大規模言語モデルを用いたSwitching機構付き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

Fetch-and-Carry Tasks by Domestic Service Robots Based on Multimodal Retrieval Models with Switching Mechanism Using Large Language Models

是方 諒介 Ryosuke Korekata　兼田 寛大 Kanta Kaneda　長嶋 隼矢 Shunya Nagashima　今井 悠人 Yuto Imai　杉浦 孔明 Komei Sugiura

慶應義塾大学
Keio University

In this study, we aim to develop a domestic service robot (DSR) that carries an everyday object to a piece of furniture by retrieving images of target objects and receptacles from collected images of an environment, based on an open-vocabulary instruction. We propose a multimodal model that retrieves both target objects and receptacles individually using a single model based on the switching mechanism via large language models. The experimental results show that our method outperformed baseline methods on the newly-built datasets in terms of the standard metrics. Furthermore, our method achieved task success rates of more than 80% in the physical experiments.

1. はじめに

高齢化が進行する現代社会において，在宅介助者不足の解決策の一つとして生活支援ロボットが注目されている [Yamamoto 19]．ロボットに対して自然言語で家事タスクを指示可能であれば利便性が向上するが，ロボットが人間の指示文を理解する能力はいまだ不十分である．

本研究では，open-vocabulary なユーザ指示文に基づき，事前に収集された環境中の画像群から対象物体および配置目標を検索することで，日常物体を指定された家具へ運搬する Physical Object Retrieval with Fetch-and-Carry (POR-FC) タスクに焦点を当てる．例えば，“Could you carry the wooden utensils on the shelf to the table with the banana on it?” という指示文がロボットに与えられた場合を想定する．このとき，ロボットは図 1 に示された環境において事前に収集された画像群から，対象物体および配置目標としてそれぞれ「棚にある木製のキッチン用品」および「バナナの置かれたテーブル」を上位に検索することが求められる．そのうえで，ユーザが選択した画像に基づいて対象物体を配置目標まで運搬することが期待される．本問題設定においては，画像の提示枚数を限定することでユーザの認知的負荷を軽減することが可能であるため，正しい画像を上位にランク付けすることが重要である．

人間が与える open-vocabulary な指示文は複雑性や曖昧性を含む場合が多くロボットが対象物体および配置目標を特定することは難しい課題である．実際に，参照表現理解を含む Vision-and-Language Navigation (VLN) タスクの標準ベンチマークである REVERIE データセット [Qi 20] において，人間の成功率は 77.84%であった一方，最先端の手法 (e.g., [Sigurdsson 23]) は 43%未満しか達成できていない．

指示文に基づく fetch-and-carry タスクは POR-FC タスクと関連が深く，これまで幅広く研究されてきた [Iocchi 15, Korekata 23]．しかし，open-vocabulary な指示文に基づく物体操作タスクを検索設定において扱う既存研究は限定的であり，それらは対象物体および配置目標の両方に対応していない (e.g., [Kaneda 24])．これらの手法を単純に POR-FC タスクに適用する場合，対象物体および配置目標のそれぞれに特化した別々のモデルを訓練する必要があるため非効率的である．

連絡先: 是方諒介，慶應義塾大学，神奈川県横浜市港北区日吉 3-14-1，rkorekata@keio.jp

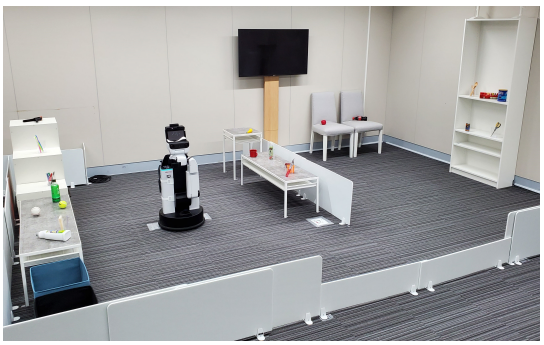

図 1: 実機実験環境

そこで，本研究では単一モデルで対象物体および配置目標を個別に検索可能な手法を提案する．モードトークンおよび large language model (LLM) による表現特定を活用することで，予測対象に応じた埋め込み空間の切り替えを可能にする．

