ACL 2025 VIENNA
JULY 27 - AUGUST 1

Disambiguating Reference in Visually Grounded Dialogues through Joint Modeling of Textual and Multimodal Semantic Structures

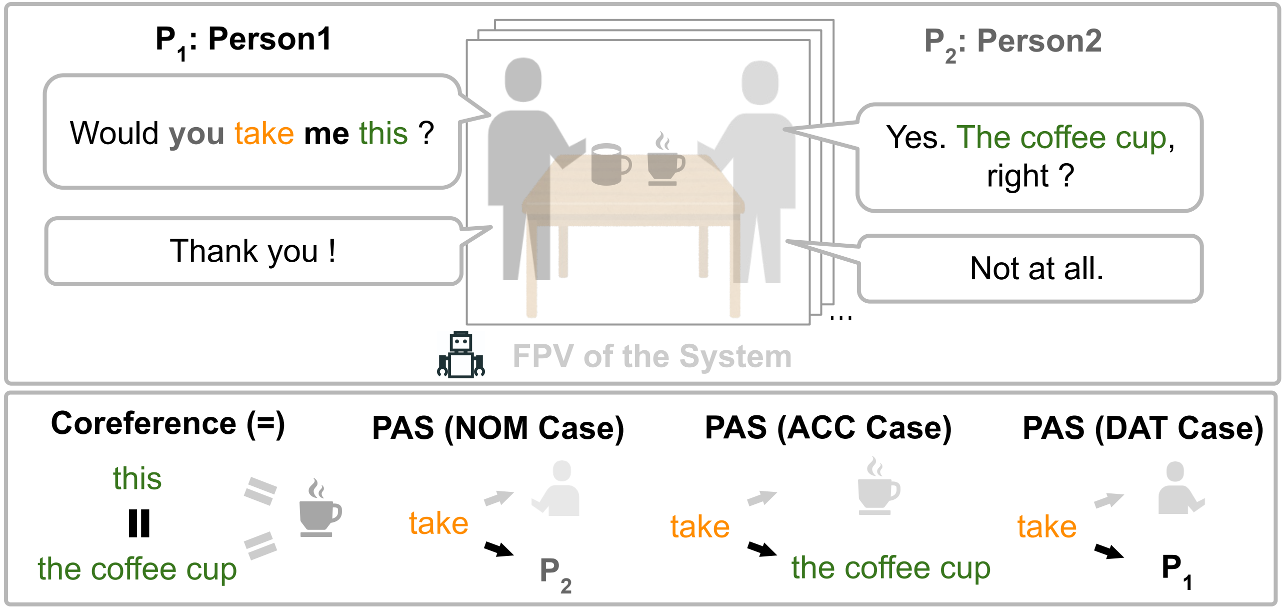
Shun Inadumi [1,2], Nobuhiro Ueda [3,†] and Koichiro Yoshino [4,2,1]

1. NAIST, 2. RIKEN GRP, 3. Kyoto University, 4. Science Tokyo

[†]Currently at NEC Corporation.

Understanding References in **Dialogue Text**

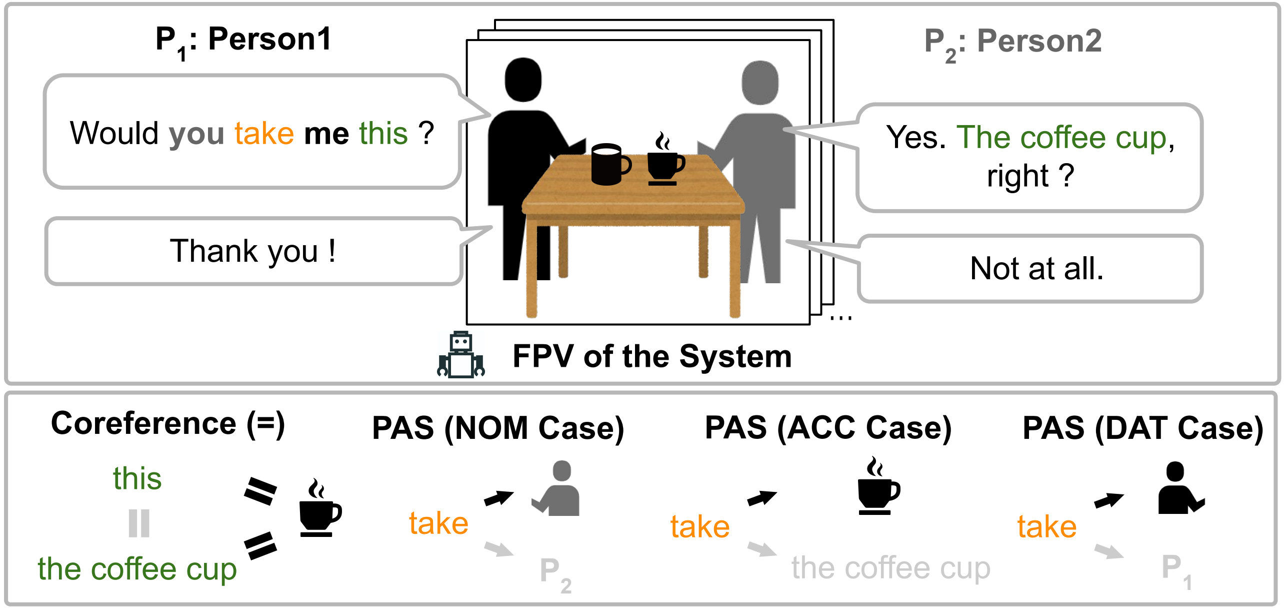
- **TRR: T**extual **R**eference **R**esolution:
  - Identify textual reference relations between phrases
  - TRR consists of coreference resolution, predicate-argument structure analysis and bridging anaphora resolution.



