Artificial Intelligence and Social Computing, Vol. 163, 2025, 168–176
https://doi.org/10.54941/ahfe1006048

Assessment of the Capabilities of Multimodal Large Language Models in Locating and Resolving Ambiguities During Human-Robot Teaming

William Valentine and Michael Wollowski

Department of Computer Science and Software Engineering, Rose-Hulman Institute of Technology, Terre Haute, IN 47803, USA

ABSTRACT

Human-robot teaming is bound by the quality of communication. Our work studies the quality of ambiguity identification and resolution by Multimodal Large Language Models (MMLLMs) towards creating a clear context for teams. We present images and associated language input to several publicly available MMLLMs. We developed a benchmark of images with associated ambiguous queries to replicate a teaming context with a human collaborator. We evaluated the performance of several MMLLMs on this benchmark to assess their capabilities in identifying and resolving ambiguities. We created a testing framework in which the MMLLM processes commands accompanied by an image and then evaluates the model's performance in detecting and resolving ambiguities. We prepared a benchmark containing 20 images from AI2-THOR's environments: 5 images from bathrooms, five from bedrooms, five from Kitchens, and five from Living Rooms. For each image, we created about 10 prompts that contained one to three ambiguities. There are 229 total ambiguous prompts across all 20 images. To create a shared context between our human and robot collaborators, our system provides a picture that captures the viewpoint of the robot as well as a query provided by the human collaborator. The chosen MMLLM processes this information and outputs both portions of the query that are ambiguous as well as suggestions for clarification. A corrected version of the prompt may then be sent to a planner or a system that provides actionable commands. To evaluate each MMLLM's performance, we compare the ambiguities identified by the model with the expected ambiguities from the datasets. We use vector embeddings and cosine similarity to determine matches. We found an 81% accuracy for the top-performing MMLLM.

Keywords: Human-AI collaboration, Ambiguity resolution, Multimodal large language models

INTRODUCTION

Human robot teaming is bound by the quality of communication. Our work studies the quality of ambiguity identification and resolution by Multimodal Large Language Models (MMLLMs). We present images and associated language input to several publicly available MMLLMs. We developed a benchmark of images with associated ambiguous queries. We evaluated

the performance of several MMLLMs on this benchmark, to assess their capabilities of identifying and resolving ambiguities. We found an impressive 81% accuracy for the top-performing MMLLM.

Overall, the main contributions of this work are as follows:

- We developed a new benchmark using a multi-modal input of images of an environment and commands related to items within that image.
- We present an algorithm to successfully achieve high performance on our benchmark (80%) in locating ambiguities in commands.
- We developed a system to use visual information to resolve ambiguities found in a given instruction with a 81% accuracy of correct identification and resolution of ambiguity.

RELATED WORK

Embodied Multi-Agent Collaboration

Prior work argues that embodied visual AI is currently similar to Computer vision shortly before the advent of visual recognition ecosystems (ImageNet, Pascal, COCO), as there are many efficient and accurate tools for embodied systems but not overwhelming interest in causing large breakthroughs (Deitke et al., 2020; Ehsani et al., 2021; Kolve et al., 2017). These embodied frameworks provide simulations of environments with physical counterparts to allow for simulation to real transfer. Because of this, a virtual robot can be tested in a one-to-one simulated environment and its knowledge can be used in a physical counterpart, and the knowledge it learned in the simulation will be usable in a physical model. A recent emphasis in the literature has focused on simulated household environments to focus on creating robots that can act as home assistants (Shen et al., 2021; Li et al., 2021; Savva et al., 2019; Szot et al., 2024; Puig et al., 2018). We select AI2-THOR to be able to accurately model a household environment from which to conduct our resolution experiments.

Figure 1: These environments AI2-THOR's environment simulation (Kolve et al., 2017). These images are all part of the data provided each MMLLM during the benchmark.

LLMS & Embodied Agents

Prior work has sought to exploit LLMs as planners that are more capable of complex generation strategies to generate instructions for robots (Song et al., 2023; Zhang et al., 2024). The primary reason LLMs have attracted interest as planners is that LLMs have demonstrated flexibility and effectiveness when complex instructions into smaller, actionable steps. As this is the case, we seek to use an LLM and provide it with visual information about the environment in order to generate smaller actionable steps towards resolving ambiguities. Prior work has attempted this, but not towards the goal of ambiguity resolution, instead opting for direct creation of movements for the robot as the LLM serves as a planner (Lu et al., 2019; Zheng et al., 2024).

