YouRefIt: Embodied Reference Understanding with Language and Gesture

Yixin Chen[1], Qing Li[1], Deqian Kong[1], Yik Lun Kei[1],
Song-Chun Zhu[2,3,4], Tao Gao[1], Yixin Zhu[2,3], Siyuan Huang[1]

[1] University of California, Los Angeles [2] Beijing Institute for General Artificial Intelligence
[3] Peking University [4] Tsinghua University

https://yixchen.github.io/YouRefIt

Abstract

We study the machine's understanding of ***embodied reference****: One agent uses both language and gesture to refer to an object to another agent in a shared physical environment. Of note, this new visual task requires understanding multimodal cues with perspective-taking to identify which object is being referred to. To tackle this problem, we introduce* ***YouRefIt****, a new crowd-sourced dataset of embodied reference collected in various physical scenes; the dataset contains 4,195 unique reference clips in 432 indoor scenes. To the best of our knowledge, this is the first embodied reference dataset that allows us to study referring expressions in daily physical scenes to understand referential behavior, human communication, and human-robot interaction. We further devise two benchmarks for image-based and video-based embodied reference understanding. Comprehensive baselines and extensive experiments provide the very first result of machine perception on how the referring expressions and gestures affect the embodied reference understanding. Our results provide essential evidence that gestural cues are as critical as language cues in understanding the embodied reference.*

1. Introduction

Human communication [51] relies heavily on establishing common ground [50, 48] by referring to objects in a shared environment. This process usually takes place in two forms: language (abstract symbolic code) and gesture (unconventionalized and uncoded). In the computer vision community, efforts of understanding reference have been primarily devoted in the first form through an artificial task, Referring Expression Comprehension (REF) [64, 17, 63, 32, 60, 56, 57], which localizes a particular object in an image with a natural language expression generated by the annotator. Evidently, the second form, gesture, has been left almost untouched. Yet, this nonverbal (gesture) form is more profound in the communication literature compared to the pure verbal (language) form with ample evolutionary evidence [1, 36, 14]; it is deeply rooted in human cognition development [29, 30] and learning process [7], and tightly coupled with the language development [23, 6, 18].

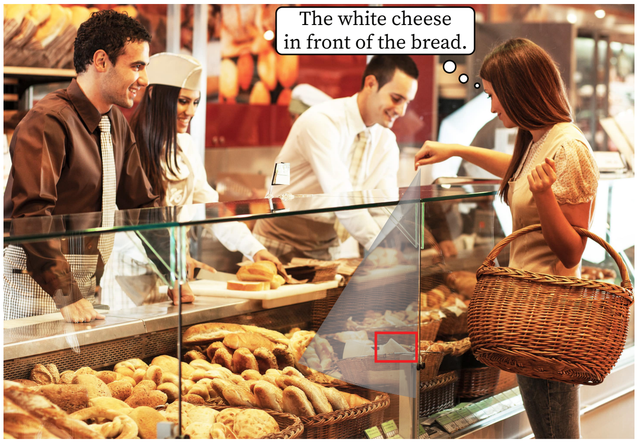


Figure 1: Imagine you walk into a bakery for your favorite. To precisely express your intent, you point to it and produce an utterance "the white cheese in front of the bread." This multimodal communicative act is produced by assuming it can be properly understood by the staff, whose embodiment differs in the shared physical environment. Such a daily deictic-interaction scenario illustrates the significance of visual perspective-taking in embodied reference.

Fundamentally, most modern literature deviates from the natural setting of reference understanding in daily scenes, which is **embodied**: An agent refers to an object to another in a *shared* physical space [39, 54, 11], as exemplified by Fig. 1. Embodied reference possesses two distinctive characteristics compared to REF. First, it is **multimodal**. People often use both natural language and gestures when referring to an object. The gestural component and language component are semantically coherent and temporally synchronous to coordinate with one another, creating a concise and vivid message [22] while elucidating the overloaded meaning if only one modality is presented [20]. Second, recognizing embodied reference requires visual **perspective-taking** [25, 2, 39], the awareness that others see things from different viewpoints and the ability to imagine what others see from their perspectives. It requires both the message sender and receiver to comprehend the immediate environments [11], including the relationship between the interlocutors and the relationships between objects, in the shared perceptual fields for effective communication.

To address the deficiencies in prior work and study reference understanding at a full spectrum, we introduce a new dataset, ***YouRefIt***, for embodied reference understanding. The reference instances in *YouRefIt* are crowd-sourced with diverse physical scenes from Amazon Mechanic Turk (AMT). Participants are instructed to film videos in which they reference objects in a scene to an imagined person (*i.e.*, a mounted camera) using both language and gestures. Minimum requirements of the scenes, objects, and words are imposed to ensure the naturalness and the variety of collected videos. Videos are segmented into short clips, with each clip containing an exact one reference instance. For each clip, we annotate the reference target (object) with a bounding box. We also identify **canonical frames** in a clip: They are the "keyframes" of the clip and contain sufficient information of the scene, human gestures, and referenced objects that can truthfully represent the reference instance. Fine-grained semantic parsing of the transcribed sentences is further annotated to support a detailed understanding of the sentences. In total, the *YouRefIt* dataset includes 4,195 embodied reference instances from 432 indoor scenes.

To measure the machine's ability in Embodied Reference Understanding (ERU), we devise two benchmarks on top of the proposed *YouRefIt* dataset. (i) **Image ERU** takes a canonical frame and the transcribed sentence of the reference instance as the inputs and predicts the bounding box of the referenced object. Image ERU adopts the settings from the well-studied REF but is inherently more challenging and holistic due to its requirement on a joint and coherent understanding of human gestures, natural language, and objects in the context of human communication. (ii) **Video ERU** takes the video clip and the sentence as the input, identifies the canonical frames, and locates the reference target within the clip. Compared to Image ERU, Video ERU takes one step further and manifests the most natural human-robot communication process that requires distinguishing the initiation, the canonical frames, and the ending of a reference act while estimating the reference target in a temporal order.

Incorporating both language and gestural cues, we formulate a new multimodal framework to tackle the ERU tasks. In experiments, we provide multiple baselines and ablations. Our results reveal that models with explicit gestural cues yield better performance, validating our hypothesis that gestural cues are as critical as language cues in resolving ambiguities and overloaded semantics with cooperation (perspective-taking) in mind [20, 19, 39, 65, 68], echoing a recent finding in the embodied navigation task [54]. We further verify that temporal cues are essential in canonical frame detection, necessitating understanding embodied reference in dynamic and natural sequences.