Example of the system analyzes a two-person dialogue

Understanding References in **Visually Grounded Dialogues**

- **MRR: M**ultimodal **R**eference **R**esolution [Ueda+, 2024] :
  - Identify phrase-to-objects reference relations
  - Direct reference and Indirect reference

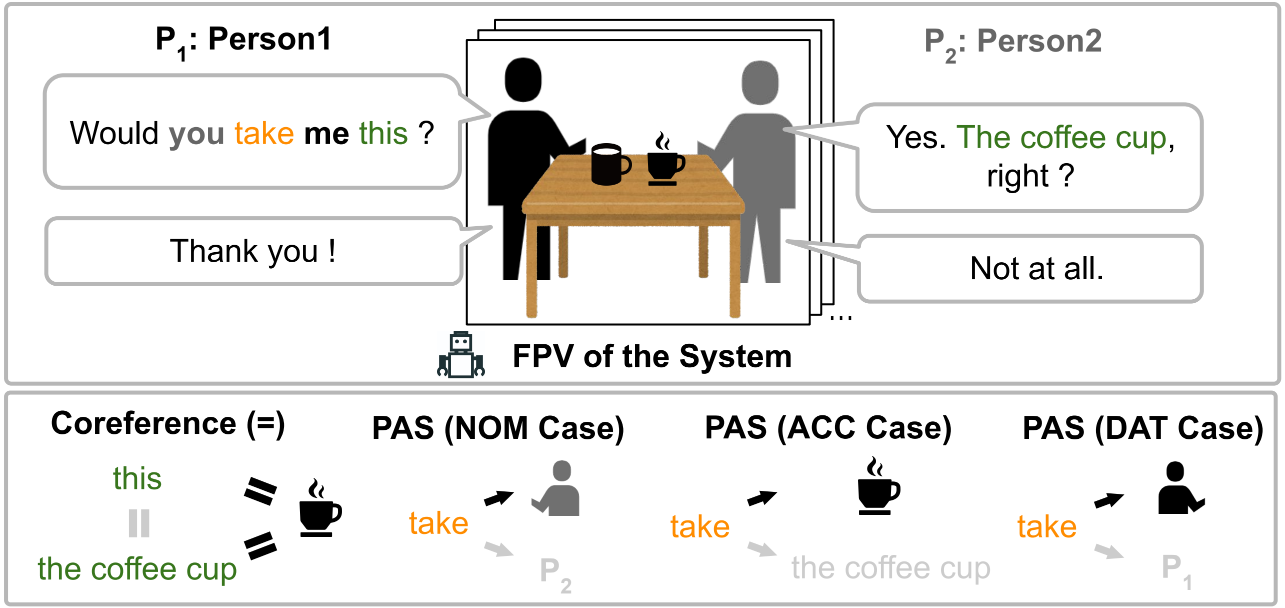


Example of the system analyzes a two-person dialogue

Understanding References in **Visually Grounded Dialogues**

- **Multimodal Reference R**
  - Identify phrase-to-obje
  - Direct reference and ir

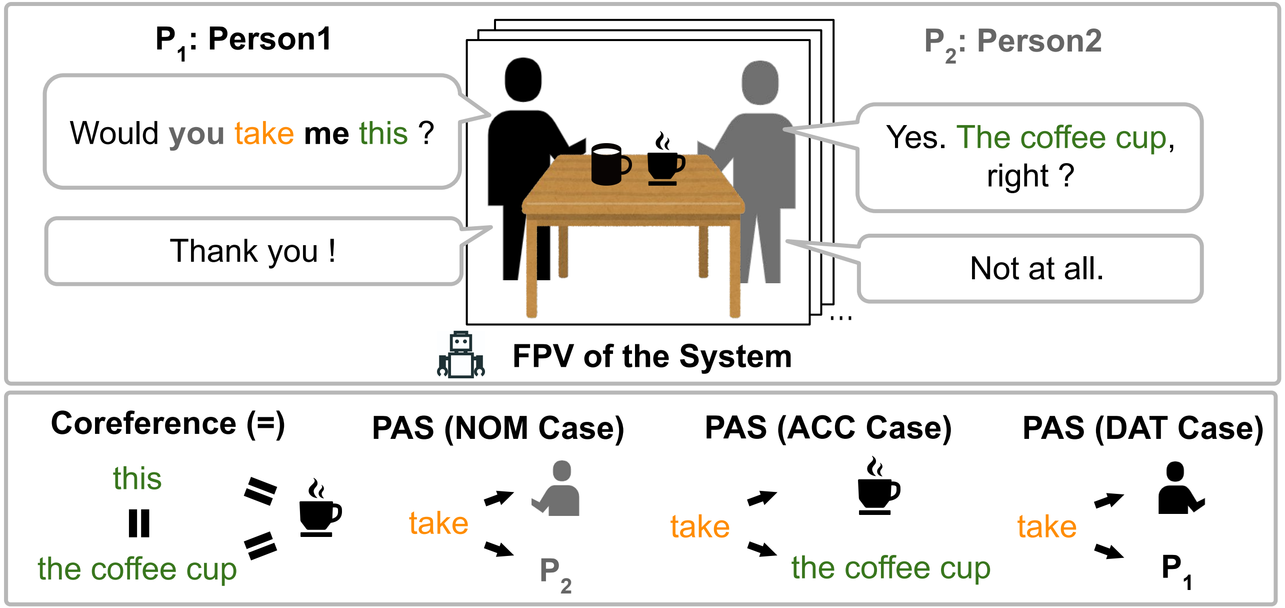
Phrase grounding [Plummer+, 2015] refers to the task of predicting only this type of reference.



Example of the system analyzes a two-person dialogue

Why is MRR important?

