Divide and Conquer: Rethinking Ambiguous Candidate Identification in Multimodal Dialogues with Pseudo-Labelling

Bhathiya Hemanthage[1,2] **Christian Dondrup**[1] **Hakan Bilen**[2] **Oliver Lemon**[1,3]

[1]Heriot-Watt University [2]University of Edinburgh [3]Alana AI

{hsb2000, c.dondrup, o.lemon}@hw.ac.uk {h.bilen}@ed.ac.uk

Abstract

Ambiguous Candidate Identification (ACI) in multimodal dialogue is the task of identifying all potential objects that a user's utterance could be referring to in a visual scene, in cases where the reference cannot be uniquely determined. End-to-end models are the dominant approach for this task, but have limited real-world applicability due to unrealistic inference-time assumptions such as requiring predefined catalogues of items. Focusing on a more generalized and realistic ACI setup, we demonstrate that a modular approach, which first emphasizes language-only reasoning over dialogue context before performing vision-language fusion, significantly outperforms end-to-end trained baselines. To mitigate the lack of annotations for training the language-only module (student), we propose a pseudo-labelling strategy with a prompted Large Language Model (LLM) as the teacher.

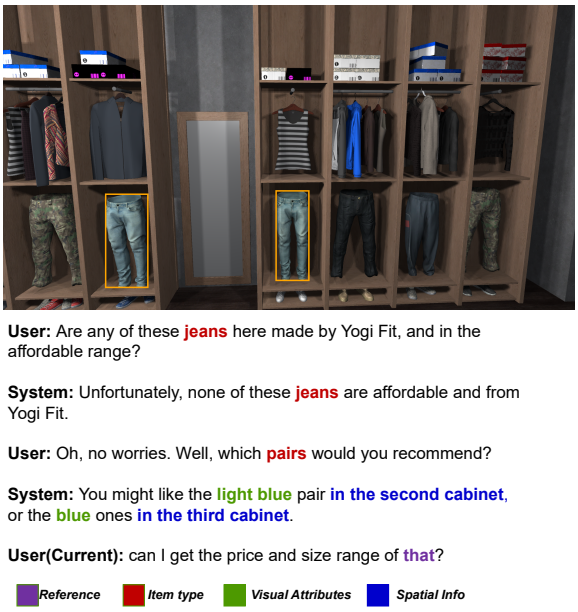

Figure 1: Example for ACI task in MM-Dialogues from SIMMC2. User reference related phrases are colored. Bounding boxes to be predicted are marked in orange.

1 Introduction

In multimodal dialogues (MM-Dialogue), Ambiguous Candidate Identification (ACI) (Kottur et al., 2021) aims to detect all the probable objects in a visual scene that are referred to by a given user utterance, where the reference cannot be uniquely identified. ACI is crucial for resolving ambiguities in multimodal conversational systems, as humans often generate ambiguous referring expressions due to factors like brevity, context dependence, and unintentional ambiguity.

Current state-of-the-art ACI models (Chen et al., 2023; Long et al., 2023) make two key unrealistic assumptions during inference. First, they assume the availability of a predefined catalog of items that may appear in a scene, and that this catalog remains fixed from training to inference. Second, they frame ACI as a candidate selection problem, where ground-truth bounding boxes for all objects are provided during inference. These assumptions limit the generalizability of these models to handle objects not seen during training, which is crucial for real-world multimodal dialogue systems. To bridge this gap, we reformulate the ACI task as a direct coordinate prediction problem, moving away from candidate selection and eliminating the reliance on predefined catalogs. This reformulation aims to improve the applicability of ACI models to more realistic and dynamic multimodal dialogue setting.

We introduce a novel approach to this more challenging reformulation of the ACI task, as illustrated in Figure 2. Our method decomposes the ACI task into two distinct stages. In the first, Dialogue Reference Extraction (DREx), we extract linguistic information on *item types, visual attributes, and spatial information* related to any object reference made in the last user utterance. It is important to note that while the focus is on the most recent user utterance, the extraction process considers the en-

Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 222–227
September 18–20, 2024.

