

A Methodology for Evaluating Multimodal Referring Expression Generation for Embodied Virtual Agents

Anonymous Author(s)*

ABSTRACT

Robust use of definite descriptions in a situated space often involves recourse to both verbal and non-verbal modalities. For IVAs, virtual agents designed to interact with humans, the ability to both recognize and generate non-verbal and verbal behavior is a critical capability. To assess how well an IVA is able to deploy multimodal behaviors, including language, gesture, and facial expressions, we propose a methodology to evaluate the agent’s capacity to generate object references in a situational context, using the domain of multimodal referring expressions as a use case. Our contributions include: 1) developing an embodied platform to collect human referring expressions while communicating with the IVA. 2) comparing human and machine-generated references in terms of evaluable properties using subjective and objective metrics. 3) reporting preliminary results from trials that aimed to check whether the agent can retrieve and disambiguate the object the human referred to, if the human has the ability to correct misunderstanding using language, deictic gesture, or both; and human ease of use while interacting with the agent.

CCS CONCEPTS

• **Human-centered computing** → **HCI design and evaluation methods**; • **Computing methodologies** → **Natural language generation**.

KEYWORDS

Embodied agents, non-verbal behaviours, multimodality, referring expression generation

ACM Reference Format:

Anonymous Author(s). 2023. A Methodology for Evaluating Multimodal Referring Expression Generation for Embodied Virtual Agents. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym ’XX)*. ACM, New York, NY, USA, 10 pages. <https://doi.org/XXXXXXXX.XXXXXXX>

1 INTRODUCTION

Recent achievements in generative language modeling, of which OpenAI’s ChatGPT is an exemplar, have demonstrated remarkable abilities in producing topically coherent, grammatically correct, and contextually appropriate text. Prior to the generative AI boom, language models such as BERT [10] and GPT-2 [54] achieved state

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference acronym ’XX, June 03–05, 2018, Woodstock, NY

© 2023 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00
<https://doi.org/XXXXXXXX.XXXXXXX>

of the art results on various language processing tasks. It may be tempting, therefore, to believe that language generation for conversational agents (CAs) is a solved problem. However, a common critique of large language models (LLMs) is that they lack *grounding* or *understanding*. Bender and Koller [4] argue that learning only from the textual form does not provide information about the “meaning” connecting utterance to communicative intent.

Humans, meanwhile, communicate in multiple non-verbal modalities, and mix these fluently with verbal modalities. A telling example is the ability of a human to answer a question like “what am I pointing at?” with appropriate situational context, which even a multimodal LLM like GPT-4 cannot. Given the recent developments in language modeling, we can expect the ability to fluently mix and match modalities to be a critical capability in the next generation of CAs. As interactive agents become more sophisticated, and see and interpret both visual and linguistic context concurrently, users will expect them to behave more like humans.

Agent embodiment is one channel to provide information needed to enable CAs to understand language in context. If one modality (e.g., language) is not communicative, another modality (e.g., gesture) can be used to disambiguate or correct the failure. As objects in a shared situated context provide anchors for the construction of common ground between interlocutors [7, 50, 51], a valuable use case to understand multimodal language use in context is **multimodal referring expressions** (MREs) that exploit information about both object characteristics and locations [8]. It is therefore necessary to come up with principled strategies to evaluate mixed-modality referring expression generation systems.

In this paper, we propose a methodology to carefully evaluate generation of multimodal referring expressions by a particular class of CAs, namely embodied interactive virtual agents (IVAs), with the goal of aiding the development of IVAs that interact with humans with symmetrical, bidirectional use of non-verbal and verbal behavior. Our novel contributions are:

- An embodied virtual agent testbed with an IVA who uses gesture and language [26, 40] to elicit MREs from humans;
- Establishing bidirectional and symmetric communication between humans and IVAs using verbal and non-verbal behavior synthesis;
- Evaluation metrics thereof that apply to both humans and IVAs, combining qualitative and quantitative metrics;
- Analysis of preliminary data gathered from interactions with our test agent.

2 RELATED WORK

The psycholinguistic literature shows the impact of deictic gesture on the successful communication of intent and reference for both speakers and hearers [17, 41]. Nonetheless, much earlier work in the area of referring expression (RE) generation has focused on linguistic description, such as relative and absolute properties of objects

(e.g., size and color) [16, 61], spatial references [12, 32, 37], and relational episodic descriptions [13]. Where non-verbal information, such as deictic gesture, is considered, much prior work focuses on RE comprehension rather than generation, e.g., [5, 35, 52, 57], and additionally typically lacks features related to agent embodiment [22, 23]. Where generation is addressed [13], it is often separated from comprehension. As such, we seek to build and evaluate models for generating MREs that are fluent and clear, and symmetric and bidirectional in the context they exploit when compared to human-generated REs. Doing so requires developing evaluation metrics that indicate when IVA-generated non-verbal behavior provides a meaningful boost in communicative capability compared to verbal behavior only.

Datasets. A number of datasets and corpora exist of human-generated descriptions of target objects in visual scenes, including Bishop [18], Drawer [63], GRE3D3 [64], TUNA [16], RS-VS [37], and recent corpora by Kunze et al. [32] and Doğan et al. [12]. Other RE corpora collected for the purpose of training comprehension models fall into three categories—verbal references only [6, 9, 20, 39, 42, 45, 67], gestures only [56, 58, 59], and embodied multimodal REs including language and gesture [30, 55].

Metrics. Correspondence between human corpora and machine generated references can be measured either by automatic metrics or human judgments. Overlap in the properties of human and machine descriptions can be computed according to Dice Coefficient [11], MASI [44], Levenshtein Distance [34], BLEU [43], ROUGE [36], CIDER [62], or METEOR [2]. 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, 32], naturalness is evaluated by (1) metrics such as error rate, identification time, and reading time [3, 29] or (2) human ranking of generated references for objects in a set of images or videos [12, 30, 32].

Prior work on embodied agents argues for the role of embodiment in representing the salient content of objects in a scene [49], in contributing to mutual understanding [25], and in evaluating the outputs of interactive systems [31]. Relatedly, Kozierok et al. [21] 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 the use of MREs, and a proposal for evaluating how non-verbal strategies complement verbal strategies for situated meaning [53].

In the remainder of this paper, we will discuss the platform we use to collect and generate MREs in a human-agent interaction (Sec. 3), specify the evaluation metrics we propose to use (Sec. 4), present preliminary results of initial data collected according to the proposed evaluation (Sec. 5), and discuss future directions (Sec. 6).

