YouRefIt: Embodied Reference Understanding with Language and Gesture

Yixin Chen, Qing Li, Deqian Kong, Yik Lun Kei, Song-Chun Zhu, Tao Gao, Yixin Zhu and Siyuan Huang

Embodied Reference

- Referential behavior is a typical form of human communication, which acts as the first step to understand the surrounding world by establishing joint attention and common ground with other agents.
- Embodied reference: An agent refers to an object to another agent in a shared physical space.

Embodied Reference

Key difference with Referential Expression Understanding (REF):

- The reference participants and referred object are in the **same shared physical space**.

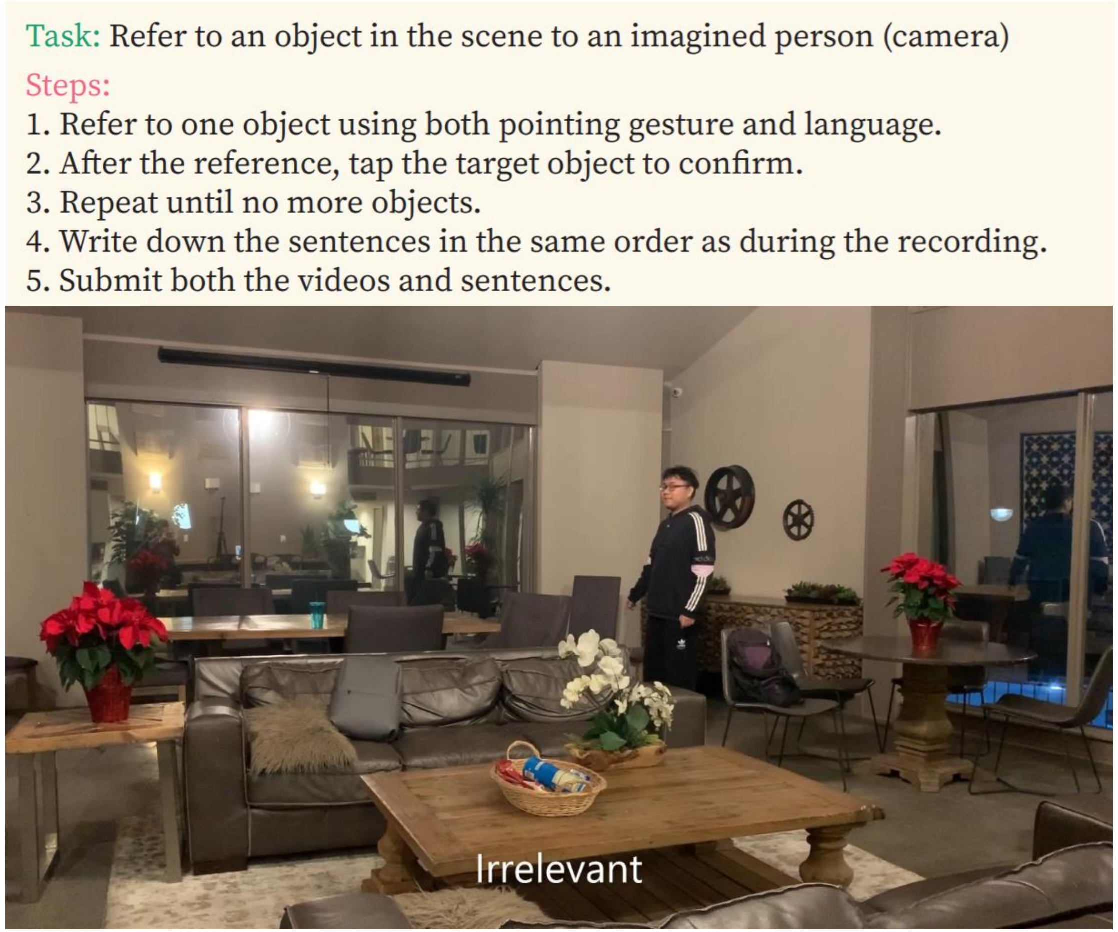
The white phone on the table

The picture on the wall

- Referrer will use both gestural and verbal information for reference.
- Embodied reference involves **visual perspective-taking**, i.e., the awareness that other people see things from different viewpoints and the ability to imagine what others see from their perspectives.
- Previous REF task takes images from Internet (MSCOCO/Flickr) or simulation(CLEVR). There's a natural domain gap compared with daily life picture.

