Multimodal Referring Expression Generation for Human-Computer Interaction

Nada Alalyani[1][0000−0002−9253−416X] and Nikhil Krishnaswamy[2][0000−0001−7878−7227]

Colorado State University, USA
{nadahass,nkrishna}@colostate.edu

Abstract. Using both verbal and non-verbal modalities in generating definite descriptions of objects and locations is a critical human capability in collaborative interactions. Despite recent advancements in AI, embodied interactive virtual agents (IVAs) are not equipped to intelligently mix modalities to communicate their intents as humans do, which hamstrings naturalistic multimodal HCI. We introduce **SCMRE**, a corpus designed for training generative AI systems in multimodal HCI, focusing on multimodal referring expressions. Our contributions include: 1) Developing an interactive virtual agent (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) Using LLaMA 2-13B to conduct parameter-efficient finetuning 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.

Keywords: Embodied agents · non-verbal behaviours · multimodality · referring expression generation

1 Introduction

As human-computer interaction (HCI) systems become more advanced and sophisticated, there is an increasing expectation for them to behave more like humans in integrating modalities to communicate their intents. Humans fluently communicate in various non-verbal modalities with verbal modalities, a capability that even advanced multimodal models are unable to achieve [34]. While modern chatbots, powered by generative large language models (LLMs) such as

OpenAI's ChatGPT, have demonstrated remarkable abilities in generating coherent and context-relevant text, learning from and generating text alone fails to demonstrate an understanding of the meaning that connects utterance to communicative intent [6].

Agent embodiment provides a structure to demonstrate language understanding in context [25]. If a particular mode of expression, such as language, is inadequately communicative, another mode, such as gesture, can be used to disambiguate intents and targets. AI advancements have developed language models, e.g., GPT-4, that enable humans to interact with computers multimodally [34], but to date embodied interactive virtual agents (IVAs) cannot typically intelligently mix modalities to communicate their intents as humans do, which hamstrings naturalistic multimodal HCI. Due to the fact that objects within a shared situated context as anchors for establishing mutual understanding between interlocutors, *Multimodal Referring Expressions* (MREs), leveraging information about both object characteristics and locations, have emerged as a valuable case study for understanding multimodal language use in context [26, 32].

In this paper, we present SCMRE, a Situated Corpus of Multimodal Referring Expressions, and leverage it to train and evaluate generative AI models for embodied HCI. Our aim is advancing the development of IVAs capable of utilizing non-verbal and verbal behavior bidirectionally and symmetrically in interactions with humans. 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.

2 Related Work

Recent advancements in embodied HCI indicate the potential for enabling human-like interactions with users [14, 22]. Nonetheless, it is argued that HCI systems lack of bidirectional and symmetrical recognition and generation of multimodal communication mechanisms [50]. Therefore, IVAs, such as the Diana system [27, 28] built on then VoxWorld platform [29, 30] to support embodied HCI in recognizing both virtual and physical environments [50–52], enabling collaboration with humans in task-based interactions. Embodiment plays a significant role in representing and interpreting objects in a scene [53], in mutual understanding [26], and in evaluating the outputs of interactive systems [1, 33]. This emphasizes the importance of IVAs in not solely recognizing but also generating multimodal communication, particularly in the domain of referring expressions (REs).

Referring Expression Generation. Despite the significant contribution of deictic gesture to the successful communication of intent, [17, 46], early RE generation research prioritized linguistic descriptions, including object properties [16, 63] and spatial references [12, 35, 43]. Non-verbal cues like deictic gesture were more explored in RE comprehension [39, 54, 58]. Agent embodiment features were rarely integrated into generation [23, 24], with most studies treating generation and comprehension separately [13].

Multimodal Generative LLMs. Recent AI advances have led to the development of multimodal foundation models (MFMs) for multimodal generation [68]. Multimodal transformers, such as CLIP [55], ViLBERT [41], VisualBERT [38], SimVLM [67], BLIP-2 [37] and Flamingo [2], process inputs from various modalities like text, images, and point clouds. Other models focus on processing video, audio, or 3D data understanding [3, 19, 70]. These models are pre-trained on large multimodal datasets containing images, audios and language.

Datasets. Various datasets contain human-generated descriptions of objects in visual scenes, such as Bishop [18], Drawer [64], GRE3D3 [65], TUNA [16], RS-VS [43], and other recent collections [35, 12]. Other datasets focus on verbal references only [45, 10, 9], gestures only [57, 59, 60], or embodied multimodal referring expressions comprehension [56, 32]. These multimodal expressions are generated either by simulators, such as VoxSim [32], and CAESAR [21], or by humans referring to images [57] or outdoor objects [8].

Metrics. Overlap in the properties of human and machine descriptions can been computed according to Dice Coefficient [11], MASI [48], Levenshtein Distance [36], BLEU [47], ROUGE [40], or METEOR [4]. Alternatively, human judges can evaluate generated REs according to adequacy of reference or naturalness. While adequacy is evaluated by object identification tasks [12, 13, 15, 35], naturalness is evaluated by (1) metrics such as error rate, identification time, and reading time [5] or (2) human ranking of generated references for objects in images or videos [1, 12, 31, 35].

In this study, we developed an IVA to elicit MREs from humans in real-time interaction, trained a MRE generative model focusing on gesture and language, and evaluated how non-verbal strategies complement verbal strategies for situated HCI, both quantitatively and qualitatively.

3 SCMRE Dataset

This section outlines the collection process of SCMRE, aimed at developing generative models for multimodal HCI combining both language and gestures. It covers the IVA development, participants recruitment, human-IVA collaboration, and data statistics.