This paper makes three major contributions. (i) We collect the first video dataset in physical scenes, *YouRefIt*, to study the reference understanding in an *embodied* fashion. We argue this is a more natural setting than prior work and, therefore, further understanding human communications and multimodal behavior. (ii) We devise two benchmarks, Image ERU and Video ERU, as the protocols to study and evaluate the embodied reference understanding. (iii) We propose a multimodal framework for ERU tasks with multiple baselines and model variants. The experimental results confirm the significance of the joint understanding of language and gestures in embodied reference.

2. Related Work

Our work is related to two topics in modern literature: (i) Referring Expression Comprehension (REF) studied in the context of Vision and Language, and (ii) reference recognition in the field of Human-Robot Interaction. Below, we compare our work with prior arts on these two topics.

2.1. Referring Expression Comprehension (REF)

REF is a visual grounding task. Given a natural language expression, it requires an algorithm to locate a particular object in a scene. Several datasets, including both images of physical scenes [21, 64, 35, 38, 8, 4, 5] and synthetic images [31], have been constructed by asking annotators or algorithms to provide utterances describing regions of images. To solve REF, researchers have attempted various approaches [60, 32, 56, 57]. Representative methods include (i) localizing a region by reconstructing the sentence using an attention mechanism [42], (ii) incorporating contextual information to ground referring expressions [66, 64], (iii) using neural modular networks to better capture the structured semantics in sentences [17, 63], and (iv) devising a one-stage approach [59, 58]. In comparison, our work fundamentally differs from REF at two levels.

Task-level REF primarily focuses on building correspondence between visual cues and verbal cues (natural language). In comparison, the proposed ERU task mimics the minimal human communication process in an embodied manner, which requires a mutual understanding of both verbal and nonverbal messages signaled by the sender. Recognizing references in an embodied setting also introduces new challenges, such as visual perspective-taking [12]: The referrers need to consider the perception from the counterpart's perspective for effective verbal and nonverbal communication, requiring a more holistic visual scene understanding both geometrically and semantically. In this paper, to study the reference understanding that echoes the above characteristics, we collect a new dataset containing natural reference scenarios with both language and gestures.

Model-level Since previous REF approaches are only capable of comprehending communicative messages in the form of natural language and mostly ignore the gestural cues, it is insufficient in the ERU setting or to be applied in our newly collected dataset. To tackle this deficiency, we design a principled framework to combine verbal (natural language) and nonverbal (gestures) cues. The proposed framework outperforms prior single-modality methods, validating the significant role of the gestural cue in addition to the language cue in embodied reference understanding.

Table 1: **Comparisons between the proposed *YouRefIt* and other reference datasets. Lang.** and **Gest.** denote whether language or gesture is used when referring to objects, and **Embo.** denotes whether referrers are embodied in the scenes where reference happens.

Datasets	**Lang.**	**Gest.**	**Embo.**	**Type**	**Source**	**No. of images**	**No. of instances**	**No. of object categories**	**Ave. sent. length**
PointAt [44]	✗	✓	✓	image	lab	220	220	28	-
ReferAt [43]	✓	✓	✓	video	lab	-	242	28	-
IPO [46]	✗	✓	✓	image	lab	278	278	10	-
IMHF [47]	✗	✓	✓	image	lab	1716	1,716	-	-
RefIt [21]	✓	✗	✗	image	image CLEF	19,894	130,525	238	3.61
RefCOCO [64]	✓	✗	✗	image	MSCOCO	19,994	142,209	80	3.61
RefCOCO+ [64]	✓	✗	✗	image	MSCOCO	19,992	141,564	80	3.53
RefCOCOg [35]	✓	✗	✗	image	MSCOCO	26,711	104,560	80	8.43
Flickr30k entities [38]	✓	✗	✗	image	Flickr30K	31,783	158,915	44,518	-
GuessWhat? [8]	✓	✗	✗	image	MSCOCO	66,537	155,280	-	-
Cops-Ref [4]	✓	✗	✗	image	COCO/Flickr	75,299	148,712	508	14.40
CLEVR-Ref+ [31]	✓	✗	✗	image	CLEVR	99,992	998,743	3	22.40
YouRefIt	✓	✓	✓	video	crowd-sourced	497,348	4,195	395	3.73

2.2. Reference in Human-Robot Interaction

The combination of verbal and nonverbal communication for reference is one of the central topics in Human-Robot Interaction. Compared with REF, this line of work focuses on more natural settings but with specialized scenarios. One stream of work emphasizes pointing direction and thus are not object-centric while missing language reference: The Innsbruck Pointing at Objects dataset [46] investigates two types of pointing gestures with index finger and tool, and the Innsbruck Multi-View Hand Gesture Dataset [47] records hand gestures in the context of human-robot interaction in close proximity. The most relevant prior arts are ReferAt [43] and PointAt [44], wherein participants are tasked to point at various objects with or without linguistic utterance. Some other notable literature includes (i) a robotics system that allows users to combine natural language and pointing gestures to refer to objects on a display [24], (ii) experiments that investigate the semantics and pragmatics of co-verbal pointing through computer simulation [33], (iii) deictic interaction with a robot when referring to a region using pointing and spatial deixis [15], and (iv) effects of various referential strategies, including talk-gesture-coordination and handshape, for robots interacting with humans when guiding attentions in museums [37].