3 METHODOLOGY

First, we develop an interactive virtual agent system for an object identification task that interprets human language and simulated gesture inputs, and responds with language and animated gestures. We then proposed metrics to address the fluency and clarity of referring expressions used. Since our goal is to create symmetric,

bidirectional communication between humans and agents, these metrics may apply to either human or agent behaviors, and we compare the use of verbal and non-verbal modalities. We then analyze preliminary data for indications of where human and agent use of different modalities aids communication, for the purposes of assessing the contribution of non-verbal behavior to the interaction.

3.1 Interactive Virtual Agent (IVA) Development

The *Diana* system [26, 47] was 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 existing system into a standalone version 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 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 Fig. 1, 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 the gesture-semantic notion of a *pointing cone* [24], the center of deixis fluctuates within a circle of radius $\pm 0.3m$ around the mouse location and the size of the deictic reticle (see Fig. 1) 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.

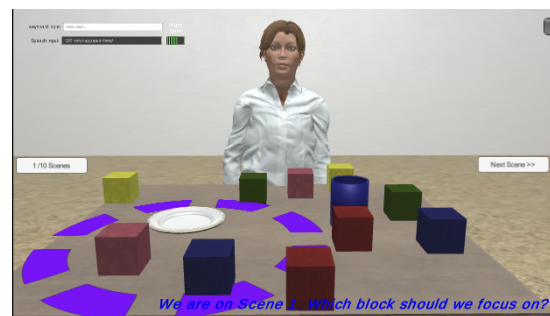


Figure 1: Experimental Diana System: the purple circle indicates where the user is pointing. Without disambiguation, any object within the pointing circle is a potential candidate for a deixis-only RE. Diana’s utterances both appear on screen and are spoken aloud via TTS.

Table 1: 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.).

Speech Prompt	PLF	Verbal	Non-Verbal	RE Type
Pick up that red block	<i>grasp(that(red(block)))</i>	✓	✓	(Att_RE)
Put this block to the right of the blue block	<i>put(this(block), right(the(blue(block))))</i>	✓	✓	(Trans_RE)
Grasp the green block beside the plate	<i>grasp(the(green(beside_adj(plate(block))))</i>	✓	-	(Rel_RE)
Lift the block you just put down	<i>lift(the(put_adj(block)))</i>	✓	-	(Hist_RE)
Take this block and put it there	<i>take(this(block)) + Put({0}, {1})*</i>	✓	✓	(Comp_RE)

Interpreting Verbal and Non-Verbal Expressions. Multimodal referring expressions can be considered special cases of *gesture utterances* as specified in [48], 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. Therefore, our motivation for developing a bidirectional evaluation scheme is to create methodologies for evaluating combined verbal and non-verbal behavior that apply equally well to human and IVA behaviors.

We follow an analysis of the EGGNOG dataset, a collection of human-human interactions in a Blocks World domain [65], wherein human-generated verbal REs are expected to fall into three complex categories, potentially involving both verbal and non-verbal content: *Attributive REs*, which describe object properties; *Relational REs*, which describe objects in relation to each other; and *Historical REs*, which describe objects already mentioned or interacted with. All three of these may be aligned with deictic gesture, but in different ways. To replicate these exhibited interpretive capabilities, we first developed four main algorithms to interpret verbal REs: (1) *<ParsingToPLF>* recursively follows a set of rules, using the Stanford CoreNLP dependency tree [38] to compose linguistic constituents into a predicate logic format (PLF). Table 1 shows the PLFs of different speech inputs and whether they need to be accompanied by non-verbal information for a complete interpretation. Multimodal references are interpreted with respect to the VoxML modeling language [33, 46] and the scene in the VoxWorld simulation platform [27, 28]. (2) The *<RelationalRE>* algorithm leverages spatial relations between objects that are tracked by the VoxWorld platform using calculi such as RCC-3D [1]. The interpreter extracts mentioned objects, localizes the target relative to other objects, and acts upon it as shown by command #7 in Fig. 2. (3) The *<HistoricalRE>* algorithm processes those sub-predicates that indicate actions that have previously been taken in the dialogue, e.g., in Fig. 2 #9, by extracting objects that were the subjects of previous actions. (4) The *<AttributiveRE>* algorithm interprets attributes of the object (in this setting, color or demonstratives). There are two subcategories of attributive REs: *Transitive REs*, an attributive RE embedded within a transitive clause (see Table 1); and *Compound REs* involving multiple events, as shown by Fig. 2 #1. A simple, non-transitive, non-compound attributive RE is shown in Fig. 2 #3. This utterance also shows how demonstratives may involve the human deictically disambiguating the target. Deictic

inputs can also be sequential pointing prompts with no speech, e.g., Fig. 2 #13–16, in which the human points to *BlueBlock* then to the *GreenBlock* to instruct Diana to put the blue block on the green block.

```

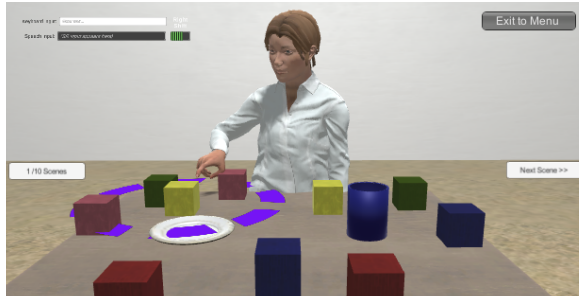
HUMAN: [1] Take the red block and put it to the left of the plate.
DIANA: [2] Which red block?
HUMAN: [3] This one. [Points to a red block]
DIANA: [4] OK! [Acts on the red block]
HUMAN: [5] Grasp this pink block and put it on the red block.
[points to a pink block]
DIANA: [6] Do you mean the red block I just put down? [Points
to a red block]
HUMAN: [7] No, the red block behind the blue block.
DIANA: [8] OK! [Acts on the red block]
HUMAN: [9] Pick up the block that you just grasped.
DIANA: [10] I've just grasped the pink block (that is) on the
red block. Is this what you mean? [points to a pink block]
HUMAN: [11] Yes.
DIANA: [12] OK! [Acts on the pink block]
HUMAN: [13] [Points to a blue block]
DIANA: [14] [Points to the same blue block]
HUMAN: [15] [Points over the green block]
DIANA: [16] [Puts the blue block on the green block]

```

Figure 2: Sample dialogue: the interaction from 1–12 is multimodal (co-gestural speech) and from 13–16 is unimodal (deictic gesture only).

