Locating and Resolving Ambiguities during Human-Robot Teaming Using Multimodal LLMs

William Valentine, Michael Wollowski
Rose-Hulman Institute of Technology, IN, USA

Introduction

In most Human-Robot Interaction (HRI) environments, communication between the human and robot is essential. However, communication is often one of the most highly complex and challenging problems within HRI.

We seek to use a variant of large language models, multimodal LLMs, to help improve the quality of communication by resolving instructional ambiguities.

Introduction

We are defining an instructional ambiguity as a confusing part of an instruction caused by **lack of information**. For the purposes of our work, these ambiguities must also be possible to resolve (make clearer).

This is because there are many forms ambiguity and some of them are not really “resolvable”.

What is a Multimodal Large Language Model?

A Multimodal Large Language Model (MMLLM) tokenizes **both** images and text into transformer embeddings. This process enables these models to answer visual reasoning questions.

"Pick that thing up"

Multimodal Data

0.062, 0.098, -0.041, 0.029,
0.133, -0.067, 0.088, 0.010,
0.079, 0.027, -0.091, 0.056,
-0.018, -0.120, -0.023, -0.019,
0.039, -0.102, 0.044, 0.063

Transformer Embeddings

Rose-Hulman Institute of Technology
DEPARTMENT OF COMPUTER SCIENCE AND SOFTWARE ENGINEERING

Multimodal LLMs, Ambiguity Resolution, and Robotics

Recent work uses LLMs to break down complex instructions into actionable steps for robots (Song et al., 2023; Zhang et al., 2024). **Prior work focuses on direct robot control**, not ambiguity resolution (Lu et al., 2019; Zheng et al., 2024).

Ambiguity resolution is a long-standing problem but it is underexplored with LLMs (Pramanick et al., 2022; Doğan et al., 2022). Traditionally, ambiguity resolution is an Natural Language Processing (NLP) task.

However, **we converted it** into a MMLLM task by adding images to aid in resolution of ambiguities.

Embodied Collaboration

Our area of application is focused on robotics. Cohesive and fluid communication is essential in **Human-Robot Interaction and Collaborative tasks**.

Towards this goal we use a **3D robotics simulation platform** to model a robot in household environments. Simulated environments enable simulation-to-reality transfer and faster prototyping. Recent focus: household environments for developing home assistant robots (Shen et al., 2021; Savva et al., 2019).

We select AI2-THOR as our our 3D embodied platform due to its realistic household simulations. Household environments (kitchens, living rooms) are filled with **potential examples of confusing communication**. We capture images of the environments from the robot's view-point.

Dataset

We collected **20 images (~256x256)** from the 3D simulation tool AI2-THOR collected from **five** environments.

We collect four images from each environment: bathrooms, bedrooms, kitchens, and living rooms.

There are 10 ambiguous instructions per image, totaling **229 instructions** with 1–2 ambiguities per instruction.

Example (1 ambiguity):

Instruction: *"Point at the **small blue thing** on the counter."*
Resolution: "Point at the **toaster** on the counter."

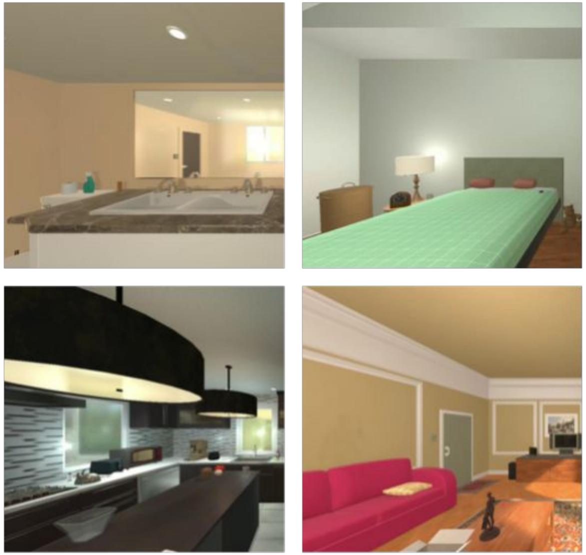
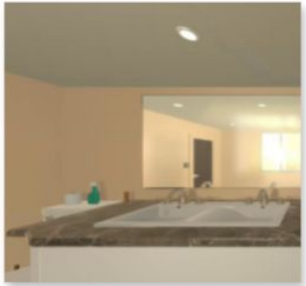


Figure: Examples of images from the dataset.

Rose-Hulman Institute of Technology
DEPARTMENT OF COMPUTER SCIENCE AND SOFTWARE ENGINEERING

Example of our approach



"Pick up the green thing"

"Pick up the green thing"

spray bottle

"Pick up the spray bottle"

Instruction → Does the instruction have ambiguities?