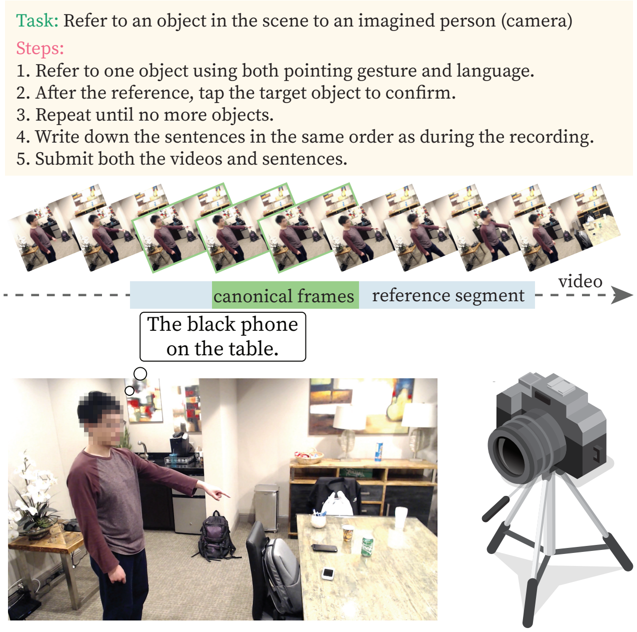
Although related, the above literature is constrained in lab settings with limited sizes, scenarios, and expressions, thus insufficient for solving the reference understanding in natural, physical scenarios with both vision and language. In comparison, crowd-sourced by AMT, our dataset is much more diverse in environment setting, scene appearance, and types of utterance. Our dataset also collects videos instead of static images commonly used in prior datasets, opening new venues to study dynamic and evolutionary patterns that occurred during natural human communications.

3. The *YouRefIt* Dataset

To study the embodied reference understanding, we introduce a new dataset named *YouRefIt*, a video collection of people referring to objects with both natural language and gesture in indoor scenes. Table 1 tabulates a detailed comparison between *YouRefIt* against twelve existing reference understanding datasets. Compared to existing datasets collected either in laboratories or from the Internet (MSCOCO/Flickr) or simulators (CLEVR), *YouRefIt* has a clear distinction: It contains videos crowd-sourced by AMT, and thus the reference happens in a more natural setting with richer diversity. Compared with the datasets on referring expression comprehension, the referrers (human) and the receivers (camera) in our dataset share the same physical environment, with both language and gesture allowed for referring to objects; the algorithm ought to understand from an embodiment perspective to tackle this problem. Next, we discuss the data collection and annotation process details, followed by a comprehensive analysis.

3.1. Data Collection

Our dataset was collected via AMT; see the data collection process in Fig. 2. Workers were asked to record a video containing actions of referring to objects in the scene to an imagined person (*i.e.*, the camera) using both natural languages (sentences) and pointing gestures. Most videos were collected in indoor scenes, such as offices, kitchens, and living rooms. Unlike existing datasets in which objects are usually put on a table with a clean background, all the objects in our collected videos were placed at their natural positions. Each video also included more than ten objects in the scene to avoid trivial scenarios and increase the reference difficulty. The camera was set up such that the referrer and all referred objects are within the field of view.



Task: Refer to an object in the scene to an imagined person (camera)

Steps:

1. Refer to one object using both pointing gesture and language.
2. After the reference, tap the target object to confirm.
3. Repeat until no more objects.
4. Write down the sentences in the same order as during the recording.
5. Submit both the videos and sentences.

Figure 2: **Illustration of the dataset collection procedure.** Participants were asked to film a series of reference tasks to an imaged person (*i.e.*, the camera) following the instructions.

When referring to a specific object, participants were instructed to use arbitrary natural languages and gestures freely. However, they were also required to avoid potential ambiguities, such that the observer would be able to uniquely identify the referred object by merely observing the reference behaviors. After reference actions were finished, participants were instructed to tap the referred object; this extra step helps annotate the referred target. In addition to the voices recorded in the video, participants were also asked to write down the sentences after the recording.

3.2. Data Annotation

The annotation process takes two stages: (i) annotation of temporal segments, canonical frames, and referent bounding boxes, and (ii) annotation of sentence parsing. Please refer to the *supplementary material* for more details of the data post-processing and annotation process.

Segments Since each collected video consists of multiple reference actions, we first segment the video into clips; each contains an exact one reference action. A segment is defined from the start of gesture movement or utterance to the end of the reference, which typically includes the raise of hand and arm, pointing action, and reset process, synchronized with its corresponding language description.

Canonical Frames In each segment, the annotators were asked to annotate further the canonical moments, which contain the "keyframes" that the referrer holds the steady pose to indicate what is being referred clearly. Combined with natural language, it is sufficient to use any canonical frame to localize the referred target.

Bounding Boxes Recall that participants were instructed to tap the referred objects after each reference action. Using this information, bounding boxes of the referred objects were annotated using Vatic [53], and the tapping actions were discarded. The object color and material were also annotated if identifiable. The taxonomy of object color and material is adopted from Visual Genome dataset [26].

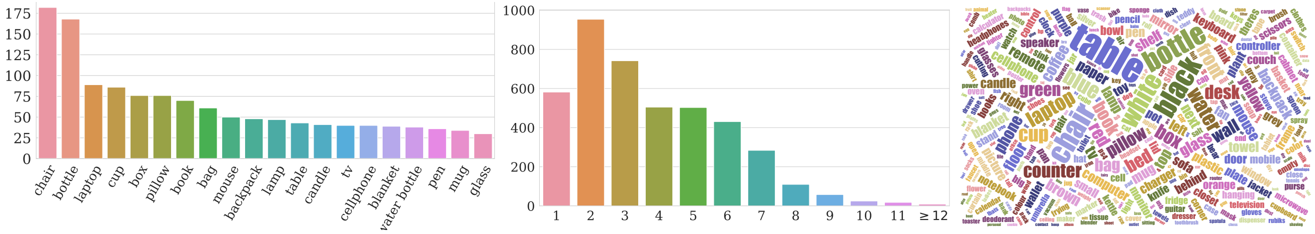
Sentence Parsing Given the sentence provided by the participants who performed reference actions, AMT annotators were asked to refine the sentence further and ensure it matches the raw audio collected from the video. We further provided more fine-grained parsing results of the sentence for natural language understanding. AMT annotators annotated target, target-attribute, spatial-relation, and comparative-relation. Take "The largest red bottle on the table" as an example: "the bottle" will be annotated as the target, "red" as target-attribute, "on the table" as spatial-relation, and "largest" as comparative-relation. For each relation, we further divided them into "relation" (*e.g.*, "on") and "relation-target" (*e.g.*, "the table").

3.3. Dataset Statistics

In total, *YouRefIt* includes 432 recorded videos and 4,195 localized reference clips with 395 object categories. We retrieved 8.83 hours of video during the post-processing and annotated 497,348 frames. The total duration of all the reference actions is 3.35 hours, with an average duration of 2.81 seconds per reference. Each reference process was annotated with segments, canonical frames, bounding boxes of the referred objects, and sentences with semantic parsing. All videos were collected with synchronized audio. We also included the body poses and hand keypoints of the participants extracted by the OpenPose [3].

