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

Shun Inadumi [1,2], Nobuhiro Ueda [3,*], Koichiro Yoshino [4,2,1],
[1] Nara Institute of Science and Technology, [2] Guardian Robot Project, RIKEN
[3] Kyoto University, [4] Institute of Science Tokyo
inazumi.shun.in6@naist.ac.jp
ueda@nlp.ist.i.kyoto-u.ac.jp koichiro@c.titech.ac.jp

Abstract

Multimodal reference resolution, including phrase grounding, aims to understand the semantic relations between mentions and real-world objects. Phrase grounding between images and their captions is a well-established task. In contrast, for real-world applications, it is essential to integrate textual and multimodal reference resolution to unravel the reference relations within dialogue, especially in handling ambiguities caused by pronouns and ellipses. This paper presents a framework that unifies textual and multimodal reference resolution by mapping mention embeddings to object embeddings and selecting mentions or objects based on their similarity.[1] Our experiments show that learning textual reference resolution, such as coreference resolution and predicate-argument structure analysis, positively affects performance in multimodal reference resolution. In particular, our model with coreference resolution performs better in pronoun phrase grounding than representative models for this task, MDETR and GLIP. Our qualitative analysis demonstrates that incorporating textual reference relations strengthens the confidence scores between mentions, including pronouns and predicates, and objects, which can reduce the ambiguities that arise in visually grounded dialogues.

1 Introduction

Understanding what mentions refer to objects in visually grounded dialogues is key to realizing a system that can collaborate with users in the real world, including robots and embodied agents (Yu et al., 2019; Kottur et al., 2021; Wu et al., 2023; Ueda et al., 2024). Recent studies have focused on identifying the objects referred to by mentions, as exemplified by "the coffee cup" and "this cup" in Figure 1, which are called direct references.

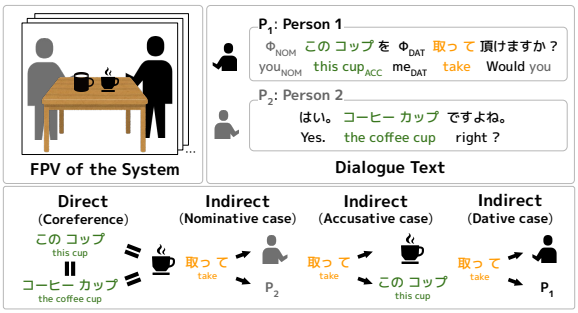


Figure 1: Example of textual and multimodal reference resolutions in the system analyzes a two-person dialogue, "Would you take me this cup? — Yes. the coffee cup, right?," from its first-person view. Japanese omits the **subject** and **object** of the **predicate** "take."

Among these studies, multimodal reference resolution is a task that identifies semantic structures (Fillmore, 1968; Clark, 1975), not only direct references but also indirect references between mentions and objects (Ueda et al., 2024). Figure 1 illustrates how the system understands the objects referred to indirectly by "take." In this example, the system can identify the objects referred to by "this cup" and "the coffee cup" through both direct and indirect references, inferred from the predicate-argument structures of "take." This allows the system to understand "who does what to whom," "when," and "where" in a dialogue by relating events to objects.

Pronouns and ellipses are typical instances of direct and indirect reference. These are frequently used in dialogue and pose challenges for multimodal reference resolution (Ueda et al., 2024). Phrase grounding between images and their captions is a well-established task (Kazemzadeh et al., 2014; Plummer et al., 2017); however, identifying direct references is still insufficient in visually grounded dialogues.[2] Figure 1 illustrates an exam-

[*] Currently at NEC Corporation.

[1] The code is publicly available at https://github.com/SInadumi/mmrr.

[2] Example of Table 3 shows the performance of GLIP (Li et al., 2022) for phrase grounding in Japanese dialogue. The

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11183–11198

ple where pronouns and ellipses of topical terms, such as subjects and objects, present difficulties for such a framework.

To address these challenges, we aim to improve the performance of multimodal reference resolution, thereby disambiguating references in visually grounded dialogues. Previous work on phrase grounding in dialogues (Das et al., 2017) shows that coreference resolution improves pronoun grounding performance (Yu et al., 2022; Lu et al., 2022). Furthermore, joint modeling of textual references[3] can improve performance in textual reference resolution tasks (Shibata and Kurohashi, 2018; Omori and Komachi, 2019; Ueda et al., 2020). Inspired by these works, we hypothesize that introducing linguistic features of textual references can also benefit multimodal reference resolution in dialogues. As "this cup" and "the coffee cup" refer to the same coffee cup, resolving the coreference between the mentions may help determine the direct and indirect reference.

In this study, we propose a framework for the joint modeling of all references. Our framework unifies textual reference resolution — including coreference resolution and predicate-argument structure analysis (Iida et al., 2007) — and multimodal reference resolution, by following recent advances in multimodal representation learning (Gupta et al., 2020; Radford et al., 2021; Li et al., 2022). We map mention embeddings to object embeddings and select mentions or objects based on their similarity. Our multimodal reference resolution model explicitly addresses indirect references and handles ellipses.

Quantitative results using our framework show the effectiveness of training textual reference relations in improving the performance in analyzing direct (§ 4.2) and indirect (§ 4.3) references. Our findings suggest that textual reference resolution positively contributes to multimodal reference resolution. In particular, our model with coreference resolution performs better in pronoun phrase grounding than representative models for this task, MDETR (Kamath et al., 2021) and GLIP (Li et al., 2022).

Our qualitative analysis shows that incorporating textual reference relations strengthens the confidence scores between mentions, including pronouns and predicates, and objects. These findings are also consistent with our quantitative results. For example, in Figure 1, we observe an increase in the confidence scores for predicting the objects referred to by mentions such as "this cup" and "take." Thus, this study demonstrates that textual reference resolution can reduce the ambiguities in visually grounded dialogues, particularly those caused by pronouns and ellipses.

Japanese case marker	**Case and anaphora relations (Abbreviations)**
"*ga*" (ガ)	Nominative case (**NOM**)
"*wo*" (ヲ)	Accusative case (**ACC**)
"*ni*" (ニ)	Dative case (**DAT**)
"*de*" (デ)	Instrumental case (**INS**)
	Locative case (**LOC**)
"*no*" (ノ)	Bridging anaphora

Table 1: Types of indirect references: We show the corresponding cases (Fillmore, 1968) and bridging anaphora (Clark, 1975) based on Japanese case markers.

2 Preliminaries

A dataset for multimodal reference resolution in two-party dialogues, the J-CRe3[4] (Ueda et al., 2024), which this study employs for experiments. The following describes reference resolution (§ 2.1.1), which consists of textual reference resolution (TRR, § 2.1.2) and multimodal reference resolution (MRR, § 2.1.3).

2.1 Task Settings

2.1.1 Reference Resolution

Given a text $\mathbf{T}$ and a sequence of images $\mathbf{V} = \{\mathbf{I}_1 \cdots \mathbf{I} \cdots\}$ corresponding to $\mathbf{T}$, reference resolution identifies the reference relations that exist between mentions and objects or events they refer to. This task consists of TRR, which analyzes between mentions, and MRR, which analyzes between mentions and objects.

As shown in Figure 1, many instances of reference relations connected by direct reference are clarified by the chain of cases between the predicate and its arguments. In addition to direct references (marked by "="), we define five labels (i.e., case) to represent the types of semantic connections in indirect references (Table 1). Let L denote the set of six types of all reference relations l.

Recall@1 is 0.377, significantly worse than that for general Japanese captions.

[3]It is a general term that encompasses coreference, case relations (Fillmore, 1968) in predicate-argument structures, and bridging anaphora (Clark, 1975).

[4]J-CRe3: Japanese Conversation Dataset for Real-world Reference Resolution

2.1.2 Textual Reference Resolution

Given a text $\mathbf{T}$, TRR identifies phrases that have a reference relation with another phrase. We refer to such phrases as mentions. We use the term TRR to refer collectively to coreference resolution, predicate-argument structure (PAS) analysis (Iida et al., 2007), and bridging anaphora (BA) resolution (Poesio and Vieira, 1998; Kobayashi and Ng, 2020; Yu and Poesio, 2020).

2.1.3 Multimodal Reference Resolution

Given a text $\mathbf{T}$ and an image $\mathbf{I}$, MRR identifies objects in $\mathbf{I}$ that have reference relations with mentions in $\mathbf{T}$. Specifically, MRR involves an object detection process that estimates up to q tuples of bounding boxes $\mathbf{O}$ and object feature series $\mathbf{X}$ for $\mathbf{I}$, denoted as $(\mathbf{O}, \mathbf{X}) = \{(o_1, \mathbf{x}_1) \cdots (o_q, \mathbf{x}_q)\}$. Based on $\mathbf{T}$ and $\mathbf{X}$, we identify elements in $\mathbf{O}$ that have reference relations with a mention. Identifying only direct references ("=") to objects from a mention is called phrase grounding (Kazemzadeh et al., 2014; Plummer et al., 2017).