Generating Verbal and Non-Verbal Expressions. 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 [66]. 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. (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 Fig. 3. (2) She directly acts on all aforementioned verbal prompts (e.g., multimodal commands in Fig. 2, #1–12) by either disambiguating candidate target objects or carrying out the requested action in the virtual space. (3) She also acts on non-verbal prompts (e.g., unimodal commands in Fig. 2 from 13–16) by performing the denoted actions after the human specifies the focus and target locations. (4) As shown in Fig. 4, 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,

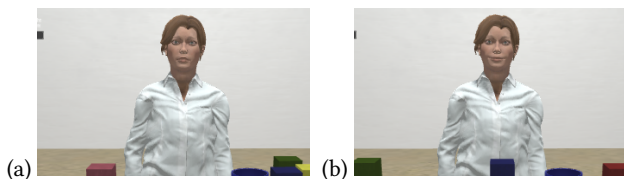
349 becomes a question of correctly generating the content of an ut-
 350 terance, movement through space of a gesture, or specific facial
 351 expression at the right time, to serve a communicative purpose.
 352



353
354
355
356
357
358
359
360
361
362
363
364 **Figure 3: Generating deictic gestures. Diana will respond to**
 365 **what she interprets the RE as referring to by pointing to**
 366 **it, which can be used to assess the correctness of her object**
 367 **grounding depending on which object the human actually**
 368 **intended to reference.**

371 4 EVALUATION

372 With the goal to enable bidirectional communication between ma-
 373 chines and humans using multimodal referring expressions as a
 374 testbed use case, specific evaluable properties must be enumerated
 375 to demonstrate where a fully-symmetrical system is more success-
 376 ful than one that maintains communicative asymmetry between the
 377 two interlocutors. The key research question with evaluation is: *do*
 378 *the metrics used clearly establish whether both interlocutors are able*
 379 *to extract the communicative intents of the others from their behav-*
 380 *ior?* Therefore, good metrics will answer if the non-verbal behavior
 381 generation methods used for an IVA is effectively contributing to
 382 the human interlocutor’s understanding, as defined as the ability
 383 to extract communicative intent from utterances and actions. We
 384 consider properties that are related to deictic and linguistic context
 385 awareness, as used in the evaluation of human-machine collabora-
 386 tion [21], and propose quantitative and qualitative metrics that
 387 assess the following properties of multimodal RE usage in a task-
 388 based environment: 1) efficient and collaborative task completion,
 389 2) software reliability and consistency, 3) ability of humans and
 390 machines to understand diverse communications, and 4) agent con-
 391 tribution of meaningful content. The version of the Diana system
 392 described above is presented to human subjects to collect samples
 393



394
395
396
397
398
399
400
401
402 **Figure 4: Diana’s facial expressions. (a) Confusion (e.g., undo-**
 403 **ing an action or responding to a negative acknowledgment).**
 404 **(b) Joy (e.g., welcoming users at the beginning of interactions**
 405 **or responding to a positive acknowledgment).**

407 of bidirectional collaborations and evaluate successful multimodal
 408 communication strategies for RE generation using both logged
 409 interactions and human judgments.
 410

411 4.1 Human-Machine Collaboration Data 412 Collection

413 During a single human-agent interaction session, the participant
 414 views 10 scenes containing 10 randomly-placed target objects to
 415 be referenced. Referencing is considered successful when Diana is
 416 able to ground the human’s MRE to the same object as the human
 417 intends to describe. The IVA’s and participant’s utterances, non-
 418 verbal behavior, and actions are logged (e.g., Fig. 5) for analysis and
 419 future training and evaluating of multimodal referring expression
 420 generation models.
 421

422 4.2 Evaluation Metrics

423 To evaluate the success of the IVA w.r.t. the key characteristics of
 424 human-machine collaboration from Sec. 4, we define 19 metrics as
 425 follows:
 426

- 427 (1) Multimodal Prompt Completion Efficiency (MPCE).
- 428 (2) Linguistic Prompt Completion Efficiency (LPCE).

429 The difference in target identification and the related task comple-
 430 tion times when using multimodal REs vs. verbal only REs indicates
 431 the increase in RE effectiveness when using multimodal generation
 432 vs. linguistic generation methods only.
 433

- 434 (3) Human-machine completion efficiency (HMCE): Time taken to
 435 complete the task. Since the task as a whole is normalized (an
 436 object referencing with 10 scenes each containing 10 objects),
 437 completion time can be directly related to referring strategies
 438 used by each interlocutor.
- 439 (4) Machine Appropriate Response Success Rate (MARSR): Rate
 440 of IVA responses to human prompts that are not followed by a
 441 negative response (e.g., no, nevermind).
- 442 (5) Proceed Without Reset (PWR): Rate of interactions that proceed
 443 without resets.
- 444 (6) Machine Interpretation of Human Communication (MIHC):
 445 Rate of correctly executed prompts.
- 446 (7) Machine Interpretation of Relational REs (MIRRE): Rate of cor-
 447 rectly executed relational prompts.
- 448 (8) Machine Interpretation of Historical REs (MIHRE): Rate of cor-
 449 rectly executed historical prompts.
- 450 (9) Human Interpretation Efficiency of Machine Communication
 451 (HIEMC): Time from generation of machine’s reference to target
 452 identification by human.
- 453 (10) Agent Pointing Success Rate (APSR): Rate of agent successfully
 454 pointing out the target object.
- 455 (11) Mutual Contribution Success Rate (MCSR): Difference between
 456 number of verbose human turns and verbose agent turns (“ver-
 457 bose” being defined as a meaningful contribution beyond posi-
 458 tive or negative acknowledgement or disambiguatory question—
 459 in our MRE use case this typically means a distinct referring
 460 expression).
- 461 (12) Machine-generated referring expressions (MGRE): Rate of machine-
 462 generated referring expressions compared to total utterances/
 463 discourse moves.
 464

- 465 (13) Recognition of Previously Mentioned Entities (RPME): Rate of
- 466 of previously mentioned entities grounded at the end of each
- 467 discourse move.
- 468 (14) Machine Historical Referencing Success (MHRS): Rate of histor-
- 469 ical references generated by the agent relative to total number
- 470 of generated REs.
- 471 (15) Machine Relational Referencing Success (MRRS): Rate of re-
- 472 lational references generated by the machine relative to total
- 473 number of generated REs.

474 The above metrics 1–15 are all calculated directly from data
 475 logged during human-agent interactions. The following metrics are
 476 collected *post facto* from the judgments of 3rd-party evaluators (see
 477 Sec. 6.1).

- 478 (16) Machine Object Identification Success Rate (MOISR): Rate of
- 479 correctly identified objects (by machine).
- 480 (17) Human Object Identification Success Rate (HOISR): Rate of
- 481 correctly identified objects (by humans).
- 482 (18) Machine References Fluency Rate (MRF): Rate of top-rated
- 483 machine references according to 3rd-party human judgments.
- 484 (19) Human References Fluency Rate (HRFR): Rate of the top-rated
- 485 human references according to 3rd-party human judgments.

