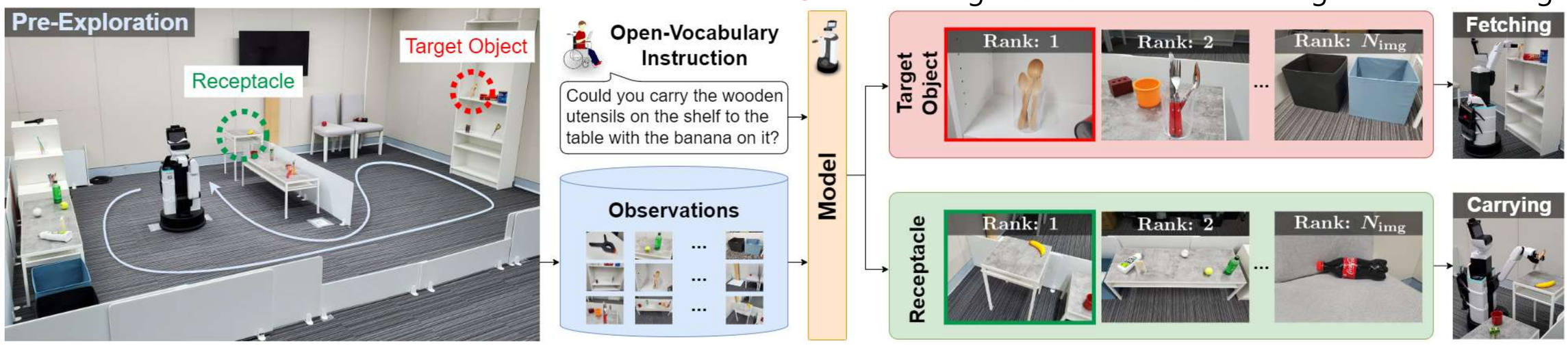
Multimodal Ranking for Target Objects and Receptacles Based on Open-Vocabulary Instructions

Ryosuke Korekata, Kanta Kaneda, Shunya Nagashima, Yuto Imai, Komei Sugiura (Keio University)

Abstract: Domestic service robots (DSRs)

Task	Image retrieval-based open-vocabulary fetch-and-carry
Novelty	1. LLM-based Task Paraphraser 2. Segment Anything Region Encoder
Results	1. Outperformed the baseline methods on the novel dataset 2. Achieved a success rate of 82% in the physical experiments



Related work: Open-vocabulary manipulation

RREx-BoT [Sigurdsson+, IROS23]	Vision-and-language navigation based on the images collected through pre-exploration
MultiRankIt [Kaneda+, RA-L24]	Object fetching tasks based on the human-in-the-loop setting
OVMM [Yenamandra+, CoRL23]	Open-vocabulary mobile manipulation tasks SOTA method achieved SR of only 10%

☹ Few existing methods handle the image retrieval setting

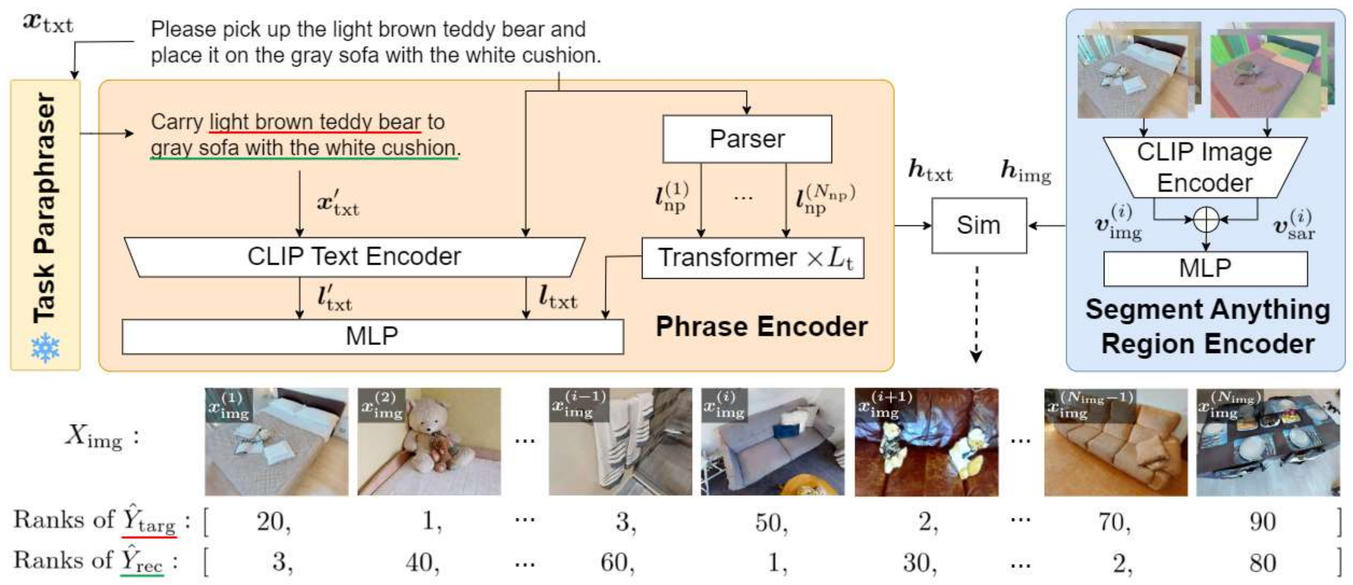
Method

■ Retrieves images of both target objects and receptacles using multimodal foundation models

1. Task Paraphraser:
   Paraphrases instructions including redundancy into standardized format using LLM

2. Phrase Encoder:
   Obtains fine-grained text features from instructions using CLIP [Radford+, ICML21]

3. Segment Anything Region Encoder:
   Enhances visual features regarding shape and contour of objects by utilizing SAM [Kirillov+, ICCV23]