2.2 J-CRe3 Annotation

J-CRe3 (Ueda et al., 2024) is a dataset of real-world interactions during collaborative work between a master and an assistive robot, including first-person video of the robot, third-person video, and audio/transcription of their dialogue. We use first-person videos, dialogue transcriptions, and annotations for reference resolution, including textual reference relations, mention-to-object direct/indirect reference relations, and bounding boxes.

Ambiguous expressions, including pronouns and ellipses, frequently occur in spoken language (Piantadosi et al., 2012), with ellipses occurring especially in Japanese (Seki et al., 2002), Chinese (Kong and Zhou, 2010), and Korean (Park et al., 2015). In the ellipses, for example, indirect references from predicates to objects can exist without explicitly mentioning those objects. Following the previous work (Ueda et al., 2024), we refer to these instances as zero references, as in the case of zero anaphora in text (Sasano et al., 2008). While datasets addressing pronouns have existed in the past (Das et al., 2017; Kottur et al., 2021; Wu et al., 2023; Goel et al., 2023b), J-CRe3 differs in that it also includes zero references.[5]

[5]Although the work of Oguz et al. (2023, 2024) addresses noun phrase ellipsis, J-CRe3 is the only dataset that provides annotations for the explicit estimation of zero references from predicates to objects.

3 Methodology

3.1 Motivation

Visually grounded dialogue datasets, including J-CRe3, have limited training data, especially in Japanese and minor languages. For this reason, state-of-the-art models such as MDETR (Kamath et al., 2021) and GLIP (Li et al., 2022), which require large image-text pairs for training, are difficult to train on Japanese datasets only.

The previous work proposed a method for handling MRR that treats phrase grounding and TRR independently before combining their results (Ueda et al., 2024). However, this method is unable to account for zero references.

To address these problems, we design an MRR model inspired by weakly supervised phrase grounding models (Gupta et al., 2020; Goel et al., 2023a), which can be trained with limited data. Our unified framework integrates learning and analysis processes for TRR and MRR using mention and object embeddings. In particular, our MRR model explicitly considers indirect references between mentions and objects, with reference to the existing Japanese text analyzer (Ueda et al., 2023).

The difficulty of MRR lies in the need for the system to also recognize references for pronouns and ellipses. In our integrated framework, we expect TRR to enhance MRR by supplementing information for ambiguous direct and indirect references, such as from pronouns and ellipses.

3.2 Overview of Our Unified Framework

Figure 2 shows the overview of our unified framework for reference resolution. Our MRR model uses a frozen object detector (Zhou et al., 2022) while fine-tuning a text encoder (Devlin et al., 2019) and a fusion module (Liu et al., 2024) that integrates text $\mathbf{T}$ with an object feature series $\mathbf{X}$. We train two models separately: one for TRR (§ 3.2.1) and the other for MRR (§ 3.2.2). They share the text encoder weights.

3.2.1 Textual Reference Resolution Model

Our TRR model for analyzing all textual reference relations in Figure 2 is based on the similarity of embeddings between mentions.

Given a dialogue text $\mathbf{T}$, the text encoder outputs subword embeddings $\mathbf{T}' \in \mathbb{R}^{p \times d_T}$ of input length p and dimension d_T. The extended representation $\widehat{\mathbf{T}}$ for $\mathbf{T}'$ is as follows:

$$\widehat{\mathbf{T}} = \mathbf{T}'\mathbf{W}_{T1} \in \mathbb{R}^{p \times d_T \times |L|}, \quad (1)$$

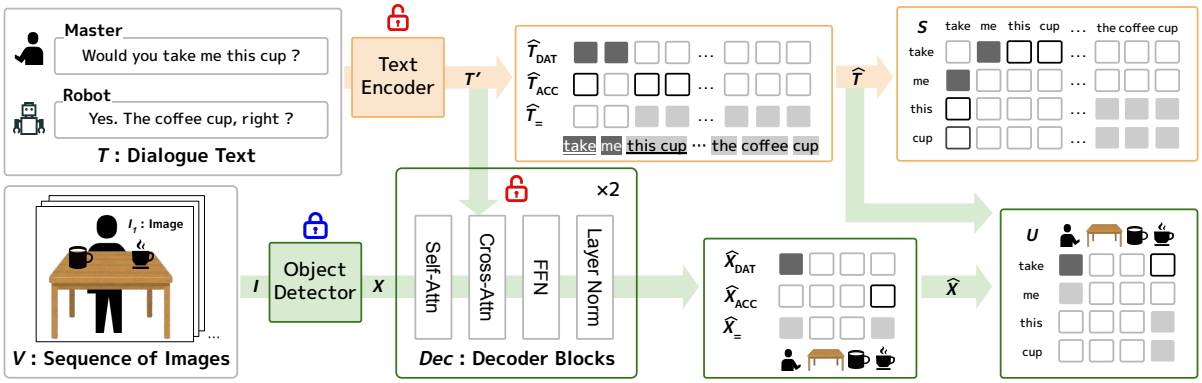


Figure 2: Overview of our framework for J-CRe3: The orange and green indicate the processing flows for TRR and MRR, respectively.

where, $\mathbf{W}_{T1} \in \mathbb{R}^{d_T \times d_T \times |L|}$ is trainable parameters. Our model calculates the dot product of each $\widehat{\mathbf{T}}$ per a relation $l \in L$ and the similarity matrix $\mathbf{S}_l$ as follows:

$$\mathbf{S}_l = \widehat{\mathbf{T}}_l \widehat{\mathbf{T}}_l^\top \in \mathbb{R}^{p \times p}. \quad (2)$$

We use $\mathbf{S}_l$ to select a mention that the other mention refers to.

The embeddings $\mathbf{T}'$ are at a subword level, but a mention is at a basic phrase level.[6] To link these different units, we use the first subword of a mention as its main representation during learning and inference. This step is the same as for our MRR model described below.

3.2.2 Multimodal Reference Resolution Model

Our MRR model for analyzing direct and indirect reference relations in Figure 2 is based on the similarity of mention and object embeddings.

Given a dialogue text $\mathbf{T}$ and an image $\mathbf{I}$, which is an element of frames $\mathbf{V}$, a text encoder and an object detector output $\mathbf{T}'$ and $(\mathbf{O}, \mathbf{X})$, respectively. Then, our model uses a single linear layer and aligns the dimension d_T of $\mathbf{T}'$ and the dimension d_O of $\mathbf{X}$ to d_S. The extended representation $\widehat{\mathbf{T}}$ and $\widehat{\mathbf{X}}$ for $\mathbf{T}'$ and $\mathbf{X}$ as follows:

$$\widehat{\mathbf{T}} = \mathbf{T}'\mathbf{W}_{T2} \in \mathbb{R}^{p \times d_S \times |L|}, \quad (3)$$

$$\widehat{\mathbf{X}} = Dec(\mathbf{X}, \mathbf{T}')\mathbf{W}_O \in \mathbb{R}^{q \times d_S \times |L|}, \quad (4)$$

$$Dec(\mathbf{X}, \mathbf{T}') \in \mathbb{R}^{q \times d_S},$$

where $(\mathbf{W}_{T2}, \mathbf{W}_O) \in \mathbb{R}^{d_S \times d_S \times |L|}$ are trainable parameters and $Dec(\cdot)$ is two decoder blocks, which uses cross-attention to condition $\mathbf{X}'$ on $\mathbf{T}'$ (Liu et al., 2024). Our model calculates the dot product of each $\widehat{\mathbf{T}}$ and $\widehat{\mathbf{X}}$ per a relation $l \in L$ and the similarity matrix $\mathbf{U}_l$ as follows:

$$\mathbf{U}_l = \widehat{\mathbf{X}}_l \widehat{\mathbf{T}}_l^\top \in \mathbb{R}^{p \times q}. \quad (5)$$

We use the similarity matrix $\mathbf{U}_l$ to select elements of $\mathbf{O}$ from a mention.

We consider frames $\mathbf{V}$ to be a sequence of one-second intervals extracted from a video from the start and end times of an utterance in a text $\mathbf{T}$. Following previous work (Gupta et al., 2020; Goel et al., 2023a), we use pooled features (Lin et al., 2017; Anderson et al., 2018) from the region proposal network (Ren et al., 2015) as object feature series $\mathbf{X}$.

