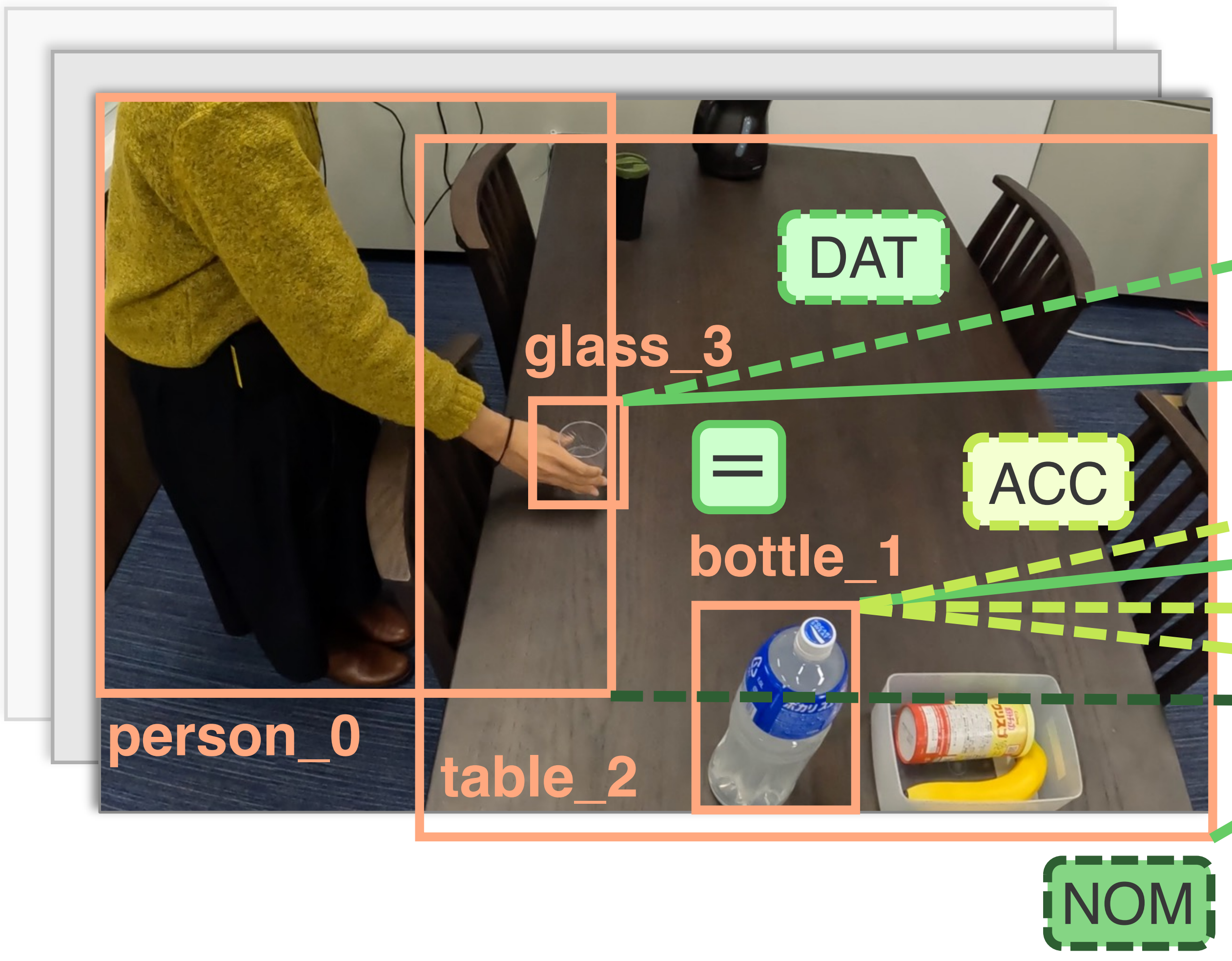
J-CRe3: A Japanese Conversation Dataset for Real-world Reference Resolution

Nobuhiro Ueda[1,2], Hideko Habe[2], Yoko Matsui[2], Akishige Yuguchi[3,2], Seiya Kawano[2,4], Yasutomo Kawanishi[2,4], Sadao Kurohashi[1,2,5], and Koichiro Yoshino[2,4]

[1]Kyoto University, [2]Guardian Robot Project, RIKEN, [3]Tokyo University of Science, [4]Nara Institute of Science and Technology, and [5]National Institute of Informatics

Overview: We built a real-world conversation dataset with dense multimodal reference tags.