3.1 Development of the Interactive Virtual Agent (IVA)

We developed a standalone version of the Diana system [27, 50], a virtual agent designed for task-based interactions with humans using live gestures and speech.

In this implementation, humans interact with randomly positioned objects, providing both verbal (relational, historical) and non-verbal (deictic), references in response to Diana's prompts. As depicted in Figure 1*a*, Diana asks questions such as "Which object should we focus on?" while the human points using the mouse/trackpad, with the purple reticle fluctuating in location and size to approximate the noise inherent in live deictic gesture detection, as in the original Diana system. We created algorithms to parse and interpret human-generated multimodal referring expressions, including *attributive REs*, which describe objects properties, *relational REs*, which define objects by their relations to other objects, and *historical REs*, which uses previous events to describe objects, aligning them with deictic gestures, as shown in Figure 1*b-f*. Diana generates verbal and non-verbal behaviors, e.g., in Figure 1*h*, to enhance social fluency [66], using text-to-speech and animation for gestures, confirming understanding, responding to prompts, and displaying emotions. This system improves naturalistic human-computer interaction by accurately integrating speech and gestures. Further details can be found in [1, 50].

3.2 Human-IVA Collaboration Data Collection

To investigate human MRE generation, we organized human-IVA interaction sessions, consisting of 24 participants from Colorado State University (CSU)'s Computer Science Department. Participants, aged 18-35 (mean = 27, SD = 4.21) and fluent in English, included both males and females with diverse native languages. The study was approved by CSU's IRB. Participants received compensation in the form of Amazon gift cards or extra course credit. Each participant downloaded the IVA executable and engaged in an object identification task across 10 scenes, using language, deixis, or both to identify 10 target blocks per scene. Successful referencing occurred when Diana correctly identified the intended object. During the interaction, the IVA's and participants' movements were logged, including parameters outlined in [25].

3.3 Data Statistics

The SCMRE corpus is organized by incorporating each generated event, including actions and referring expressions, as a distinct sample. As shown in Table 1, the elicitation process resulted in a total of 10,408 events, including 7,681 pointing-only references, 551 transitive attributive events, 641 attributive events, 369 relational events, 27 historical events, 453 non-executed events, and 686 non-referencing events—which include 428 undoing events, 118 refusal events, and 117 affirmative events. In terms of modalities used by humans, 575 events were generated multimodally by mixing deixis and language, 7,681 events were generated using pointing-only, and 2,152 events were generated using speech-only. The number of events generated by each participant varied from 258 to 801 (mean = 444, SD = 171). Additionally, the data includes 194 recorded videos spanning approximately 36 hours, ranging from 24 minutes to 4 hours (mean = 01:27:52, SD = 0.04). The IVA, *Diana*, responded to each human-generated event, totaling 10,408 IVA responses. She generated 5,271 multimodal actions for 539 multimodal events, 3,628 pointing-only events, 686 non-referencing events,

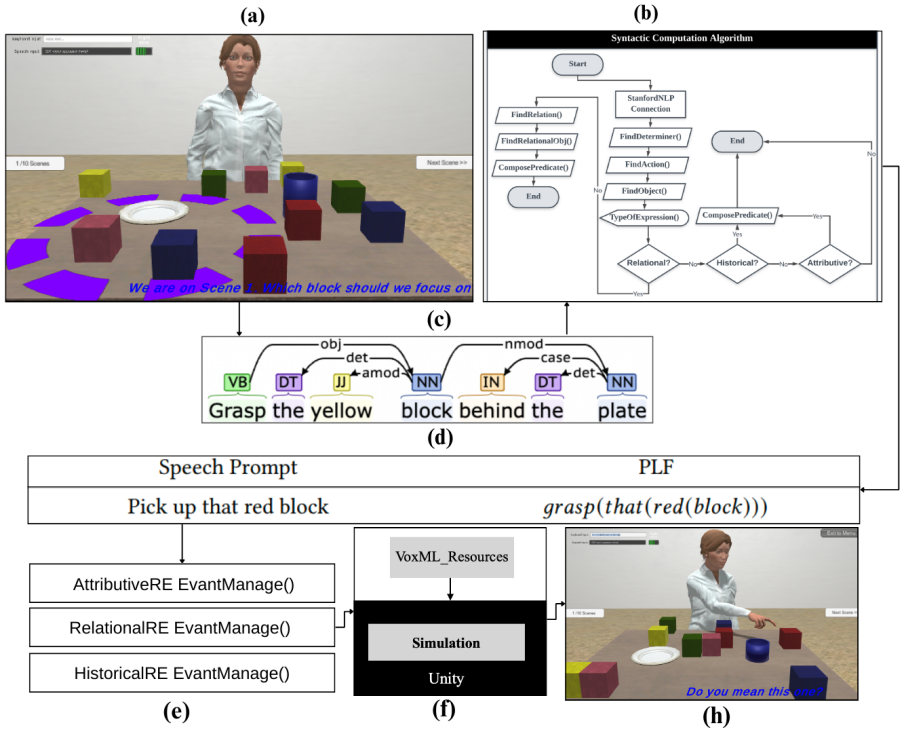

Fig. 1. Human-IVA interaction. a) Diana asks: "Which object should we focus on?" with a fluctuating purple circle indicating the pointing gesture; b) Speech parsing using Stanford CoreNLP [44]; c) Syntactic transformation of speech to Predicate Logic Format (PLF); d) Example of speech converted to PLF; e) Interpretation algorithms for complex MREs; f) Simulation of PLFs using VoxWorld platform; g) Diana acts on human prompts.

and 418 speech-only events. Moreover, she reacted unimodally: using deictic gestures for 4,053 pointing events to confirm understanding and using language to request more information for 1,084 events.