3.2.3 Loss Functions

Using the softmax cross entropy used on phrase grounding (Li et al., 2022) and Japanese text analysis (Ueda et al., 2020), we define the loss functions for as follows:

$$\mathcal{L}_S = \sum_{l \in L} loss\{\mathbf{S}_l; \mathbf{S}_{(l,ground)}\}, \quad (6)$$

$$\mathcal{L}_U = \sum_{l \in L} loss\{\mathbf{U}_l; \mathbf{U}_{(l,ground)}\}, \quad (7)$$

where $\mathcal{L}_S$ corresponds to TRR model, and $\mathcal{L}_U$ corresponds to MRR model. Here, $\mathbf{S}_{(l,ground)} \in \{0,1\}^{p \times p}$ and $\mathbf{U}_{(l,ground)} \in \{0,1\}^{p \times q}$ represent matrices of positive examples of $\mathbf{S}_l$ and $\mathbf{U}_l$ in a reference relation l.

4 Experiments

4.1 Settings

Compared Models We assume the MRR-only model is a baseline model (§ 3.2.2). In our experiments, we first train a TRR model by only

[6] This phrase consists of one content word and zero or more function words.

coreference resolution, predicate-argument structure (PAS) analysis and bridging anaphora (BA) resolution, or TRR. Then, we leverage this text encoder to train the baseline model. This approach allows us to investigate how TRR benefits phrase grounding and MRR. As TRR models, we use our TRR model (§ 3.2.1) and a Japanese text analyzer, KWJA (Ueda et al., 2023). Both models can handle TRR well, but this study focuses on results in MRR.

For phrase grounding comparison with the MRR models, we fine-tune MDETR (Kamath et al., 2021) and GLIP (Li et al., 2022), representative models for this task on Japanese datasets. For MRR comparison, we consider a method that combines phrase grounding outputs from GLIP with TRR outputs from KWJA (Ueda et al., 2024), denoted as GLIP + KWJA. Specifically, GLIP performs phrase grounding on both text and image and outputs only i) direct reference relations, while KWJA performs TRR on text only and outputs ii) all textual reference relations. We derive indirect reference relations by aligning the i) and ii) relations.

Dataset We use J-CRe3 and Flickr30k-Ent-JP (Nakayama et al., 2020) as Japanese datasets to fine-tune our MRR model, MDETR, and GLIP. J-CRe3 contains 93 dialogues and a total of 11,062 images, with each dialogue containing 10 to 16 utterances. Flickr30k-Ent-JP includes images and corresponding Japanese captions with direct references between mentions and objects, totaling 31,783 captions and 63,566 images.

We pre-train MDETR and GLIP using Visual Genome (Krishna et al., 2017), GQA (Hudson and Manning, 2019), and Flickr30k-Ent-JP, with these weights serving as initial values for fine-tuning on the Japanese datasets. In contrast, the MRR models do not undergo pre-training and are instead fine-tuned directly on these datasets.

To train TRR models, we use J-CRe3 as well as a corpus of web documents (Hangyo et al., 2012), Wikipedia, and blog posts annotated with textual reference relations. These datasets contain 6,542 documents and dialogues for the training set.

Evaluation Metrics We use Recall@k (R@k; $k = \{1, 5, 10\}$) to evaluate phrase grounding and MRR. An MRR model, MDETR, and GLIP predict bounding boxes and their confidence scores for each mention. Recall@k is the percentage of times that ground truth boxes are among the top k predicted boxes with the highest confidence scores. We consider predicted boxes to match ground truth boxes if their Intersection-over-Union is 0.5 or greater.

Models	Text Encoder	Object Detector	Others	Total
Ours	339M	—	27M	366M
GLIP	278M	31M	92M	401M
MDETR	278M	8M	20M	306M

Table 2: Number of trainable parameters in the models: While the total number of parameters in our models remains mostly unchanged between MRR and phrase grounding, task-specific components such as W_{T2} and W_O differ depending on the task.

Implementation Details Our framework uses Japanese DeBERTa-v2-large (He et al., 2021b) as a text encoder, and Detic (Zhou et al., 2022) with Swin-Transformer (Liu et al., 2021) as its backbone for an object detector. MDETR and GLIP use mDeBERTa-v3-base (He et al., 2021a) as a text encoder. Table 2 shows the number of trainable parameters for the MRR models developed in our framework, GLIP, and MDETR.

We set the maximum length of the subword embeddings $\mathbf{T}'$ to $p = 256$ and the dimension to be $d_T = d_O = d_S = 1,024$. The MRR models output the maximum value of the predicted bounding boxes O for the two datasets: $q = 128$ for the J-CRe3 and $q = 256$ for the Flickr30k-Ent-JP.

For the TRR, each training instance consists of three sentences, shifting by one sentence at a time. For phrase grounding and MRR, the instance unit definitions vary by dataset. In J-CRe3, each training instance comprises three utterances, shifting by one utterance at a time, paired with an image; evaluation is performed on individual utterance–image pairs.[7] In Flickr30k-Ent-JP, each instance, used for training and evaluation, consists of up to five captions paired with an image.

We fine-tuned the TRR and MRR models using AdamW (Loshchilov and Hutter, 2019) with a learning rate of 5e-5, weight decay of 0.01, and 1,000 warmup steps and trained for 16 epochs with a batch size of 16 and 32. MDETR and GLIP are also fine-tuned with the same settings, except for 2 epochs and a batch size of 4 and 16. We performed the TRR and MRR models experiments with 4×RTX 3090s in 6 hours and MDETR and GLIP experiments with 2×RTX A6000s in 2 days.

[7] See Appendix C for ablation study results on the utterance length of $\mathbf{T}$ during evaluation of phrase grounding and MRR.

Models	Overall (996)			Nouns (671 / 996)			Pronouns (120 / 996)		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Coreference Resolution (Coref.) ⇒ **Phrase Grounding** (PG)									
Baseline †	0.342	0.567	0.649	0.344	0.574	0.664	0.277	0.527	0.641
w/ KWJA †	0.315	0.542	0.640	0.308	0.549	0.657	0.283	0.527	0.633
w/ Ours †	0.348	0.584	0.681	0.339	0.567	0.663	**0.361**	**0.683**	**0.772**
GLIP	**0.445**	**0.745**	**0.808**	**0.454**	**0.752**	**0.816**	0.241	0.650	0.733
MDETR	0.348	0.468	0.510	0.365	0.481	0.523	0.133	0.208	0.250

Models	R@1	R@5	R@10
Coref. ⇒ **PG**			
Baseline †	0.558	0.735	0.767
w/ KWJA †	0.559	0.733	0.767
w/ Ours †	0.560	0.733	0.767
GLIP	**0.822**	**0.951**	**0.970**
MDETR	0.769	0.896	0.924

Table 3: The results of phrase grounding: For models with †, we report the average of 3 randomly seeded training and evaluation iterations of MRR. The highlighted items indicate where the MRR models with coreference resolution show improvements or deteriorations compared to the baseline. **Left**: The results of J-CRe3: The numbers in parentheses indicate positive instances. **Right**: The results of Flickr30k-Ent-JP.

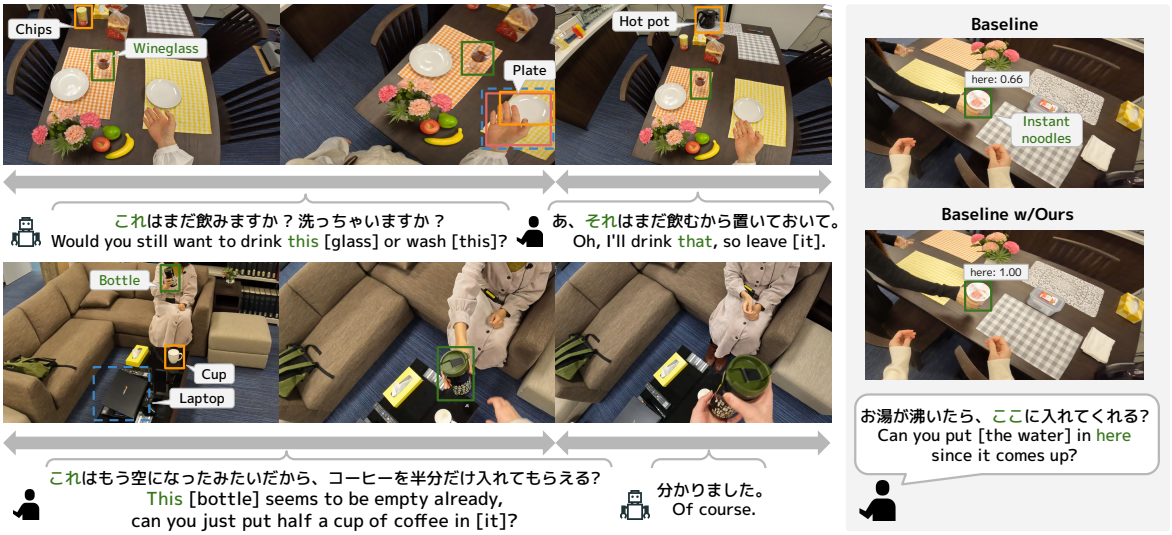


Figure 3: Examples of phrase grounding: The green mentions and objects are targets for grounding, and the mentions in square brackets are omitted in Japanese. **Left**: We show the Recall@1 errors of models in orange (GLIP), blue dashed (Baseline), and red (Baseline w/ Ours), while only incorrectly predicted bounding boxes are shown, as correct predictions are omitted. **Right**: We also show confidence scores.

