**YouRefIt: Embodied Reference Understanding with Language and Gesture**

Yixin Chen¹, Qing Li¹, Deqian Kong², YiK Lun Kei¹, Song-Chun Zhu²,³,⁴, Tao Gao¹, Yixin Zhu²,³, Siyuan Huang¹

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

### Key Features

- **Multimodal**: People often use both natural language and gestures when referring to an object.
- **Perspective-taking**: Embodied reference requires the awareness that others see things from different viewpoints and the ability to imagine what others see from their perspectives.

### Contribution & Discovery

- We crowd-source the first video dataset in physical scenes, **YouRefIt**, to study the reference understanding in an embodied setting.
- We devise two benchmarks. Image ERU and Video ERU, as the protocols to study and evaluate the embodied reference understanding.
- 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.

### YouRefIt Dataset

We introduce a new dataset named **YouRefIt**, a video collection of people referring to objects with both natural language and gesture in indoor scenes.

- **YouRefIt** contains videos crowd-sourced by Amazon Mechanical Turk (AMT), and thus the reference happens in a more natural setting with richer diversity.
- The referrers (human) and the receivers (camera) in YouRefIt share the same physical environment, with both language and gesture allowed for referring to objects.
- YouRefIt includes 432 recorded videos and 4,195 localized reference clips with 395 object categories.

### YouRefIt Dataset

- **Language-ERU**: 142.4 ± 2.3 frames
- **Feature-ERU**: 39.2 ± 2.1 frames
- **Video-E RU**: 21.4 ± 2.2 frames

### 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 gesture cues.

### Video ERU

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.

### Result

- **Image ERU**: Results indicate that the canonical frames can provide sufficient language and gestural cues for clear reference, however temporal information can improve the performance of canonical frame detection.
- **Video ERU**: Results confirm the significance of the joint understanding of language and gestures in embodied reference.