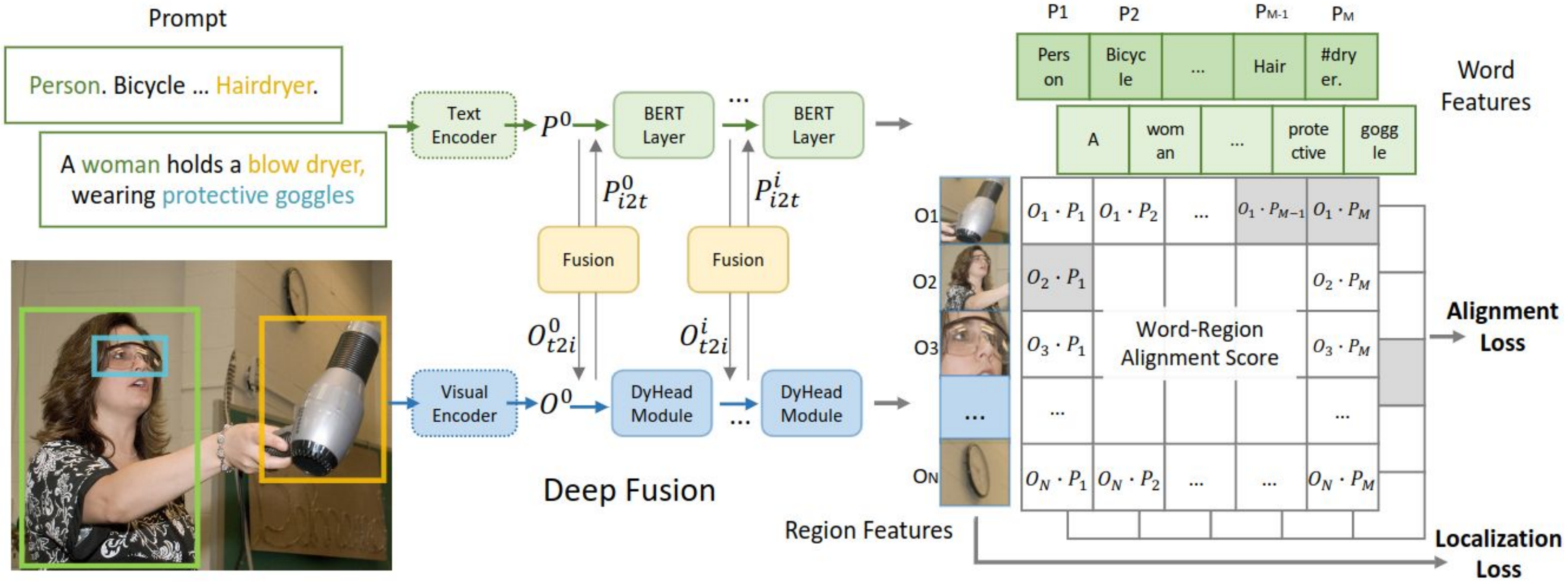
MRR enables systems to understand **dialogue events**
— such as "who does what to whom"— linked to **real-world objects**.



Example of the system analyzes a two-person dialogue

Limitations of Existing Models

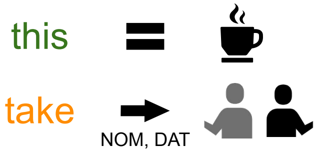
- GLIP (Grounded Language-Image Pretraining) [Li+, 2022] :
  A phrase grounding model trained on large-scale **image-caption pairs** with direct references.



Cited from [Li+, 2022].

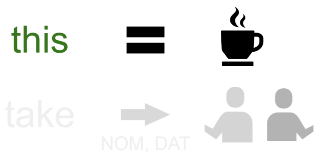
Limitations of Existing Models

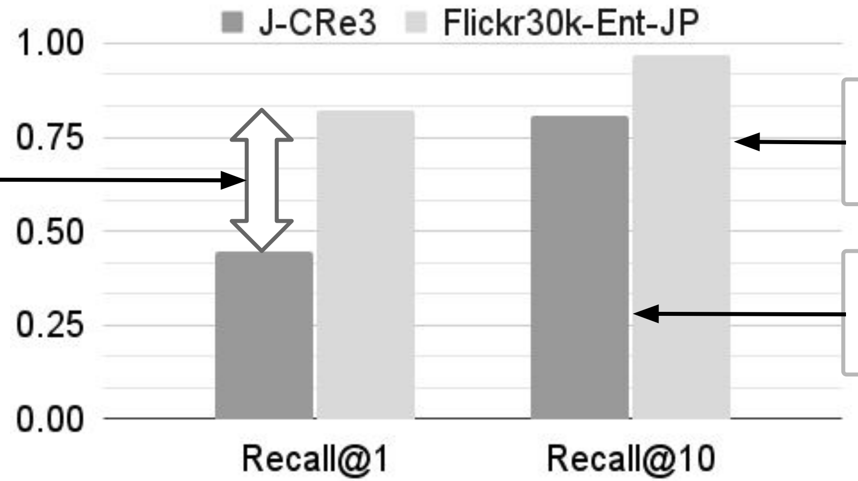
- In dialogue parsing, GLIP struggles with:
    - Resolving direct references made via **pronouns**
    - Parsing indirect references involving **ellipses**



Limitations of Existing Models

- In dialogue parsing, GLIP struggles with:
  - Resolving direct references made via **pronouns**
  - Parsing indirect references involving **ellipses**





Performance gap

Japanese Captions
[Nakayama+, 2020]

Japanese Dialogues
[Ueda+, 2024]