Speaker	Utterance
Master	Pour
	the coke
	here.
Robot	What about
	the sports drink?
Master	Leave (it) (here).
	(I'll) drink (it)
	later.

NOM	ACC	DAT	Coref
Robot	the coke	here	
Robot	the sports drink		
Robot	the sports drink		
Master	the sports drink		

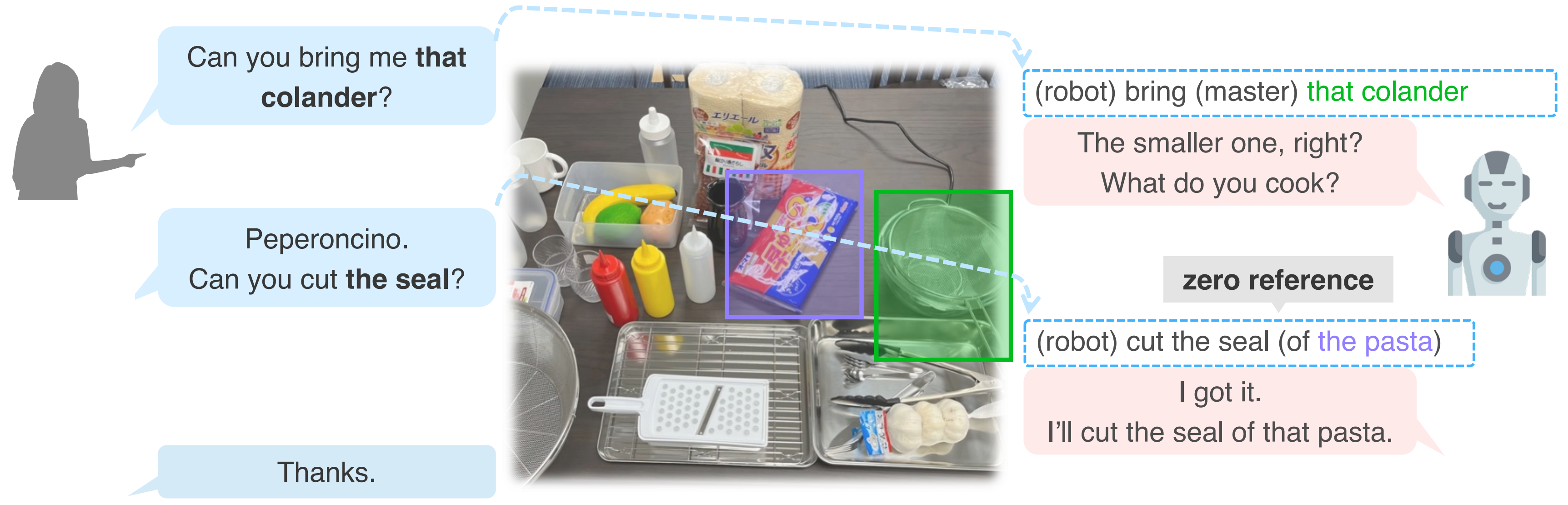
Object bounding boxes & instance IDs

Text-to-object references

Textual references

Background

In **collaborative real-world conversations**, utterances often refer to objects, and **grounding referential phrases** is essential for human-assisting systems.



Can you bring me **that colander**?

(robot) bring (master) that colander

The smaller one, right? What do you cook?

Peperoncino. Can you cut **the seal**?

zero reference

(robot) cut the seal (of the pasta)

I got it. I'll cut the seal of that pasta.

Thanks.

In conversational texts, referential phrases are often omitted (called **zero references**), which makes it hard for conventional tasks (e.g., phrase grounding) to solve.