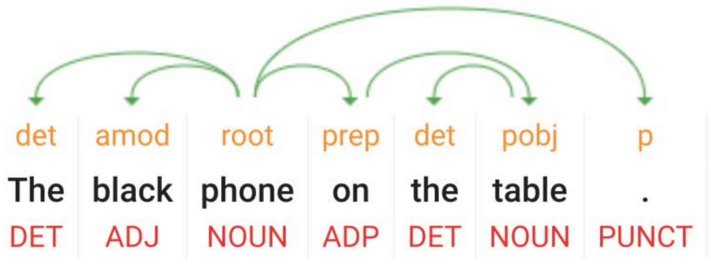
Data Collection

- YouRefIt dataset is collected using the Amazon Mechanic Turk (AMT) platform



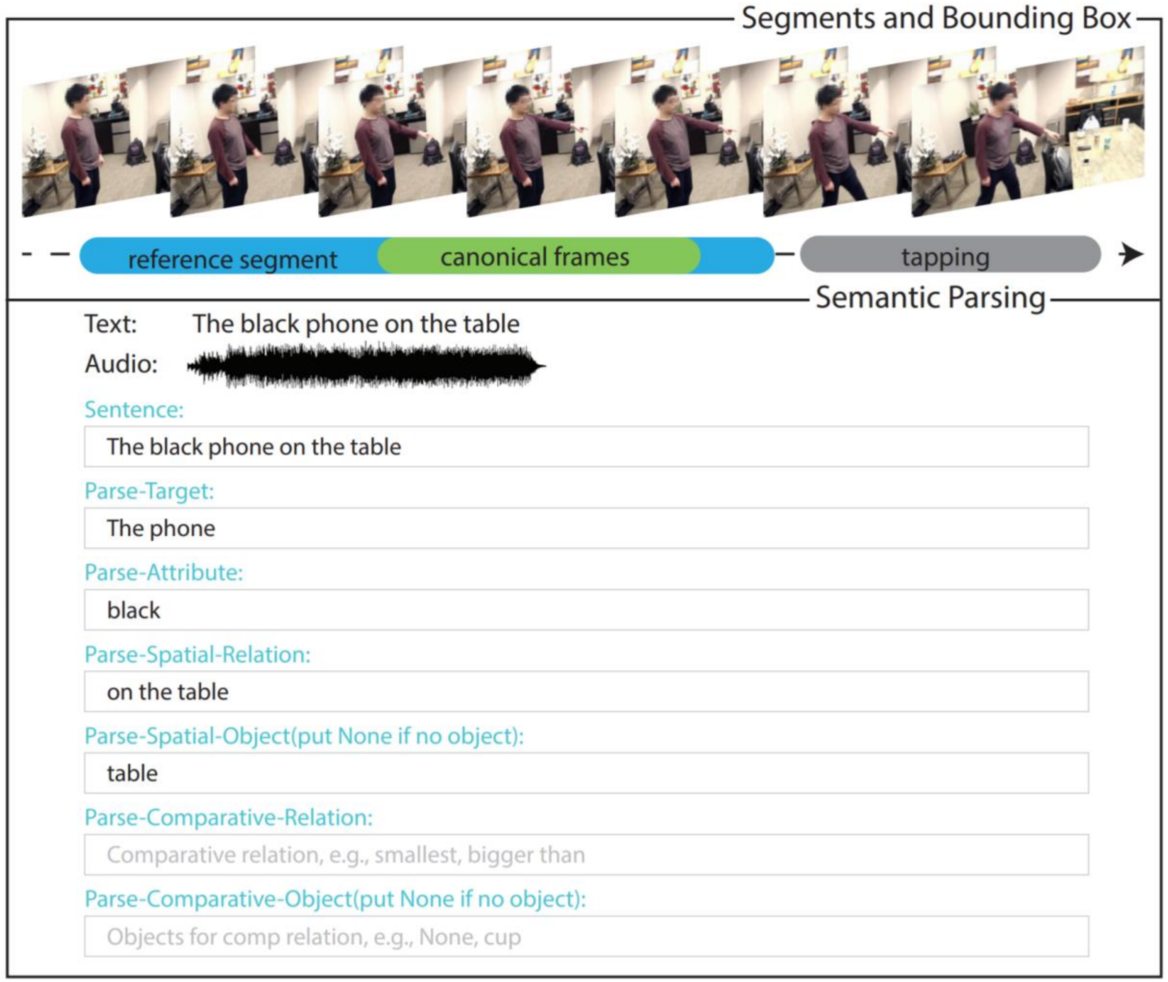
Data Annotation

- Reference segments