Object Categories Fig. 3a shows the frequencies of the top-20 referred object categories, which roughly follow the Zipf's law [69]. Since most videos were shot in indoor scenes, the most frequently referred are daily objects, such as "chair," "bottle," and "cup."

Reference Sentence Fig. 3c shows the word cloud of sentences after removing the stop words. Interestingly, the most frequent word is "table," which is not even in the top-5 referred objects. A further inspection implies that the "table" is the most frequently used relational object while referring to objects by natural languages. Fig. 3b shows the distribution of sentence lengths with an average of 3.73. We observe that the sentences in *YouRefIt* are much shorter than those of language-only reference datasets (*e.g.*, 8.43 for RefCOCOg and 14.4 for Cops-Ref). This discrepancy implies that while naturally referring to objects, humans prefer a multimodal communication pattern that combines gestures with fewer words (compared to using a single modality) to minimize the cognitive load [49].



(a) The frequency of the top-20 referred objects.

(b) The distribution of sentence lengths.

(c) The language wordle.

Figure 3: Statistics of the *YouRefIt* dataset.

4. Embodied Reference Understanding (ERU)

In this section, we benchmark two tasks of embodied reference understanding on the *YouRefIt* dataset, namely, Image ERU and Video ERU. The first benchmark evaluates the performance of understanding embodied reference based on the canonical frame, whereas the second benchmark emphasizes how to effectively recognize the canonical moments and reference targets simultaneously in a video sequence. Below, we describe the detailed settings, baselines, analyses, and ablative studies in the experiments.

Dataset Splits We randomly split the dataset into the training and test sets with a ratio of 7:3, resulting in 2,950 instances for training and 1,245 instances for testing.

4.1. Image ERU

Given the canonical frame and the sentence from an embodied reference instance, Image ERU aims at locating the referred object in the image through both the human language and gestural cues.

Experimental Setup and Evaluation Protocol For each reference instance, we randomly pick one frame from the annotated canonical frames. We adopt the evaluation protocol similar to the one presented in Mao *et al*. [35]: (i) predict the region referred by the given image and sentence, (ii) compute the Intersection over Union (IoU) ratio between the ground-truth and the predicted bounding box, and (iii) count it as correct if the IoU is larger; otherwise wrong. We use accuracy as the evaluation metric. Following object detection benchmark [13], we report the results under three different IoUs: 0.25, 0.5, and 0.75.

We also evaluate on subsets with various object sizes, *i.e*., *small*, *medium* and *large*. Object size is estimated using the ratio between the area of the ground-truth object bounding box and the area of the image. The size thresholds are 0.48% and 1.76% based on the size distribution in the dataset; see the size distribution in *supplementary material*.

Methods We devise a novel multimodal framework for Image ERU that leverages both the language and gestural cues; see Fig. 4. At a high level, our framework includes both the visual and language encoder, similar to prior REF models [59, 58, 34], as well as explicitly extracted gesture features. We utilize the features from three modalities to effectively predict the target bounding box.

Specifically, we use Darknet-53 [40] pre-trained on COCO object detection [28] as the visual encoder. The textual encoder is the uncased base version of BERT [9] followed by two fully connected layers. We incorporate two types of gestural features: (i) the Part Affinity Field (PAF) [3] heatmap, and (ii) the pointing saliency heatmap. Inspired by the visual saliency prediction, we train MSI-Net [27] on the *YouRefIt* dataset to predict the salient regions by considering both the latent scene structure and the gestural cues, generating more accurate guidance compared to the commonly used Region of Interests (RoIs); see some examples of predicted salient regions in Fig. 5. We aggregate the visual feature and PAF heatmaps by max-pooling and concatenation, fusing them with textual features by updating text-conditional visual features attended to different words through a sub-query module [58]. Following convolution blocks, the saliency map feature is concatenated with the text-conditional visual feature as the high-level guidance to predict anchor boxes and confidence scores; we use the same classification and regression loss as in Yang *et al*. [59] for anchor-based bounding box prediction.

Baselines and Ablations We first evaluate the Image ERU performance on FAOA [59] and ReSC [58], originally designed for the REF task. We also design baselines to test the gestural cues in a two-stage architecture, similar to MAttNet [63]. We generate the RoIss by Region Proposal Network from Faster R-CNN [41] pre-trained on the MSCOCO dataset. To score the object proposal, we test two categories of heatmaps that reflect the gestural cues. (i) By pointing heatmap from the primary pointing direction characterized by arm, hand, and index finger. Following Fan *et al*. [10], we generate the pointing heatmap by a Gaussian distribution to model the variation of a pointing ray w.r.t. the primary pointing direction. We choose 15° and 30° as the standard deviations (*i.e*., $\text{RPN}_{\text{pointing15}}$ and $\text{RPN}_{\text{pointing30}}$). (ii) By pointing saliency map (*i.e*., $\text{RPN}_{\text{saliency}}$). The scores are computed according to the heatmap of average density.

We design ablation studies from two aspects: data and architecture. For the **data-wise** ablation, we first evaluate the MattNet, FAOA, and ReSC models pre-trained on the REF datasets RefCOCO, RefCOCO+, and RefCOCOg, where the references are not embodied. Therefore, these three pre-trained models neglect the human gestural cues. Next, for a

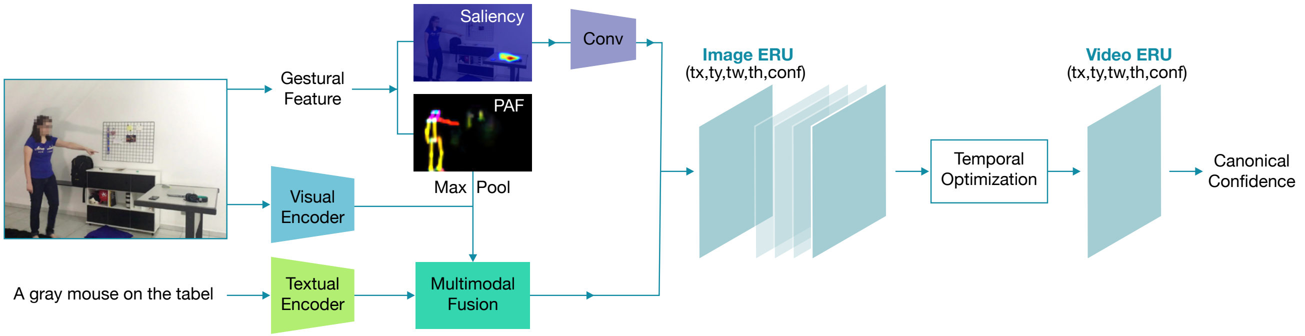


Figure 4: The proposed multimodal framework for the ERU task that incorporates both human gestural cues and language cues.