LLMS & Ambiguities

Resolving ambiguities using LLMs is not a novel idea, but using it towards in the area of robotics has not been explored rigorously. Resolving ambiguities in robotics has been a widely sought after goal for many years (Pramanick et al., 2022; Doğan et al., 2022; Liang, Zhang, and Fisac, Liang et al.; Brown et al., 1999). As this problem has a great deal of forms and different causation, we do not seek to find a complete and satisfactory method of resolution for all ambiguities. We seek to deal with ambiguities whose ambiguity is caused by a lack of information where an actual resolution exists. Prior work has sought to reduce ambiguities in image generation with LLMs (Mehrabi et al., 2023).

MULTIMODAL LLMS

Using multimodal input towards learning based methods has long been viewed as an area where much progress can be made (Wollowski et al., 2020). As the advent of MLLMs has occurred, focus has shifted towards feeding MLLMs visual, textual and in some cases acoustic data (Bai et al., 2024). In almost all of these cases, the introduction of a novel dataset is required to properly provide enough data to create a functional model (Lin et al., 2024; Jin et al., 2023; Han et al., 2023; Li et al., 2023; Wang et al., 2024; Zhang et al., 2023). Towards a similar goal, we introduce a dataset of ambiguities along with images of the environment needed to "solve" the ambiguity.

METHODOLOGY

This study investigates the capability of Multimodal Large Language Models (MMLLMs) to identify ambiguities in natural language commands within a visual context. The primary objective is to assess how effectively an MMLLM can parse and interpret human instructions, highlighting the unclear parts that could hinder robotic comprehension. The methodology involves creating a testing framework where the MMLLM processes commands accompanied by an image, then evaluates the model's performance in detecting and resolving ambiguities.

DATA PREPARATION

We prepared a benchmark containing 20 images of from AI2Thor's environments: 5 images from bathrooms, 5 from bedrooms, 5 from Kitchens, and 5 from Living Rooms. For each image, we created about 10 prompts that contained one to three ambiguities. There are 229 total ambiguous prompts across all 20 images. An example of an ambiguous prompt accompanying the kitchen image is: "Point at *the small blue appliance* on the counter", where the correctly resolved command is "Point at the **toaster** on the counter". Providing a more complex example, the following command contains two ambiguities, "Move towards the *soft objects* resting against *the large brown object*." A successful resolution by a model would resemble, "Move towards the **pillows** resting against the **brown bed's headboard**."

ARCHITECTURE

The first step is for an instruction to be provided to the MMLLM. The modal processes the command alongside a picture captured from the viewpoint of the AI2-THOR agent. It then will attempt to locate unclear instructions and report the portions of ambiguity in the instruction. If it locates portions of ambiguity it will try using information from the agent's viewpoint to generate resolutions for the ambiguities. A corrected version of the prompt is then sent to the AI2-THOR agent from which point a planner or similar system may try to convert the disambiguated instruction into actionable movement commands.

EVALUATION METRICS

To evaluate the MMLLM's performance, we compare the ambiguities identified by the model with the expected ambiguities from the datasets. The accuracy for each command is calculated as:

$$\text{Accuracy} = (\text{Number of Correct Matches}/\text{Total Expected Ambiguities})^{*}100$$

How a correct match is determined is by comparing the MMLLM's generated resolution with the ground-truth's resolution. We use a fuzzy comparison between the two of them in order to see if the core idea is the same rather than the exact letters used. Towards the beginning of the project we used an exact letter match and found it provided meaningless results, as many fully correct resolutions would be marked as incorrect since they missed one starting or ending letter. The accuracy is then the number of correct matches divided by the ground truth's total number of ambiguities.

RESULTS

The individual results from each command are totaled to provide an overall accuracy metric for each dataset.

In the Table 1, we chose MMLLM models based on trying to find the highest accuracy MMLLMs available. We wanted to try to provide an accurate overview of publicly available models, either via free-to-use API in

cases such as ChatGPT, Claude, and Gemini. We also desired to see how some open-source models (LLama and LLaVA) would perform. None of the models received any training or fine-tuning to be accurately compared to the performance of the "out-of-the-box" version of the model. We provided the same prompt to each model to ensure fairness of evaluation.

Table 1: Results for percentage of ambiguity resolution.