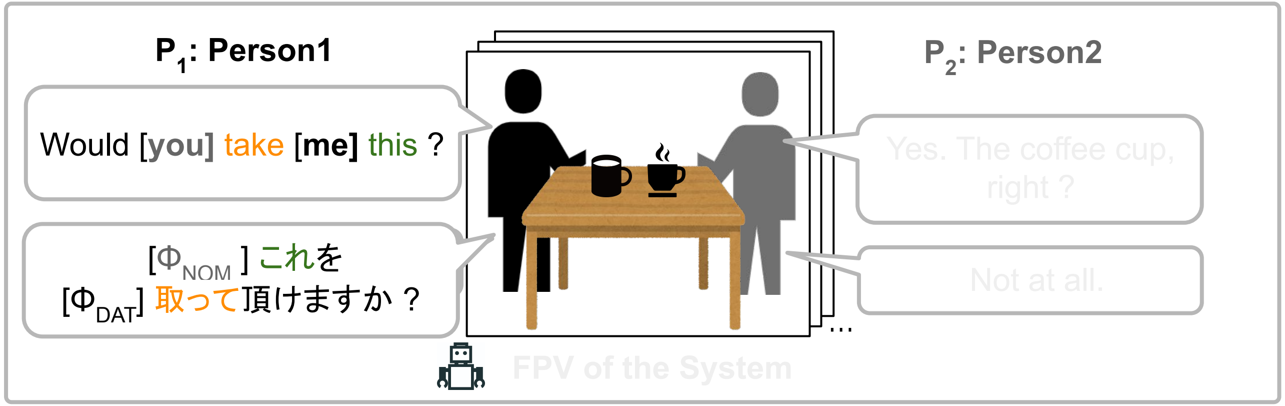
Phrase grounding results of GLIP

Limitations of Existing Models

- In dialogue parsing, GLIP struggles with:
  - ○ Resolving direct references made via **pronouns**
  - ○ Parsing indirect references involving **ellipses**

this =

take → NOM, DAT



P$_1$: Person1

Would [you] take [me] this ?

$[\Phi_{NOM}]$ これを
$[\Phi_{DAT}]$ 取って頂けますか ？

P$_2$: Person2

Yes. The coffee cup, right ?

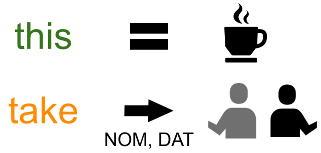
Not at all.

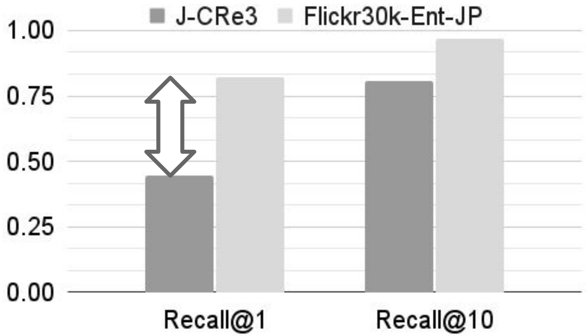
FPV of the System

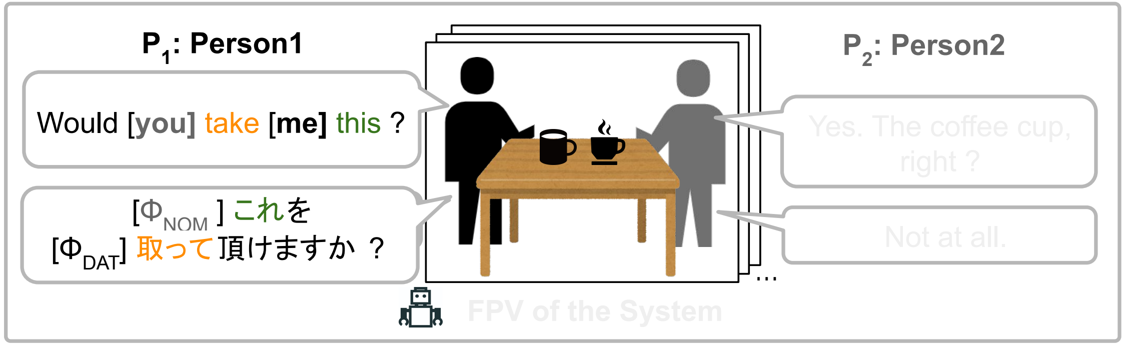
Pro-drop languages, such as Japanese, often omit **subjects** and **objects**.

Limitations of Existing Models

- In dialogue parsing, GLIP struggles with:
  - Resolving direct references made via **pronouns**
  - Parsing indirect references involving **ellipses**





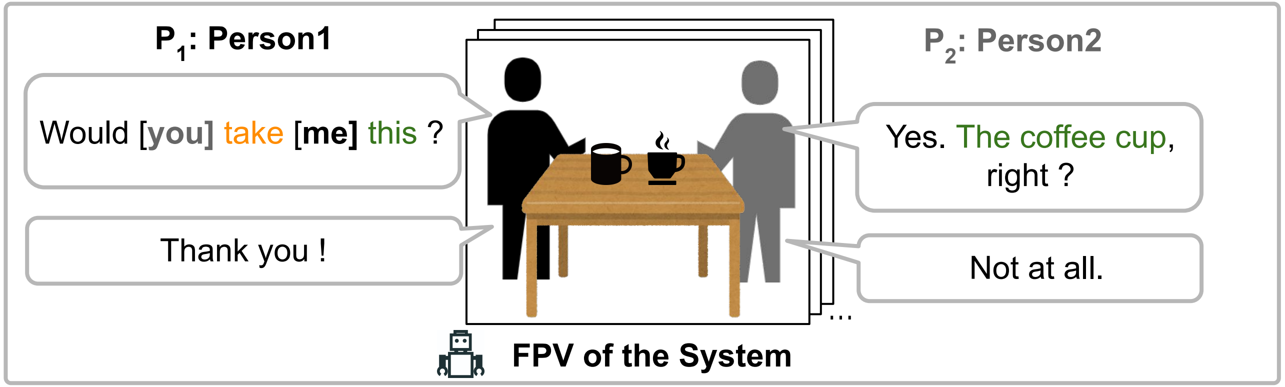

P$_1$: Person1
Would [you] take [me] this ?
[Φ$_{NOM}$] これを
[Φ$_{DAT}$] 取って頂けますか ？
P$_2$: Person2
Yes. The coffee cup, right ?
Not at all.
…
FPV of the System