486 In this paper, we include preliminary results for the following
 487 metrics: Multimodal Prompt Completion Efficiency (MPCE), Human
 488 Interpretation Efficiency of Machine Communication (HIEMC), and
 489 Agent Pointing Success Rate (APSR), in addition to the illustrations
 490 of generated referring expressions by each of the IVA and sub-
 491 ject, IVA’s ability to disambiguate, human’s ability to correct IVA’s
 492 misunderstanding, the impact of deictic gesture on interlocutors’
 493 understanding, and IVA’s dialogue history.

496 5 PRELIMINARY RESULTS

497 5.1 Automated Quantitative Evaluation

498 In a preliminary study, constituting the complete 10-scene inter-
 499 action with a sample test subject, we logged 330 different human
 500 referring expressions, including 141 pointing-only references for
 501 target object identification, 141 pointing-only references for target
 502 location identification, 33 multimodal REs, and 15 linguistic REs, as
 503 depicted in Fig. 6a. Linguistically, as shown in Fig. 6c, 84% REs are
 504 transitive attributive references (e.g., *move the red block to the plate*).
 505 Similarly, we logged 330 different machine referring expressions,
 506 including 141 pointing-only REs to the referents, 174 multimodal
 507 REs, and 15 linguistic REs, as depicted in Fig. 6b. Consequently, we
 508 used these logged data to obtain preliminary results regarding the
 509 ease of agent disambiguation, human recognition of agent intent
 510 from verbal and non-verbal behavior, and overall interaction.

511 In Fig. 5a, interlocutors’ moves, including actions, speech, and
 512 gestures, are logged with their timestamps. We see that the hu-
 513 man started pointing to the focus object (*BlueBlock1*) and moving
 514 it behind *YellowBlock1*. Logs also include the positions of each, dis-
 515 tance from agent to each, and the agent’s action after pointing to
 516 each of the two blocks. The human then used language only (“Pick
 517 up the yellow block”) to instruct Diana to pick up *YellowBlock2*.
 518 This instruction required Diana ask for disambiguation: “Which
 519 yellow block?”, as there are two yellow blocks in the scene. To
 520 disambiguate, the human uses pointing, and the object, its position,
 521 and distance are logged, along with Diana’s action. This illustrates
 522 Diana’s capability to clearly disambiguate the object the human
 523 referenced and efficiently execute the human’s prompt as shown in
 524 Fig. 7a and b, which leads to bidirectional communicative efficiency,
 525 with both human and agent combining verbal and non-verbal be-
 526 havior. When Diana has a misunderstanding, the human can correct
 527 it using language, deictic gesture, or both (Fig. 5b). Diana confirms
 528 that disambiguation was successful using deictic gesture to the
 529 correct object.

(a)	<pre> [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))) </pre>	<p>523</p> <p>524</p> <p>525</p> <p>526</p> <p>527</p> <p>528</p> <p>529</p> <p>530</p> <p>531</p> <p>532</p> <p>533</p> <p>534</p> <p>535</p>
(b)	<pre> [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. </pre>	<p>536</p> <p>537</p> <p>538</p> <p>539</p> <p>540</p> <p>541</p> <p>542</p> <p>543</p> <p>544</p> <p>545</p> <p>546</p> <p>547</p> <p>548</p> <p>549</p> <p>550</p> <p>551</p> <p>552</p> <p>553</p> <p>554</p> <p>555</p> <p>556</p> <p>557</p> <p>558</p> <p>559</p> <p>560</p> <p>561</p> <p>562</p> <p>563</p> <p>564</p> <p>565</p> <p>566</p> <p>567</p> <p>568</p> <p>569</p> <p>570</p> <p>571</p> <p>572</p> <p>573</p> <p>574</p> <p>575</p> <p>576</p> <p>577</p> <p>578</p> <p>579</p> <p>580</p>

Figure 5: (a) Trial sample of Diana’s ability to disambiguate the target; (b) Trial sample of human’s ability to correct misunderstanding.

and distance are logged, along with Diana’s action. This illustrates Diana’s capability to clearly disambiguate the object the human referenced and efficiently execute the human’s prompt as shown in Fig. 7a and b, which leads to bidirectional communicative efficiency, with both human and agent combining verbal and non-verbal behavior. When Diana has a misunderstanding, the human can correct it using language, deictic gesture, or both (Fig. 5b). Diana confirms that disambiguation was successful using deictic gesture to the correct object.

In human-human interactions, pointing reduces cognitive load [17]. Similarly, this is observed with the IVA as shown in the contingency table, Table 2. 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 disambiguation by pointing to the referents; nonetheless, using only speech for communication requires the agent to ask for additional information, i.e., gestures, to clearly identify the target and point to it as depicted in Fig. 7c. We see that a relationship exists between the modalities used and the level of ambiguity, such that use of pointing significantly reduces the ambiguity level of the prompt (p -value < 0.001 using Fisher’s exact test [14]).

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 [19]. Therefore, we integrated a dialogue history to the IVA. This stack stores all requested actions along with target objects, and accommodates interpretations of verbal, gestural, and multimodal inputs. Fig. 8 shows the number of actions in the dialogue history by the end of each scene in the preliminary data. These stored actions are available for use by both

