue and behavior generation strategies. We hope it will be beneficial for researchers to evaluate and improve the performance of IVAs in a variety of applications, ultimately leading to more intuitive and user-friendly IVAs.

Chapter 6

Conclusion

In this thesis, we confirmed the potential of integrating embodied interactive virtual agents into natural language technology, offering a pathway to systems that surpass mere language processing and approach genuine language understanding. Our focus on embodied systems is driven by their ability to simulate interactions related to those with another person. By enabling these systems to engage in interactions that imply an understanding of their conversational partner’s intents, we can make significant strides towards achieving true computational natural language understanding. Our experiments, centered on multimodal referring expression generation, serve as a representative use case. We established an IVA testbed, gathered human-generated multimodal referring expressions, fine-tuned an intelligent model for generating MREs, and developed an evaluation framework to assess multimodal behavior generation and comprehension. Through these efforts, we advance the capabilities of embodied interactive virtual agents and pave the way for more sophisticated language understanding systems.

6.1 Thesis Contributions

Here are the novel contributions of this thesis:

- The development of an embodied interactive virtual agent to elicit MREs from humans, involving the evaluation or implementation of five key components. Firstly, the "GSR" was evaluated to ensure it recognizes collaborative task speech across diverse patterns and accents. Second, the “ParsingToPLF” Algorithm was devised to compose constituents in a format that differentiates between attributive, relational, and historical descriptions. Thirdly, the “RelationalAlgorithm” leveraged previously tracked relations to extract the object associated with the received spatial description. Fourthly, the “HistoricalAlgorithm” was introduced to access the dialogue history, storing all previously executed events and identify the target ob-

jects. Lastly, the “AttributiveREAlgorithm” was designed to interpret object attributes such as color or demonstratives. Each of these components, whether dealing with attributive, relational, historical, or composed REs, may be aligned with deictic gestures, albeit in distinct manners.

- Collecting a situated corpus of multimodal referring expressions (SCMRE) collected from human-IVA interactions. We established bidirectional interactions between humans and IVAs using verbal and non-verbal behavior synthesis. The IVA was adapted to be presented to humans with a sequence of 10 scenes each includes randomly generated objects. They were asked to refer to objects using deictic gesture and speech.
- Implementing an AI model that contextually generates multimodal referring expressions in real time interactions. We trained a LLaMA2 model on SCMRE to capture language, gesture, situational information, and dialogue history. This enabled us to generate situational expressions using both speech and gesture.
- Evaluation metrics that apply to both humans and IVAs, combining qualitative and quantitative metrics. Using our the developed IVA, we analyzed IVA-system outputs and user interactions based on these metrics, demonstrating their utility in assessing the contribution of non-verbal information to bidirectional interpretation and disambiguation of object descriptions in context.

6.2 Broader Implications

In this section, we aim to contextualize our work within its impact on the research field of collaborative tasks requiring the use of MREs.

This study showcases significant advancements in multimodal human-IVA capabilities, particularly in bridging the gap between human and IVA communication in generating referring expressions. As the development of the PLF transformer, SCMRE corpus, fine-tuned generative model, and comprehensive evaluation framework have significantly enhanced the naturalness and effec-

tiveness of interactions between humans and IVAs, they are poised to aid researchers in the IVA and multimodal interaction fields to enhance and evaluate dialogue and behavior generation strategies. We expect the PLF parser and evaluation metrics will generalize to tasks situated in contexts that necessitate the utilization of MREs, such as robot instruction or situated collaborative problem solving. Moreover, SCMRE corpus serves as a benchmark for assessing or building different algorithms and systems in multimodal communication. The model also can be seamlessly integrated into IVAs for multimodal communication or further fine-tuned by researchers or developers to adapt to specific tasks or datasets, enhancing its performance and applicability in various contexts.