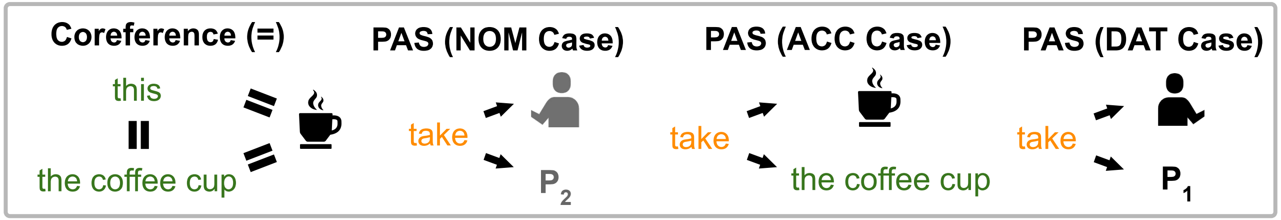
By resolving these ambiguities,
we aim to improve the understanding of real-world dialogues.

Using Textual Reference Relations 💡

By incorporating textual references, we can improve MRR performance.

e.g.) If the coffee cup = this = ☕ is known, this = ☕ can be uniquely identified.

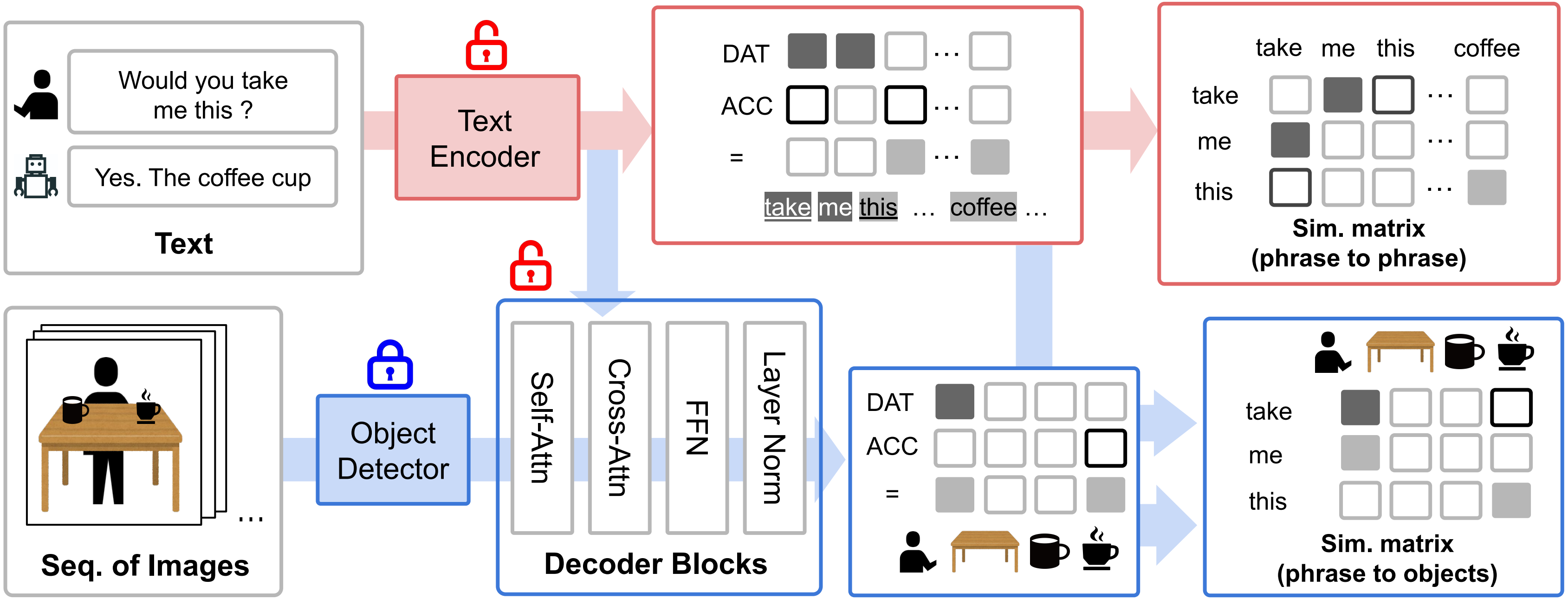




Proposed Framework 💡

We propose a framework to jointly model TRR and MRR.