- Canonical frames: “keyframes” that the referrer holds the steady pose to clearly indicate what is being referred
- Bounding boxes of target objects
- Semantic parsing

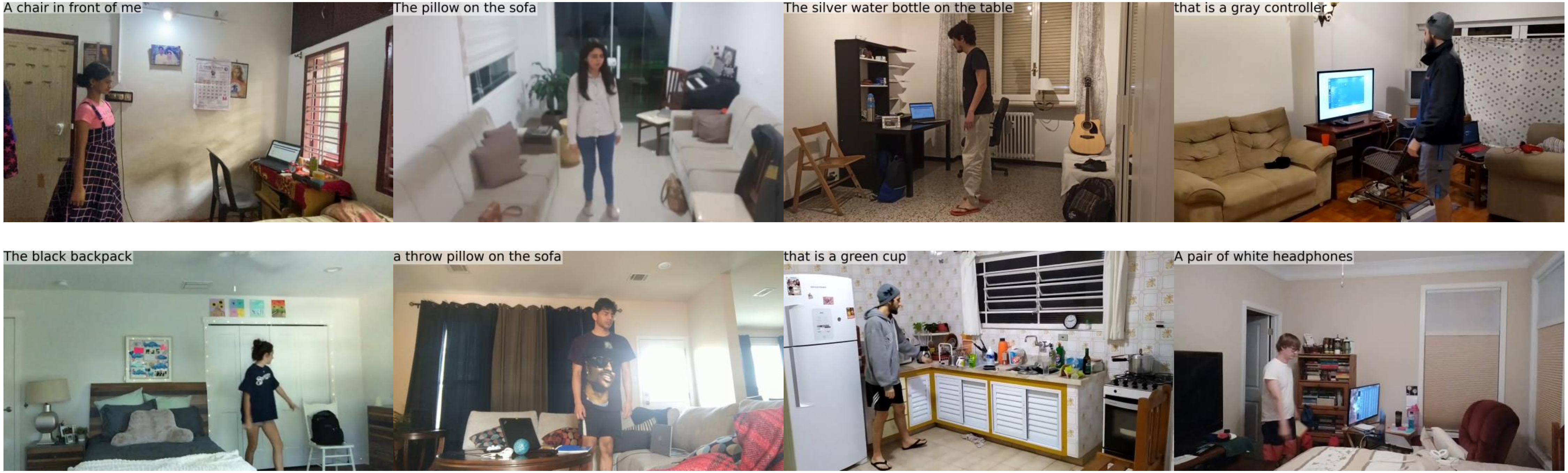


‘The black phone on the table.”