fair comparison without the gestural cues, we further generate an inpainted version of *YouRefIt*, where humans are segmented and masked by a pre-trained Mask R-CNN [16, 55], and the masked images are inpainted by DeepFill [62, 61] pre-trained on the Places2 [67] dataset; see examples in Fig. 5. After the human gestural cues are masked out, we train FAOA and ReSC on the inpainted dataset, denoted as $\text{FAOA}_{\text{inpaint}}$ and $\text{ReSC}_{\text{inpaint}}$. For the **architecture-wise** ablation, we compare two variants of our proposed full model to evaluate the contribution of different components: (i) $\text{Ours}_{\text{no\_lang}}$: without the language embedding module, and (ii) $\text{Ours}_{\text{PAF\_only}}$: with the PAF heatmap as the only gestural cue; see the *supplementary material* for more details.

Results and Discussion Table 2 tabulates the quantitative results of the Image ERU, and Fig. 5 shows some qualitative results. We categorize the models based on their information sources: *Language-only*, *Gesture-only*, and *Language + Gesture*. Below, we summarize some key findings.

1. Gestural cues are essential for embodied reference understanding. As shown in Table 2, FAOA and ReSC models show significant performance improvement when trained on the original *YouRefIt* dataset compared to that on the inpainted version. Of note, in embodied reference, the referrer will adjust their own position to ensure the referred targets are not blocked by its body, one of the main advantages introduced by perspective-taking. As such, the inpainted images always contain the reference targets with only gestural cues masked.
2. Language cues elucidate ambiguities where the gestural cues alone cannot resolve. As shown by the *Gesture-only* models, RPN+heatmap models possess ambiguities when presented with gestural cues alone; pointing gestures suppress the descriptions of target location and attend to spatial regions but are not object-centric. Without the referring expressions, the performance of $\text{Ours}_{\text{no\_lang}}$ also deteriorates compared to $\text{Ours}_{\text{Full}}$.
3. Explicit gestural features are beneficial for understanding embodied reference. $\text{Ours}_{\text{PAF\_only}}$, which incorporates PAF features that encode unstructured pairwise relationships between body parts, outperforms the original FAOA and ReSC models. By further adding the saliency heatmap, our full model $\text{Ours}_{\text{Full}}$ achieves the best performance in all baselines and ablations. Taken together, these results strongly indicate that the fusion of the language and gestural cues could be the crucial ingredient to achieving high model performance.

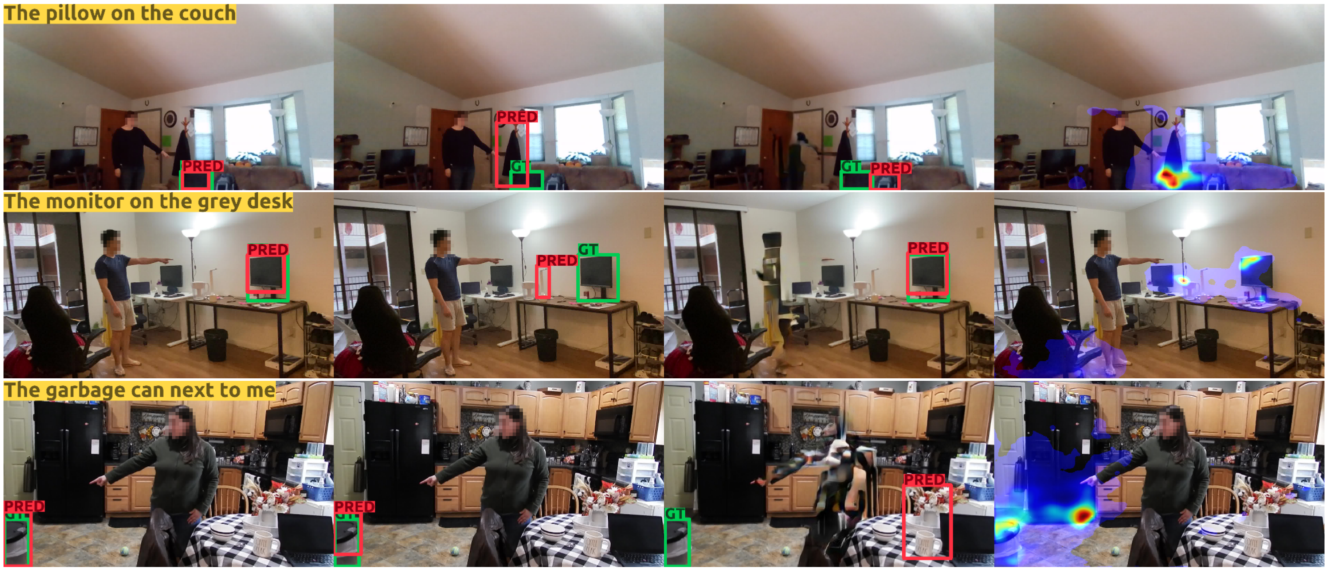
Human Performance We also conducted a human study of the embodied reference understanding task. We ask three Amazon Turkers to annotate the referred object bounding box in 1,000 images randomly sampled from the test set. We report the average accuracy under different IoUs in Table 2. Humans achieve significantly higher accuracy than all current machine learning models, demonstrating the human's outstanding capability to understand embodied references combined with language and gestural cues. The performance drops when the IoU threshold increases, especially for *small* and *medium* objects, indicating the difficulties in resolving the ambiguity in small objects.

4.2. Video ERU

Compared with Image ERU discussed above, Video ERU is a more natural and practical setting in human-robot interaction. Given a referring expression and a video clip that captures the whole dynamics of a reference action with consecutive body movement, Video ERU aims at recognizing the canonical frames and estimate the referred target at the same time.