4 MRE Generation Model

4.1 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 2. Dataset before and after preparation is publicly available in GitHub[1].

[1] `https://github.com/nadahass/SCMRE_Dataset`

Table 1. 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		416	
Total			**10,408**

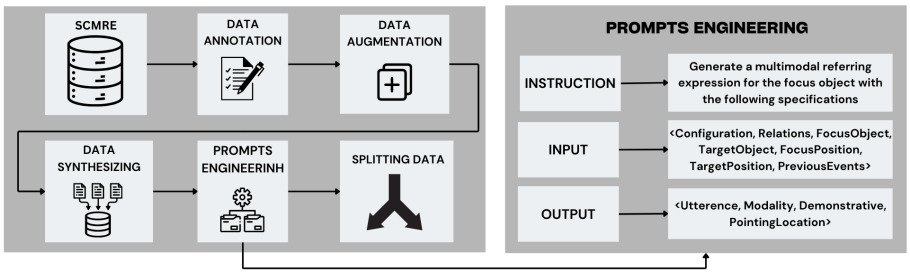

Fig. 2. The main steps of data preparation, including annotation, augmentation, synthesizing, prompting, and splitting.

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. The parameters that were logged for these prompts include timestamps, utterances, relations, configurations, and previous events. To effectively explore human referential behaviors and train our model, we included the remaining parameters: focus objects (focus of discourse), destination objects (objects to which other objects are moved), 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.

Table 2. 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**

Data Augmentation. A data augmentation method was utilized to increase both the size and diversity of SCMRE. Specifically, we employed the Synonym Augmentation technique from the NLPAug library [42] 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 [69] 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 (Table 2).

Data Synthesis. 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. Consequently, an additional 5,632 multimodal RE samples were incorporated, expanding the multimodal samples to 7,928 and increasing the total dataset size from 16,527 to 22,159, as shown in Table 2.

Prompt Engineering. We used Alpaca [61] as the basis for our MRE-generating model. Alpaca's Instruction-following models require structuring the data in a way that aligns with the model's architecture, 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 2. 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.

Data Splitting. For training experiments, we split 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 set, consisting of 20% of the total data, was reserved for evaluating

the model's performance. The testing data, comprising 20% of the validation data, was used to assess the model's generalization ability on unseen data. Table 3 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 were shuffled and the division was performed randomly.

Table 3. Training, Validation, and Testing Sets

LLaMAModels	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

4.2 Model Architecture

We used open-weight LLaMA models [62] to conduct parameter-efficient fine-tuning for generating contextually-correct and situationally-fluent referring expressions, including language and gesture. As illustrated in Figure 3, 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, $\langle Modality, Utterance, Location, Demonstratives \rangle$. $M \in \{\text{Gesture, Language, Ensemble}\}$, 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 = RedBlock, $P =< X, Y, Z >$, R = [Right(RedBlock, GreenBlock),...], $C = [< X', Y', Z' >, ...]$, H = [Put(YellowBlock),...] $\rangle$, represents the target object (the red block), the current spatial arrangement, associated relations and previous events. The corresponding output, $\langle$ multimodal, pick the red block, $< X'', Y'', Z'' >$, the $\rangle$, contains the elements of the generated multimodal referring expression. Here, this output prompts the agent to utter "pick the red block" while pointing to location $< X'', Y'', Z'' >$.

4.3 Learning Experiments

We fine-tuned multiple LLMs using Low-Rank Adapters (LoRA [20]) to enhance parameter and memory efficiency. LLaMA [62], developed by Meta AI, includes large-scale language models available in four parameter sizes: 7B, 13B, 33B, and 65B, and empirical studies indicate that even the LLaMA-13B model, with just $\frac{1}{10}$ of the parameters, surpasses GPT-3 (175B) [7] in most benchmark evaluations. For this study, we selected LLaMA-7B and LLaMA-13B 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 LLaMA 2 in this study, which for convenience is hereafter simply referred to as "LLaMA."

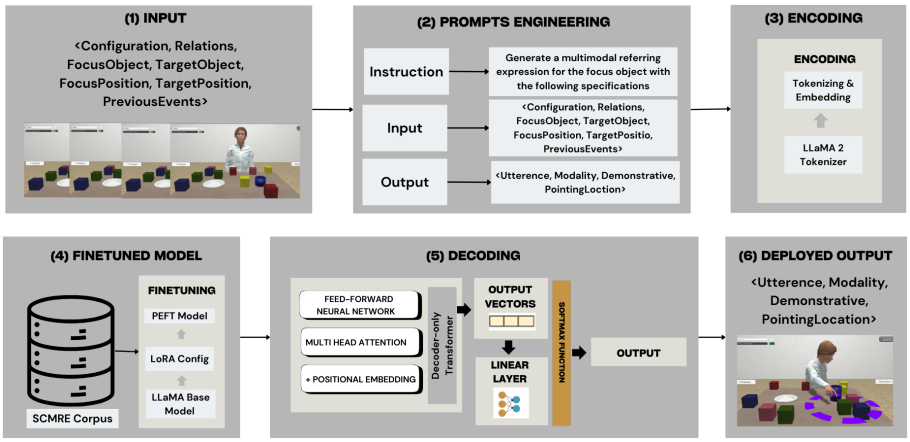

Fig. 3. The architecture of multimodal referring expressions generation model

Table 4. 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

According to the code implementation of Alpaca-LoRA, we applied patches to the LoRA modules for the key, query, and value matrices, setting their rank to 8, a scaling 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.