This evolution of cutting-edge multimodal communication capabilities, which understand and respond to human intentions with unprecedented clarity and fluency, marks a significant step forward, making interactions with IVAs more intuitive and effective in the domain of MREs. The findings of this study suggest that deploying a generative LLM within IVA systems to clearly and fluently integrate multiple modalities enhance real-time communication of intentions. This results in growing up situated task-based interaction between humans and machines by increasing the capabilities of interactive systems to collaborate with humans, such as in shared-task contexts, leading to more natural, fluent, bidirectional human-IVA interactions. This encourages the utilization of interactive systems in collaborative problem-solving, personal assistance, and educational contexts, as they proficiently interpret and generate detailed multimodal descriptions of objects within their environments.

6.3 Future Research Directions

We identified five fundamental challenges in multimodal language comprehension and generation research. Nevertheless, future research could investigate additional aspects to further progress the field. Here, we provide an overview of potential ideas for advancing multimodal IVA-human communication. It is important to acknowledge the limitations of our study, including the scope of the challenges identified and the methods used, which may not cover all relevant factors. Address-

ing these limitations in future research could offer a more comprehensive understanding and lead to significant advancements in the field.

Human-Agent Collaboration. Studying how IVAs can collaborate with humans in complex tasks is another important research direction. This involves designing interaction protocols that rely on understanding and anticipating human needs and intents. The IVA can be developed to take initiative and suggest actions based on its understanding of the context and user behavior, thereby improving the efficiency and effectiveness of human-agent collaboration. Furthermore, expanding the diversity of multimodal communication events (e.g., historical REs) in the SCMRE corpus will enhance researchers’ understanding of human referring expressions and broaden AI’s capacity to learn a wider range of human-like interactions. Improved interpretation and generation of multimodal behaviors will enhance interaction fluency, clarity, and usability, facilitating more efficient and bidirectional communication between humans and IVAs.

Improving Contextual Understanding. Enhancing the IVA’s ability to understand contextual information is still crucial. Developing advanced NLP models that handle longer contexts and incorporate multimodal information more effectively. Integrating recurrent neural networks (RNNs) can maintain and utilize long-term dialogue history, ensuring that the agent can provide more coherent and contextually appropriate responses throughout an interaction. Moreover, training a neurosymbolic model on a collaborative multimodal corpus parsed to PLFs can bridge the current parser’s limitations in parsing prompts beyond the implemented syntax. For example, the prompt “move the blue block in the corner to the pink block” results in an invalid PLF form if “in the corner” is not correctly parsed. This approach can enhance the parser’s ability to handle more complex syntactical structures.

Personalization and Adaptation. By incorporating personalization and adaptation techniques into MREG, communication systems can generate descriptions that are better suited to individual users’ needs. For example, a system may adapt the language, tone, or complexity of REs based

on the user's language proficiency, or interaction history. This can be implemented by incorporating machine learning techniques that allow the IVA to learn from user interactions over time. Reinforcement learning algorithms, for example, can help the IVA adapt its responses based on feedback from the user.

Cross-Domain Applications. Applying MREG in various domains beyond simple task-based interactions is a promising direction for future research. This involves adapting the IVA to specific domains such as education, healthcare, and customer service. Domain-specific training data can help the IVA understand and respond appropriately to different types of queries and tasks. Moreover, this will help to identify challenges unique to each domain and guide further improvements.

Multilingual and Multicultural Capabilities. Extending research to include multilingual and multicultural aspects is crucial for ensuring that IVAs can operate effectively across different languages and cultural contexts. MREG can benefit from multilingual and multicultural capabilities by incorporating linguistic and cultural diversity into the generation process. For example, a system that generates multimodal referring expressions may need to consider linguistic variations and cultural differences when selecting language and gesture combinations.