J-CRe3 Dataset

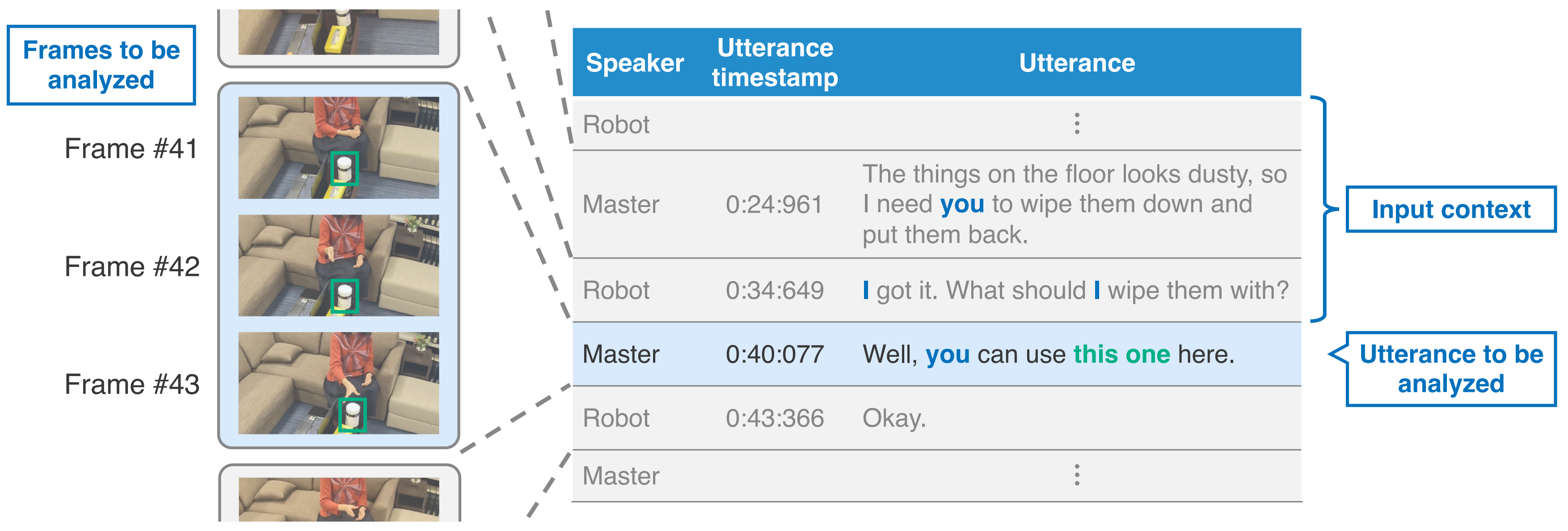
We proposed a **multimodal reference resolution** task that **comprehensively handles zero references** and built the **J-CRe3** dataset for the task.

Dataset	# Annotated images	Text type	# Dialogues	Video	Zero reference
RefCOCO (Yu et al., 2016)	20k	Referring expression	-	✗	✗
RefCOCO+ (Yu et al., 2016)	142k	Referring expression	-	✗	✗
RefCOCOg (Mao et al., 2016)	26k	Referring expression	-	✗	✗
VisualGenome (Krishna et al., 2017)	108k	Caption	-	✗	✗
Flickr30k Entities (Plummer et al., 2017)	30k	Caption	-	✗	✗
VisCoref (Yu et al., 2019)	5k	Dialogue	5,000	✗	✗
Visual Recipe Flow (Shirai et al., 2022)	6k	Cooking recipe	-	✗	✗
BioVL2 (Nishimura et al., 2021, 2022)	3k	Experimental procedure	-	✓	✗
EPIC-KITCHENS (Damen et al., 2022)	277k	Narration	-	✓	✗
RefEgo (Kurita et al., 2023)	226k	Referring expression	-	✓	✗
SIMMC 2.1 (Kottur and Moon, 2023)	2k	Dialogue	11,244	✗	✗
J-CRe3 (ours)	11k	Dialogue	93	✓	✓

Construction Procedures

1. Collect diverse dialogue scenarios through crowdsourcing
2. Record scenario-based **in-person conversation** at a living room, a dining room, and a kitchen
3. Convert the recorded **egocentric videos and dialogue audios** into image sequences (1fps) and dialogue texts
4. Densely annotate the processed images and texts with textual and visual tags (See **Overview**).