Experimental Setup and Evaluation Protocol For each reference instance, we sample image frames with 5 FPS from the original video clip. Average precision, recall, and F1-score are reported for the canonical frame detection. For referred bounding box prediction, we report the averaged accuracy in all canonical frames.

Baselines To further exploit the temporal constraints in videos, we integrate a temporal optimization module to aggregate and optimize the multimodal feature extracted from the Image ERU. We test two designs of temporal optimization module: (i) ConvLSTM: a two-layer convolutional Long Short-Term Memory [45], and (ii) Transformer: a three-layer Transformer encoder [52] with four attention heads in each layer. After the temporal optimization mod-



(a) Ours$_{\text{Full}}$ (b) Ours$_{\text{no\_lang}}$ (c) ReSC$_{\text{inpaint}}$ (d) Saliency Map

Figure 5: **Qualitative results in Image ERU of representative models with various information sources and pointing saliency map.** Green/red boxes are the predicted/ground-truth reference targets. Sentences used during the references are shown at the top-left corner.

Table 2: Comparisons of Image ERU performances on the *YouRefIt* dataset.

Model	**IoU=0.25**				**IoU=0.5**				**IoU=0.75**			
	all	*small*	*medium*	*large*	*all*	*small*	*medium*	*large*	*all*	*small*	*medium*	*large*
Language-only												
MAttNet$_{\text{pretrain}}$	14.2	2.3	4.1	34.7	12.2	2.4	3.8	29.2	9.1	1.0	2.2	23.1
FAOA$_{\text{pretrain}}$	15.9	2.1	9.5	34.4	11.7	1.0	5.4	27.3	5.1	0.0	0.0	14.1
FAOA$_{\text{inpaint}}$	23.4	14.2	23.6	32.1	16.4	9.0	17.9	22.5	4.1	1.4	4.7	6.2
ReSC$_{\text{pretrain}}$	20.8	3.5	17.5	40.0	16.3	0.5	14.8	36.7	7.6	0.0	4.3	17.5
ReSC$_{\text{inpaint}}$	34.3	20.3	38.9	44.0	25.7	8.1	32.4	36.5	9.1	1.1	10.1	16.0
Gesture-only												
RPN+Pointing$_{15}$	15.3	10.5	16.9	18.3	10.2	7.2	12.4	11.0	6.5	3.8	9.1	6.6
RPN+Pointing$_{30}$	14.7	10.8	17.0	16.4	9.8	7.4	12.4	9.8	6.5	3.8	8.9	6.8
RPN+Saliency[27]	27.9	29.4	34.7	20.3	20.1	**21.1**	26.8	13.2	12.2	**10.3**	**17.9**	8.6
Ours$_{\text{no\_lang}}$	41.4	29.9	48.3	46.3	30.6	17.4	37.0	37.4	10.8	1.7	13.9	16.6
Language + Gesture												
FAOA[59]	44.5	30.6	48.6	54.1	30.4	15.8	36.2	39.3	8.5	1.4	9.6	14.4
ReSC[58]	49.2	32.3	54.7	60.1	34.9	14.1	42.5	47.7	10.5	0.2	10.6	20.1
Ours$_{\text{PAF\_only}}$	52.6	35.9	60.5	61.4	37.6	14.6	49.1	49.1	12.7	1.0	16.5	20.5
Ours$_{\text{Full}}$	**54.7**	**38.5**	**64.1**	**61.6**	**40.5**	16.3	**54.4**	**51.1**	**14.0**	1.2	17.2	**23.3**
Human	94.2±0.2	93.7±0.0	92.3±1.3	96.3±1.7	85.8±1.4	81.0±2.2	86.7±1.9	89.4±1.7	53.3±4.9	33.9±7.1	55.9±6.4	68.1±3.0

ule, we use the features of each frame to predict canonical frames and anchor bounding boxes simultaneously.

We further design a third *Frame-based* baseline that learns from the individual frame by adding two fully connected regression layers on top of our model in Image ERU. This *Frame-based* model takes all sampled frames from the video clip during training and testing.

During training, we add a binary cross-entropy loss for canonical frame detection on top of the loss function for bounding box prediction in the Image ERU framework. Please refer to the *supplementary material* for more details.

Results and Discussion Table 3 shows quantitative results of predicting reference targets with the ground-truth canonical frames given a video. We observe that the frame-based method and the temporal optimization methods reach similar performance, comparable to the model that only trained on selected canonical frames (*i.e.*, Ours$_{\text{Full}}$). This result indicates that the canonical frames can indeed provide sufficient language and gestural cues for clear reference purposes, and the temporal models may be distracted from non-canonical frames. This observation aligns with the settings of previous REF tasks. Meanwhile, as shown in Table 4 and Fig. 7, temporal information can significantly improve the performance of canonical frame detection; both the *ConvLSTM* and the *Transformer* model outperform the *Frame-based* method by a large margin. These results indicate the significance of distinguishing various stages of reference behaviors, *e.g.*, initiation, canonical moment, and ending, for better efficacy in embodied reference understanding. Fig. 6 shows some qualitative results.

Table 3: Video ERU performance comparisons on the *YouRefIt* dataset.

Model	**IoU=0.25**				**IoU=0.5**				**IoU=0.75**			
	all	*small*	*medium*	*large*	*all*	*small*	*medium*	*large*	*all*	*small*	*medium*	*large*
Frame-based	**55.2**	42.3	**58.9**	**64.8**	**41.7**	**22.7**	53.4	**48.8**	16.9	1.6	21.8	**27.0**
Transformer	52.3	40.2	55.6	58.3	38.8	21.2	54.1	47.1	13.9	1.5	20.8	22.7
ConvLSTM	54.8	**43.1**	57.5	60.0	39.3	22.5	**54.8**	46.7	**17.3**	**1.8**	**24.3**	25.5
$\text{Ours}_{\text{Full}}$	54.7	38.5	64.1	61.6	40.5	16.3	54.4	51.1	14.0	1.2	17.2	23.3