Bibliography

- [1] Emily M Bender and Alexander Koller. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, 2020.
- [2] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.
- [3] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [5] Chu-Cheng Lin, Waleed Ammar, Chris Dyer, and Lori Levin. Unsupervised POS induction with word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1311–1316, Denver, Colorado, May–June 2015. Association for Computational Linguistics.
- [6] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

- [7] Roghayeh Barmaki and Charles E Hughes. Embodiment analytics of practicing teachers in a virtual immersive environment. *Journal of Computer Assisted Learning*, 34(4):387–396, 2018.
- [8] Johnathan Mell, Jonathan Gratch, Tim Baarslag, Reyhan Aydoğan, and Catholijn M Jonker. Results of the first annual human-agent league of the automated negotiating agents competition. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 23–28, 2018.
- [9] Stephanie Carnell, Benjamin Lok, Melva T James, and Jonathan K Su. Predicting student success in communication skills learning scenarios with virtual humans. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 436–440, 2019.
- [10] Adam Kendon. *Gesture: Visible action as utterance*. Cambridge University Press, 2004.
- [11] Alex Lascarides and Matthew Stone. A formal semantic analysis of gesture. *Journal of Semantics*, page ffp004, 2009.
- [12] brian ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander T Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Mengyuan Yan, Noah Brown, Michael Ahn, Omar Cortes, Nicolas Sievers, Clayton Tan, Sichun Xu, Diego Reyes, Jarek Rettinghouse, Jornell Quiambao, Peter Pastor, Linda Luu, Kuang-Huei Lee, Yuheng Kuang, Sally Jesmonth, Nikhil J. Joshi, Kyle Jeffrey, Rosario Jauregui Ruano, Jasmine Hsu, Keerthana Gopalakrishnan, Byron David, Andy Zeng, and Chuyuan Kelly Fu. Do as i can, not as i say: Grounding language in robotic affordances. In Karen Liu, Dana Kulic, and Jeff Ichnowski, editors, *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 287–318. PMLR, 14–18 Dec 2023.

- [13] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.
- [14] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In Jie Tan, Marc Toussaint, and Kouros Darvish, editors, *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR, 06–09 Nov 2023.
- [15] Isaac Wang, Jesse Smith, and Jaime Ruiz. Exploring virtual agents for augmented reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.
- [16] Mary Ellen Foster. Enhancing human-computer interaction with embodied conversational agents. In *Universal Access in Human-Computer Interaction. Ambient Interaction: 4th International Conference on Universal Access in Human-Computer Interaction, UAHCI 2007 Held as Part of HCI International 2007 Beijing, China, July 22-27, 2007 Proceedings, Part II 4*, pages 828–837. Springer, 2007.

- [17] Aleksandra Kalinowska, Patrick M Pilarski, and Todd D Murphey. Embodied communication: How robots and people communicate through physical interaction. *Annual Review of Control, Robotics, and Autonomous Systems*, 6:205–232, 2023.
- [18] James Pustejovsky and Nikhil Krishnaswamy. Embodied human-computer interactions through situated grounding. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–3, 2020.
- [19] Susan Goldin-Meadow. The role of gesture in communication and thinking. *Trends in cognitive sciences*, 3(11):419–429, 1999.
- [20] David McNeill. So you think gestures are nonverbal? *Psychological review*, 92(3):350, 1985.
- [21] Albert Gatt and Kees Van Deemter. Lexical choice and conceptual perspective in the generation of plural referring expressions. *Journal of Logic, Language and Information*, 16(4):423–443, 2007.
- [22] Kees Van Deemter. Generating referring expressions that involve gradable properties. *Computational Linguistics*, 32(2):195–222, 2006.
- [23] Ielka Van Der Sluis and Emiel Krahmer. Generating referring expressions in a multimodal context: An empirically oriented approach. In *Computational Linguistics in the Netherlands 2000*, pages 158–176. Brill Rodopi, 2001.
- [24] Lars Kunze, Tom Williams, Nick Hawes, and Matthias Scheutz. Spatial referring expression generation for hri: Algorithms and evaluation framework. In *2017 AAAI Fall Symposium Series*, 2017.
- [25] Fethiye Irmak Doğan, Sinan Kalkan, and Iolanda Leite. Learning to generate unambiguous spatial referring expressions for real-world environments. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4992–4999. IEEE, 2019.