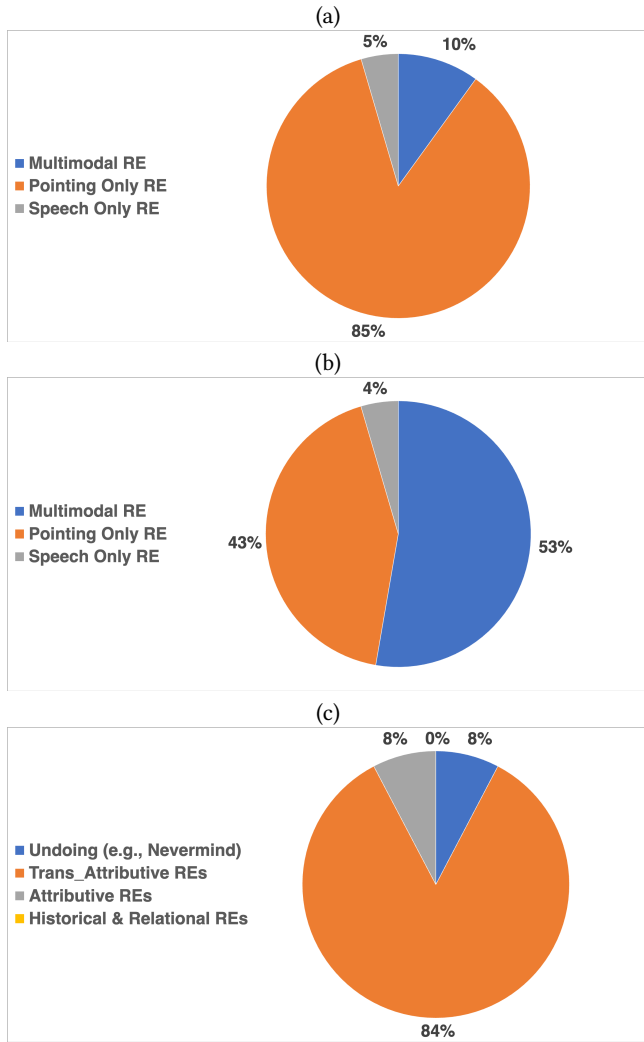


Figure 6: Preliminary results on (a) Human generated REs (b) Diana generated REs: categories and quantity (c) Categories of human verbal REs.

Table 2: Contingency table of human RE ambiguity and modalities used: # ambiguous REs by modality type

Modality	Did Agent Disambiguate?	
	No	Yes
Multimodal RE	15	0
Pointing Only RE	141	0
Speech Only RE	0	33
<i>p</i> -value	< 2.2e - 16	

humans and the IVA to refer to objects that may have previously been interacted with, as described in Sec. 3.1.

Table 3 shows how the IVA’s dialogue history is constructed and revisited to understand the human’s intents within a shared space.

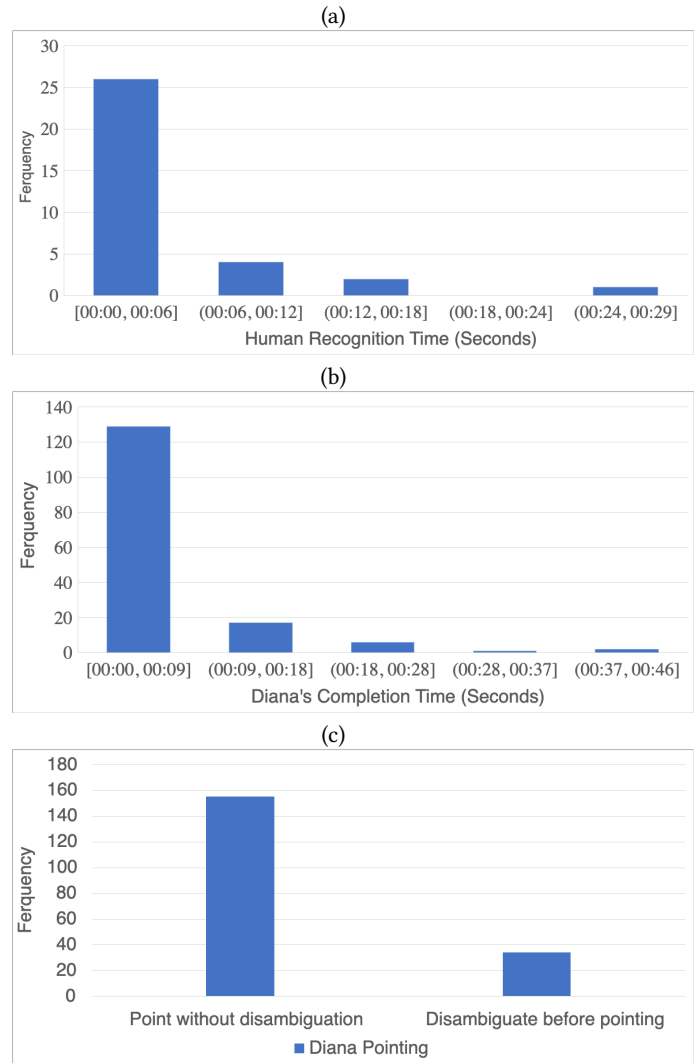
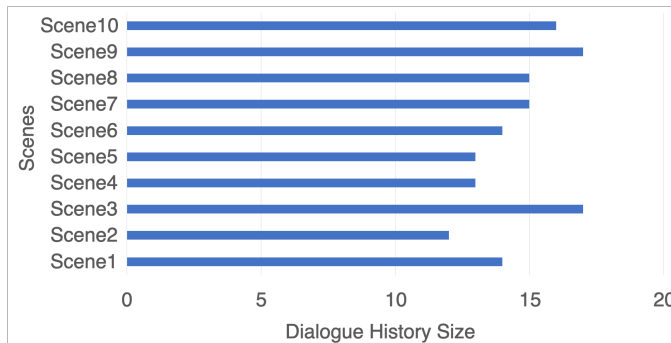


Figure 7: (a) Human Interpretation Efficiency of Machine Communication (Metric #3: HIEMC); (b) Multimodal Prompt Completion Efficiency (Metric #1: MPCE) by Diana; (c) Agent Pointing Success Rate (Metric #10: APSR).

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 3, #1–3. If the human uses a mention of a previously executed action to indicate an object as in Table 3, #4 (“grasp the block you just slid”), the IVA visits the dialogue history to 1) retrieve the most recently referenced object that is relevant to the provided action (in this case, *GreenBlock2*, as it satisfies the *adj_slid*(·) predicate), 2) push the new most recent action and referent onto the stack for future retrieval if necessary.

Table 3: 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 put put	GreenBlock2 RedBlock1 GreenBlock1
3	Multimodal	<i>slide(GreenBlock2; left(the(plate)))</i>	slide put put	GreenBlock2 RedBlock1 GreenBlock1
2	Pointing Only	<i>put(RedBlock1; left(the(plate)))</i>	put put	RedBlock1 GreenBlock1
1	Pointing Only	<i>put(GreenBlock1; < 0.5919505; 1.12487; -0.3801433 >)</i>	put	GreenBlock1

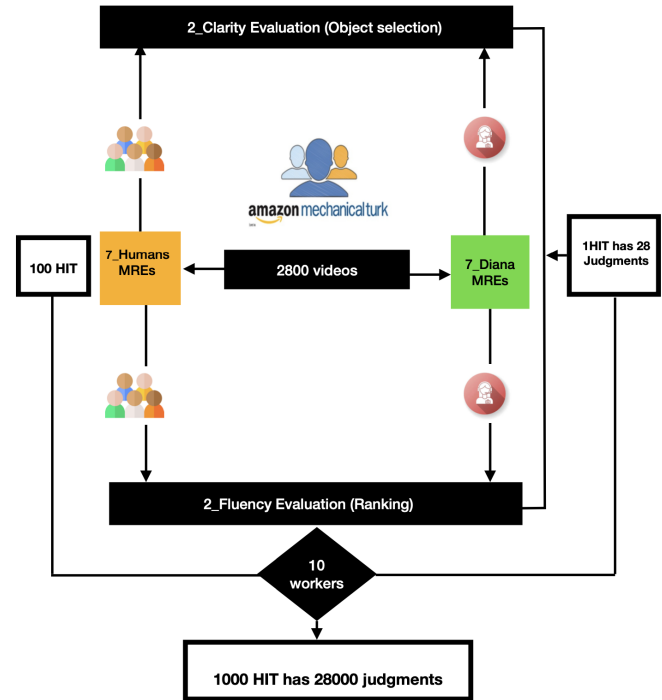
**Figure 8: IVA's dialogue history length at end of each scene.**