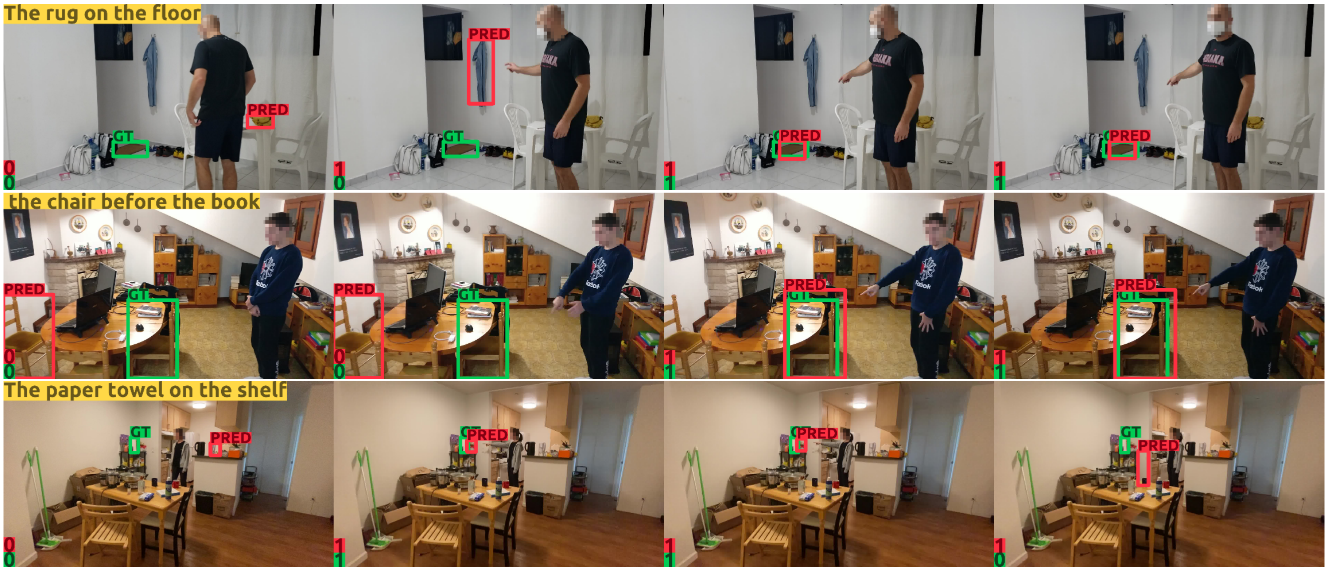


Figure 6: **Qualitative results in Video ERU of the ConvLSTM model.** Each row represents four selected frames from one reference clip. Green/red boxes indicate the predicted/ground-truth reference targets. 0 denotes non-canonical frame, and 1 canonical frame.

Table 4: Canonical frame detection performance.

Method	Avg. Prec	Avg. Rec	Avg. F1
Frame-based	31.9	37.7	34.5
Transformer	35.1	**44.2**	39.1
ConvLSTM	**57.0**	37.9	**45.4**

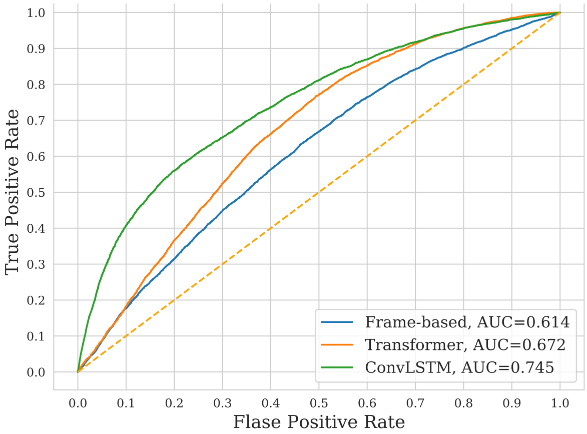


Figure 7: ROC Curve for canonical frame detection.

5. Conclusion and Future Work

We present the novel problem of embodied reference understanding. Such a setting with both language and gestural cues is more natural for understanding human communication in our daily activities. To tackle this problem, we crowd-source the *YouRefIt* dataset and devise two benchmarks on images and videos. We further propose a multimodal framework and conduct extensive experiments with ablations. The experimental results provide strong empirical evidence that language and gestural coordination is critical for understanding human communication.

Our work initiates the research on embodied reference understanding and can be extended to many aspects. For example, the difficulty in resolving reference ambiguity within a single-round communication, even for humans, calls for studying embodied reference using multi-round dialogues. Human-robot interaction may benefit from referential behavior generation by considering scene contexts. We hope our work can inspire more future work on these promising directions, focusing on understanding human communication from multimodal (verbal/nonverbal) inputs.

References

[1] Michael A Arbib, Katja Liebal, and Simone Pika. Primate vocalization, gesture, and the evolution of human language. *Current anthropology*, 49(6):1053–1076, 2008. 1

[2] C Daniel Batson, Shannon Early, and Giovanni Salvarani. Perspective taking: Imagining how another feels versus imaging how you would feel. *Personality and social psychology bulletin*, 23(7):751–758, 1997. 1

[3] Zhe Cao, Gines Hidalgo Martinez, Tomas Simon, Shih-En Wei, and Yaser A Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019. 4, 5

[4] Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K Wong, and Qi Wu. Cops-ref: A new dataset and task on compositional referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3

[5] Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency. Refer360: A referring expression recognition dataset in 360: A referring expression recognition dataset in 360 images images. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. 2

[6] Cristina Colonnesi, Geert Jan JM Stams, Irene Koster, and Marc J Noom. The relation between pointing and language development: A meta-analysis. *Developmental Review*, 30(4):352–366, 2010. 1

[7] Susan Wagner Cook, Zachary Mitchell, and Susan Goldin-Meadow. Gesturing makes learning last. *Cognition*, 106(2):1047–1058, 2008. 1

[8] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5

[10] Lifeng Fan, Yixin Chen, Ping Wei, Wenguan Wang, and Song-Chun Zhu. Inferring shared attention in social scene videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5

[11] Lifeng Fan, Shuwen Qiu, Zilong Zheng, Tao Gao, Song-Chun Zhu, and Yixin Zhu. Learning triadic belief dynamics in nonverbal communication from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[12] Adam D Galinsky, William W Maddux, Debra Gilin, and Judith B White. Why it pays to get inside the head of your opponent: The differential effects of perspective taking and empathy in negotiations. *Psychological Science*, 19(4):378–384, 2008. 2