2. 問題設定

本論文では，ユーザから与えられた fetch-and-carry に関する指示文をもとに対象物体および配置目標の画像を検索し，運搬を行うタスクを POR-FC タスクと定義する．入力は指示文および屋内環境で撮影された画像群であり，出力はそれぞれ対象物体および配置目標に関してランク付けされた 2 つの画像リストである．本タスクは，画像検索および動作実行という 2 つのサブタスクから構成される．画像検索においては，対象物体画像および配置目標画像がそれぞれ出力される画像リストにおいて上位にランク付けされることが望ましい．また，動作実行においてはロボットが対象物体を正確に把持し，指定された配置目標まで運ぶことが期待される．なお，対象物体および配置目標はユーザが選択した画像から特定される．

本研究では，屋内環境の画像は事前の探索によって収集済みであることを前提とする．また，ロボットの移動，物体把持，および物体配置に関する軌跡生成はヒューリスティックな手法に基づくものとする．

3. 提案手法

図 2 に，提案手法のモデル構造を示す．提案手法は，Task Paraphraser (TP)，Switching Phrase Encoder (SPE)，および Segment Anything Region Encoder (SARE) という主に 3 つのモジュールから構成される．

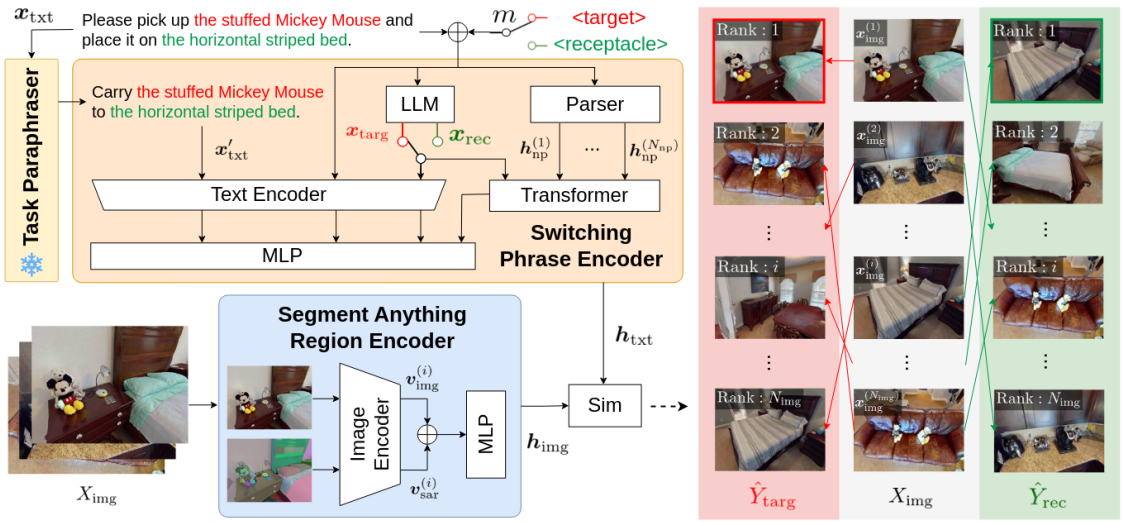

図 2: 提案手法のモデル構造．“MLP”，“Sim”，および “⊕” はそれぞれ多層パーセプトロン，コサイン類似度，および連結を示す．

3.1 入力

モデルへの入力は，$\boldsymbol{x} = \{m, \boldsymbol{x}_{\text{txt}}, X_{\text{img}}\}$ である．ここで，$m \in \{\langle\text{target}\rangle, \langle\text{receptacle}\rangle\}$，$\boldsymbol{x}_{\text{txt}} \in \{0,1\}^{V\times L}$，$X_{\text{img}} = \{\boldsymbol{x}_{\text{img}}^{(i)}\}_{i=1}^{N_{\text{img}}}$，および $\boldsymbol{x}_{\text{img}}^{(i)} \in \mathbb{R}^{3\times W\times H}$ はそれぞれ，モードトークン，トークナイズされた指示文，収集済みの画像群，および幅 W で高さ H の画像を示す．また，V，L，および N_{img} はそれぞれ語彙サイズ，最大トークン長，および画像数を示す．

3.2 Task Paraphraser

TP モジュールは，LLM [gpt] を用いて $\boldsymbol{x}_{\text{txt}}$ を POR-FC タスクに適した標準形 $\boldsymbol{x}'_{\text{txt}}$ に言い換える役割を果たす．open-vocabulary な指示文はしばしば冗長性や文法的な誤りを含むため，関連のある表現が不明瞭な場合がある．本モジュールは，そのような指示文を統一的に扱うことを可能にする．