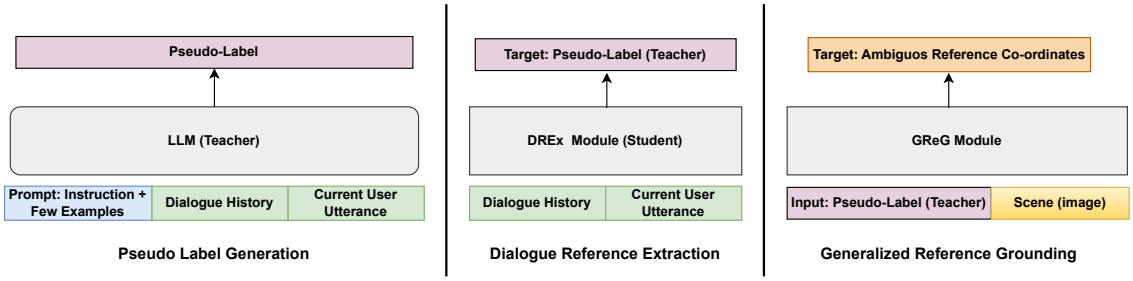

Figure 2: **Training** setting of the proposed modular approach. Pseudo-Labels generated by prompted LLM are used as a target for training the DREx module and as an input for training the Generalized Reference Grounding (GReG) module. During inference, references extracted by the student model are used

tire dialogue history to ensure comprehensive contextual understanding. Subsequently, in the second stage, Generalized Reference Grounding (GReG), we predict the visual coordinates for these extracted references.

Modular vs. End-to-end models Although end-to-end modeling with multimodal fusion has demonstrated significant advancements in various visual-language grounding tasks, including phrase grounding (Plummer et al., 2015), referring expression comprehension(REC) (Yu et al., 2016; Nagaraja et al., 2016), and open vocabulary object detection(Gu et al., 2021), we argue that a modular approach presents several advantages for the more complex ACI task. Firstly, decoupling reference extraction from visual grounding promotes explicit text-only reasoning over the dialogue context, which is crucial for the ACI task. Secondly, the modular approach mitigates the challenges posed by lengthy language contexts in vision-language fusion by presenting the grounding model with only the essential linguistic information.

Despite the advantages, a key challenge of the modular approach is the lack of annotated data for training separate modules. Specifically, the SIMMC2.1 dataset used in our experiments lacks annotations for DREx. To address this, we propose a semi-supervised learning (SSL) setup where pseudo-labels generated by prompting a Large Language Model (LLM) serve as training targets.

2 Related Work

Ambiguous Candidate Identification is first introduced as part of the SIMMC2.1 (Kottur et al., 2021) multi-modal, task-oriented dialogue dataset. In the original evaluation setup proposed for SIMMC2.1, ACI assumes a pre-defined set of items and ground-truth bounding boxes for candidate objects. Due to these (unrealistic) assumptions, models that leverage significant visual semantic information in a symbolic form (Chen et al., 2023; Long et al., 2023) have achieved strong performance on the SIMMC2.1 ACI task despite their limited visual-language grounding capabilities. For example, (Long et al., 2023) represented each catalogue item using a unique token and encoded all ground-truth bounding boxes of candidate objects.

Pseudo Labeling (Lee et al., 2013) is an established method in Semi-Supervised Learning (SSL) (Van Engelen and Hoos, 2020), which aims to generate (pseudo-)labels for unlabeled data to guide the learning process. Typically, pseudo-labels are generated by a teacher model trained on limited labeled data. The emergence of LLMs that can be prompted to generate labels with very few examples has further reduced the labeled data requirement in language modeling tasks (Wang et al., 2021; Ding et al., 2022; Mishra et al., 2023). However, to the best of our knowledge, this is the first study to investigate the use of LLMs as pseudo-label generators for a multimodal dialogue task.