Canonical Frames



Dataset Sample

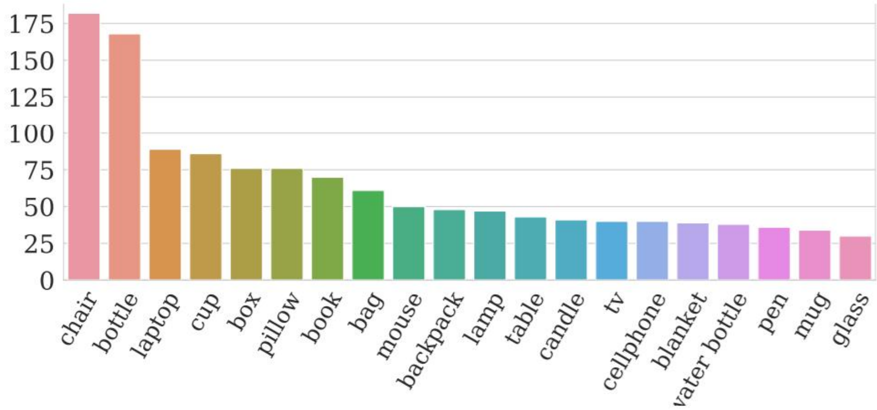


Statistics

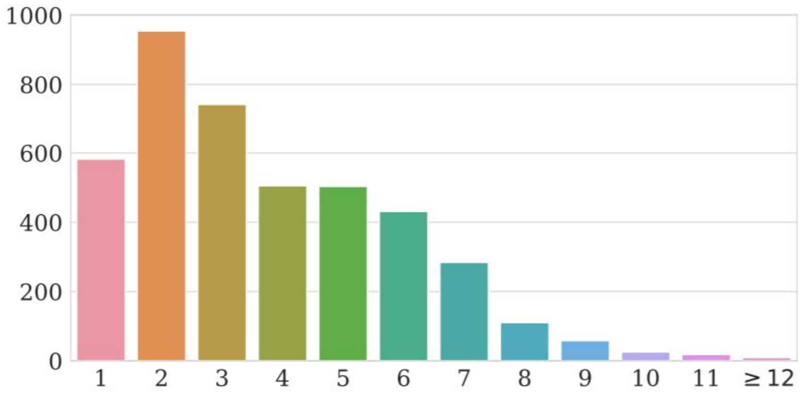
Datasets	**Lang.**	**Gest.**	**Embo.**	**Type**	**Source**	**No. of images**	**No. of instances**	**No. of object categories**	**Ave. sent. length**
PointAt [44]	✗	✓	✓	image	lab	220	220	28	-
ReferAt [43]	✓	✓	✓	video	lab	-	242	28	-
IPO [46]	✗	✓	✓	image	lab	278	278	10	-
IMHF [47]	✗	✓	✓	image	lab	1716	1,716	-	-
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

Statistics

- We retrieved 8.83 hours of video during the post-processing and annotated 497,348 frames.
- In total, YouRefIt includes 432 recorded videos and 4,195 localized reference clips with 395 object categories.
- The total duration of all the reference actions is 3.35 hours, with an average duration of 2.81 seconds per reference.



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



(b) The distribution of sentence lengths.

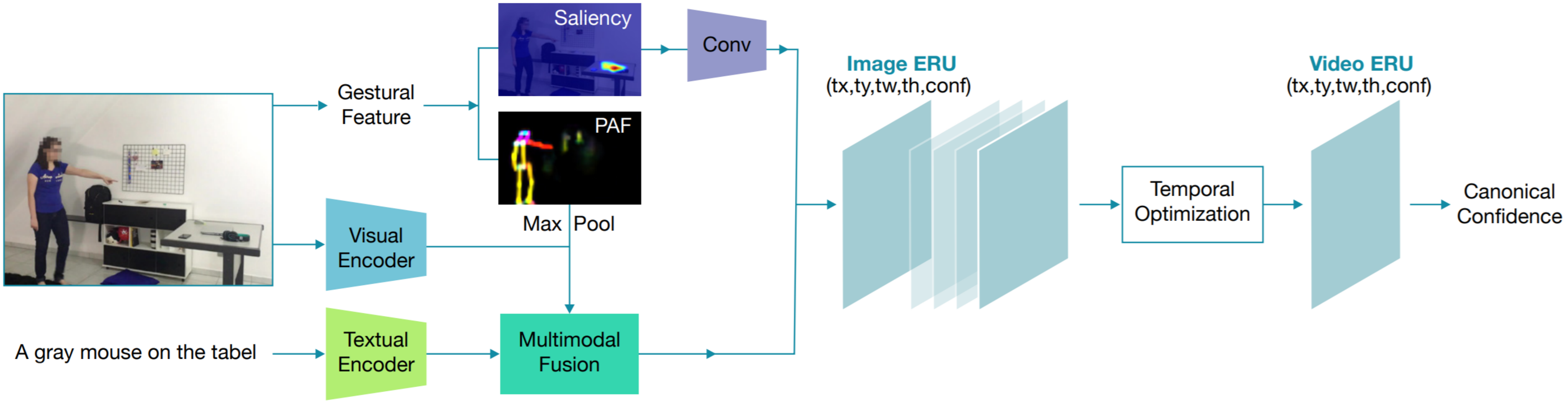
(c) The language wordle.