4.2 Experiments on Phrase Grounding

Main Results Table 3 shows the results of phrase grounding. In the overall evaluation of J-CRe3, including noun phrases and pronouns, our MRR model with coreference resolution using our TRR model (Baseline w/ Ours) outperforms the baseline and MDETR in terms of Recall@5 and 10. While our MRR model performs slightly worse than the baseline for nouns, it considerably outperforms all other models, including GLIP, for pronouns.

Based on J-CRe3 results, we compare the baseline, Baseline w/ Ours, and an MRR model using KWJA as a TRR model (Baseline w/ KWJA). When our TRR model is used for phrase grounding, it improves performance for pronouns while minimizing performance degradation for nouns compared to KWJA. Our unified framework highlights the effectiveness of coreference resolution in phrase grounding for visually grounded dialogues.

Performance on Flickr30k-Ent-JP In the evaluation on Flickr30k-Ent-JP, GLIP shows the highest performance, and no change in Recall@k due to coreference resolution was observed for either Baseline w/ Ours or Baseline w/ KWJA. Unlike MDETR and GLIP, the MRR models use a frozen object detector, including Baseline w/ Ours. Thus, the upper bound of Recall@k for object detection depends on the detector. The actual upper bound is 0.799, which is the limiting value of Recall@k for MRR models.

Qualitative Analysis Figure 3 shows the examples of phrase grounding results in Table 3. As shown on the left side in Figure 3, our MRR model has fewer Recall@1 errors for pronouns such as "this" ("これ") and "it" ("それ") compared to the baseline and GLIP. Since all models, including our MRR model, rely on inference from a single image, they exhibited inconsistencies in predictions. For

Models	NOM (2,053)			ACC (915)			DAT (1,074)			INS–LOC (139)			Bridging (163)		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Textual Reference Resolution (TRR) ⇒ **Multimodal Reference Resolution** (MRR)															
Baseline †	0.568	0.735	0.763	0.229	0.505	0.606	0.559	0.726	0.749	0.124	0.350	0.465	0.378	0.564	0.662
w/ KWJA †	0.574	**0.756**	**0.785**	**0.240**	0.506	0.601	**0.582**	**0.748**	**0.779**	0.199	0.431	**0.561**	**0.411**	**0.613**	**0.680**
w/ Ours †	**0.585**	0.745	0.773	0.230	**0.520**	0.607	0.576	0.735	0.772	0.172	0.424	0.532	0.386	0.588	0.662
GLIP + KWJA	0.060	0.111	0.118	0.190	0.386	0.420	0.065	0.079	0.081	**0.273**	**0.510**	0.539	0.226	0.288	0.294

Table 4: The results of indirect references in MRR on the J-CRe3: The highlighted items indicate where the MRR models with TRR show improvements or deteriorations compared to the baseline. See the caption of Figure 3 for † and parentheses.

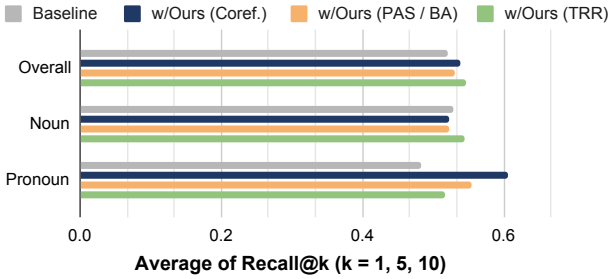


Figure 4: Ablation study results of Baseline w/ Ours in phrase grounding: We compare the improvements achieved by coreference resolution (Coref.), predicate-argument structure analysis and bridging anaphora resolution (PAS / BA), and textual reference resolution (TRR). See Table 5 for detailed results.

example, some errors involved estimating an object like "plate" for a mention of "this."

The right side of Figure 3 shows the examples of the baseline and our MRR model. Both models accurately predict the pronoun "here" ("ここ"), but their confidence scores for the object "instant noodle" differ: the baseline assigned a score of 0.66, whereas our model assigned a score of 1.00. Thus, our qualitative analysis demonstrates that incorporating coreference relations strengthens the confidence of pronoun-to-object predictions.

Comparison of Textual Reference Relations We discuss the benefits of incorporating textual reference relations into phrase grounding using the baseline and our MRR model. Figure 4 shows the ablation results for MRR models with coreference resolution, PAS analysis and BA resolution or TRR, in phrase grounding.

Incorporating textual reference relations improves pronoun performance compared to the baseline, regardless of the relation type, with coreference relations being the most effective. Furthermore, incorporating all textual reference relations yields the best performance for nouns. These results demonstrate the benefits of jointly modeling direct references, coreference, and case and bridging anaphora references, in phrase grounding for visually grounded dialogues.

4.3 Experiments on Multimodal Reference Resolution

Main Results Table 4 shows the results of indirect references in MRR. Our MRR model with TRR (Baseline w/ Ours) consistently outperforms the baseline and GLIP + KWJA in terms of Recall@10, regardless of the TRR models. However, GLIP + KWJA achieves the highest Recall@*1* and 5 for instrumental and locative cases (INS-LOC). Since GLIP + KWJA cannot handle zero references, the performance of INS-LOC in J-CRe3 probably depends on phrase grounding performance.

Qualitative Analysis Figure 5 shows the examples of indirect reference results in MRR of Table 4. Here, we focus on the Baseline w/ KWJA, which performs well in Table 4.

The left side of Figure 5 shows an example of zero references with two objects, "the apple" and "the banana," which are targets of accusative case relations (ACC), referred to by the mentions "peel" ("むいちゃおう") and "cut" ("カットしよう"). Compared to the baseline, the Baseline w/ KWJA correctly analyzes these objects.

The right side of Figure 5 shows an ACC and a bridging anaphora as examples of indirect reference results. Both models correctly analyzed the referenced object ("onion") from mentions, but the confidence scores were higher for Baseline w/ KWJA. Our qualitative analysis suggests that textual reference relations strengthen the prediction of objects for predicates and anaphoric mentions in learning MRR.

Comparison of Textual Reference Relations We discuss the benefits of incorporating textual reference relations into MRR using the baseline and our MRR model. Figure 6 shows the ablation

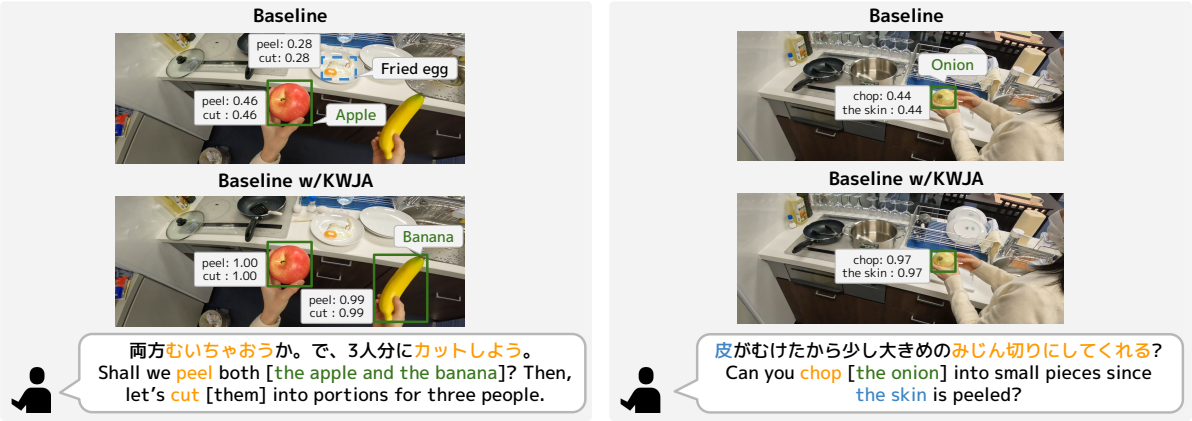


Figure 5: Examples of indirect references in MRR: The orange and blue mentions are targets for the accusative case and bridging anaphora, respectively. The green mentions correspond to referring objects. See the caption of Figure 3 for square brackets and green and blue dashed boxes.