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 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 lists the hyper-parameters, training sets and training time that are related to each fine-tuned model.

4.4 Results

Loss Entropy. The loss curve for LLaMA-13B in Figure 4*a*, trained for 4,430 steps (1 epoch), shows faster convergence and achieves lower loss values compared to LLaMA-7B in Figure 4*c*, 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*b*,*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.

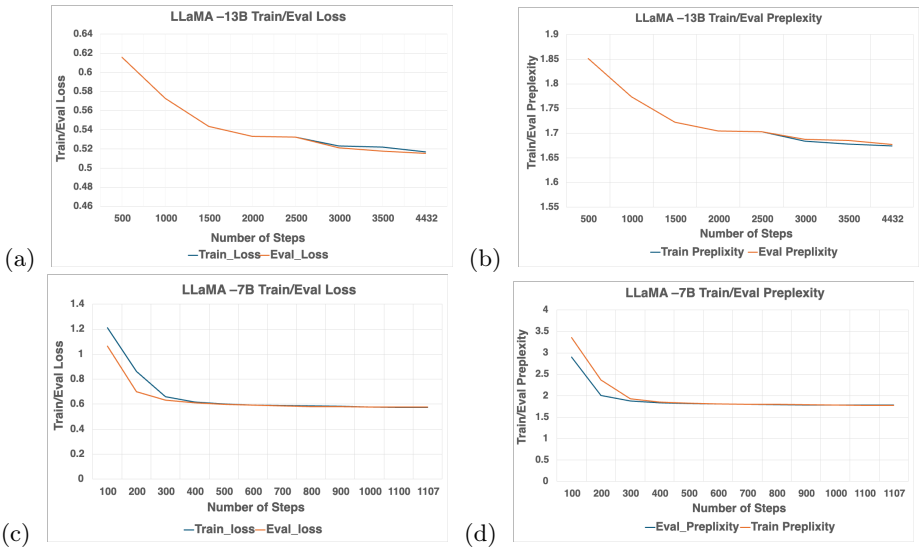

Fig. 4. The loss (a) and perplexity (b) of LLaMA-13B after one epoch of training. The loss (c) and perplexity (d) of LLaMA-7B after 1,107 steps of training

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), 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 performance in mixing modalities for generating referring expressions as depicted in Figure 5*d*. 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 5*e*: 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 5*a*, they account for 96% of outputs. On the LLaMA-7B model (Figure 5*b*), they account for 55.13%, and on the enhanced dataset with fewer steps (Figure 5*c*), they account for at 54.31%.

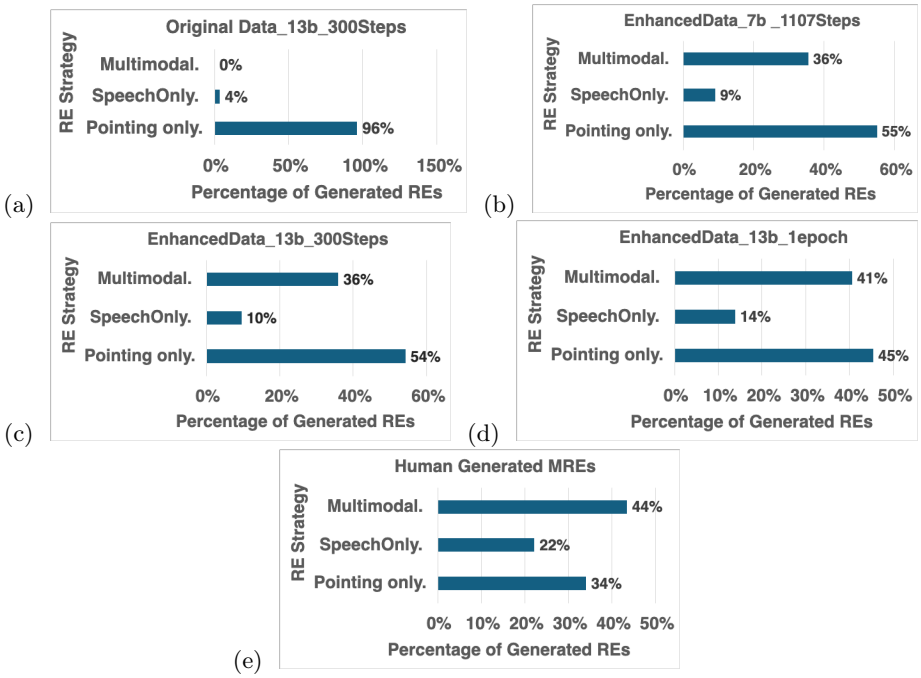

Fig. 5. Quantities of Human and LLM Generated Pointing, Linguistic and Multimodal Referring Expressions.

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 maintain semantic similarity to the ground truth MRE. Semantic similarity must be attained at both the speech and position levels. 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 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.