Visual-Language Grounding generally seeks to identify regions within an image corresponding to a linguistic query. Two distinct tasks within this field are REC (Yu et al., 2016; Nagaraja et al., 2016) and phrase grounding (Plummer et al., 2015). REC specifically targets the identification of a single region that optimally corresponds to a given linguistic expression, where phrase grounding typically focuses on grounding zero, one, or many regions matching with simpler noun phrases. Recent visual language pre-trained (VLP) models (such as Kamath et al. (2021); Yan et al. (2023); Peng et al. (2023) have shown their capability in both tasks through task-specific fine-tuning.

3 Methodology

This section first outlines our proposed modules for reformulated ACI in multimodal dialogues. Then we discuss the training and inference procedures.

3.1 Reformulated Task Definition

Given a multi-turn dialogue (D) between a user and an assisting agent (System), accompanied by an image (I) of a scene in which the dialogue is grounded, Ambiguous Candidate Identification (ACI) aims to generate image bounding boxes that tightly encompass each potential item that may have been referred to by the user in their last utterance.

3.2 Proposed Modules

As illustrated in Figure 2, our method consists of: Dialogue Reference Extractor (DREx) and a Generalized Referring Expression Grounder (GReG). Intuitively, we breakdown the ACI task into modules, where each individual module can benefit from the existing work in Dialogue Systems or Visual-Language Grounding.

Dialogue Reference Extraction: The primary objective of this module is to extract any item references made by the user in their last utterance. The module analyzes all previous turns in the dialogue and extracts three types of information: (1) the types of items referenced (e.g., jeans, sofa), (2) the visual attributes of the items, such as color, size, and pattern, and (3) the spatial information pertaining to the items (e.g., behind the rack). Importantly, while it considers the entire dialogue history, the Dialogue Reference Extraction (DREx) module only extracts item references relevant to the current user turn and disregards references to items from previous turns. Output of the module may consist of multiple items as shown in Figure 2.

Generalized Reference Grounding Taking the extracted references for a particular dialogue turn with the grounded scene image I as inputs, the GReG module predicts the bounding box coordinates for each of the matching items.

3.3 Training and Inference Procedure

In the training phase, for a given multimodal dialogue (D, I), we first generate pseudo-labels using a prompted LLM, henceforth referred to as the teacher model. These pseudo-labels produced by the teacher model serve two purposes. Primarily, they are used as targets to train the DREx module, which acts as the student model. Secondly, the pseudo-labels are also used as the inputs to the GReG module during training. In the inference phase, we use the trained student model to extract the references and use as input to the GReG module.

4 Experiments

4.1 Dataset

We conduct experiments using the SIMMC2.1 (Kottur et al., 2021) dataset, a collection of multimodal task oriented dialogues with each utterance grounded in a scene co-observed by conversational agent and the user. Dialogues emulate a shopping experience between agent and user in fashion and furniture domains. While the entire SIMMC2.1 dataset consists of 117,236 utterances across 11,244 dialogues, a subset of 5593 (Train:4239, val: 414, Test:940) utterances from 5259 dialogues provide annotations for the ACI task.

4.2 Evaluation Metrics

We report standard Pascal VOC AP scores along with the Object-F1 score, as outlined in SIMMC2.1 (Kottur et al., 2021). However, the Object-F1 score in SIMMC2.1 ACI is defined for a candidate selection setting, where each object within a scene is symbolically defined (e.g. O32). For our reformulated setting, we compute the Object-F1 using an Intersection over Union (IoU) threshold of 0.5.

Mean-F1: The Object-F1 score is derived from the aggregate of True Positives (TPs), False Positives(FPs), and False Negatives (FNs) across the dataset, inherently favoring samples containing a larger number of targets. To capture this bias, we also report the mean-F1 score, by calculating the F1 score separately for each sample and then averaging these scores. In scenarios where no ground-truth targets are present, the F1 is 1 if, and only if, no bounding boxes are predicted; otherwise 0.

4.3 Experiment Setup

Prompted LLM (Teacher): For all our experiments, we use ChatGPT-4 as the as the teacher model. For each of the ACI samples, we generate pseudo-labels by presenting the current user utterance along with the dialogue history.