			[b]		
MMLLM Name	**Bathroom**	**Bedroom**	**Kitchen**	**Living Room**	**Average**
GPT-4o	84.058	**75.471**	**81.132**	**85.185**	**81.659**
Claude 3.5 Sonnet	72.464	52.830	**81.132**	66.667	68.559
Gemini 1.5 Flash	**85.507**	67.925	77.358	64.815	74.672
LLama 3.2-11B-Vision	63.768	24.528	37.736	48.148	44.978
LLaVA 1.5-7b-hf	76.811	45.283	62.963	49.020	59.912

DISCUSSION

Overall Performance

From Table 1, GPT-4o achieves the best average accuracy (approximately 81%) across all contexts, consistently outperforming the other models. Gemini 1.5 Flash also demonstrates competitive accuracy, particularly on the challenging prompts, matching GPT-4o's performance in that category. In contrast, LLaVA 1.5-7b-hf shows the lowest performance in each context, suggesting that it struggles with parsing the instructions and correlating them accurately with the visual context.

Error Analysis

To better understand the nature of the misclassifications:

Over-Detection of Ambiguities: Some models, particularly Claude 3.5 Sonnet, tend to label standard action words such as *"pick up"* as ambiguous, thereby reducing their final accuracy scores.

Under-Detection of Ambiguities: When multiple objects share similar properties (e.g., multiple "red" items), models such as LLaVA 1.5-7b-hf sometimes fail to identify each potential ambiguity or do not cross-reference the images thoroughly.

Misalignment with Visual Context: Open-source models (e.g., LLaMA 3.2-11B-Vision) are more prone to mismatches between textual references and the available objects in the image, suggesting they struggle with robustly leveraging visual embeddings.

Resolution Quality

While *identifying* ambiguities is one task, providing effective *resolutions* is another. Even in instances where the models correctly detect multiple ambiguities, the proposed resolution can be incomplete. For example, if the command is to *"Grab the smaller green bottle on the counter"* and there are two green bottles of similar size, an ideal resolution might specify *"Grab the green bottle on the counter next to the sink"*. However, some models merely replace *"smaller"* with *"bottle with the narrower neck"*, potentially reducing clarity further. GPT-4o generally offers the most precise alternative descriptions by leveraging fine-grained attributes it identifies in the image.

Summary of Findings

These results underscore that while state-of-the-art closed-source MMLLMs show promise in ambiguity resolution (with performance exceeding 80%), improvement opportunities remain for open-source models. Increased training data, better alignment strategies, or specialized fine-tuning could significantly elevate their capacity to locate and resolve ambiguities within visual contexts.

Performance Comparison Between Models

Across the results from unique MMLLMs, GPT-4o consistently outperforms or performs at the same level as all other models. This indicates that GPT-4o (average accuracy of 81%) would be the best candidate for the ambiguity resolution using images. The open-source models (LLama, LLaVA) tended to perform worse when a particular instruction had multiple ambiguities. The non-open-source models tended to overestimate the number of ambiguities in a particular instruction and occasionally suggest more complex resolutions than were necessary. An example of this is given the instruction "pick up the red vegetable" ambiguities would be detected for "pick" and "vegetable", even though the only real ambiguity is vegetable if there are multiple red vegetables.

Consequence of the Accuracy of the Approach

The accuracy of the approach is over 80% so as a result it seems the developed has viability as a resolution approach. More importantly than that, this indicates that ambiguities that would not have a resolution method with text alone may soon have a way to be resolved. Imagine your ambiguity is centered entirely around the color of paint for a robot to use. There is no way to resolve the ambiguity other than to directly ask the human which may cause annoyance and slow down the robot's ability to actually perform the task.

CONCLUSION

In this work, we studied the use of MMLLMs to resolve ambiguities in queries issued to a robot, when accompanied by a picture of the environment. Publicly available models, without additional training, are doing well when it comes to identifying and resolving ambiguities. We achieved an accuracy

of over 80%. This suggests that MMLLMs are a viable tool in human-machine teaming. Future work should further explore off-the-shelf as well as customized use of MMLMMs in a variety of contexts.

ACKNOWLEDGMENT

We would like to acknowledge the Rose Research Fellows Program and Rose-Hulman's Department of Computer Science and Software Engineering for supporting our work.

REFERENCES

Bai, T., H. Liang, B. Wan, L. Yang, B. Li, Y. Wang, B. Cui, C. He, B. Yuan, and W. Zhang (2024). A survey of multimodal large language model from a data-centric perspective. *ArXiv abs/2405.16640*.