- [26] Aly Magassouba, Komei Sugiura, and Hisashi Kawai. Multimodal attention branch network for perspective-free sentence generation. In *Conference on Robot Learning*, pages 76–85. PMLR, 2020.
- [27] Rui Fang, Malcolm Doering, and Joyce Y Chai. Embodied collaborative referring expression generation in situated human-robot interaction. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 271–278, 2015.
- [28] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547, 2019.
- [29] Xinghang Li, Di Guo, Huaping Liu, and Fuchun Sun. Reve-ce: Remote embodied visual referring expression in continuous environment. *IEEE Robotics and Automation Letters*, 7(2):1494–1501, 2022.
- [30] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991, 2020.
- [31] Mohit Shridhar, Dixant Mittal, and David Hsu. Ingress: Interactive visual grounding of referring expressions. *The International Journal of Robotics Research*, 39(2-3):217–232, 2020.
- [32] Yang Li, Xiaoxue Chen, Hao Zhao, Jiangtao Gong, Guyue Zhou, Federico Rossano, and Yixin Zhu. Understanding embodied reference with touch-line transformer. In *ICLR*, 2023.
- [33] Emiel Kraemer and Ielka van der Sluis. A new model for generating multimodal referring expressions. In *Proceedings of the ENLG*, volume 3, pages 47–54, 2003.

- [34] Alfred Kranstedt, Stefan Kopp, and Ipke Wachsmuth. Murml: A multimodal utterance representation markup language for conversational agents. In *AAMAS'02 Workshop Embodied conversational agents-let's specify and evaluate them!*, 2002.
- [35] Alfred Kranstedt and Ipke Wachsmuth. Incremental generation of multimodal deixis referring to objects. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*, 2005.
- [36] Jingyu Wu, Shi Chen, Shuyu Gan, Weijun Li, Changyuan Yang, and Lingyun Sun. Cultural self-adaptive multimodal gesture generation based on multiple culture gesture dataset. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3538–3549, 2023.
- [37] Geunmo Kim, Jaewoong Yoo, and Hyedong Jung. Co-speech gesture generation via audio and text feature engineering. In *Companion Publication of the 25th International Conference on Multimodal Interaction*, pages 174–178, 2023.
- [38] Rui Fang, Malcolm Doering, and Joyce Chai. Collaborative models for referring expression generation in situated dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.
- [39] Mengwei Xu, Wangsong Yin, Dongqi Cai, Rongjie Yi, Daliang Xu, Qipeng Wang, Bingyang Wu, Yihao Zhao, Chen Yang, Shihe Wang, et al. A survey of resource-efficient llm and multimodal foundation models. *arXiv preprint arXiv:2401.08092*, 2024.
- [40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

- [41] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic vision-linguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.
- [42] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.
- [43] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.
- [44] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.
- [45] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.
- [46] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755, 2020.
- [47] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems*, 33:25–37, 2020.

- [48] Lei Han, Tian Zheng, Lan Xu, and Lu Fang. Occuseg: Occupancy-aware 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2940–2949, 2020.
- [49] Peter Gorniak and Deb Roy. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, 21:429–470, 2004.
- [50] Jette Viethen and Robert Dale. Algorithms for generating referring expressions: do they do what people do? In *Proceedings of the fourth international natural language generation conference*, pages 63–70, 2006.
- [51] Jette Viethen and Robert Dale. The use of spatial relations in referring expression generation. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 59–67, 2008.
- [52] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4185–4194, 2019.
- [53] Alessandro Moschitti, Bo Pang, and Walter Daelemans. Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.
- [54] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.
- [55] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016.

- [56] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.
- [57] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.
- [58] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512, 2017.
- [59] Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K Wong, and Qi Wu. Cops-ref: A new dataset and task on compositional referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10086–10095, 2020.
- [60] Boris Schauerte, Jan Richarz, and Gernot A Fink. Saliency-based identification and recognition of pointed-at objects. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4638–4643. IEEE, 2010.
- [61] Dadhichi Shukla, Ozgur Erkent, and Justus Piater. Probabilistic detection of pointing directions for human-robot interaction. In *2015 international conference on digital image computing: techniques and applications (DICTA)*, pages 1–8. IEEE, 2015.
- [62] Dadhichi Shukla, Özgür Erkent, and Justus Piater. A multi-view hand gesture rgb-d dataset for human-robot interaction scenarios. In *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)*, pages 1084–1091. IEEE, 2016.