DREx (Student) Module: Parallels can be drawn between Dialogue State Tracking (DST) in text-only dialogues and DREx, by considering item type, visual attributes, and position as the slots to

Grounding Model	Pseudo-Label	Val AP	Val Object-F1	Val Mean-F1	Test AP	Test Object-F1	Test Mean-F1
Student- Baseline Comparison							
MDETR (Baseline)	None	18.43	30.40	34.85	17.39	28.59	34.85
MDETR(Modular)	Student	31.76	40.29	44.99	31.56	40.08	46.88
- *Student-Baseline Diff*	N/A	+13.33	+9.89	+11.14	+14.17	+11.49	+12.03
UNINEXT (Baseline)	None	44.85	61.69	56.18	38.97	54.09	52.57
UNINEXT(Modular)	Student	48.63	64.47	55.75	43.17	57.33	54.45
- *Student-Baseline Diff*	N/A	+3.78	+2.78	-0.43	+4.20	+3.24	+1.88
Student- Teacher Comparison							
MDETR	Teacher	36.59	43.41	45.32	39.26	43.80	48.28
- *Student-Teacher Diff*	N/A	-4.83	-3.12	-0.33	-8.70	-3.72	-1.40
UNINEXT	Teacher	59.07	71.23	58.96	56.35	67.28	57.92
- *Student-Teacher Diff*	N/A	-10.44	-6.76	-3.21	-0.60	-9.95	-3.47

Table 1: Top: Comparison of pseudo-labelling based modular approach for ACI against end-to-end trained baselines. Bottom: Comparison of performance with student(DREx) labels replaced by labels from teacher(LLM).

be tracked. Inspired by the success of end-to-end language models in DST in text-only dialogues (Peng et al., 2020; Hosseini-Asl et al., 2020; Ham et al., 2020), we train a GPT2-based simple language model (with only 124M parameters) for the DREx task.

GReG Module Leveraging the similarity of the GReG task with visual grounding, we experiment with two different VLP models: MDETR (Kamath et al., 2021) and UNINEXT (Yan et al., 2023), both of which are capable of grounding (multiple) object regions based on a language queries.

Baselines: We use MDETR and UNINEXT models fine-tuned in an end-to-end manner as two baselines. (More details in Appendix A.)

5 Results and Discussion

Firstly we compare the results of our modular approach against respective end-to-end trained baselines. Results in Table 1 show that the our approach outperforms respective baselines by significant margins, across all metrics in the test set, showcasing the effectiveness of the proposed approach.

Furthermore, the gain in performance is considerably higher for MDETR compared to UNINEXT. This is likely due to the poor performance of the MDETR-Baseline in handling long dialogue context. MDETR relies on box-token contrastive alignment loss for vision-language grounding, which struggles with aligning long dialogue with images, resulting in a diluted loss signal. However, when pseudo-labels with shortened context are used, a significant improvement is observed. This is in contrast to UNINEXT, which does not use any alignment losses.

Secondly, we assess the robustness of the student model in comparison to the teacher model. For this experiment, we generated pseudo-labels for the validation and test splits using teacher model. The performance on the ACI task, when pseudo-labels from the teacher are presented to the GReG module, is shown in the bottom part of 1. The results suggest that there is potential for further improvements with a better student model.

6 Conclusion

In multimodal dialogues, identifying ambiguous candidates is critical due to prevalent non-deterministic references. We introduce a modular strategy that simplifies ACI into two tasks, each task leveraging existing methodologies from text-only dialogues and visual-language grounding. To address the scarcity of annotations for training the reference extraction module, which emphasizes intra-language reasoning, we employ a pseudo-labelling technique where a prompted LLM serves as the teacher. Our experiments with a simple auto-regressive language model as student and two distinct grounding techniques confirm the effectiveness of our approach compared to traditional end-to-end training.

Although our work focuses on ACI in multimodal dialogues, the general approach of modularization with LLM-based pseudo-labelling can be extended to other complex multimodal tasks with long language context, such as interactive task completion (Padmakumar et al., 2022; Gao et al., 2023). Broadly speaking, the emergence of LLMs would provide an opportunity for more explainable modular approaches for tasks requiring substantial intra-language reasoning.