6 FUTURE EVALUATION

A larger study is preparation with a goal to collect data from roughly 150 participants who use REs of different types and strategies while collaborating with Diana to perform the task described above. Each participant views 10 scenes to refer to 10 randomly placed target objects, resulting in a total of 15,000 samples and recorded videos. Recorded video will consist of screen captures showing the human instructions as they are rendered in the scene, but direct video of the participants will not be collected. The gathered data will then be used to train generative models (e.g., fine-tuning an open-source large language model such as LLaMA [60] or similar) to produce contextually correct and situationally fluent REs that combine language and gesture. These REs will be evaluated according to the metrics discussed above, as well as human judgments as described below.

6.1 Human Evaluation

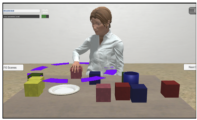
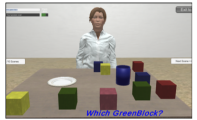
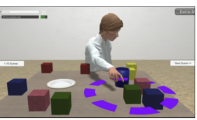
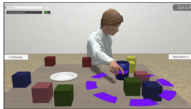
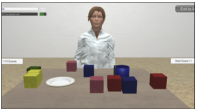
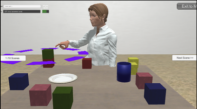

To evaluate the success of multimodal referring expression generation (MREG) models, two human-based experiments will be conducted using crowdsourcing platforms such as Amazon Mechanical Turk (AMT). We propose two primary criteria to assess how generative modules imbued with situational awareness and the ability to prompt non-verbal behavior could be compared with humans' generation capabilities. Criterion 1: how well the agent-generated strategies *qualitatively* compared to humans-generated strategies, as evaluated using a preference ordering method; Criterion 2: how well the agent-generated multimodal references *quantitatively* compared to humans-generated multimodal references, as evaluated using task completion. Fig. 9 shows the MREG evaluation framework including the design, participants and procedures.

**Figure 9: Crowdsourcing framework for evaluating multimodal referring expression generation models.**

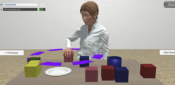



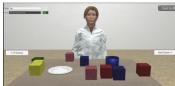


6.2 Study Design

Human MREs will be selected from the data gathered according to the strategy outlined in Sec. 4.1. These will be compared with REs generated by the virtual agent when driven by a generative model trained over the human data. A total of 2,800 videos (7 references \times 10 blocks \times 20 configurations \times 2 agents—human and Diana) will be collected. The 7 referencing strategies for each target object will use pointing only once, speech only three times, and a multimodal ensemble three times. This follows the pattern established for data collection in Krishnaswamy and Putejovsky's EMRE dataset [30] which allows for variability in the language used in linguistic or multimodal REs. Videos will be used in a set of AMT human intelligence tasks (HITs), where each HIT will involve workers rating 28 videos for *both* fluency and clarity, including 7 machine generated REs and 7 human REs, for a total of 100 HITs. Each HIT will be completed by 10 workers, for a total of 1,000 HITs and 28,000 individual judgments (2,000 for each individual RE in the dataset). Recruited

813 **How fluent and clear are these human based descriptions?**
 814 **Insert the arrow above the referent and rate the fluency from 1-5**
 815

816 Gesture only	816 Speech only Pick up the green block	816 Gesture + Speech Pick up this green block
		
817 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5	817 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5	817 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
823 Gesture + Speech Pick up the green block behind the cup	823 Speech only Pick up the blue block behind the cup	823 Gesture + Speech Pick up the block you just put down
		
824 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5	824 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5	824 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
831 Speech only Pick up the block you just put down		
		
832 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5		

839 **How fluent are these Diana based descriptions?**
 840 **Insert the arrow above the referent and rate the fluency from 1-5**

841 Gesture only Human: Points to pink block Diana: points to the block	841 Speech only Human: Pick up the green block Diana: which green block?	841 Gesture + Speech Human: Pick up the green block Diana: Is this the green block that you mean?
		
842 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5	842 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5	842 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
848 Gesture + Speech Human: Pick up the green block Diana: Do you mean this green block behind the cup?	848 Speech only Human: Pick up the blue block Diana: is it the one behind the cup?	848 Gesture + Speech Human: Pick up the green block Diana: this is what I just put down do you mean it?
		
849 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5	849 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5	849 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
854 Speech only Human: Pick up the green block Diana: Is it what I've just put down and on the red block?		
		
855 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5		

862 **Figure 10: Each set in the HIT includes two tasks for quantitative and qualitative evaluation of human REs and IVA REs.**

867 workers will be fluent English speakers between 18 and 60 years
 868 old and be given 15 minutes for each task while being compensated
 869 for their time via the platform.

871 Each HIT will require workers to evaluate 2 sets of 14 videos
 872 according to both the aforementioned criteria (Sec. 6.1). Each set
 873 will contain 7 videos of human REs and 7 of machine-generated REs.
 874 Workers will be informed whether the descriptions are generated
 875 by humans or by the embodied agent. As shown in Fig. 10, first
 876 participants will be asked to rate the “fluency” of each description
 877 in the video using a Likert-type scale (from 5—best—to 1—worst).
 878 Then they will be asked to locate the target object that is mentioned
 879 by the video, which will be compared to the actual object that was
 880 intended to be referenced, as stored in the dataset. This assesses
 881 the correctness of the referring expression: does a human listener
 882 correctly retrieve the object that was intended to be referenced,
 883 and how do verbal and non-verbal signals each contribute to the
 884 ability to correctly retrieve the object from the referring expression
 885 provided?