Brown, F., A. Agah, J. Gauch, T. Schreiber, and S. Speer (1999). Ambiguity resolution in natural language understanding, active vision, memory retrieval, and robot reasoning and actuation. In *IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 99CH37028)*, Volume 6, pp. 988–993 vol. 6.

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

Do˘gan, F. I., I. Torre, and I. Leite (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.

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

Han, T., M. Bain, A. Nagrani, G. Varol, W. Xie, and A. Zisserman (2023, June). AutoAD: Movie Description in Context. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA, pp. 18930–18940. IEEE Computer Society.

Jin, P., R. Takanobu, C. Zhang, X. Cao, and L. Yuan (2023). Chat-univi: Unified visual representation empowers large language models with image and video understanding. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13700–13710.

Kolve, E., R. Mottaghi, D. Gordon, Y. Zhu, A. Gupta, and A. Farhadi (2017). AI2-THOR: An interactive 3d environment for visual AI. *CoRR abs/1712.05474*. Li, C., F. Xia, R. Mart'ın-Mart'ın, M. Lingelbach, S. Srivastava, B. Shen, K. Vainio, C. Gokmen, G. Dharan, T. Jain, A. Kurenkov, C. K. Liu, H. Gweon, J. Wu, L. Fei-Fei, and S. Savarese (2021). igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. *CoRR abs/2108.03272*.

Li, J., X. Wang, X. Wu, Z. Zhang, X. Xu, J. Fu, P. Tiwari, X. Wan, and B. Wang (2023). Huatuo-26m, a large-scale chinese medical qa dataset. *ArXiv abs/2305.01526*.

Liang, K., Z. Zhang, and J. F. Fisac. Introspective planning: Aligning robots' uncertainty with inherent task ambiguity. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Lin, B., Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin, and L. Yuan (2024, November). Video-LLaVA: Learning united visual representation by alignment before projection. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, pp. 5971–5984. Association for Computational Linguistics.

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

Mehrabi, N., P. Goyal, A. Verma, J. Dhamala, V. Kumar, Q. Hu, K.-W. Chang, R. Zemel, A. Galstyan, and R. Gupta (2023, July). Resolving ambiguities in text-to-image generative models. In A. Rogers, J. Boyd-Graber, and 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.

Pramanick, P., C. Sarkar, S. Banerjee, and B. Bhowmick (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. Puig, X., K. Ra, M. Boben, J. Li, T. Wang, S. Fidler, and A. Torralba (2018). Virtualhome: Simulating household activities via programs. *CoRR abs/1806.07011*.

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

Shen, B., F. Xia, C. Li, R. Mart'ın-Mart'ın, L. Fan, G. Wang, C. P'erez-D'Arpino, S. Buch, S. Srivastava, L. Tchapmi, M. Tchapmi, K. Vainio, J. Wong, L. Fei-Fei, and S. Savarese (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.

Song, C. H., B. M. Sadler, J. Wu, W.-L. Chao, C. Washington, and Y. Su (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.

Szot, A., A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. Chaplot, O. Maksymets, A. Gokaslan, V. Vondrus, S. Dharur, F. Meier, W. Galuba, A. Chang, Z. Kira, V. Koltun, J. Malik, M. Savva, and D. Batra (2024). Habitat 2.0: Training home assistants to rearrange their habitat. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Wang, Y., Y. He, Y. Li, K. Li, J. Yu, X. Ma, X. Li, G. Chen, X. Chen, Y. Wang, P. Luo, Z. Liu, Y. Wang, L. Wang, and Y. Qiao (2024). Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations*.

Wollowski, M., T. Bath, S. Brusniak, M. Crowell, S. Dong, J. Knierman, W. Panfil, S. Park, M. Schmidt, and A. Suvarna (2020). Chapter 20 - constructing mutual context in human-robot collaborative problem solving with multimodal input. In W. F. Lawless, R. Mittu, and D. A. Sofge (Eds.), *Human-Machine Shared Contexts*, pp. 399–420. Academic Press.

Zhang, H., X. Li, and L. Bing (2023, December). Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In Y. Feng and E. Lefever (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Singapore, pp. 543–553. Association for Computational Linguistics.

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

Zheng, S., Jiazheng Liu, Y. Feng, and Z. Lu (2024). Steve-eye: Equipping LLM-based embodied agents with visual perception in open worlds. In *The Twelfth International Conference on Learning Representations*.