YouRefIt: Embodied Reference Understanding with Language and Gesture

Yixin Chen¹, Qing Li¹, Deqian Kong¹, Yik Lun Kei¹, Song-Chun Zhu²,³,⁴, Tao Gao¹, Yixin Zhu²,³, Siyuan Huang¹

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].

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 ref-
ference understanding at a full spectrum, we introduce a new dataset, \textit{YouRefIt}, for embodied reference understanding. The reference instances in \textit{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 \textit{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 \textit{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, \textit{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.

\textbf{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.

\textbf{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.
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) deixic 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 handshake, 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.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Lang.</th>
<th>Gest.</th>
<th>Embo.</th>
<th>Type</th>
<th>Source</th>
<th>No. of images</th>
<th>No. of instances</th>
<th>No. of object categories</th>
<th>Ave. sent. length</th>
</tr>
</thead>
<tbody>
<tr>
<td>PointAt [44]</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>image</td>
<td>lab</td>
<td>220</td>
<td>220</td>
<td>28</td>
<td>-</td>
</tr>
<tr>
<td>ReferAt [43]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>video</td>
<td>lab</td>
<td>-</td>
<td>242</td>
<td>28</td>
<td>-</td>
</tr>
<tr>
<td>IPO [46]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>image</td>
<td>lab</td>
<td>278</td>
<td>278</td>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td>IMHF [47]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>image</td>
<td>lab</td>
<td>1716</td>
<td>1,716</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RefIt [21]</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>image</td>
<td>image CLEF</td>
<td>19,894</td>
<td>130,525</td>
<td>238</td>
<td>3.61</td>
</tr>
<tr>
<td>RefCOCO [64]</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>image</td>
<td>MSCOCO</td>
<td>19,994</td>
<td>142,209</td>
<td>80</td>
<td>3.61</td>
</tr>
<tr>
<td>RefCOCO+ [64]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>image</td>
<td>MSCOCO</td>
<td>19,992</td>
<td>141,564</td>
<td>80</td>
<td>3.53</td>
</tr>
<tr>
<td>RefCOCOg [35]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>image</td>
<td>MSCOCO</td>
<td>26,711</td>
<td>104,560</td>
<td>80</td>
<td>8.43</td>
</tr>
<tr>
<td>Flickr30k entities [38]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>image</td>
<td>Flickr30K</td>
<td>31,783</td>
<td>158,915</td>
<td>44,518</td>
<td>-</td>
</tr>
<tr>
<td>GuessWhat? [8]</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>image</td>
<td>MSCOCO</td>
<td>66,537</td>
<td>155,280</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CLEVR-Ref+ [31]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>image</td>
<td>CLEVR</td>
<td>99,992</td>
<td>998,743</td>
<td>3</td>
<td>22.40</td>
</tr>
</tbody>
</table>

YouRefIt ✓ ✓ ✓ video crowd-sourced 497,348 4,195 395 3.73

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.
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].
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 RoI loss 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., RPNpointing15 and RPNpointing30). (ii) By pointing saliency map (i.e., RPNsaliency). 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
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 FAOAinpaint and ReSCinpaint. For the architecture-wise ablation, we compare two variants of our proposed full model to evaluate the contribution of different components: (i) Ours\textsubscript{no lang}: without the language embedding module, and (ii) Ours\textsubscript{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 performances of Ours\textsubscript{no lang} also deteriorates compared to Ours\textsubscript{Full}.

3. **Explicit gestural features are beneficial for understanding embodied reference.** Ours\textsubscript{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 Ours\textsubscript{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-
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.

<table>
<thead>
<tr>
<th>Model</th>
<th>IoU=0.25</th>
<th>IoU=0.5</th>
<th>IoU=0.75</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>all</td>
<td>small</td>
<td>medium</td>
</tr>
<tr>
<td>Language-only</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAAttNet[pretrain]</td>
<td>14.2</td>
<td>2.3</td>
<td>4.1</td>
</tr>
<tr>
<td>FAOA[pretrain]</td>
<td>15.9</td>
<td>2.1</td>
<td>9.5</td>
</tr>
<tr>
<td>FAOA[inpaint]</td>
<td>23.4</td>
<td>14.2</td>
<td>23.6</td>
</tr>
<tr>
<td>ReSC[pretrain]</td>
<td>20.8</td>
<td>3.5</td>
<td>17.5</td>
</tr>
<tr>
<td>ReSC[inpaint]</td>
<td>34.3</td>
<td>20.3</td>
<td>38.9</td>
</tr>
<tr>
<td>Gesture-only</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPN+Pointing[15]</td>
<td>15.3</td>
<td>10.5</td>
<td>16.9</td>
</tr>
<tr>
<td>RPN+Pointing[40]</td>
<td>14.7</td>
<td>10.8</td>
<td>17.0</td>
</tr>
<tr>
<td>RPN+Saliency[27]</td>
<td>27.9</td>
<td>29.4</td>
<td>34.7</td>
</tr>
<tr>
<td>Ours[all]</td>
<td>41.4</td>
<td>29.9</td>
<td>48.3</td>
</tr>
<tr>
<td>Language + Gesture</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAOA[59]</td>
<td>44.5</td>
<td>30.6</td>
<td>48.6</td>
</tr>
<tr>
<td>ReSC[58]</td>
<td>49.2</td>
<td>32.3</td>
<td>54.7</td>
</tr>
<tr>
<td>Ours[pred, only]</td>
<td>52.6</td>
<td>35.9</td>
<td>60.5</td>
</tr>
<tr>
<td>Ours[PAF]</td>
<td>54.7</td>
<td>38.5</td>
<td>64.1</td>
</tr>
<tr>
<td>Human</td>
<td>94.2±0.2</td>
<td>93.7±0.0</td>
<td>92.3±1.3</td>
</tr>
</tbody>
</table>

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., OursFull). 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.

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., OursFull). 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>IoU=0.25</th>
<th>IoU=0.5</th>
<th>IoU=0.75</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>all</td>
<td>small</td>
<td>medium</td>
</tr>
<tr>
<td>Frame-based</td>
<td>55.2</td>
<td>42.3</td>
<td>58.9</td>
</tr>
<tr>
<td>Transformer</td>
<td>52.3</td>
<td>40.2</td>
<td>55.6</td>
</tr>
<tr>
<td>ConvLSTM</td>
<td>54.8</td>
<td>43.1</td>
<td>57.5</td>
</tr>
<tr>
<td>Ours</td>
<td>Full</td>
<td>54.7</td>
<td>38.5</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Avg. Prec</th>
<th>Avg. Rec</th>
<th>Avg. F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame-based</td>
<td>31.9</td>
<td>37.7</td>
<td>34.5</td>
</tr>
<tr>
<td>Transformer</td>
<td>35.1</td>
<td>44.2</td>
<td>39.1</td>
</tr>
<tr>
<td>ConvLSTM</td>
<td>57.0</td>
<td>37.9</td>
<td>45.4</td>
</tr>
</tbody>
</table>

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


[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