- [63] Boris Schauerte and Gernot A Fink. Focusing computational visual attention in multimodal human-robot interaction. In *International conference on multimodal interfaces and the workshop on machine learning for multimodal interaction*, pages 1–8, 2010.
- [64] Nikhil Krishnaswamy and James Pustejovsky. Generating a novel dataset of multimodal referring expressions. In *Proceedings of the 13th International Conference on Computational Semantics-Short Papers*, pages 44–51, 2019.
- [65] Md Mofijul Islam, Reza Mirzaiee, Alexi Gladstone, Haley Green, and Tariq Iqbal. Caesar: An embodied simulator for generating multimodal referring expression datasets. *Advances in Neural Information Processing Systems*, 35:21001–21015, 2022.
- [66] Yixin Chen, Qing Li, Deqian Kong, Yik Lun Kei, Song-Chun Zhu, Tao Gao, Yixin Zhu, and Siyuan Huang. Yourefit: Embodied reference understanding with language and gesture. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1385–1395, 2021.
- [67] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.
- [68] Rebecca Passonneau. Measuring agreement on set-valued items (masi) for semantic and pragmatic annotation. 2006.
- [69] Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union, 1966.
- [70] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
- [71] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 human language technology conference*

- of the North American chapter of the association for computational linguistics*, pages 150–157, 2003.
- [72] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.
- [73] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.
- [74] Albert Gatt, Anja Belz, and Eric Kow. The tuna-reg challenge 2009: Overview and evaluation results. Association for Computational Linguistics, 2009.
- [75] Anja Belz and Albert Gatt. Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proceedings of ACL-08: HLT, Short Papers*, pages 197–200, 2008.
- [76] Nikhil Krishnaswamy and James Pustejovsky. An evaluation framework for multimodal interaction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.
- [77] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.
- [78] Robyn Kozierok, John Aberdeen, Cheryl Clark, Christopher Garay, Bradley Goodman, Tonia Korves, Lynette Hirschman, Patricia L McDermott, and Matthew W Peterson. Assessing open-ended human-computer collaboration systems: applying a hallmarks approach. *Frontiers in artificial intelligence*, 4:670009, 2021.

- [79] Francis Quek, David McNeill, Robert Bryll, Susan Duncan, Xin-Feng Ma, Cemil Kirbas, Karl E McCullough, and Rashid Ansari. Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 9(3):171–193, 2002.
- [80] James Pustejovsky and Nikhil Krishnaswamy. Multimodal semantics for affordances and actions. In *International Conference on Human-Computer Interaction*, pages 137–160. Springer, 2022.
- [81] Nikhil Krishnaswamy and Nada Alalyani. Embodied multimodal agents to bridge the understanding gap. In *Proceedings of the First Workshop on Bridging Human-Computer Interaction and Natural Language Processing*, pages 41–46, Online, April 2021. Association for Computational Linguistics.
- [82] Nikhil Krishnaswamy and James Pustejovsky. The role of embodiment and simulation in evaluating hci: Experiments and evaluation. In *International Conference on Human-Computer Interaction*, pages 220–232, 2021.
- [83] Nikhil Krishnaswamy and James Pustejovsky. VoxSim: A visual platform for modeling motion language. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. ACL, 2016.
- [84] Nikhil Krishnaswamy, Pradyumna Narayana, Rahul Bangar, Kyeongmin Rim, Dhruva Patil, David McNeely-White, Jaime Ruiz, Bruce Draper, Ross Beveridge, and James Pustejovsky. Diana’s world: A situated multimodal interactive agent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13618–13619, 2020.
- [85] Nada Alalyani and Nikhil Krishnaswamy. A methodology for evaluating multimodal referring expression generation for embodied virtual agents. In *Companion Publication of the 25th International Conference on Multimodal Interaction*, pages 164–173, 2023.
- [86] Isaac Wang, Mohtadi Ben Fraj, Pradyumna Narayana, Dhruva Patil, Gururaj Mulay, Rahul Bangar, J Ross Beveridge, Bruce A Draper, and Jaime Ruiz. Eggnog: A continuous, multi-