$\boldsymbol{x}'_{\text{txt}}$ は LLM を用いて対象物体および配置目標に関連する表現を特定することにより得られる．例えば，$\boldsymbol{x}_{\text{txt}}$ が “Could you, if you does not mind, to pick up the cardboard box in the room and move it over towards the couch next to the fireplace in the room?” であるとき，$\boldsymbol{x}'_{\text{txt}}$ は “Carry the cardboard box to the couch next to the fireplace.” と出力される．なお，3.3 節で後述されるように，$\boldsymbol{x}'_{\text{txt}}$ は SPE モジュールの補助入力として使用される．

3.3 Switching Phrase Encoder

SPE モジュールは，m および LLM による表現特定を利用し，予測対象に応じて言語特徴量の埋め込み空間を切り替える役割を果たす．POR-FC タスクでは，1 つの指示文から対象物体および配置目標の両方を予測する必要がある．しかし，それぞれに特化したモデルを別々に訓練することは非効率的である．このため，単一モデルでの訓練および推論を可能にする Switching 機構を採用する．提案手法には，m によって定まる T モードおよび R モードという 2 つのモードが存在する．各モードでは，それぞれ対象物体画像および配置目標画像が上位にランク付けされることが期待される．なお，X_{img} はモードにかかわらず同一である．

本モジュールへの入力は，m，$\boldsymbol{x}_{\text{txt}}$，および $\boldsymbol{x}'_{\text{txt}}$ で構成される．まず，モデルを条件付けるために m を $\boldsymbol{x}_{\text{txt}}$ の先頭に連結する．TP モジュールと同様に LLM を用いて，対象物体に関する表現 $\boldsymbol{x}_{\text{targ}}$ および配置目標に関する表現 $\boldsymbol{x}_{\text{rec}}$ を特定する．予測対象に無関係な表現の影響を抑えるため，モードに応じて適切な方を $\boldsymbol{x}_{\text{p}}$ とする．次に，一般的な parser を用いて $\boldsymbol{x}_{\text{txt}}$ から名詞句 $\{\boldsymbol{x}_{\text{np}}^{(i)}\}_{i=1}^{N_{\text{np}}}$ を抽出し，複数の参照表現を含む指示文から複数の粒度の言語特徴量を得る．ここで，N_{np} は名詞句の最大数を示す．事前学習済みの CLIP text encoder [Radford 21] を用いて $\boldsymbol{x}_{\text{txt}}$，$\boldsymbol{x}'_{\text{txt}}$，$\boldsymbol{x}'_{\text{p}}$，および $\{\boldsymbol{x}_{\text{np}}^{(i)}\}_{i=1}^{N_{\text{np}}}$ からそれぞれ言語特徴量 $\boldsymbol{l}_{\text{txt}} \in \mathbb{R}^{768}$，$\boldsymbol{l}'_{\text{txt}} \in \mathbb{R}^{768}$，$\boldsymbol{l}_{\text{p}} \in \mathbb{R}^{768}$，および $\{\boldsymbol{l}_{\text{np}}^{(i)} \in \mathbb{R}^{768}\}_{i=1}^{N_{\text{np}}}$ を抽出する．最後に，出力 $\boldsymbol{h}_{\text{txt}} \in \mathbb{R}^{768}$ は以下のように得られる．

$$\boldsymbol{h}_{\text{txt}} = \text{MLP}([\boldsymbol{l}_{\text{p}}; \boldsymbol{l}_{\text{txt}}; \boldsymbol{l}'_{\text{txt}}; \text{Transformer}([\boldsymbol{l}_{\text{p}}; \boldsymbol{l}_{\text{np}}^{(1)}; \ldots; \boldsymbol{l}_{\text{np}}^{(N_{\text{np}})}])])$$

ここで，MLP(·) および Transformer(·) はそれぞれ多層パーセプトロンおよび transformer encoder [Vaswani 17] を示す．

3.4 Segment Anything Region Encoder

SARE モジュールは，X_{img} およびセグメンテーションマスク重畳画像の視覚特徴量を基盤モデルを用いて並列に抽出する．画像全体から単純に特徴量を抽出する既存手法は，色やテクスチャが類似した物体を誤認識することがある．そこで，セグメンテーションマスクに関連する補助的な画像を導入することで物体の形状や輪郭に関連する視覚特徴量を扱う．