887 7 CONCLUSION

888 As interactive agents become more widespread in everyday use,
 889 developers will need principled ways of evaluating their behavior.
 890 Modern generative large language models already demand new
 891 methods of evaluation beyond metrics such as accuracy, precision,
 892 and recall on benchmark datasets. Factors such as fluency, reliability,
 893 correctability, and ease of use must be taken into account. This
 894 is doubly the case when non-linguistic modalities are involved, as
 895 would be the case with *embodied* IVAs. In this paper, we proposed
 896 a quantitative and qualitative evaluation framework to assess the
 897 quality of generated multimodal referring expressions, including
 898 language, gesture, and actions grounded in a shared virtual environ-
 899 ment. We developed an instance of an IVA for an object referencing
 900 task designed to elicit multimodal referring expressions from hu-
 901 man interlocutors and developed a set of metrics for evaluating
 902 the quality of referring expressions that apply equally to those pro-
 903 duced by both humans and humanoid IVAs using combined verbal
 904 and non-verbal information. We showed preliminary results from
 905 naive users of the experimental platform, and analyzed system out-
 906 puts based on a subset of our proposed metrics to showcase their
 907 utility for evaluating the contribution of non-verbal information
 908 toward bidirectional interpretation and disambiguation of definite
 909 descriptions of objects in context. We also detailed how our pre-
 910 liminary study will be expanded and scaled up. Our framework
 911 targets both timing and fluency of the interaction and proposes
 912 a set of qualitative and quantitative metrics that we hope will be
 913 beneficial for researchers in the IVA and multimodal interaction
 914 communities to assess dialogue and behavior generation strategies
 915 for multimodal interaction systems.

918 REFERENCES

919 [1] Julia Albath, Jennifer L Leopold, Chaman L Sabharwal, and Anne M Maglia. 2010.
 920 RCC-3D: Qualitative Spatial Reasoning in 3D. In *CAINE*. 74–79.
 921 [2] Satyanjee Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for
 922 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*. 65–72.
 923 [3] Anja Belz and Albert Gatt. 2008. Intrinsic vs. extrinsic evaluation measures for
 924 referring expression generation. In *Proceedings of ACL-08: HLT, Short Papers*.
 925 197–200.
 926 [4] Emily M Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning,
 927 form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5185–5198.

- [5] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snaveley, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12538–12547.
- [6] Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K Wong, and Qi Wu. 2020. 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*. 10086–10095.
- [7] Herbert H Clark, Robert Schreuder, and Samuel Buttrick. 1983. Common ground at the understanding of demonstrative reference. *Journal of verbal learning and verbal behavior* 22, 2 (1983), 245–258.
- [8] Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive science* 19, 2 (1995), 233–263.
- [9] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multimodal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5503–5512.
- [10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- [11] Lee R Dice. 1945. Measures of the amount of ecologic association between species. *Ecology* 26, 3 (1945), 297–302.
- [12] Fethiye Irmak Doğan, Sinan Kalkan, and Iolanda Leite. 2019. Learning to generate unambiguous spatial referring expressions for real-world environments. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 4992–4999.
- [13] Rui Fang, Malcolm Doering, and Joyce Y Chai. 2015. Embodied collaborative referring expression generation in situated human-robot interaction. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. 271–278.
- [14] Ronald Aylmer Fisher et al. 1936. Statistical methods for research workers. *Statistical methods for research workers*. 6th Ed (1936).
- [15] Albert Gatt, Anja Belz, and Eric Kow. 2009. The TUNA-REG Challenge 2009: Overview and evaluation results. Association for Computational Linguistics.
- [16] Albert Gatt and Kees Van Deemter. 2007. Lexical choice and conceptual perspective in the generation of plural referring expressions. *Journal of Logic, Language and Information* 16, 4 (2007), 423–443.
- [17] Susan Goldin-Meadow. 1999. The role of gesture in communication and thinking. *Trends in cognitive sciences* 3, 11 (1999), 419–429.
- [18] Peter Gorniak and Deb Roy. 2004. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research* 21 (2004), 429–470.
- [19] Herbert P Grice. 1975. Logic and conversation. In *Speech acts*. Brill, 41–58.
- [20] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 787–798.
- [21] Robyn Kozierek, John Aberdeen, Cheryl Clark, Christopher Garay, Bradley Goodman, Tonia Korves, Lynette Hirschman, Patricia L McDermott, and Matthew W Peterson. 2021. Assessing open-ended human-computer collaboration systems: applying a hallmarks approach. *Frontiers in artificial intelligence* 4 (2021), 670009.
- [22] Emiel Krahmer and Ielka van der Sluis. 2003. A new model for generating multimodal referring expressions. In *Proceedings of the ENLG*, Vol. 3. 47–54.
- [23] Alfred Kranstedt, Stefan Kopp, and Ipke Wachsmuth. 2002. Murml: A multimodal utterance representation markup language for conversational agents. In *AAMAS'02 Workshop Embodied conversational agents-let's specify and evaluate them!*
- [24] Alfred Kranstedt, Andy Lücking, Thies Pfeiffer, Hannes Rieser, and Ipke Wachsmuth. 2006. 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*. Springer, 300–311.
- [25] Nikhil Krishnaswamy and Nada Alalayani. 2021. Embodied Multimodal Agents to Bridge the Understanding Gap. In *Proceedings of the First Workshop on Bridging Human-Computer Interaction and Natural Language Processing*. Association for Computational Linguistics, Online, 41–46. <https://aclanthology.org/2021.hcinlp-1.7>
- [26] Nikhil Krishnaswamy, Pradyumna Narayana, Rahul Bangar, Kyeongmin Rim, Dhruva Patil, David McNeely-White, Jaime Ruiz, Bruce Draper, Ross Beveridge, and James Pustejovsky. 2020. Diana's World: A Situated Multimodal Interactive Agent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13618–13619.
- [27] Nikhil Krishnaswamy, William Pickard, Brittany Cates, Nathaniel Blanchard, and James Pustejovsky. 2022. The VoxWorld platform for multimodal embodied agents. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 1529–1541.
- [28] Nikhil Krishnaswamy and James Pustejovsky. 2016. VoxSim: A Visual Platform for Modeling Motion Language. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. ACL.
- [29] Nikhil Krishnaswamy and James Pustejovsky. 2018. An evaluation framework for multimodal interaction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- [30] Nikhil Krishnaswamy and James Pustejovsky. 2019. Generating a novel dataset of multimodal referring expressions. In *Proceedings of the 13th International Conference on Computational Semantics-Short Papers*. 44–51.
- [31] Nikhil Krishnaswamy and James Pustejovsky. 2021. The Role of Embodiment and Simulation in Evaluating HCI: Experiments and Evaluation. In *International Conference on Human-Computer Interaction*. 220–232.
- [32] Lars Kunze, Tom Williams, Nick Hawes, and Matthias Scheutz. 2017. Spatial referring expression generation for hri: Algorithms and evaluation framework. In *2017 AAAI Fall Symposium Series*.
- [33] Kiyong Lee, Nikhil Krishnaswamy, and James Pustejovsky. 2023. An Abstract Specification of VoxML as an Annotation Language. In *Workshop on Interoperable Semantic Annotation (ISA-19)*. 66.
- [34] Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, Vol. 10. Soviet Union, 707–710.
- [35] Xinghang Li, Di Guo, Huaping Liu, and Fuchun Sun. 2022. ReVe-ce: Remote embodied visual referring expression in continuous environment. *IEEE Robotics and Automation Letters* 7, 2 (2022), 1494–1501.
- [36] Chin-Yew Lin and Eduard Hovy. 2003. 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*. 150–157.
- [37] Aly Magassouba, Komei Sugiura, and Hisashi Kawai. 2020. Multimodal attention branch network for perspective-free sentence generation. In *Conference on Robot Learning*. PMLR, 76–85.
- [38] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Baltimore, Maryland, 55–60. <https://doi.org/10.3115/v1/P14-5010>
- [39] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 11–20.
- [40] David G McNeely-White, Francisco R Ortega, J Ross Beveridge, Bruce A Draper, Rahul Bangar, Dhruva Patil, James Pustejovsky, Nikhil Krishnaswamy, Kyeongmin Rim, Jaime Ruiz, et al. 2019. User-aware shared perception for embodied agents. In *2019 IEEE International Conference on Humanized Computing and Communication (HCC)*. IEEE, 46–51.
- [41] David McNeill. 1985. So you think gestures are nonverbal? *Psychological review* 92, 3 (1985), 350.
- [42] Alessandro Moschitti, Bo Pang, and Walter Daelemans. 2014. 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)*.
- [43] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.
- [44] Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. (2006).
- [45] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*. 2641–2649.
- [46] James Pustejovsky and Nikhil Krishnaswamy. 2016. VoxML: A Visualization Modeling Language. *Proceedings of LREC* (2016).
- [47] James Pustejovsky and Nikhil Krishnaswamy. 2020. Embodied human-computer interactions through situated grounding. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–3.
- [48] James Pustejovsky and Nikhil Krishnaswamy. 2021. Embodied human computer interaction. *KI-Künstliche Intelligenz* 35, 3-4 (2021), 307–327.
- [49] James Pustejovsky and Nikhil Krishnaswamy. 2022. Multimodal semantics for affordances and actions. In *International Conference on Human-Computer Interaction*. Springer, 137–160.
- [50] James Pustejovsky, Nikhil Krishnaswamy, and Tuan Do. 2017. Object Embodiment in a Multimodal Simulation. In *AAAI Spring Symposium: Interactive Multi-sensory Object Perception for Embodied Agents*.
- [51] James Pustejovsky, Nikhil Krishnaswamy, Bruce Draper, Pradyumna Narayana, and Rahul Bangar. 2017. Creating common ground through multimodal simulations. In *Proceedings of the IWCS workshop on Foundations of Situated and Multimodal Communication*.