- modal data set of naturally occurring gestures with ground truth labels. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 414–421. IEEE, 2017.
- [87] Andrew Cameron Morris, Viktoria Maier, and Phil Green. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Eighth International Conference on Spoken Language Processing*, 2004.
- [88] Charles Spearman. The proof and measurement of association between two things. 1961.
- [89] James Pustejovsky and Nikhil Krishnaswamy. VoxML: A visualization modeling language. *Proceedings of LREC*, 2016.
- [90] Alfred Kranstedt, Andy Lücking, Thies Pfeiffer, Hannes Rieser, and Ipke Wachsmuth. Deixis: How to determine demonstrated objects using a pointing cone. In *Gesture in Human-Computer Interaction and Simulation: 6th International Gesture Workshop, GW 2005, Berder Island, France, May 18-20, 2005, Revised Selected Papers 6*, pages 300–311. Springer, 2006.
- [91] Hannah VanderHoeven, Nathaniel Blanchard, and Nikhil Krishnaswamy. Robust motion recognition using gesture phase annotation. In Vincent G. Duffy, editor, *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management*, pages 592–608, Cham, 2023. Springer Nature Switzerland.
- [92] Hannah VanderHoeven, Nathaniel Blanchard, and Nikhil Krishnaswamy. Point target detection for multimodal communication. In Vincent G. Duffy, editor, *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management*, pages 356–373, Cham, 2024. Springer Nature Switzerland.
- [93] James Pustejovsky and Nikhil Krishnaswamy. Embodied human computer interaction. *KI-Künstliche Intelligenz*, 35(3-4):307–327, 2021.

- [94] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, June 2014. Association for Computational Linguistics.
- [95] Julia Albath, Jennifer L. Leopold, Chaman L. Sabharwal, and Anne M. Maglia. RCC-3D: Qualitative spatial reasoning in 3D. In *CAINE*, pages 74–79, 2010.
- [96] Herbert P Grice. Logic and conversation. In *Speech acts*, pages 41–58. Brill, 1975.
- [97] Isaac Wang, Pradyumna Narayana, Dhruva Patil, Gururaj Mulay, Rahul Bangar, Bruce Draper, Ross Beveridge, and Jaime Ruiz. EGGNOG: A continuous, multi-modal data set of naturally occurring gestures with ground truth labels. In *To appear in the Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition*, 2017.
- [98] Karl Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.
- [99] Alan Agresti. *Categorical data analysis*, volume 792. John Wiley & Sons, 2012.
- [100] Artur Garcez, Tarek R Besold, L d Raedt, Peter Földiak, Pascal Hitzler, Thomas Icard, Kai-Uwe Kühnberger, Luis C Lamb, Risto Miikkulainen, and Daniel L Silver. Neural-symbolic learning and reasoning: contributions and challenges. 2015.
- [101] Tarek R Besold, Artur d’Avila Garcez, Sebastian Bader, Howard Bowman, Pedro Domingos, Pascal Hitzler, Kai-Uwe Kühnberger, Luis C Lamb, Daniel Lowd, Priscila Machado Vieira Lima, et al. Neural-symbolic learning and reasoning: A survey and interpretation. *arXiv preprint arXiv:1711.03902*, 2017.

- [102] Yoshua Bengio. The consciousness prior. *arXiv preprint arXiv:1709.08568*, 2017.
- [103] Edward Ma. Nlp augmentation. <https://github.com/makcedward/nlpaug>, 2019.
- [104] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2019.
- [105] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- [106] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.
- [107] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [108] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [109] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.
- [110] Ronald A Fisher. Xv.—the correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1919.

- [111] Erik Frøkjær, Morten Hertzum, and Kasper Hornbæk. Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 345–352, 2000.
- [112] Zijian Lew, Joseph B Walther, Augustine Pang, and Wonsun Shin. Interactivity in online chat: Conversational contingency and response latency in computer-mediated communication. *Journal of Computer-Mediated Communication*, 23(4):201–221, 2018.