Evaluating Existing Models: Settings

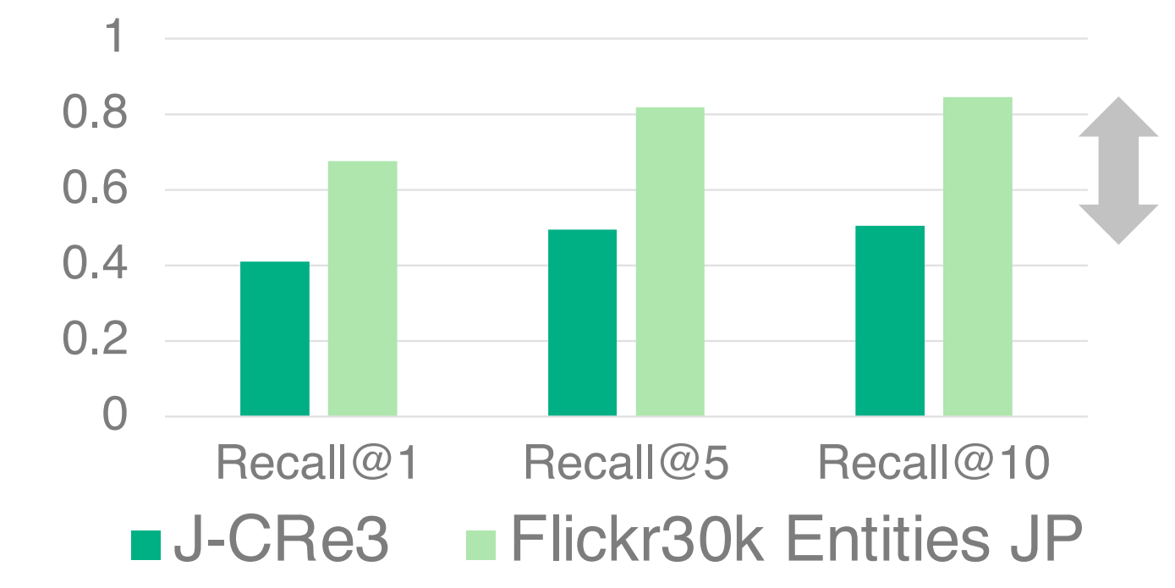


Speaker	Utterance timestamp	Utterance
Robot		⋮
Master	0:24:961	The things on the floor looks dusty, so I need **you** to wipe them down and put them back.
Robot	0:34:649	**I** got it. What should **I** wipe them with?
Master	0:40:077	Well, **you** can use **this one** here.
Robot	0:43:366	Okay.
Master		⋮

Evaluating Existing Models : Results

Text-to-object reference

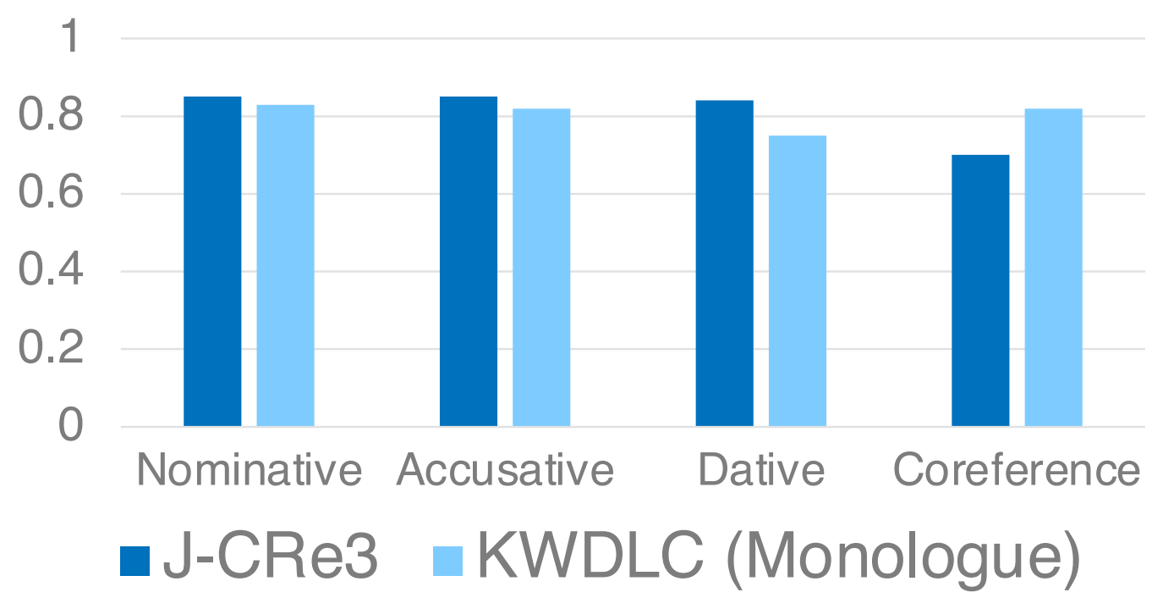
Model: fine-tunied MDETR (Kamath et al., 2020)



Much lower performance due to textual and visual context understanding J-CRe3 requires