YES → Generate Resolutions for Ambiguities using Images → Replace unclear portions of the instruction with the resolution → Return improved instructions

NO

Results

The accuracy of each MMLLM in correctly locating **and** providing a solution for all ambiguities in an instruction. We select all of these models based on them being "**free-to-use**".

MMLLM Name	Bathroom	Bedroom	Kitchen	Living Room	Overall
GPT-4o	84.06	**75.47**	**81.13**	**85.19**	**81.66**
Claude 3.5 Sonnet	72.46	52.83	**81.13**	66.67	68.56
Gemini 1.5 Flash	**85.51**	67.93	77.36	64.82	74.67
LLaMA 3.2-11B-Vision	63.77	24.53	37.74	48.15	44.98
LLaVA 1.5-7b-hf	76.81	45.28	62.96	49.02	59.91

Discussion

GPT-4o achieves the highest average accuracy (**~81%**) across **all** room types, consistently outperforming other models.

Gemini 1.5 Flash performs competitively, especially on challenging prompts, approaching and exceeding GPT-4o's accuracy in some cases.

LLaVA 1.5-7b-hf shows the lowest performance, indicating difficulty in parsing and grounding instructions in visual context.

Future Work

Our results indicate that MMLLMs are capable of resolving some kinds of instructional ambiguity with high accuracy.

Future work would involve creating a larger dataset with many more examples of ambiguities and images (at least 10K images) to fine tune an MMLLM on.

Future work should also use images from the real-world along with images form simulated environments. This would help to demonstrate the effectiveness of different ambiguity resolution approaches in practice.

References

Deitke et al., 2020
Deitke, M., Han, W., Herrasti, A., Kembhavi, A., Kolve, E., Mottaghi, R., Salvador, J., Schwenk, D., VanderBilt, E., Wallingford, M., Weihs, L., Yatskar, M., & Farhadi, A. (2020). *RoboTHOR: An open simulation-to-real embodied AI platform*. CoRR abs/2004.06799.

Ehsani et al., 2021
Ehsani, K., Han, W., Herrasti, A., VanderBilt, E., Weihs, L., Kolve, E., Kembhavi, A., & Mottaghi, R. (2021). *ManipulaTHOR: A framework for visual object manipulation*. CoRR abs/2104.11213.

Shen et al., 2021
Shen, B., Xia, F., Li, C., Martín-Martín, R., Fan, L., Wang, G., Pérez-D'Arpino, C., Buch, S., Srivastava, S., Tchapmi, L., Tchapmi, M., Vainio, K., Wong, J., Fei-Fei, L., & Savarese, S. (2021). *iGibson 1.0: A simulation environment for interactive tasks in large realistic scenes*. In 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 7520–7527. IEEE Press.

Savva et al., 2019
Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., & Batra, D. (2019). *Habitat: A platform for embodied AI research*. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9338–9346.

Song et al., 2023
Song, C. H., Sadler, B. M., Wu, J., Chao, W.-L., Washington, C., & Su, Y. (2023). *LLM-Planner: Few-shot grounded planning for embodied agents with large language models*. In 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2986–2997.

References

Zhang et al., 2024
Zhang, Y., Yang, S., Bai, C., Wu, F., Li, X., Li, X., & Wang, Z. (2024). *Towards efficient LLM grounding for embodied multi-agent collaboration*. ArXiv abs/2405.14314.

Lu et al., 2019
Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). *ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks*. Red Hook, NY, USA: Curran Associates Inc.

Zheng et al., 2024
Zheng, S., Liu, J., Feng, Y., & Lu, Z. (2024). *Steve-Eye: Equipping LLM-based embodied agents with visual perception in open worlds*. In The Twelfth International Conference on Learning Representation.

Pramanick et al., 2022
Pramanick, P., Sarkar, C., Banerjee, S., & Bhowmick, B. (2022). *Talk-to-resolve: Combining scene understanding and spatial dialogue to resolve granular task ambiguity for a collocated robot*. Robotics and Autonomous Systems, 155, 104183.

Doğan et al., 2022
Doğan, F. I., Torre, I., & Leite, I. (2022). *Asking follow-up clarifications to resolve ambiguities in human-robot conversation*. In 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 461–469.

Mehrabi et al., 2023
Mehrabi, N., Goyal, P., Verma, A., Dhamala, J., Kumar, V., Hu, Q., Chang, K.-W., Zemel, R., Galstyan, A., & Gupta, R. (2023, July). *Resolving ambiguities in text-to-image generative models*. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada, pp. 14367–14388. Association for Computational Linguistics.