本モジュールは X_{img} を入力とし，各画像 $\boldsymbol{x}_{\text{img}}^{(i)}$ の視覚特徴量 $\boldsymbol{h}_{\text{img}} \in \mathbb{R}^{768}$ を出力する．まず，SAM [Kirillov 23] により得られたセグメンテーションマスクを $\boldsymbol{x}_{\text{img}}^{(i)}$ に重畳することで $\boldsymbol{x}_{\text{sar}}^{(i)} \in \mathbb{R}^{3\times W\times H}$ を得る．事前学習済みの CLIP image encoder (ViT-L/14) を用いて，$\boldsymbol{x}_{\text{img}}^{(i)}$ および $\boldsymbol{x}_{\text{sar}}^{(i)}$ からそれぞれ視覚特徴量 $\boldsymbol{v}_{\text{img}}^{(i)} \in \mathbb{R}^{768}$ および $\boldsymbol{v}_{\text{sar}}^{(i)} \in \mathbb{R}^{768}$ を抽出する．次に，これらを連結して多層パーセプトロンに入力することで，$\boldsymbol{h}_{\text{img}}$ を得る．最後に，$\boldsymbol{h}_{\text{txt}}$ および $\boldsymbol{h}_{\text{img}}$ の類似度スコアをコサイン類似度に基づき算出する．モデル全体の出力は，類似度スコアに関して降順に X_{img} をランク付けすることで得られる．ただし，$\hat{Y}_{\text{targ}}$ および $\hat{Y}_{\text{rec}}$ はそれぞれ入力時に $m = \langle\text{target}\rangle$ および $m = \langle\text{receptacle}\rangle$ と指定された合計 2 回の推論によって得られる．なお，損失関数は [Kaneda 24] と同様である．

4. シミュレーション実験

4.1 データセット

我々の知る限り，POR-FC タスクの標準的なデータセットは存在しない．VLN タスク (e.g., [Qi 20]) や我々のタスクと関連の深い LTRPO タスク (e.g., [Kaneda 24]) の標準データセットは，対象物体を配置目標へ運搬するタスクを考慮しておらず適していない．さらに，これらの既存のデータセットの多くは

表 1: テスト集合における提案手法およびベースライン手法の定量的比較

[%]	手法	予測対象		HM3D-FC				MP3D-FC			
		対象物体	配置目標	MRR↑	R@5↑	R@10↑	R@20↑	MRR↑	R@5↑	R@10↑	R@20↑
(i)	[Radford 21]	✓	✓	10.8	13.7	24.9	49.5	15.0	14.6	28.5	59.9
(ii-a)	[Kaneda 24]	✓		20.5 ±2.3	30.1 ±3.4	48.2 ±1.4	73.2 ±2.8	26.7 ±2.4	35.9 ±4.0	52.8 ±5.3	71.1 ±2.7
(ii-b)			✓	19.8 ±1.1	27.1 ±3.2	49.1 ±5.9	74.6 ±3.1	16.4 ±1.6	23.3 ±2.0	39.7 ±5.3	60.1 ±3.7
(iii)	提案手法	✓	✓	**32.0** ±0.5	**47.7** ±1.4	**67.9** ±0.8	**87.3** ±1.1	**36.8** ±1.5	**46.5** ±2.8	**63.5** ±2.8	**76.3** ±1.5

Matterport3D (MP3D) [Chang 18] データセットに基づいて構築されており，環境数が数十程度と多様性に欠ける．これに対し，Habitat-Matterport 3D (HM3D) [Ramakrishnan 21] は建物レベルの環境を数百程度含む大規模なデータセットである．しかし，HM3D データセットに対して人間がアノテーションした自然言語指示文を含む標準的なデータセットは存在しない．そこで，本研究では HM3D データセットおよび MP3D データセットの両方から収集した画像に対して新たに指示文を付与することで POR-FC タスクのための新規データセットである Learning-To-Rank in Real Indoor Environments for Fetch-and-Carry (LTRRIE-FC) を構築した．

マップが与えられた HM3D の連続空間から画像を収集するため，ロボットが環境で探索動作を行うことを模倣したシミュレータを用いた．ロボットは，格子状に設けられた各 viewpoint をランダムに選択することで画像を収集した．各 viewpoint において，カメラの姿勢は物体や家具が多く写るように設定された．MP3D データセットからの画像収集方法は [Kaneda 24] と同様である．なお，open-vocabulary な物体検出器である Detic [Zhou 22] を用いて，収集画像から対象物体画像および配置目標画像として適切な画像を抽出した．