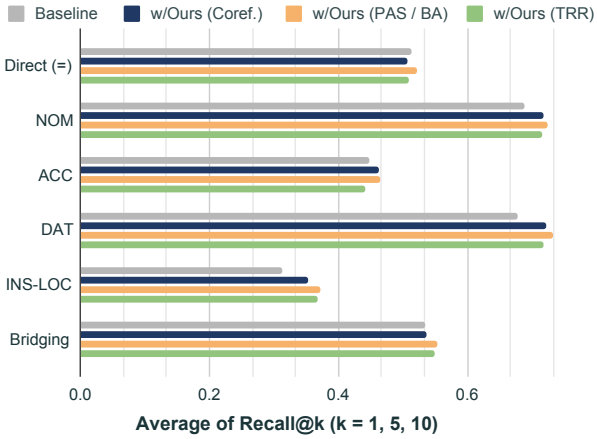


Figure 6: Ablation study results of Baseline w/ Ours in MRR: See the caption of Figure 4 for a detailed comparison setting and Tables 6 and 7 for detailed results.

results for MRR models with coreference resolution, PAS analysis and BA resolution or TRR in MRR.

The results for direct references in MRR show that our model did not outperform the baseline, regardless of whether coreference resolution, PAS analysis and BA resolution, or TRR were included. We speculate that this is because the expressive power of our model was primarily allocated to analyzing indirect references in MRR.

Our model, with PAS analysis and BA resolution, achieved the best performance on the results for indirect references in MRR, although coreferences sometimes hindered its performance. Thus, incorporating such textual reference relations is beneficial for analyzing indirect references in the MRR of dialogues.

5 Related Work and Discussion

Existing phrase grounding models can be broadly divided into two architectures:

i) Freeze the weights of the object detector and use the bounding boxes of the detection results as pseudo-labels, often called weakly supervised phrase grounding (Rohrbach et al., 2016; Datta et al., 2019; Wang et al., 2020; Gupta et al., 2020; Goel et al., 2023a).

ii) Incorporate object detection into the training model and dynamically detect boxes according to input text (Kamath et al., 2021; Li et al., 2022; Liu et al., 2024).

Our MRR model follows the approach i) to reduce the learning costs, while we chose the approach ii) as the comparison model for their higher phrase grounding performance.

Previous studies have explored the benefits of coreference and anaphora resolution for pronoun phrase grounding in English (Yu et al., 2022; Lu et al., 2022; Oguz et al., 2023) and multilingual settings (Oguz et al., 2024), focusing on dialogue and procedural texts. These findings are partially consistent with our results on phrase grounding (§ 4.2). However, our study goes further by comprehensively investigating indirect references between mentions and objects (§ 4.3) to apply these findings to real-world systems, such as assistive robots. The ability to identify the object referred to by a predicate is crucial for task planning in collaborative robots, as it directly contributes to their success rate (Migimatsu and Bohg, 2022; Shirai

et al., 2024; Han et al., 2024). Our findings are, therefore, naturally related to this field.

Our proposed framework provides an approach to MRR by explicitly handling zero references, a capability lacking in methods that combine separate phrase grounding and TRR models (Ueda et al., 2024), such as GLIP + KWJA used in our experiment. While recent vision-and-language models based on large language models (OpenAI, 2024) can potentially achieve comparable analysis through prompting techniques (Yang et al., 2023), their application to videos such as J-CRe3 remains challenging due to the time-consuming and high computational costs.

6 Conclusion

This paper has presented a unified framework for textual and multimodal reference resolution to disambiguate references in visually grounded dialogues. Our results showed that incorporating textual reference relations improved performance in multimodal reference resolution, including phrase grounding. In particular, our model with coreference resolution outperformed representative models on phrase grounding for pronouns. Further analysis demonstrated that incorporating textual reference relations strengthens the confidence scores for pronouns, predicates, and anaphoric mentions of objects. In future research, we plan to explore the cross-lingual applicability of our framework to languages other than Japanese. Furthermore, we will improve the multimodal reference resolution model by augmenting the multimodal references and applying this to assistive robots.

Acknowledgments

This work was supported by RIKEN Junior Research Associate Program and JST, PRESTO Grant Number JPMJPR24TC, Japan.

Limitations

Data We acknowledge that our experimental results are limited to two-party Japanese dialogues. Specifically, for indirect references, the nominative and dative case objects referred to by anaphoric mentions are usually either the master or the robot (See § 2.2). Future research should explore other languages and multi-party dialogue settings to investigate the applicability of our framework beyond these constraints.

Models Our MRR model is constrained by the frozen object detector, which results in lower Recall@k scores for noun phrase grounding compared to GLIP. Additionally, all models rely on inference from a single image, which makes it challenging to maintain prediction consistency across different visual contexts.

To address these issues, future research will first focus on integrating the object detection process to enhance the MRR model performance further. This improvement will involve data augmentation techniques, such as leveraging large language models to generate visually grounded dialogues and reference relations by combining existing images or videos. Moreover, exploring video-based architectures and incorporating a first-person view of the system are expected to improve prediction consistency and resolve ambiguities in visually grounded dialogues by leveraging user movements across sequential frames.

Experiments Further analysis — including a detailed analysis of textual indirect references — is needed to fully understand the effect of TRR on MRR. Moreover, investigating the correlation between TRR and MRR performance could provide valuable insights, though this would require multiple training and evaluation iterations of TRR and MRR models.

Ethical Consideration

This study primarily utilized publicly available datasets, such as Flickr30k-Ent-JP and J-CRe3, to prevent ethical concerns. However, J-CRe3 contains videos of identifiable individuals who participated in the data collection process. Therefore, to safeguard their privacy, any use of models trained on J-CRe3 should be used with caution, particularly when intended for commercial use.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086.

Herbert H. Clark. 1975. Bridging. In *Theoretical Issues in Natural Language Processing*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and

Dhruv Batra. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 326–335.

Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. 2019. Align2Ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2601–2610.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, volume 1, pages 4171–4186.

Charles J Fillmore. 1968. The case for case. *Universals in Linguistic Theory*, pages 21–119.

Arushi Goel, Basura Fernando, Frank Keller, and Hakan Bilen. 2023a. Semi-supervised multimodal coreference resolution in image narrations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 11067–11081.

Arushi Goel, Basura Fernando, Frank Keller, and Hakan Bilen. 2023b. Who are you referring to? coreference resolution in image narrations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15247–15258.

Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. 2020. Contrastive learning for weakly supervised phrase grounding. In *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, pages 752–768.

Muzhi Han, Yifeng Zhu, Song-Chun Zhu, Ying Nian Wu, and Yuke Zhu. 2024. InterPreT: Interactive predicate learning from language feedback for generalizable task planning. In *Proceedings of Robotics: Science and Systems (RSS)*.

Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. 2012. Building a diverse document leads corpus annotated with semantic relations. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation (PACLIC)*, pages 535–544.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. DeBERTaV3: improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. arXiv:2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6700–6709.

Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. 2007. Annotating a Japanese text corpus with predicate-argument and coreference relations. In *Proceedings of the Linguistic Annotation Workshop*, pages 132–139.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. MDETR - modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1780–1790.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798.

Hideo Kobayashi and Vincent Ng. 2020. Bridging resolution: A survey of the state of the art. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 3708–3721.

Fang Kong and Guodong Zhou. 2010. A tree kernel-based unified framework for Chinese zero anaphora resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 882–891. Association for Computational Linguistics.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4903–4912.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32—-73.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10965–10975.

Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer*

Vision and Pattern Recognition (CVPR), pages 2117–2125.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2024. Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. In *Proceedings of the 18th European Conference on Computer Vision (ECCV)*, pages 38–55.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*.

Panzhong Lu, Xin Zhang, Meishan Zhang, and Min Zhang. 2022. Extending phrase grounding with pronouns in visual dialogues. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7614–7625.

Toki Migimatsu and Jeannette Bohg. 2022. Grounding predicates through actions. In *Proceedings of the 2022 International Conference on Robotics and Automation (ICRA)*, pages 3498–3504.

Hideki Nakayama, Akihiro Tamura, and Takashi Ninomiya. 2020. A visually-grounded parallel corpus with phrase-to-region linking. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pages 4204–4210.

Cennet Oguz, Pascal Denis, Simon Ostermann, Emmanuel Vincent, Natalia Skachkova, and Josef Van Genabith. 2024. MMAR: Multilingual and multimodal anaphora resolution in instructional videos. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1618–1633.

Cennet Oguz, Pascal Denis, Emmanuel Vincent, Simon Ostermann, and Josef van Genabith. 2023. Find-2-Find: Multitask learning for anaphora resolution and object localization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8099–8110.

Hikaru Omori and Mamoru Komachi. 2019. Multi-task learning for Japanese predicate argument structure analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, volume 1, pages 3404–3414.

OpenAI. 2024. GPT-4o system card. arXiv:2410.21276.

Arum Park, Seunghee Lim, and Munpyo Hong. 2015. Zero object resolution in Korean. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, pages 439–448, Shanghai, China.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123(1):74–93.

Massimo Poesio and Renata Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pages 8748–8763.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Proceedings of the 28th Advances in Neural Information Processing Systems (NIPS)*, volume 28, pages 91–99.

Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of textual phrases in images by reconstruction. In *Proceedings of the 14th European Conference on Computer Vision (ECCV)*, pages 817–834.

Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. 2008. A fully-lexicalized probabilistic model for Japanese zero anaphora resolution. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, pages 769–776.

Kazuhiro Seki, Atsushi Fujii, and Tetsuya Ishikawa. 2002. A probabilistic method for analyzing Japanese anaphora integrating zero pronoun detection and resolution. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*.

Tomohide Shibata and Sadao Kurohashi. 2018. Entity-centric joint modeling of Japanese coreference resolution and predicate argument structure analysis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 579–589.

Keisuke Shirai, Cristian C. Beltran-Hernandez, Masashi Hamaya, Atsushi Hashimoto, Shohei Tanaka, Kento Kawaharazuka, Kazutoshi Tanaka, Yoshitaka Ushiku, and Shinsuke Mori. 2024. Vision-language interpreter for robot task planning. In *Proceedings of the*

2024 IEEE International Conference on Robotics and Automation (ICRA), pages 2051–2058.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

Nobuhiro Ueda, Hideko Habe, Akishige Yuguchi, Seiya Kawano, Yasutomo Kawanishi, Sadao Kurohashi, and Koichiro Yoshino. 2024. J-CRe3: A Japanese conversation dataset for real-world reference resolution. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 9489–9502.

Nobuhiro Ueda, Daisuke Kawahara, and Sadao Kurohashi. 2020. BERT-based cohesion analysis of Japanese texts. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 1323–1333.

Nobuhiro Ueda, Kazumasa Omura, Takashi Kodama, Hirokazu Kiyomaru, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi. 2023. KWJA: A unified Japanese analyzer based on foundation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 3, pages 538–548.

Qinxin Wang, Hao Tan, Sheng Shen, Michael Mahoney, and Zhewei Yao. 2020. MAF: Multimodal alignment framework for weakly-supervised phrase grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2030–2038.

Te-Lin Wu, Satwik Kottur, Andrea Madotto, Mahmoud Azab, Pedro Rodriguez, Babak Damavandi, Nanyun Peng, and Seungwhan Moon. 2023. SIMMC-VR: A task-oriented multimodal dialog dataset with situated and immersive VR streams. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 6273–6291.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. arXiv:2310.11441.

Juntao Yu and Massimo Poesio. 2020. Multitask learning-based neural bridging reference resolution. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 3534–3546.

Xintong Yu, Hongming Zhang, Ruixin Hong, Yangqiu Song, and Changshui Zhang. 2022. VD-PCR: Improving visual dialog with pronoun coreference resolution. *Pattern Recognition*, 125:108540.

Xintong Yu, Hongming Zhang, Yangqiu Song, Yan Song, and Changshui Zhang. 2019. What you see is what you get: Visual pronoun coreference resolution in dialogues. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5123–5132.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. 2022. Detecting twenty-thousand classes using image-level supervision. In *Proceedings of the 17th European Conference on Computer Vision (ECCV)*, pages 350–368.

A Resources

A.1 Data

- **Flickr30kEnt-JP** (Nakayama et al., 2020): `https://github.com/nlab-mpg/Flickr30kEnt-JP`.
- **J-CRe3** (Ueda et al., 2024): `https://github.com/riken-grp/J-CRe3`.
- **Visual Genome** (Krishna et al., 2017): `https://homes.cs.washington.edu/~ranjay/visualgenome`.
- **GQA** (Hudson and Manning, 2019): `https://cs.stanford.edu/people/dorarad/gqa`.
- **Kyoto University Web Document Leads Corpus** (Hangyo et al., 2012): `https://github.com/ku-nlp/KWDLC`.
- **Wikipedia Annotated Corpus**: `https://github.com/ku-nlp/WikipediaAnnotatedCorpus`.
- **Annotated FKC Corpus**: `https://github.com/ku-nlp/AnnotatedFKCCorpus`.

A.2 Model

- **Japanese DeBERTa-v2-large** (He et al., 2021b): `https://huggingface.co/ku-nlp/deberta-v2-large-japanese`.
- **mDeBERTa-v3-base** (He et al., 2021a): `https://huggingface.co/microsoft/mdeberta-v3-base`.
- **Detic** (Zhou et al., 2022): `https://github.com/facebookresearch/Detic`; we used `Detic_LCOCOI21k_CLIP_SwinB_896b32_4x_ft4x_max-size` as a frozen object detector.

- **MDETR** (Kamath et al., 2021): https://github.com/ashkamath/mdetr.
- **GLIP** (Li et al., 2022): https://github.com/microsoft/GLIP.

A.3 Software

- **KWJA** (Ueda et al., 2023): https://github.com/ku-nlp/kwja.
- **Multi-modal Reference Resolution** (Ueda et al., 2024): https://github.com/riken-grp/multimodal-reference.

B Detailed Results in Multimodal Reference Resolution

We present detailed quantitative evaluation results for phrase grounding and MRR using the MRR models in Tables 5, 6, and 7.

C Ablation Study on Utterance Length

Since the evaluation of the models in Tables 3, 4, 5, 6 and 7 is based on single utterances ($\mathbf{T}$), we further conduct an ablation study on utterance length to investigate its effect on model performance. Using longer utterances as input is expected to provide richer textual reference cues — such as coreferences and predicate-argument structures — than a single utterance, potentially leading to improved model performance.

C.1 Results in Phrase Grounding

Figure 7c shows that increasing the input utterance length $\mathbf{T}$, which also increases the number of coreference relations contained in $\mathbf{T}$, consistently improves pronoun performance across all models. We observe that GLIP suffers a decrease in noun performance as the utterance length increases (Figure 7b), resulting in an overall performance decrease (Figure 7a). In contrast, the MRR models based on our proposed framework (the baseline and Baseline w/ Ours) maintain stable noun performance regardless of utterance length (Figure 7b).

C.2 Results in Multimodal Reference Resolution

Figure 8 shows that increasing the input utterance length $\mathbf{T}$ improves performance in all models for direct references, as well as for several types of indirect references, including nominal cases (NOM), accusative cases (ACC), and bridging anaphora. In contrast, longer utterances did not improve performance on dative cases (DAT) and instrumental and locative cases (INS-LOC).

While Figure 8a shows that longer utterances lead to improved performance on direct references in MRR, Figure 7a shows that no such improvement is observed in phrase grounding. A factor contributing to the improvement observed in MRR is the increase in case relations and bridging anaphora, which are considered during evaluation. This suggests that these types of textual reference cues can support the resolution of direct references, especially in longer utterances. These findings are consistent with the trends observed in Figure 6 and Table 5.

D Analysis of Confidence Score Averages

We analyze confidence score averages to compare the baseline model with two MRR models — one using KWJA (Ueda et al., 2023) and the other using our TRR model (§ 3.2.1) — to assess how the presence and type of TRR models affect model confidence.

Table 9 shows that the average confidence scores, computed over Top-k and all predictions, exhibit a consistent rightward shift in distribution across the models — Baseline w/ Ours, Baseline w/ KWJA, and Baseline, in that order. This trend holds for both phrase grounding and MRR. This quantitative result aligns with the trends observed in our qualitative analysis (Figures 3 and 5) and suggests that incorporating textual reference, particularly in Baseline w/ Ours, which incorporates our TRR model, tends to produce higher confidence scores in predictions.

However, higher confidence scores do not always translate into better performance. For example, as shown in Table 3, Baseline w/ Ours improves phrase grounding accuracy over the baseline, whereas Baseline w/ KWJA performs worse. Table 4 shows that Baseline w/ Ours and Baseline w/ KWJA provide little improvement for indirect references of accusative cases (ACC) in MRR. Collectively, these observations imply that the MRR models with TRR may suffer from overconfidence. To address this issue, regularization strategies such as label smoothing (Szegedy et al., 2016) may help calibrate confidence scores.