- 1045 [52] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua
1046 Shen, and Anton van den Hengel. 2020. Reverie: Remote embodied visual refer-
1047 ring expression in real indoor environments. In *Proceedings of the IEEE/CVF*
1048 *Conference on Computer Vision and Pattern Recognition*. 9982–9991.
- 1049 [53] Francis Quek, David McNeill, Robert Bryll, Susan Duncan, Xin-Feng Ma, Cemil
1050 Kirbas, Karl E McCullough, and Rashid Ansari. 2002. Multimodal human dis-
1051 course: gesture and speech. *ACM Transactions on Computer-Human Interaction*
1052 *(TOCHI)* 9, 3 (2002), 171–193.
- 1053 [54] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya
1054 Sutskever, et al. 2019. Language models are unsupervised multitask learners.
1055 *OpenAI blog* 1, 8 (2019), 9.
- 1056 [55] Boris Schauerte and Gernot A Fink. 2010. Focusing computational visual at-
1057 tention in multi-modal human-robot interaction. In *International conference on*
1058 *multimodal interfaces and the workshop on machine learning for multimodal*
1059 *interaction*. 1–8.
- 1060 [56] Boris Schauerte, Jan Richarz, and Gernot A Fink. 2010. Saliency-based identi-
1061 fication and recognition of pointed-at objects. In *2010 IEEE/RSJ International*
1062 *Conference on Intelligent Robots and Systems*. IEEE, 4638–4643.
- 1063 [57] Mohit Shridhar, Dixant Mittal, and David Hsu. 2020. INGRESS: Interactive visual
1064 grounding of referring expressions. *The International Journal of Robotics Research*
1065 39, 2-3 (2020), 217–232.
- 1066 [58] Dadhichi Shukla, Ozgur Erkent, and Justus Piater. 2015. Probabilistic detection of
1067 pointing directions for human-robot interaction. In *2015 international conference*
1068 *on digital image computing: techniques and applications (DICTA)*. IEEE, 1–8.
- 1069 [59] Dadhichi Shukla, Özgür Erkent, and Justus Piater. 2016. A multi-view hand
1070 gesture rgb-d dataset for human-robot interaction scenarios. In *2016 25th IEEE*
1071 *international symposium on robot and human interactive communication (RO-*
1072 *MAN)*. IEEE, 1084–1091.
- 1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
- [60] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne
Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal
Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [61] Kees Van Deemter. 2006. Generating referring expressions that involve gradable
properties. *Computational Linguistics* 32, 2 (2006), 195–222.
- [62] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider:
Consensus-based image description evaluation. In *Proceedings of the IEEE confer-*
ence on computer vision and pattern recognition. 4566–4575.
- [63] Jette Viethen and Robert Dale. 2006. Algorithms for generating referring ex-
pressions: do they do what people do?. In *Proceedings of the fourth international*
natural language generation conference. 63–70.
- [64] Jette Viethen and Robert Dale. 2008. The use of spatial relations in referring
expression generation. In *Proceedings of the Fifth International Natural Language*
Generation Conference. 59–67.
- [65] Isaac Wang, Mohtadi Ben Fraj, Pradyumna Narayana, Dhruva Patil, Gururaj
Mulay, Rahul Bangar, J Ross Beveridge, Bruce A Draper, and Jaime Ruiz. 2017.
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). IEEE, 414–421.
- [66] Isaac Wang, Jesse Smith, and Jaime Ruiz. 2019. Exploring virtual agents for
augmented reality. In *Proceedings of the 2019 CHI Conference on Human Factors*
in Computing Systems. 1–12.
- [67] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg.
2016. Modeling context in referring expressions. In *European Conference on*
Computer Vision. Springer, 69–85.

Received 21 July 2023