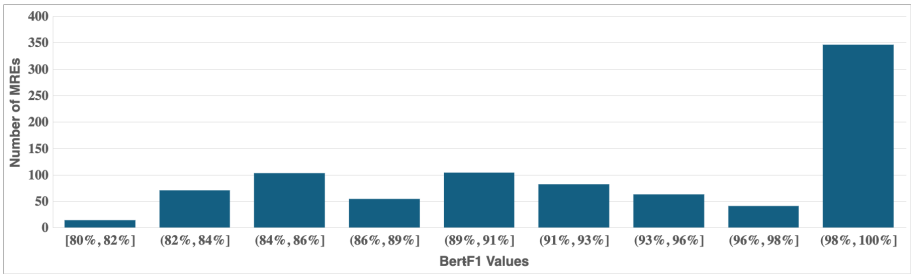

Fig. 6. The distribution of BERT-F1 similarity results between human-generated and LLM-generated MREs

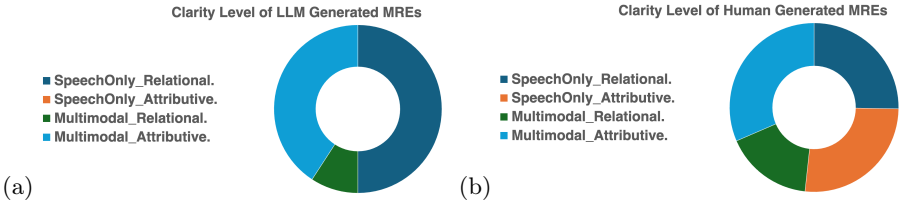

Fig. 7. Clarity level of the (a) LLM-generated and (b) human generated referring expressions

Clarity of Generated MREs. Based on the significant reduction in ambiguity levels observed when humans used co-gestural referring expressions (REs) while interacting with the IVA (see Section 5.1), we evaluate whether LLM-generated references maintain this level of clarity. We compared human and LLM-generated REs based on the information provided to communicate their intents. Figure 7 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 8 presents examples of all combinations of RE strategies for both humans and the fine-tuned model when referring to the same target objects in identical situations.

Correctness of Generated Positions. The fine-tuned LLaMA-13B, trained for one epoch, achieved remarkable performance, with an average accuracy, precision,

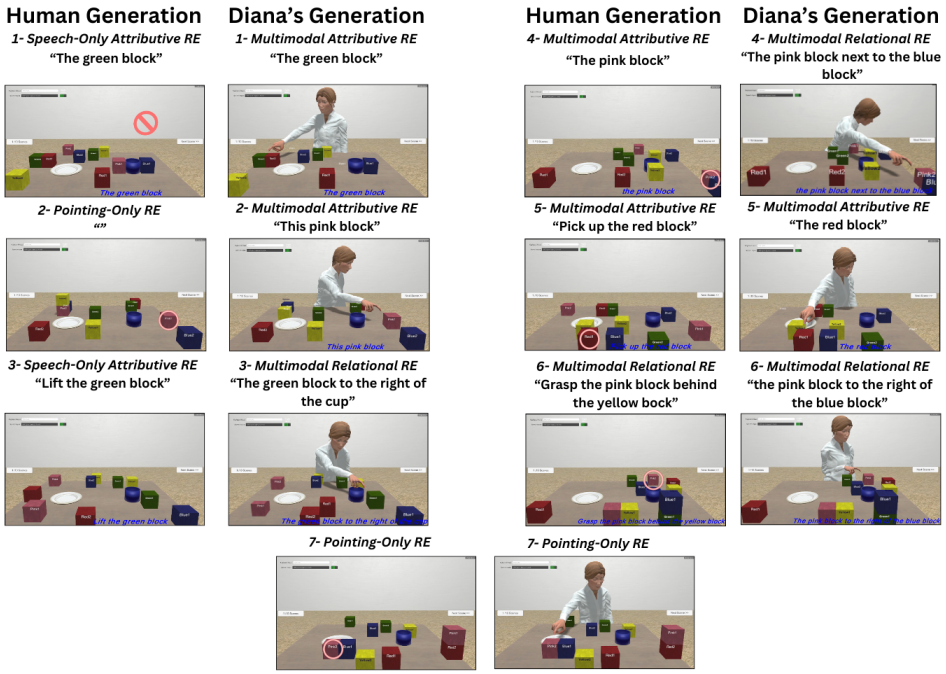

Fig. 8. Comparing human and IVA-generated REs for identical configurations.

recall, and F1-score of 99% for correctly generated positions. Performance 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 9.

5 Evaluation

To enable bidirectional communication between IVAs and humans using multimodal referring expressions, we previously proposed quantitative and qualitative metrics to assess if an IVA's non-verbal behavior generation aids human understanding. These metrics cover the following aspects: task completion efficiency, software reliability, understanding diverse communications, and meaningful content contribution by the agent as detailed in [1].

5.1 Automated Quantitative Evaluation

Using quantitative metrics in [1], we assessed *Multimodal Prompt Completion Efficiency* (MPCE) and *Linguistic Prompt Completion Efficiency* (LPCE) by measuring differences in target identification and task completion times for multimodal versus verbal-only REs. *Human Interpretation Efficiency of Machine Communication* (HIEMC) measured the time from the machine's reference generation to human target identification, and *Agent Pointing Success Rate* (APSR)