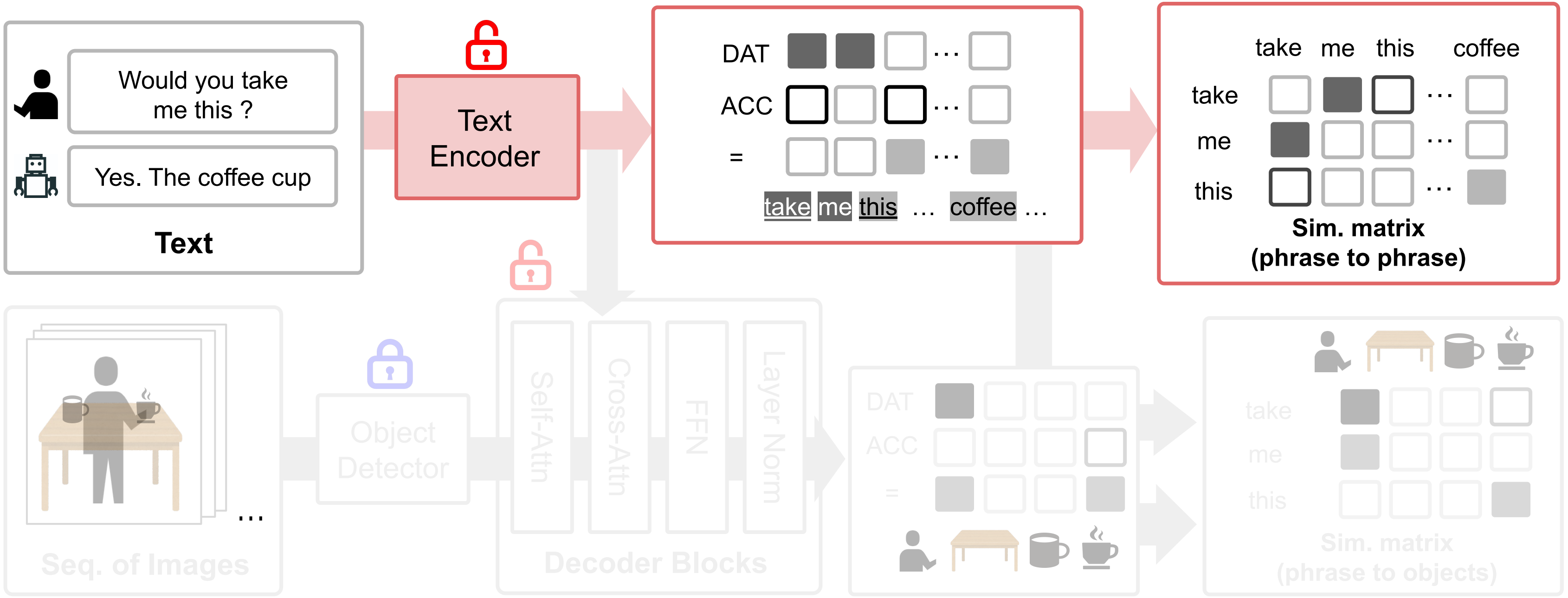
Learning to align phrase embeddings with object features



Step1: Textual Reference Resolution

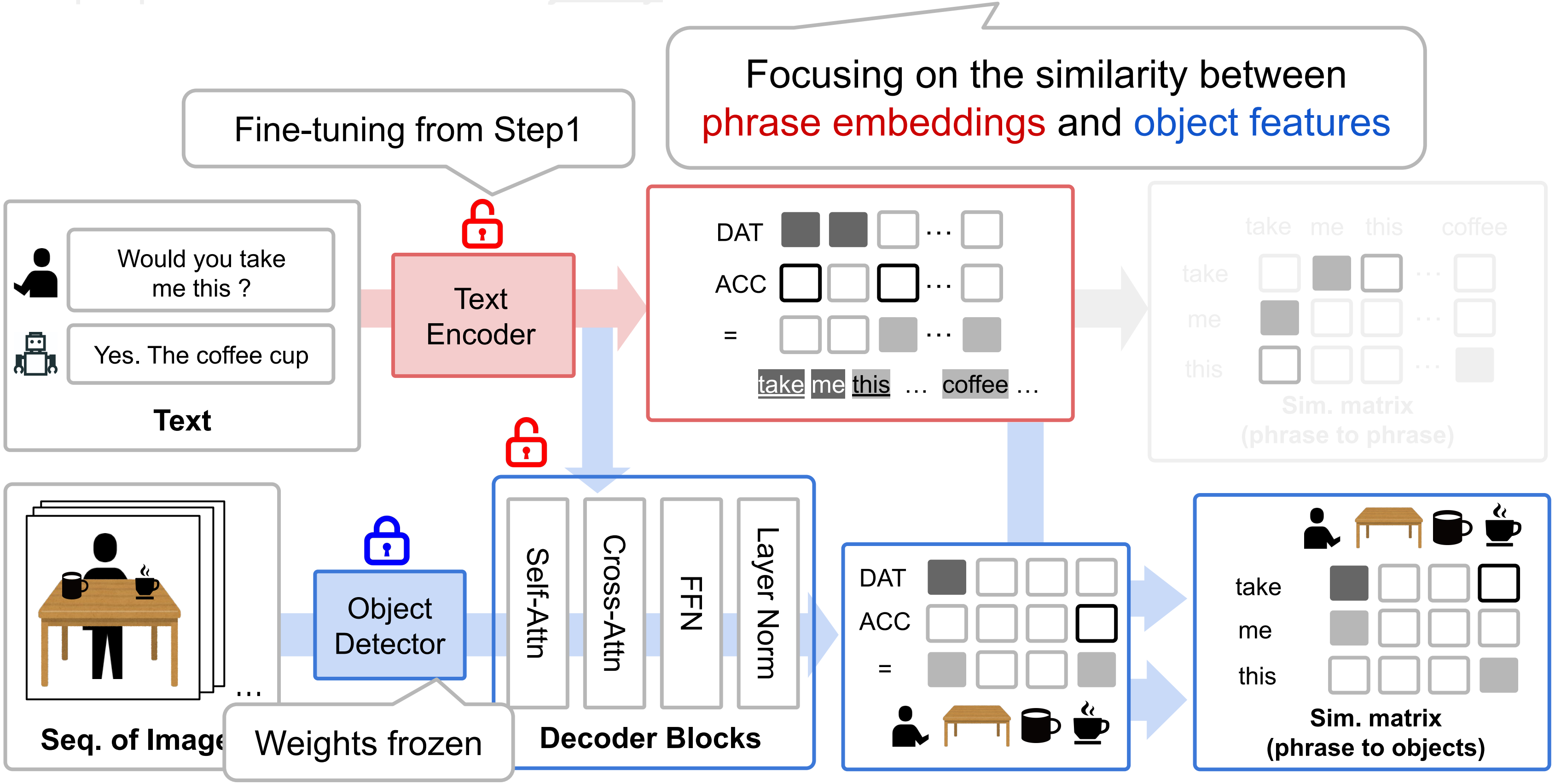
We propose a framework to jointly model TRR and MRR.