Models		**Overall** (996)			**Nouns** (671 / 996)			**Pronouns** (120 / 996)		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Baseline	μ	0.342	0.567	0.649	0.344	0.574	0.664	0.277	0.527	0.641
	$\pm\sigma$	0.008	0.011	0.023	0.006	0.014	0.005	0.020	0.091	0.118
Baseline	**Coreference Resolution ⇒ Phrase Grounding**									
w/ KWJA	μ	0.315	0.542	0.640	0.308	0.549	0.657	0.283	0.527	0.633
	$\pm\sigma$	0.024	0.008	0.031	0.029	0.009	0.011	0.036	0.054	0.094
w/ Ours	μ	0.348	0.584	0.681	0.339	0.567	0.663	**0.361**	**0.683**	**0.772**
	$\pm\sigma$	0.020	0.028	0.016	0.029	0.023	0.010	0.070	0.071	0.050
Baseline	**PAS Analysis** and **BA Resolution ⇒ Phrase Grounding**									
w/ KWJA	μ	0.258	0.476	0.566	0.265	0.485	0.575	0.230	0.491	0.594
	$\pm\sigma$	0.115	0.141	0.122	0.135	0.158	0.151	0.070	0.136	0.067
w/ Ours	μ	0.338	0.580	0.672	0.329	0.571	0.667	0.311	0.613	0.738
	$\pm\sigma$	0.014	0.014	0.020	0.016	0.024	0.014	0.026	0.048	0.048
Baseline	**Textual Reference Resolution ⇒ Phrase Grounding**									
w/ KWJA	μ	0.325	0.549	0.627	0.340	0.570	0.656	0.302	0.550	0.597
	$\pm\sigma$	0.026	0.026	0.036	0.009	0.005	0.004	0.047	0.038	0.050
w/ Ours	μ	0.347	**0.600**	**0.689**	0.345	**0.597**	**0.690**	0.258	0.597	0.694
	$\pm\sigma$	0.016	0.016	0.013	0.021	0.038	0.020	0.036	0.037	0.066

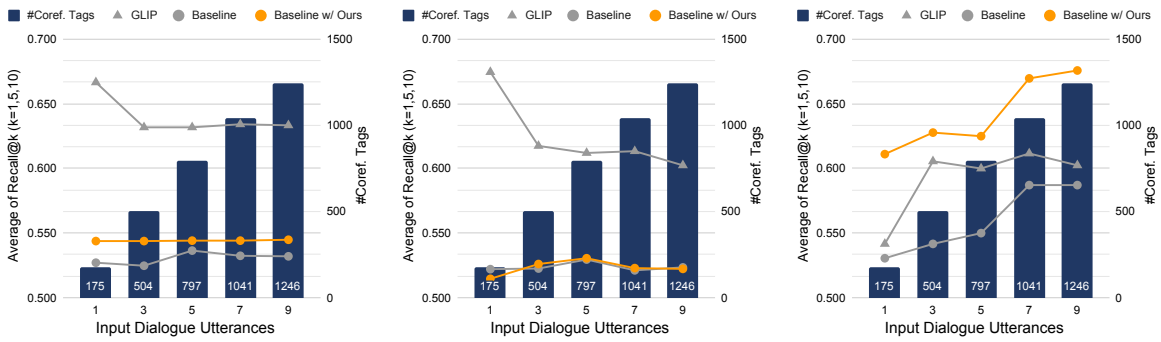
Table 5: Detail results of phrase grounding on the J-CRe3: We report the average (μ) and standard deviation ($\pm\sigma$) of 3 randomly seeded training and evaluation iterations.

Models		**NOM** (2,053)			**ACC** (915)			**DAT** (1,074)			**INS–LOC** (139)			**Bridging** (163)		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Baseline	μ	0.568	0.735	0.763	0.229	0.505	0.606	0.559	0.726	0.749	0.124	0.350	0.465	0.378	0.564	0.662
	$\pm\sigma$	0.014	0.030	0.034	0.007	0.027	0.017	0.008	0.031	0.037	0.039	0.047	0.004	0.028	0.053	0.061
Baseline	**Coreference Resolution ⇒ Multimodal Reference Resolution**															
w/ KWJA	μ	0.542	0.697	0.726	0.261	0.492	0.583	0.576	0.725	0.753	0.177	0.414	0.537	0.390	0.578	0.650
	$\pm\sigma$	0.052	0.124	0.144	0.010	0.052	0.096	0.054	0.101	0.121	0.018	0.008	0.020	0.039	0.044	0.090
w/ Ours	μ	0.572	0.772	0.812	**0.267**	**0.527**	0.618	0.585	0.779	0.818	0.146	0.400	0.520	0.372	0.564	0.689
	$\pm\sigma$	0.011	0.002	0.010	0.017	0.013	0.012	0.014	0.001	0.007	0.014	0.010	0.018	0.019	0.021	0.030
Baseline	**PAS Analysis** and **BA Resolution ⇒ Multimodal Reference Resolution**															
w/ KWJA	μ	0.549	0.732	0.761	0.236	0.497	0.587	0.574	0.727	0.752	0.203	0.376	0.477	0.370	0.560	0.644
	$\pm\sigma$	0.009	0.027	0.039	0.010	0.008	0.029	0.009	0.028	0.034	0.046	0.043	0.023	0.009	0.014	0.040
w/ Ours	μ	0.575	**0.776**	**0.822**	0.253	0.519	0.620	0.585	**0.784**	**0.829**	0.194	0.410	0.508	0.370	0.595	**0.697**
	$\pm\sigma$	0.013	0.003	0.007	0.026	0.021	0.010	0.009	0.009	0.003	0.031	0.007	0.004	0.033	0.037	0.030
Baseline	**Textual Reference Resolution ⇒ Multimodal Reference Resolution**															
w/ KWJA	μ	0.574	0.756	0.785	0.240	0.506	0.601	0.582	0.748	0.779	**0.199**	**0.431**	**0.561**	**0.411**	**0.613**	0.680
	$\pm\sigma$	0.005	0.006	0.006	0.014	0.008	0.018	0.027	0.012	0.015	0.025	0.044	0.073	0.022	0.016	0.018
w/ Ours	μ	**0.585**	0.745	0.773	0.230	0.520	0.607	0.576	0.735	0.772	0.172	0.424	0.532	0.386	0.588	0.662
	$\pm\sigma$	0.006	0.017	0.023	0.013	0.032	0.034	0.009	0.024	0.028	0.004	0.025	0.007	0.023	0.049	0.026

Table 6: Detail results of indirect references in MRR on the J-CRe3: See the caption of Table 5 for μ and $\pm\sigma$.

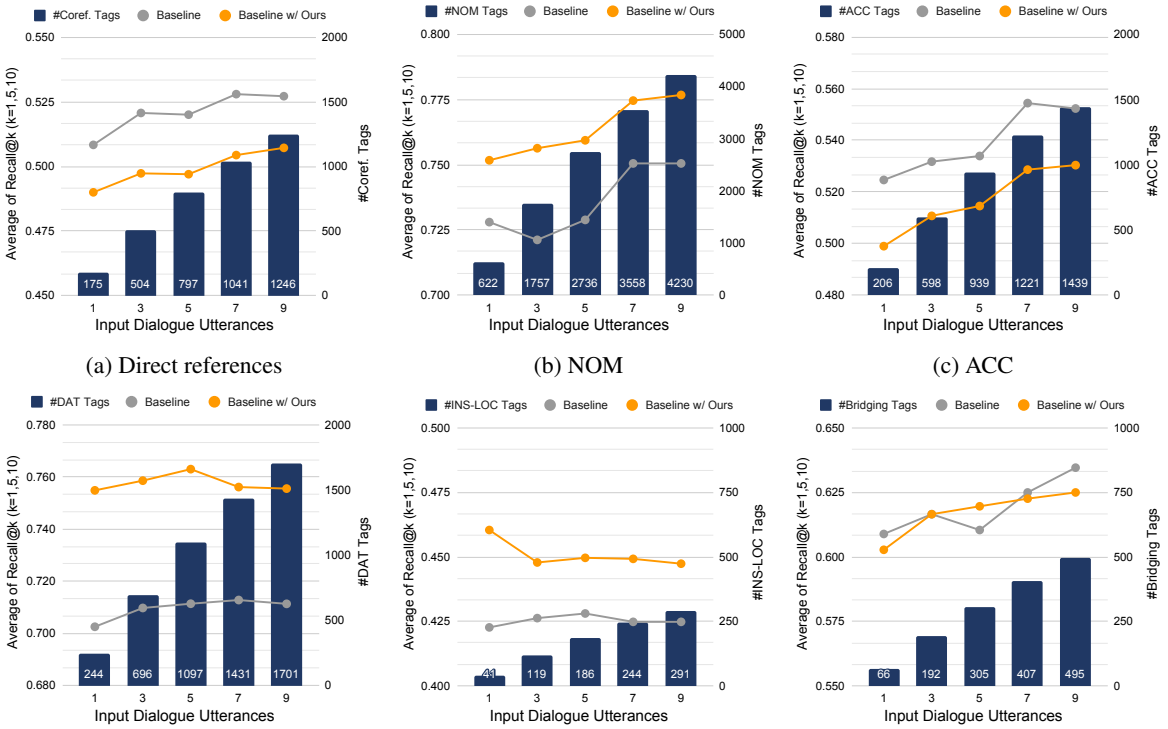
Models		**Overall** (996)			**Nouns** (671 / 996)			**Pronouns** (120 / 996)		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Baseline	μ	0.313	0.564	0.658	0.304	0.552	0.643	0.252	0.566	0.697
	$\pm\sigma$	0.013	0.026	0.011	0.016	0.036	0.022	0.012	0.014	0.026
Baseline	**Coreference Resolution ⇒ Multimodal Reference Resolution**									
w/ KWJA	μ	0.325	0.556	0.642	0.335	0.569	0.653	**0.316**	**0.583**	0.691
	$\pm\sigma$	0.018	0.034	0.072	0.046	0.007	0.031	0.072	0.068	0.101
w/ Ours	μ	0.319	0.549	0.650	0.334	0.550	0.636	0.244	0.558	**0.708**
	$\pm\sigma$	0.017	0.014	0.006	0.011	0.005	0.009	0.020	0.008	0.014
Baseline	**PAS Analysis** and **BA Resolution ⇒ Multimodal Reference Resolution**									
w/ KWJA	μ	0.295	0.534	0.624	0.304	0.545	0.634	0.280	0.513	0.636
	$\pm\sigma$	0.016	0.028	0.046	0.014	0.010	0.025	0.025	0.026	0.029
w/ Ours	μ	**0.331**	0.563	**0.667**	0.315	0.542	0.646	0.277	0.558	0.694
	$\pm\sigma$	0.014	0.009	0.005	0.017	0.012	0.009	0.004	0.050	0.025
Baseline	**Textual Reference Resolution ⇒ Multimodal Reference Resolution**									
w/ KWJA	μ	0.319	0.563	0.650	0.328	0.570	0.656	0.280	0.563	0.672
	$\pm\sigma$	0.027	0.027	0.022	0.034	0.028	0.027	0.009	0.055	0.058
w/ Ours	μ	0.325	0.557	0.644	0.304	0.543	0.631	0.305	0.577	0.683
	$\pm\sigma$	0.014	0.032	0.026	0.016	0.033	0.032	0.025	0.069	0.058