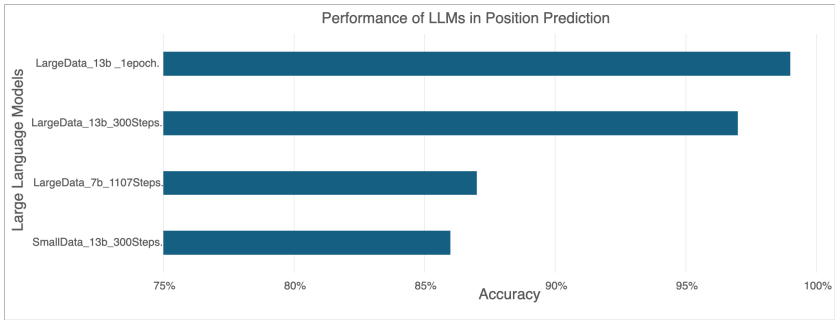

Fig. 9. The performance of LLaMA models in predicting positions of target objects

tracked the success rate of the agent pointing to the target object. The subsequent analysis of SCMRE showed that multimodal referring expressions significantly reduce prompt ambiguity (p-value $< .001$, χ^2-test) and enhance IVA response efficiency (p-value $< .001$, ANOVA test; see Figure 10*a,b*). This is evidenced by the IVA's ability to correctly identify referenced objects and execute human prompts, demonstrating bidirectional communicative efficiency (see Figure 10*c,d*). Additional quantitative metrics, along with their corresponding results, for evaluating the model's generation capability are detailed in Section 4.4. These results show the capacity of a generative model within HCI systems to contextually integrate gestures and language, thereby enhancing task-based interactions and facilitating more natural human-computer communication.

5.2 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 *Machine References Fluency Rate* (MRFR), the rate of top-rated machine references based on third-party human judgments, and *Human References Fluency Rate* (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 *Machine Object Identification Success Rate* (MOISR), the rate of correctly identified objects (by machine), and *Human Object Identification Success Rate* (HOISR), the rate of correctly identified objects (by humans), through task completion [1]. Evaluation data and results are publicly available on GitHub[2].

Study Design. We selected 50 human MREs from the SCMRE dataset. These were compared with 50 REs generated by the virtual agent 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

[2] `https://github.com/nadahass/Human-based-Evaluation-MREG.git`

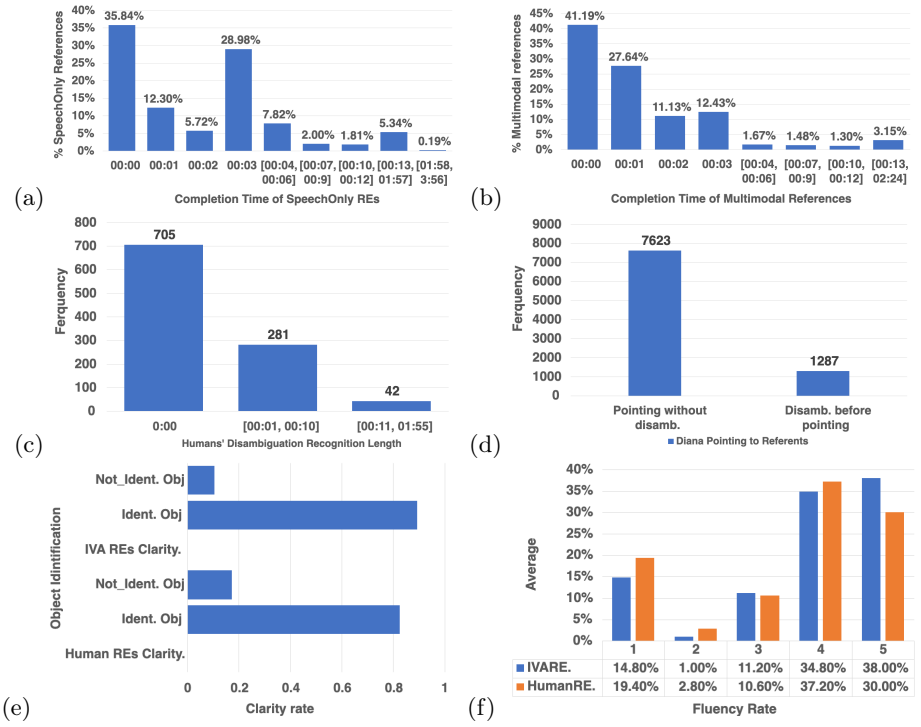

Fig. 10. Diana's completion time of (a) Speech Event (LPCE); (b) Multimodal Event (MPCE); (c) Human Interpretation Efficiency of Machine Communication (HIEMC); (d) Agent Pointing Success Rate (APSR); (e) Machine Object Identification Success Rate (MOISR) and Human Object Identification Success Rate (HOISR), and (f) Machine References Fluency Rate (MRFR) and Human References Fluency Rate (HRFR)

human and IVA generation are pointing only REs, relational speech-only REs, attributive speech-only REs, relational multimodal REs and attributive multimodal REs. Videos were used in a set of AMT human intelligence tasks (HITs), wherein workers rated 1 video for *both* fluency and clarity, including 1 machine generated RE or 1 human RE, for a total of 100 HITs. 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 1,000 individual judgments. We recruited workers fluent in English between 18 and 60 years old. They were given 1 hour per task and were compensated $0.75 per HIT.

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 accuracy rates of identifying objects referred to by humans and IVA were compared to the intended objects in the

dataset. As shown in Figure 10*e*, the overall HOISR and MOISR were 82.6% and 89.4%, respectively, demonstrating that the clarity level of IVA-generated MREs strongly competes with human-generated MREs (p-value $< 2.2e{-}16$ using Pearson's χ^2-test [49]). For the fluency task, Figure 10*f* shows MRFR of 73% and HRFR of 67% , with ratings at "4" and "5". These results indicate that both human-generated and IVA-generated MREs are perceived similarly in terms of and fluency (p-value $= 0.5529$ using Pearson's χ^2-test). This suggests that IVAs are capable of generating REs of comparable quality to those of humans.