Textual reference

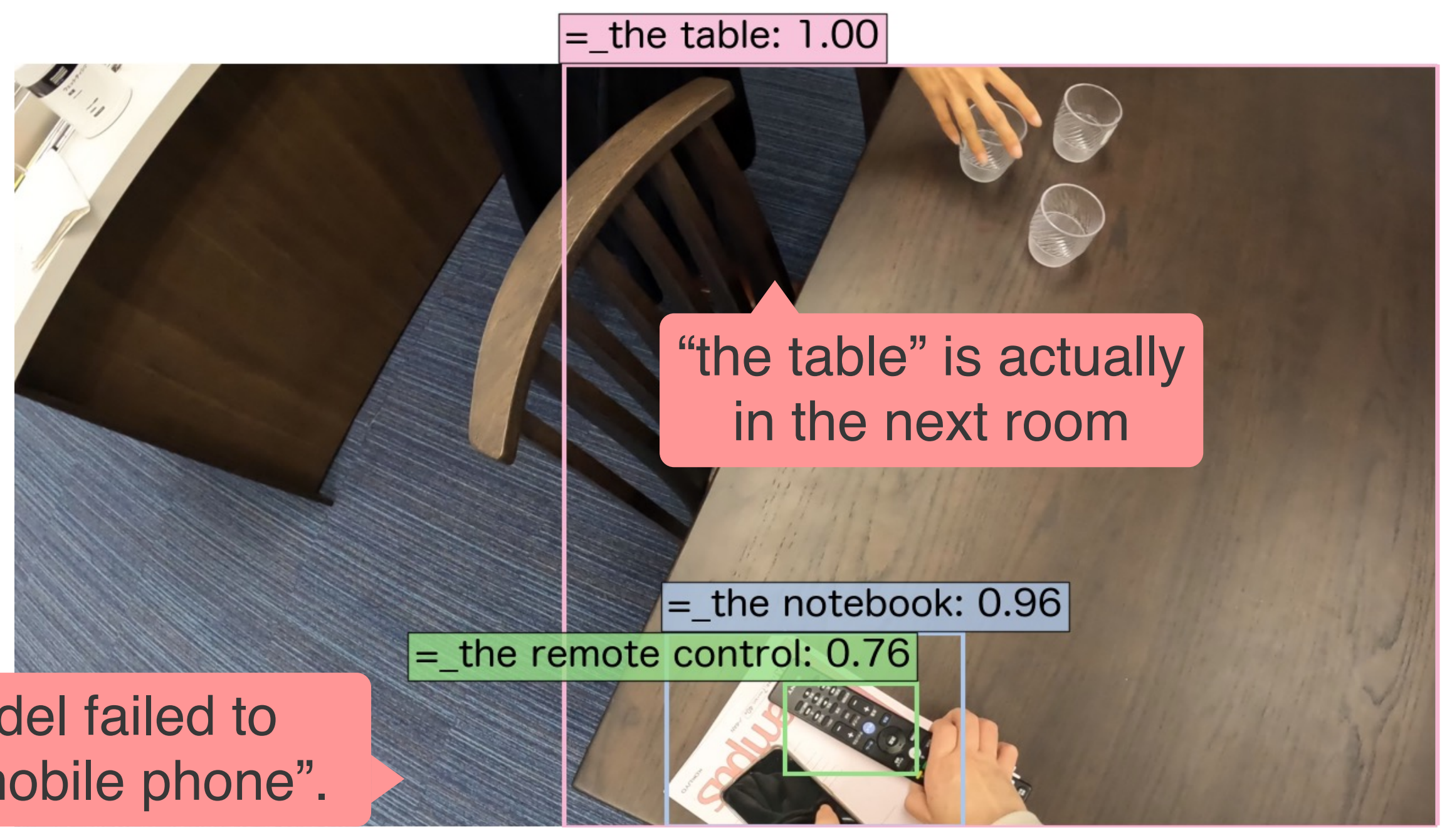
Model: fine-tunied DeBERTa (He et al., 2020)



Comparable results to those of an existing monologue corpus

Case Analysis

The notebook and mobile phone should be put on the table in the next room, and the remote control should be put on the sofa, right?



"the table" is actually in the next room

The model failed to ground "mobile phone".

Conclusion

- Our J-CRe3 dataset comprehensively handles references including zero references in real-world conversations.
- We are exploring integrated resolution models for textual and text-to-object references.

https://github.com/riken-grp/J-CRe3