Focusing on the similarity between phrase embeddings



Step2: Multimodal Reference Resolution

We propose a framework to jointly model TRR and MRR.



Focusing on the similarity between phrase embeddings and object features

Fine-tuning from Step1

Would you take me this ?

Yes. The coffee cup

Text

Text Encoder

DAT

ACC

=

take me this … coffee …

take me this coffee

take

me

this

Sim. matrix (phrase to phrase)

Seq. of Image

Object Detector

Weights frozen

Self-Attn

Cross-Attn

FFN

Layer Norm

Decoder Blocks

DAT

ACC

=

take

me

this

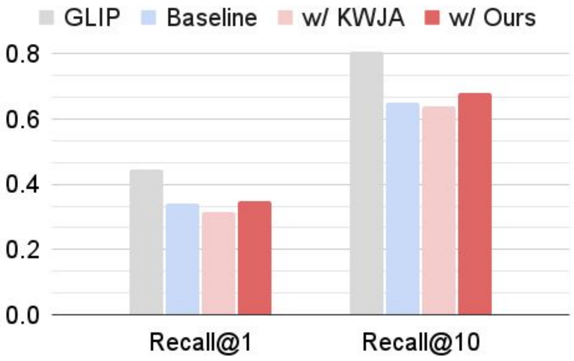
Sim. matrix (phrase to objects)

Phrase Grounding Results

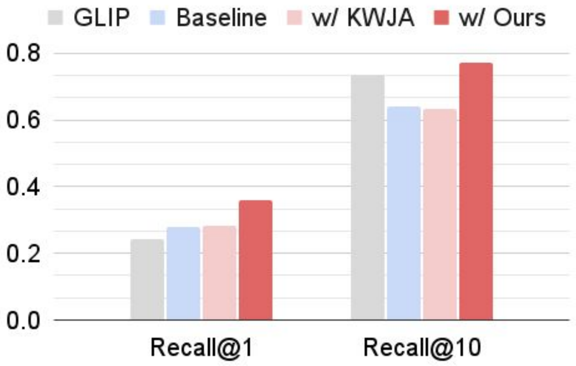
- Compared models:
  - Baseline
  - Baseline w/ Ours
  - Baseline w/ KWJA [Ueda+, 2023]
  - GLIP

Phrase grounding model with coreference resolution (fine-tuned on Japanese data [Nakayama+, 2020, Ueda+, 2024])

Pre-trained on English data [Krishna+, 2017, Hudson+, 2019]



Japanese Dialogue (Overall, 996)



Japanese Dialogue (**Pronouns**, 120/996)

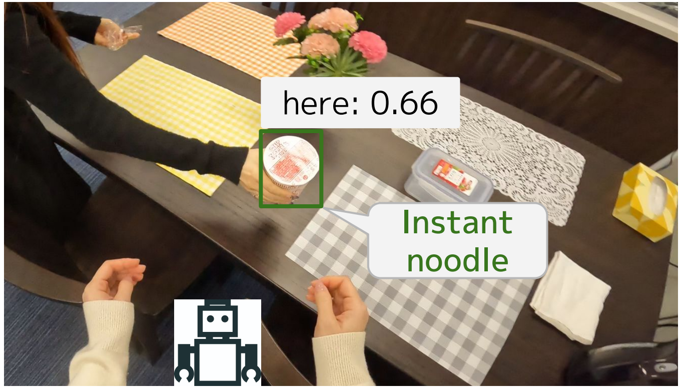
Baseline w/ Ours achieved improved **pronoun phrase grounding** through **coreference resolution**.

Phrase Grounding Results

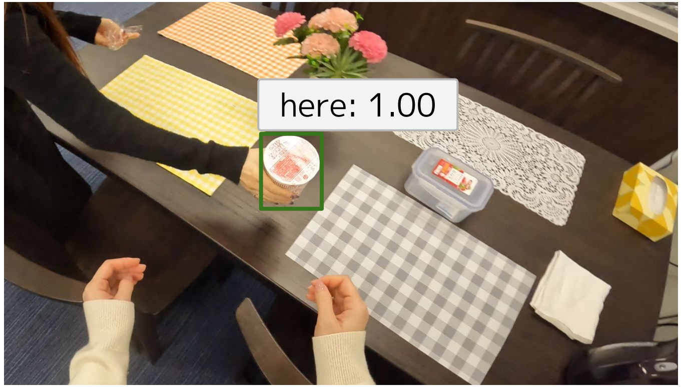
- Compared models:
  - ○ Baseline
  - ○ Baseline w/ Ours

Phrase grounding model with coreference resolution
(fine-tuned on Japanese data [Nakayama+, 2020, Ueda+, 2024])

Baseline



Baseline w/ Ours



Can you put [the water] in here
since it comes up ?

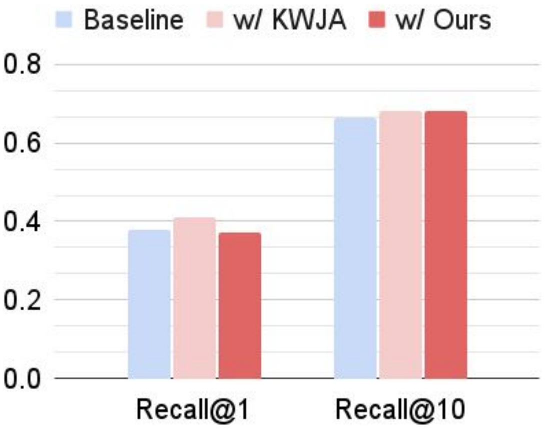
Japanese
omits [ ] phrases.