[13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 5

[14] Marta Halina, Federico Rossano, and Michael Tomasello. The ontogenetic ritualization of bonobo gestures. *Animal cognition*, 16(4):653–666, 2013. 1

[15] Yasuhiko Hato, Satoru Satake, Takayuki Kanda, Michita Imai, and Norihiro Hagita. Pointing to space: modeling of deictic interaction referring to regions. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2010. 3

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2017. 6

[17] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[18] Jana M Iverson and Susan Goldin-Meadow. Gesture paves the way for language development. *Psychological science*, 16(5):367–371, 2005. 1

[19] baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-Chun Zhu. Lemma: A multi-view dataset for learning multi-agent multi-task activities. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 2

[20] Kaiwen Jiang, Stephanie Stacy, Chuyu Wei, Adelpha Chan, Federico Rossano, Yixin Zhu, and Tao Gao. Individual vs. joint perception: a pragmatic model of pointing as communicative smithian helping. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2021. 1, 2

[21] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 2, 3

[22] Adam Kendon. *Gesture: Visible action as utterance*. Cambridge University Press, 2004. 1

[23] Sotaro Kita. *Pointing: Where language, culture, and cognition meet*. Psychology Press, 2003. 1

[24] Alfred Kobsa, Jurgen Allgayer, Carola Reddig, Norbert Reithinger, Dagmar Schmauks, Karin Harbusch, and Wolfgang Wahlster. Combining deictic gestures and

natural language for referent identification. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, 1986. 3

[25] Robert M Krauss and Susan R Fussell. Perspective-taking in communication: Representations of others' knowledge in reference. *Social cognition*, 9(1):2–24, 1991. 1

[26] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73, 2017. 4

[27] Alexander Kroner, Mario Senden, Kurt Driessens, and Rainer Goebel. Contextual encoder–decoder network for visual saliency prediction. *Neural Networks*, 129:261–270, 2020. 5, 7

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2014. 5

[29] Ulf Liszkowski, Malinda Carpenter, Anne Henning, Tricia Striano, and Michael Tomasello. Twelve-month-olds point to share attention and interest. *Developmental science*, 7(3):297–307, 2004. 1

[30] Ulf Liszkowski, Malinda Carpenter, Tricia Striano, and Michael Tomasello. 12-and 18-month-olds point to provide information for others. *Journal of cognition and development*, 7(2):173–187, 2006. 1

[31] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3

[32] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. Improving referring expression grounding with cross-modal attention-guided erasing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2

[33] Andy Lücking, Thies Pfeiffer, and Hannes Rieser. Pointing and reference reconsidered. *Journal of Pragmatics*, 77:56–79, 2015. 3

[34] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[35] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3, 5

[36] David McNeill. *How language began: Gesture and speech in human evolution*. Cambridge University Press, 2012. 1

[37] Karola Pitsch and Sebastian Wrede. When a robot orients visitors to an exhibit. referential practices and interactional dynamics in real world hri. In *Proceedings of International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2014. 3

[38] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 2, 3

[39] Shuwen Qiu, Hangxin Liu, Zeyu Zhang, Yixin Zhu, and Song-Chun Zhu. Human-robot interaction in a shared augmented reality workspace. In *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, 2020. 1, 2

[40] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 5

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(6):1137–1149, 2016. 5

[42] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016. 2

[43] Boris Schauerte and Gernot A Fink. 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*, 2010. 3

[44] Boris Schauerte, Jan Richarz, and Gernot A Fink. Saliency-based identification and recognition of pointed-at objects. In *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, 2010. 3

[45] Xingjian Shi, Zhourong Chen, Hao Wang, Dit Yan Yeung, Wai Kin Wong, and Wang Chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2015. 6

[46] Dadhichi Shukla, Ozgur Erkent, and Justus Piater. Probabilistic detection of pointing directions for human-robot interaction. In *International Conference*

on Digital Image Computing: Techniques and Applications, 2015. 3

[47] Dadhichi Shukla, Özgür Erkent, and Justus Piater. A multi-view hand gesture rgb-d dataset for human-robot interaction scenarios. In *Proceedings of International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016. 3

[48] Stephanie Stacy, Qingyi Zhao, Minglu Zhao, Max Kleiman-Weiner, and Tao Gao. Intuitive signaling through an “imagined we”. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2020. 1

[49] John Sweller, Jeroen JG Van Merrienboer, and Fred GWC Paas. Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3):251–296, 1998. 4

[50] Ning Tang, Stephanie Stacy, Minglu Zhao, Gabriel Marquez, and Tao Gao. Bootstrapping an imagined we for cooperation. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2020. 1

[51] Michael Tomasello. *Origins of human communication*. MIT press, 2010. 1

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 6

[53] Carl Vondrick, Donald Patterson, and Deva Ramanan. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision (IJCV)*, 101(1):184–204, 2013. 4

[54] Qi Wu, Cheng-Ju Wu, Yixin Zhu, and Jungseock Joo. Communicative learning with natural gestures for embodied navigation agents with human-in-the-scene. In *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, 2021. 1, 2

[55] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 6

[56] Sibei Yang, Guanbin Li, and Yizhou Yu. Cross-modal relationship inference for grounding referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2

[57] Sibei Yang, Guanbin Li, and Yizhou Yu. Graph-structured referring expression reasoning in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[58] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. *arXiv preprint arXiv:2008.01059*, 2020. 2, 5, 7

[59] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2019. 2, 5, 7

[60] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2

[61] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. *arXiv preprint arXiv:1806.03589*, 2018. 6

[62] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[63] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 5

[64] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016. 1, 2, 3

[65] Tao Yuan, Hangxin Liu, Lifeng Fan, Zilong Zheng, Tao Gao, Yixin Zhu, and Song-Chun Zhu. Joint inference of states, robot knowledge, and human (false-)beliefs. In *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2020. 2

[66] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding referring expressions in images by variational context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[67] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(6):1452–1464, 2017. 6

[68] Yixin Zhu, Tao Gao, Lifeng Fan, Siyuan Huang, Mark Edmonds, Hangxin Liu, Feng Gao, Chi Zhang, Siyuan Qi, Ying Nian Wu, Josh Tenenbaum, and Song-Chun Zhu. Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering*, 6(3):310–345, 2020. 2

[69] George Kingsley Zipf. *Human Behaviour and the Principles of Least Effort*. Addison-Wesley, 1949. 4