大規模言語モデルを用いたマルチモーダル検索モデルに基づく生活支援ロボットによる物体操作

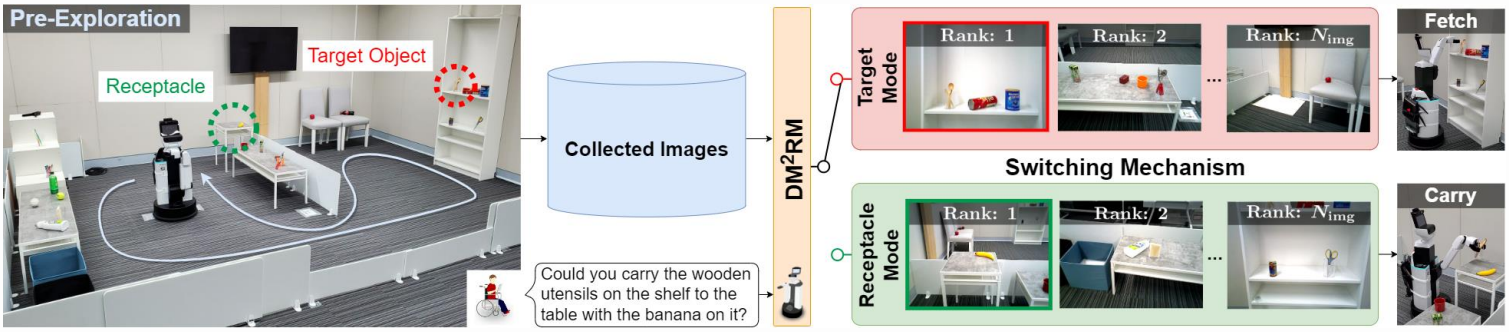
是方諒介　慶應義塾大学大学院 理工学研究科 開放環境科学専攻

概要：Open-Vocabularyな指示文に基づく物体操作

対象タスク	生活支援ロボットに**open-vocabulary**な自然言語指示文を与え，対象物体および配置目に関する画像検索に基づき物体操作
新規性	マルチモーダル基盤モデルに基づく**Switching機構**を導入し，対象物体および配置目標に関する予測を単一モデルで実現
結果	大規模環境で構築したデータセットにおいて既存手法を凌駕 **ゼロショット転移**条件の実機実験でタスク成功率82%を達成

関連研究：画像検索による物体操作を扱う手法は少数

MultiRankIt [Kaneda+, RA-L24]	自動化とオペレータによる介入を組み合わせた Human-in-the-Loop設定でのfetchタスク実行
RREx-BoT [Sigurdsson+, IROS23]	事前収集済み画像からの対象物体検索に基づく Vision-and-Language Navigation
OVMM [Yenamandra+, CoRL23]	open-vocabulary mobile manipulationタスク SOTA手法でもタスク成功率10%程度



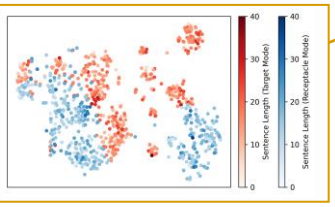
手法：Dual-Mode Multimodal Ranking Model (DM²RM)

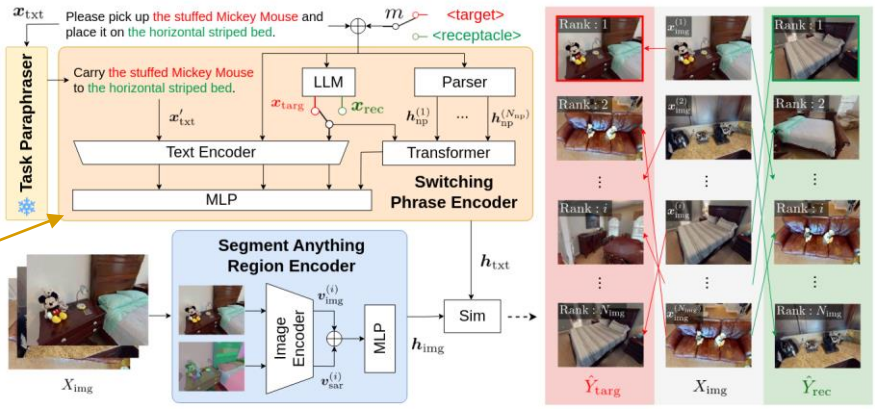
■ マルチモーダル基盤モデルを用いたSwitching機構により，対象物体および配置目標を単一モデルで検索可能

1. **Switching Phrase Encoder**：
   モードトークンおよび**大規模言語モデル**による表現特定
2. **Task Paraphraser**：
   冗長または文法誤りを含む指示文を**標準形**に変換
3. **Segment Anything Region Encoder**：
   SAM [Kirillov+, ICCV23] によるセグメンテーションマスクを重畳

✓ 言語特徴量 h_{txt} をt-SNE [Maaten+, JMLR08] により可視化

✓ モードごとにクラスタが分離





実験設定：1. 大規模屋内環境で収集したデータセット，2. 標準家庭環境 [Okada+, AR19] における実機実験

1. LTRRIE-FC：HM3D [Ramakrishnan+, NeurIPS21] を基に構築

クラウドソーシングにより，
226人のアノテータから物体操作指示文を収集

環境数	画像数	指示文数	平均文長
774	7,148	6,581	15.69

2. ユーザ指示文に基づく画像検索 + 把持・配置動作

ロボット：
HSR [Yamamoto+, ROBOMECH J.19]

物体：
YCBオブジェクト [Calli+, RAM15]

実験結果：1. 新規データセットにおいて既存手法を凌駕，2. ゼロショット転移条件でタスク成功率82%を達成

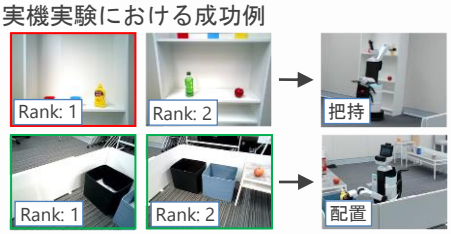
1. 定量的結果：画像検索タスクにおける標準的な評価指標を採用

手法	対象物体	配置目標	MRR [%]	Recall@10 [%]
CLIP [Radford+, ICML21]	✓	✓	10.8	24.9
MultiRankIt [Kaneda+, RA-L24]	✓		20.5 ± 2.3	48.2 ± 1.4
		✓	19.8 ± 1.1	+11.5 49.1 ± 5.9
DM²RM（提案手法）	✓	✓	**32.0 ± 0.5**	**67.9 ± 0.8** +18.8

2. 定性的結果：実機実験における成功例

■ Target モード

■ Receptacle モード



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

[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.
[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.
[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, Y. Bisk, and C. Paxton, "HomeRobot: Open-Vocabulary Mobile Manipulation," in CoRL, 2023.
[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.
[Maaten+, JMLR08] L. Van der Maaten and G. Hinton, "Visualizing Data Using t-SNE," JMLR, vol.9, no.11, 2008.
[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, A. Chang, M. Savva, Y. Zhao, and D. Batra, "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.
[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.

理工学部以外，B2以下でも研究室見学可能です．連絡先：rkorekata@keio.jp