Embodied Reference Understanding (ERU)

- Image ERU:
  - Input: one canonical frame, the transcribed sentence
  - Predicts the bounding box of the referred object

- Video ERU:
  - Input: the video of reference segment, the transcribed sentence
  - Identifies the canonical frames and predicts the target bounding box

Framework



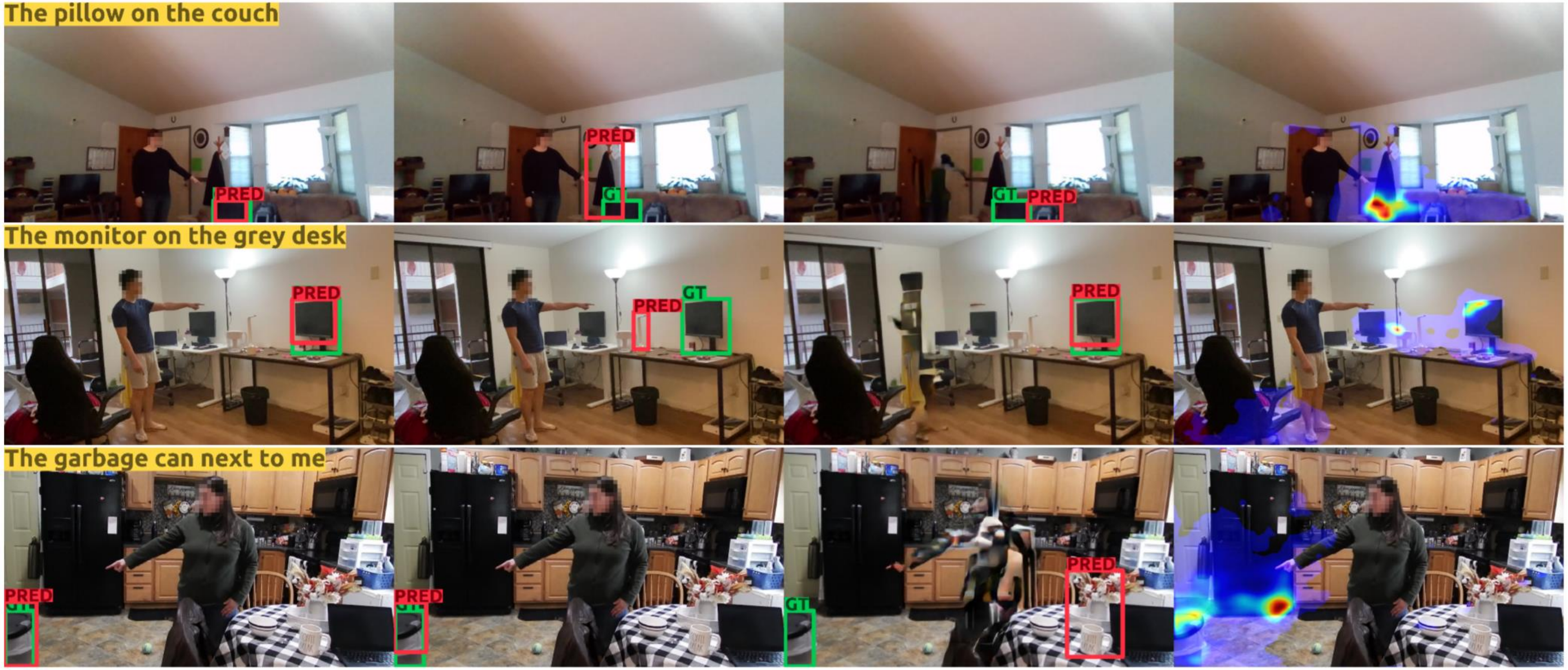
Saliency
Conv
Gestural
Feature
PAF
Image ERU
(tx,ty,tw,th,conf)
Video ERU
(tx,ty,tw,th,conf)
Visual
Encoder
Max Pool
Temporal
Optimization
Canonical
Confidence
A gray mouse on the tabel
Textual
Encoder
Multimodal
Fusion

Image ERU

- Our proposed framework, which explicitly considers all information sources (Language + Gesture) yields the best performance compared to the baseline models. Gesture information is essential in embodied reference understanding.