Table 7: Detail results of direct references in MRR on the J-CRe3: See the caption of Table 5 for μ and $\pm\sigma$.



(a) Overall (b) Nouns (c) Pronouns

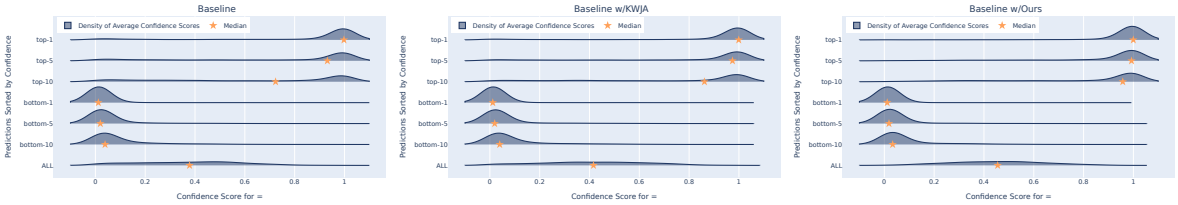
Figure 7: Ablation study results on utterance length in phrase grounding: We compare GLIP, Baseline, and Baseline w/ Ours by varying the input utterance length. Changes in the average of Recall@k ($k = \{1, 5, 10\}$) are shown in a range of 0.2.



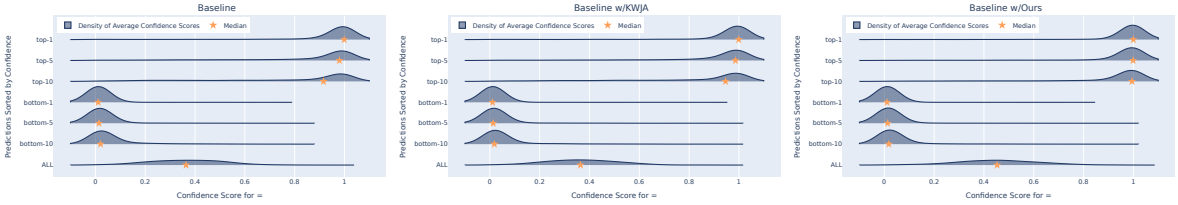
(a) Direct references (b) NOM (c) ACC

(d) DAT (e) INS-LOC (f) Bridging

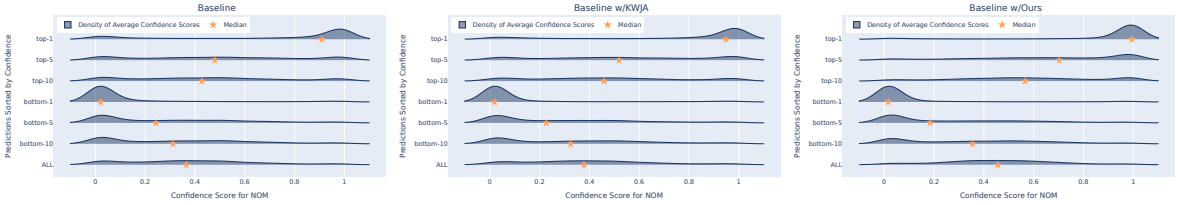
Figure 8: Ablation study results on utterance length in MRR: We compare Baseline and Baseline w/ Ours by varying the input utterance length. Changes in the average of Recall@k ($k = \{1, 5, 10\}$) are shown in a range of 0.1.



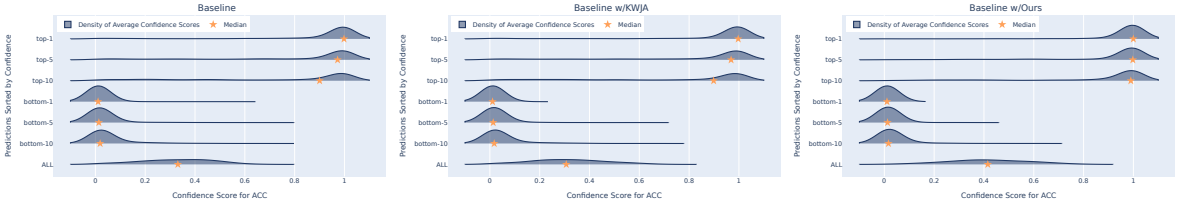
(a) Confidence score distribution in phrase grounding



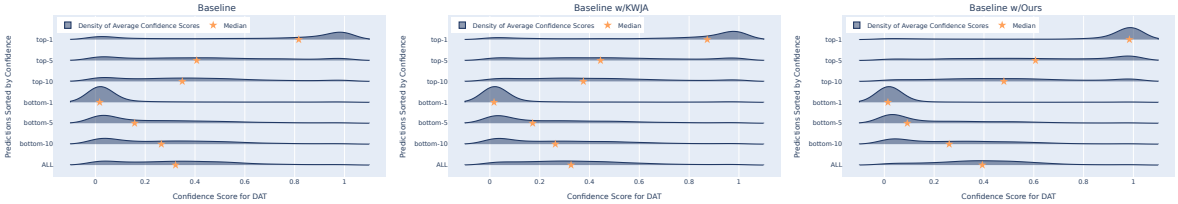
(b) Confidence score distribution in MRR (Direct references)



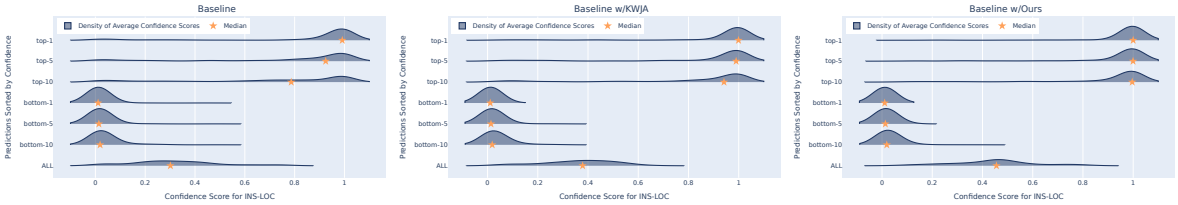
(c) Confidence score distribution in MRR (NOM)



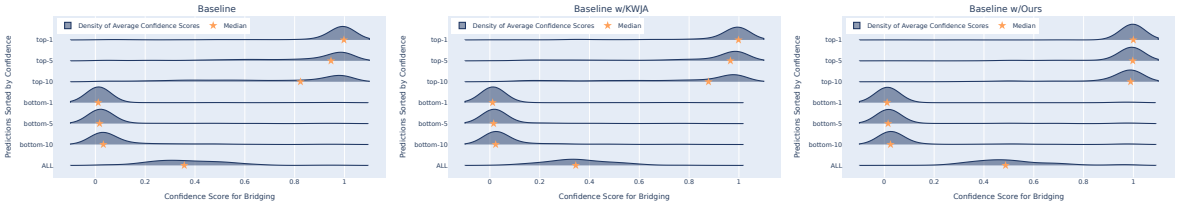
(d) Confidence score distribution in MRR (ACC)



(e) Confidence score distribution in MRR (DAT)



(f) Confidence score distribution in MRR (INS-LOC)



(g) Confidence score distribution in MRR (Bridging)

Figure 9: Violin plots of average confidence scores across the MRR models in phrase grounding and MRR: Each distribution summarizes the scores computed over Top-k, Bottom-k ($k = \{1, 5, 10\}$), and all predictions, based on model predictions from three random seeds.