Coreference resolution strengthens confidence scores
in pronoun-to-object predictions.

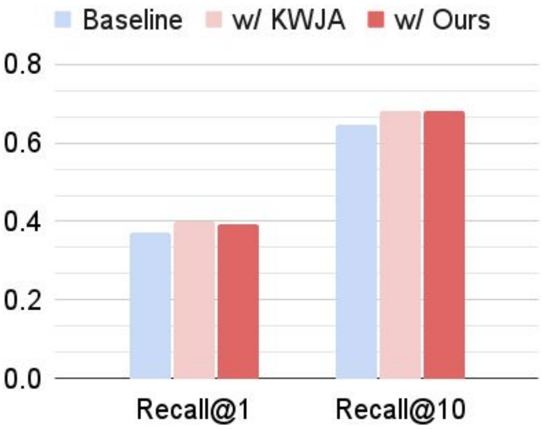
Multimodal Reference Resolution Results

- Compared models:
  - Baseline
  - Baseline w/ Ours
  - Baseline w/ KWJA [Ueda+, 2023]

MRR model with TRR
(fine-tuned on Japanese data [Ueda+, 2024])



Indirect (Predicate-argument structure)



Indirect (Bridging anaphora)

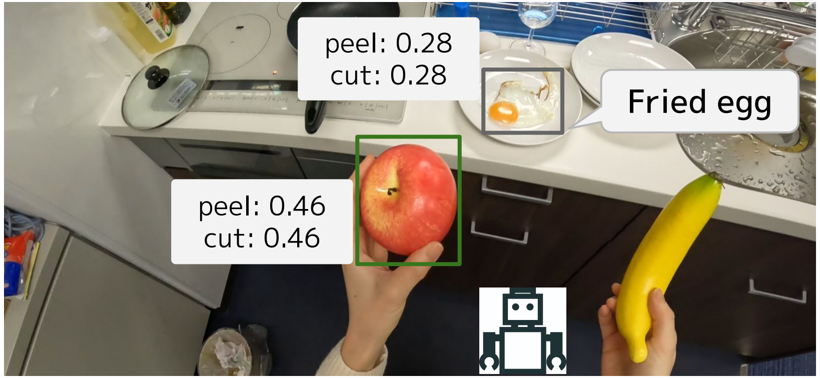
Improved indirect reference performance
through textual reference resolution.

Multimodal Reference Resolution Results

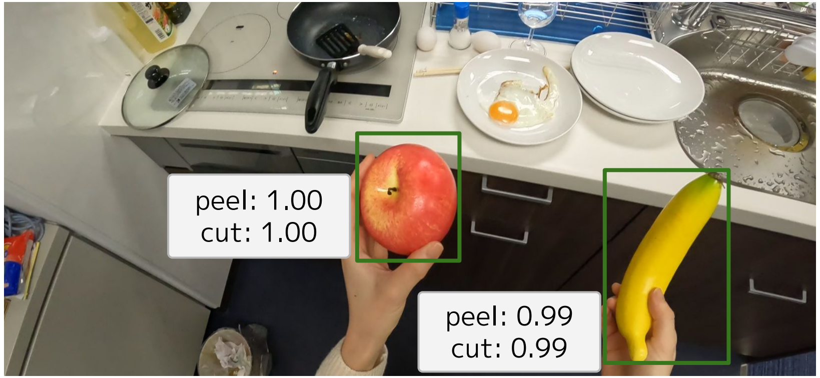
- Compared models:
  - Baseline
  - Baseline w/ KWJA [Ueda+, 2023]

MRR model with TRR
(fine-tuned on Japanese data [Ueda+, 2024])

Baseline



Baseline w/ KWJA



Shall we peel both [the apple and the banana]?
Then, let’s cut [them] into portions for three people.

Japanese
omits [ ] phrases.

TRR strengthens confidence scores for predicates.

Conclusion

Purpose :

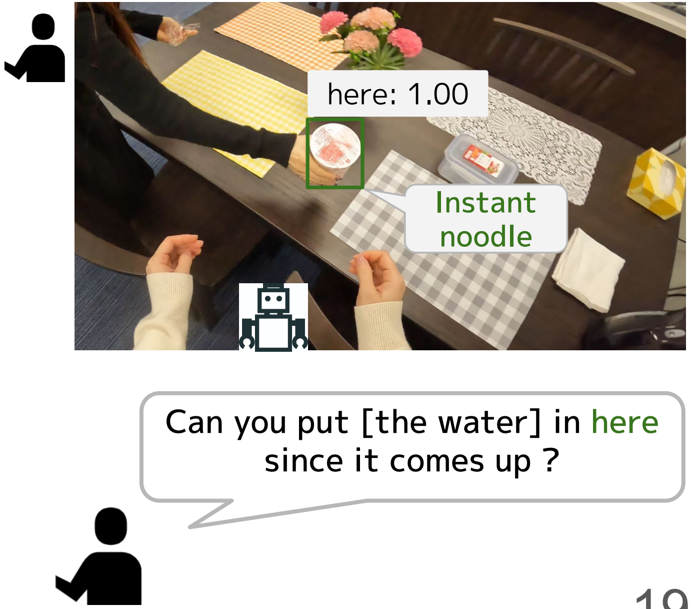
- Through **resolving ambiguities in visually-grounded dialogues**, we aim to improve the understanding of real-world dialogues.

Idea 💡 :

- We propose a framework to jointly model textual reference resolution (TRR) and multimodal reference resolution.

Main Results :

- Improved pronoun phrase grounding through coreference resolution.
- Improved indirect reference performance through TRR.


here: 1.00
Instant noodle
Can you put [the water] in here since it comes up ?