Model	IoU=0.25				IoU=0.5				IoU=0.75			
	all	*small*	*medium*	*large*	*all*	*small*	*medium*	*large*	*all*	*small*	*medium*	*large*
Language-only												
$\text{MAttNet}_{\text{pretrain}}$	14.2	2.3	4.1	34.7	12.2	2.4	3.8	29.2	9.1	1.0	2.2	23.1
$\text{FAOA}_{\text{pretrain}}$	15.9	2.1	9.5	34.4	11.7	1.0	5.4	27.3	5.1	0.0	0.0	14.1
$\text{FAOA}_{\text{inpaint}}$	23.4	14.2	23.6	32.1	16.4	9.0	17.9	22.5	4.1	1.4	4.7	6.2
$\text{ReSC}_{\text{pretrain}}$	20.8	3.5	17.5	40.0	16.3	0.5	14.8	36.7	7.6	0.0	4.3	17.5
$\text{ReSC}_{\text{inpaint}}$	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
$\text{Ours}_{\text{no\_lang}}$	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
$\text{Ours}_{\text{PAF\_only}}$	52.6	35.9	60.5	61.4	37.6	14.6	49.1	49.1	12.7	1.0	16.5	20.5
$\text{Ours}_{\text{Full}}$	**54.7**	**38.5**	**64.1**	**61.6**	**40.5**	16.3	**54.4**	**51.1**	**14.0**	1.2	17.2	**23.3**
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

Image ERU



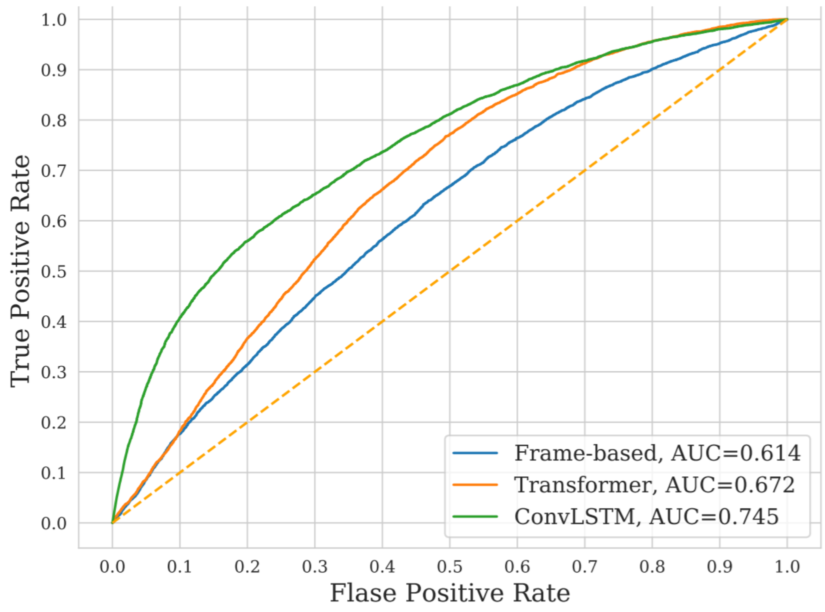
(a) $\text{Ours}_{\text{Full}}$ (b) $\text{Ours}_{\text{no\_lang}}$ (c) $\text{ReSC}_{\text{inpaint}}$ (d) Saliency Map

Video ERU

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

- Canonical frames can provide sufficient gestural and language information for clear reference

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**



- Temporal information can greatly improve performance on canonical frame detection

Video ERU

The rug on the floor
the chair before the book
The paper towel on the shelf

Future Direction

- Embodied reference in multi-round dialogues
- Referential behavior generation
- Active learning with referential interaction
- …

Thank you

Check our website at https://yixchen.github.io/YouRefIt