6 Conclusion

Given the advancements in interactive agents, there is a growing expectation that they will contribute to interactions in ways that resemble human behavior, rather than just performing tasks. This study showcases significant advancements in multimodal HCI capabilities, specifically in bridging the gap between human and IVA communication capabilities in generating referring expressions. The developed SCMRE corpus, coupled with the fine-tuned generative model and comprehensive evaluation framework, enables more effective and naturalistic interactions between humans and IVAs. Our findings demonstrate a means by which IVAs can close the gap with human in generating contextually appropriate multimodal referring expressions, which is one crucial capacity for more naturalistic HCI. Future work will focus on refining the IVA's ability to handle more complex and bidirectional interaction scenarios, enhancing real-time processing capabilities, and integrating these models into diverse application domains. Further research is also needed to explore long-term user adaptation and the IVA's ability to learn from ongoing interactions.

Acknowledgments. We express our gratitude to our reviewers for their valuable comments. Additionally, we extend our thanks to our participants for their contributions in providing the SCMRE data.

References

1. Alalyani, N., Krishnaswamy, N.: A methodology for evaluating multimodal referring expression generation for embodied virtual agents. In: Companion Publication of the 25th International Conference on Multimodal Interaction. pp. 164–173 (2023)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022)
3. Alayrac, J.B., Recasens, A., Schneider, R., Arandjelović, R., Ramapuram, J., De Fauw, J., Smaira, L., Dieleman, S., Zisserman, A.: Self-supervised multimodal versatile networks. Advances in Neural Information Processing Systems **33**, 25–37 (2020)
4. Banerjee, S., Lavie, A.: 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. pp. 65–72 (2005)
5. Belz, A., Gatt, A.: Intrinsic vs. extrinsic evaluation measures for referring expression generation. In: Proceedings of ACL-08: HLT, Short Papers. pp. 197–200 (2008)
6. Bender, E.M., Koller, A.: 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. pp. 5185–5198 (2020)
7. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
8. Chen, Y., Li, Q., Kong, D., Kei, Y.L., Zhu, S.C., Gao, T., Zhu, Y., Huang, S.: Yourefit: Embodied reference understanding with language and gesture. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1385–1395 (2021)
9. Chen, Z., Wang, P., Ma, L., Wong, K.Y.K., Wu, Q.: 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. pp. 10086–10095 (2020)
10. De Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., Courville, A.: Guesswhat?! visual object discovery through multi-modal dialogue. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5503–5512 (2017)
11. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
12. Doğan, F.I., Kalkan, S., Leite, I.: Learning to generate unambiguous spatial referring expressions for real-world environments. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4992–4999. IEEE (2019)
13. Fang, R., Doering, M., Chai, J.Y.: Embodied collaborative referring expression generation in situated human-robot interaction. In: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction. pp. 271–278 (2015)
14. Foster, M.E.: 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. pp. 828–837. Springer (2007)
15. Gatt, A., Belz, A., Kow, E.: The tuna-reg challenge 2009: Overview and evaluation results. Association for Computational Linguistics (2009)
16. Gatt, A., Van Deemter, K.: Lexical choice and conceptual perspective in the generation of plural referring expressions. Journal of Logic, Language and Information **16**(4), 423–443 (2007)
17. Goldin-Meadow, S.: The role of gesture in communication and thinking. Trends in cognitive sciences **3**(11), 419–429 (1999)
18. Gorniak, P., Roy, D.: Grounded semantic composition for visual scenes. Journal of Artificial Intelligence Research **21**, 429–470 (2004)
19. Han, L., Zheng, T., Xu, L., Fang, L.: Occuseg: Occupancy-aware 3d instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2940–2949 (2020)
20. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)

21. Islam, M.M., Mirzaiee, R., Gladstone, A., Green, H., Iqbal, T.: Caesar: An embodied simulator for generating multimodal referring expression datasets. Advances in Neural Information Processing Systems **35**, 21001–21015 (2022)
22. Kalinowska, A., Pilarski, P.M., Murphey, T.D.: Embodied communication: How robots and people communicate through physical interaction. Annual Review of Control, Robotics, and Autonomous Systems **6**, 205–232 (2023)
23. Krahmer, E., van der Sluis, I.: A new model for generating multimodal referring expressions. In: Proceedings of the ENLG. vol. 3, pp. 47–54 (2003)
24. Kranstedt, A., Kopp, S., Wachsmuth, I.: Murml: A multimodal utterance representation markup language for conversational agents. In: AAMAS'02 Workshop Embodied conversational agents-let's specify and evaluate them! (2002)
25. Krishnaswamy, N., Alalyani, N.: Embodied multimodal agents to bridge the understanding gap. In: Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing. pp. 41–46 (2021)
26. Krishnaswamy, N., Alalyani, N.: Embodied multimodal agents to bridge the understanding gap. In: Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing. pp. 41–46. Association for Computational Linguistics, Online (Apr 2021)
27. Krishnaswamy, N., Narayana, P., Bangar, R., Rim, K., Patil, D., McNeely-White, D., Ruiz, J., Draper, B., Beveridge, R., Pustejovsky, J.: Diana's world: A situated multimodal interactive agent. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 13618–13619 (2020)
28. Krishnaswamy, N., Narayana, P., Wang, I., Rim, K., Bangar, R., Patil, D., Mulay, G., Beveridge, R., Ruiz, J., Draper, B., et al.: Communicating and acting: Understanding gesture in simulation semantics. In: Proceedings of the 12th International Conference on Computational Semantics (IWCS)—Short papers (2017)
29. Krishnaswamy, N., Pickard, W., Cates, B., Blanchard, N., Pustejovsky, J.: The voxworld platform for multimodal embodied agents. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 1529–1541 (2022)
30. Krishnaswamy, N., Pustejovsky, J.: Voxsim: A visual platform for modeling motion language. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations. pp. 54–58 (2016)
31. Krishnaswamy, N., Pustejovsky, J.: An evaluation framework for multimodal interaction. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)
32. Krishnaswamy, N., Pustejovsky, J.: Generating a novel dataset of multimodal referring expressions. In: Proceedings of the 13th International Conference on Computational Semantics-Short Papers. pp. 44–51 (2019)
33. Krishnaswamy, N., Pustejovsky, J.: The role of embodiment and simulation in evaluating hci: Experiments and evaluation. In: International Conference on Human-Computer Interaction. pp. 220–232 (2021)
34. Krishnaswamy, N., Pustejovsky, J.: Affordance embeddings for situated language understanding. Frontiers in Artificial Intelligence **5**, 774752 (2022)
35. Kunze, L., Williams, T., Hawes, N., Scheutz, M.: Spatial referring expression generation for hri: Algorithms and evaluation framework. In: 2017 AAAI Fall Symposium Series (2017)
36. Levenshtein, V.I., et al.: Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics doklady. vol. 10, pp. 707–710. Soviet Union (1966)
37. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)

38. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
39. Li, X., Guo, D., Liu, H., Sun, F.: Reve-ce: Remote embodied visual referring expression in continuous environment. IEEE Robotics and Automation Letters **7**(2), 1494–1501 (2022)
40. Lin, C.Y., Hovy, E.: 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. pp. 150–157 (2003)
41. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems **32** (2019)
42. Ma, E.: Nlp augmentation. https://github.com/makcedward/nlpaug (2019)
43. Magassouba, A., Sugiura, K., Kawai, H.: Multimodal attention branch network for perspective-free sentence generation. In: Conference on Robot Learning. pp. 76–85. PMLR (2020)
44. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 55–60. Association for Computational Linguistics, Baltimore, Maryland (Jun 2014)
45. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 11–20 (2016)
46. McNeill, D.: So you think gestures are nonverbal? Psychological review **92**(3), 350 (1985)
47. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
48. Passonneau, R.: Measuring agreement on set-valued items (masi) for semantic and pragmatic annotation (2006)
49. Pearson, K.: 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)
50. Pustejovsky, J., Krishnaswamy, N.: Embodied human-computer interactions through situated grounding. In: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents. pp. 1–3 (2020)
51. Pustejovsky, J., Krishnaswamy, N.: Situated meaning in multimodal dialogue: human-robot and human-computer interactions. Traitement Automatique des Langues **61**(3), 17–41 (2020)
52. Pustejovsky, J., Krishnaswamy, N.: Embodied human computer interaction. KI-Künstliche Intelligenz **35**(3), 307–327 (2021)
53. Pustejovsky, J., Krishnaswamy, N.: Multimodal semantics for affordances and actions. In: International Conference on Human-Computer Interaction. pp. 137–160. Springer (2022)
54. Qi, Y., Wu, Q., Anderson, P., Wang, X., Wang, W.Y., Shen, C., Hengel, A.v.d.: Reverie: Remote embodied visual referring expression in real indoor environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9982–9991 (2020)

55. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
56. Schauerte, B., Fink, G.A.: Focusing computational visual attention in multi-modal human-robot interaction. In: International conference on multimodal interfaces and the workshop on machine learning for multimodal interaction. pp. 1–8 (2010)
57. Schauerte, B., Richarz, J., Fink, G.A.: Saliency-based identification and recognition of pointed-at objects. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 4638–4643. IEEE (2010)
58. Shridhar, M., Mittal, D., Hsu, D.: Ingress: Interactive visual grounding of referring expressions. The International Journal of Robotics Research **39**(2-3), 217–232 (2020)
59. Shukla, D., Erkent, O., Piater, J.: Probabilistic detection of pointing directions for human-robot interaction. In: 2015 international conference on digital image computing: techniques and applications (DICTA). pp. 1–8. IEEE (2015)
60. Shukla, D., Erkent, Ö., Piater, J.: 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). pp. 1084–1091. IEEE (2016)
61. Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford\_alpaca` (2023)
62. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models (2023). arXiv preprint arXiv:2302.13971 (2023)
63. Van Deemter, K.: Generating referring expressions that involve gradable properties. Computational Linguistics **32**(2), 195–222 (2006)
64. Viethen, J., Dale, R.: Algorithms for generating referring expressions: do they do what people do? In: Proceedings of the fourth international natural language generation conference. pp. 63–70 (2006)
65. Viethen, J., Dale, R.: The use of spatial relations in referring expression generation. In: Proceedings of the Fifth International Natural Language Generation Conference. pp. 59–67 (2008)
66. Wang, I., Smith, J., Ruiz, J.: Exploring virtual agents for augmented reality. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 1–12 (2019)
67. Wang, Z., Yu, J., Yu, A.W., Dai, Z., Tsvetkov, Y., Cao, Y.: Simvlm: Simple visual language model pretraining with weak supervision. arXiv preprint arXiv:2108.10904 (2021)
68. Xu, M., Yin, W., Cai, D., Yi, R., Xu, D., Wang, Q., Wu, B., Zhao, Y., Yang, C., Wang, S., et al.: A survey of resource-efficient llm and multimodal foundation models. arXiv preprint arXiv:2401.08092 (2024)
69. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations (2019)
70. Zhu, L., Yang, Y.: Actbert: Learning global-local video-text representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8746–8755 (2020)