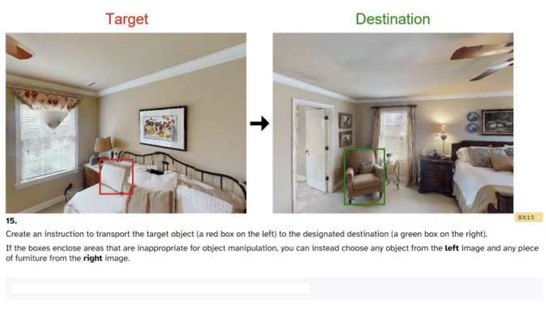


Settings: 1. Newly-built dataset, 2. Physical experiments in the standard environment [Okada+, AR19]

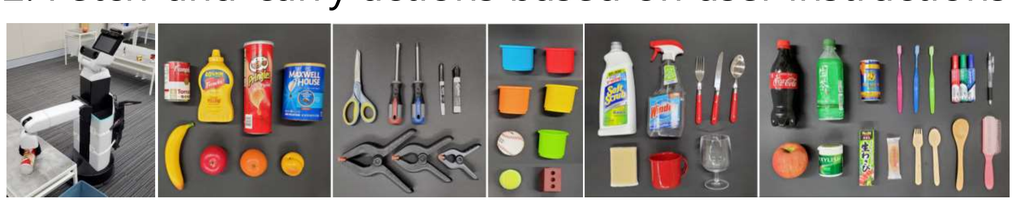
1. LTRRIE-FC dataset based on HM3D [Ramakrishnan+, NeurIPS21]

Instructions were collected by 226 annotators using a crowdsourcing service

#envs	#images	#instrs	Sentence length
774	7,148	6,581	15.69 (average)



2. Fetch-and-carry actions based on user instructions



DSR: HSR [Yamamoto+, ROBOMECH J.19]

Objects: YCB [Calli+, RAM15]

Results: 1. Outperformed the baseline methods, 2. Achieved a success rate of 82% in zero-shot setting

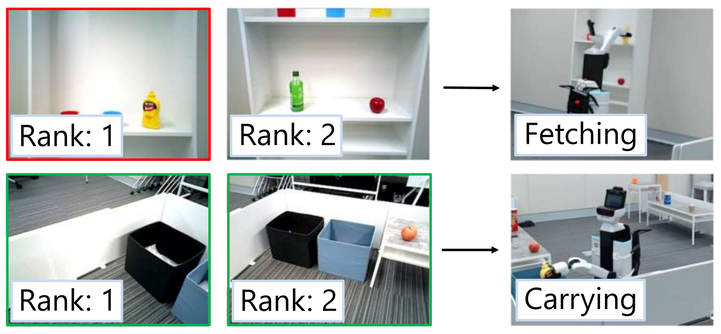
1. Quantitative results: Standard metrics for image retrieval

[%]	MRR↑	Recall@5↑	Recall@10↑
CLIP [Radford+, ICML21]	10.8	13.7	24.9
MultiRankIt [Kaneda+, RA-L24]	20.5 ± 2.3	30.1 ± 3.4	48.2 ± 1.4
	+2.0	+3.1	+4.8
Ours	**22.5** ± 1.4	**33.2** ± 1.8	**53.0** ± 2.5

2. Qualitative result: Successful sample

■ Target Object

■ Receptacle



"Can you take the mustard container on the shelf to the black box?"

[Sigurdsson+, IROS23] G. Sigurdsson, J. Thomason, G. Sukhatme, and R. Piramuthu, "RREx-BoT: Remote Referring Expressions with a Bag of Tricks," in IROS, 2023, pp.5203-5210.
[Kaneda+, RA-L24] K. Kaneda, S. Nagashima, R. Korekata, M. Kambara, and K. Sugiura, "Learning-To-Rank Approach for Identifying Everyday Objects Using a Physical-World Search Engine," IEEE RA-L, vol.9, no.3, pp.2088-2095, 2024.
[Yenamandra+, CoRL23] S. Yenamandra, A. Ramachandran, K. Yadav, A. Wang, M. Khanna, T. Gervet, T. Yang, V. Jain, A. Clegg, J. Turner, Z. Kira, M. Savva, A. Chang, D. Chaplot, D. Batra, R. Mottaghi, et al., "HomeRobot: Open-Vocabulary Mobile Manipulation," in CoRL, 2023.
[Radford+, ICML21] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in ICML, 2021, pp.8748-8763.
[Kirillov+, ICCV23] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. Berg, W. Lo, P. Dollar, and R. Girshick, "Segment Anything," in ICCV, 2023, pp.4015-4026.
[Okada+, AR19] H. Okada, T. Inamura, and K. Wada, "What Competitions were Conducted in the Service Categories of the World Robot Summit?," AR, vol.33, no.17, pp.900-910, 2019.
[Ramakrishnan+, NeurIPS21] S. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. Turner, E. Undersander, W. Galuba, A. Westbury, et al., "Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D Environments for Embodied AI," in NeurIPS, 2021.
[Yamamoto+, ROBOMECH J.19] T. Yamamoto, K. Terada, A. Ochiai, F. Saito, Y. Asahara, and K. Murase, "Development of Human Support Robot as the Research Platform of a Domestic Mobile Manipulator," ROBOMECH J., vol.6, no.1, pp.1-15, 2019.
[Calli+, RAM15] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. Dollar, "Benchmarking in Manipulation Research: Using the Yale-CMU-Berkeley Object and Model Set," IEEE RAM, vol.22, no.3, pp.36-52, 2015.