Acknowledgements

This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF) (http://www.ecdf.ed.ac.uk/).This work also used the Cirrus UK National Tier-2 HPC Service at EPCC funded by the University of Edinburgh and EPSRC (EP/P020267/1).

References

Yirong Chen, Ya Li, Tao Wang, Xiaofen Xing, Xiangmin Xu, Quan Liu, Cong Liu, and Guoping Hu. 2023. Exploring prompt-based multi-task learning for multimodal dialog state tracking and immersive multimodal conversation. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 1–8, Prague, Czech Republic. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Shafiq Joty, Boyang Li, and Lidong Bing. 2022. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*.

Qiaozi Gao, Govind Thattai, Xiaofeng Gao, Suhaila Shakiah, Shreyas Pansare, Vasu Sharma, Gaurav Sukhatme, Hangjie Shi, Bofei Yang, Desheng Zheng, et al. 2023. Alexa arena: A user-centric interactive platform for embodied ai. *arXiv preprint arXiv:2303.01586*.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2021. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*.

DongHoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. 2020. End-to-End Neural Pipeline for Goal-Oriented Dialogue Systems using GPT-2. In *ACL*, pages 583–592. Association for Computational Linguistics.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A Simple Language Model for Task-Oriented Dialogue. Cite arxiv:2005.00796Comment: 22 Pages, 2 figures, 16 tables.

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*, pages 1780–1790.

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*, pages 4903–4912, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta.

Yuxing Long, Huibin Zhang, Binyuan Hui, Zhenglu Yang, Caixia Yuan, Xiaojie Wang, Fei Huang, and Yongbin Li. 2023. Improving situated conversational agents with step-by-step multi-modal logic reasoning. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 15–24, Prague, Czech Republic. Association for Computational Linguistics.

Nishant Mishra, Gaurav Sahu, Iacer Calixto, Ameen Abu-Hanna, and Issam H Laradji. 2023. Llm aided semi-supervision for extractive dialog summarization. *arXiv preprint arXiv:2311.11462*.

Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. 2016. Modeling context between objects for referring expression understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 792–807. Springer.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2022. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2017–2025.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2020. SOLOIST: Few-shot Task-Oriented Dialog with A Single Pre-trained Auto-regressive Model. *CoRR*, abs/2005.05298.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649.

Jesper E Van Engelen and Holger H Hoos. 2020. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. 2023. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15325–15336.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer.

A Implementation Details

A.1 DREx Module (Student)

We initialized the DREx model with pretrained weights from OpenAI's GPT2(small). The Adam optimizer was used with default settings from Huggingface's AdamW implementation (learning rate = 1e-3, epsilon = 1e-6, weight decay = 0). Training was conducted over 100 epochs with 4 A100 GPUs with effective batch size of 16.

A.2 GReG Module

MDETR For both the baseline and pseudo-label experiments, we fine-tuned the MDETR ResNet101 pretrained checkpoint over a period of 50 epochs with effective batch size of 8. The learning rate was reduced by a factor of 10 after the first 30 epochs. Initial learning rates were set at $1e-5$ for the backbone and $5e-5$ for the remainder of the network.

UNINEXT For both the baseline and pseudo-label experiments, UNINEXT pretrained checkpoint with ResNet50 backbone was fine-tuned for 20 epochs with effective batch size of 16.The learning rate was reduced by a factor of 10 after the first 12 epochs. Initial learning rates was set at at $1e-4$.

B Pseudo-Label example

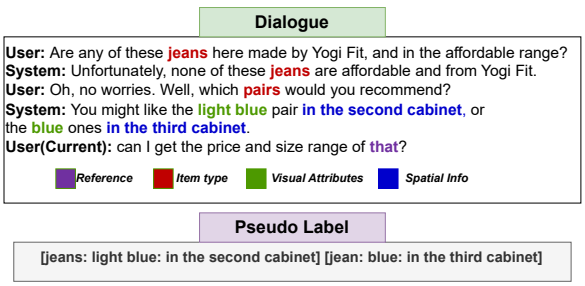

Figure 3: Sample pseudo label with the dialogue.