指示文は，クラウドソーシングサービスを用いて 226 人のアノテータから収集した．アノテータには，それぞれ対象物体および配置目標を含む 2 枚の画像を提示したうえで，対象物体を配置目標へ運搬するような指示文を付与するように求めた．

LTRRIE-FC データセットは，774 個の屋内環境から収集されたと 7148 枚の実画像および 6581 の英語による指示文から構成される．語彙数は 2491，総単語数は 103263 語，平均文長は 15.69 語である．訓練集合，検証集合，およびテスト集合にはそれぞれ 5814 個，354 個，および 413 個のサンプルが含まれる．各集合にはそれぞれ 690 個，42 個，および 42 個の環境が含まれ，互いに重複はない．なお，テスト集合は画像を収集した環境に応じて HM3D-FC および MP3D-FC という 2 つのサブ集合に分かれている．

4.2 学習設定

提案手法の訓練可能パラメータは約 71M，積和演算数は 309G であった．モデルの訓練には約 1 時間，1 つの指示文と 1 枚の画像との類似度計算には約 14.8×10^{-3} 秒を要した．エポックごとに検証集合においてモデルの MRR および Recall@10 を測定し，その和が最大であるときのテスト集合における評価を最終的な性能とした．

4.3 定量的結果

表 1 に，テスト集合における提案手法とベースライン手法との定量的比較結果を示す．表中の値は，5 回の試行における平均値および標準偏差である．各指標において，最良のスコアが太字で示されている．

本研究では，MultiRankIt [Kaneda 24] および CLIP [Radford 21] ベースライン法として用いた．MultiRankIt は POR-FC タスクと関連が深い LTRPO タスクにおいて良好な結果が得られている手法であるが，単一モデルで対象物体および配置目標の両方を扱うことができないためそれぞれについて別々のモデルを訓練した．評価指標として，Mean Reciprocal Rank (MRR) および Recall@K を用いた．これらは，画像検索タスクにおいて標準的な指標であるためである．

表 1 より，HM3D-FC テスト集合において提案手法 (iv) の MRR は 32.0%である一方，ベースライン手法 (i)，(ii-a)，および (ii-b) の MRR はそれぞれ 10.8%, 20.5%, および 19.8%であった．さらに，MP3D-FC テスト集合において提案手法 (iv)，ベースライン手法 (i)，(ii-a)，および (ii-b) の MRR は，それぞれ 36.8%，15.0%，26.7%，および 16.4%であった．したがって，提案手法はベースライン手法を主要評価指標である MRR において，HM3D-FC テスト集合で 11.5 ポイント，MP3D-FC テスト集合で 10.1 ポイント上回った．これらの性能差はすべて統計有意であった ($p < 0.01$)．

4.4 定性的結果

図 3 に，提案手法とベースライン手法の一つである MultiRankIt との定性的比較結果を示す．各モードについて，Ground Truth (GT) 画像および検索上位 3 件の画像を示す．また，対象物体画像および配置目標画像をそれぞれ赤色および緑色の枠で囲むことにより示す．

図 3 の (a) および (b) は，HM3D-FC テスト集合のサンプルである．提案手法において，$\boldsymbol{x}_{\text{targ}}$ および $\boldsymbol{x}_{\text{rec}}$ はそれぞれ “white lamp on the desk near the bed” および “white desk near the black chair” であった．本サンプルにおいてベースライン手法の MRR が 30%であったのに対し，提案手法の MRR は 100%であった．結果より，ベースライン手法は (a) および (b) の両方で，同一の無関係な画像を 1 位として誤って上位に検索してしまったことが分かる．一方，提案手法はそれぞれのモードにおいて正しい画像を 1 位として検索することに成功した．以上より，SPE モジュールにおける Switching 機構が有効に機能していたことが示唆される．

5. 実機実験

5.1 環境設定および実装

トヨタ自動車製の Human Support Robot [Yamamoto 19] を使用した．図 1 に，実験環境を示す．本環境は，屋内環境における片付けタスクに関する国際的なベンチマークである World Robot Summit 2020 Partner Robot Challenge/Real Space [wrs 20] において標準化された環境を再現したものである．また，合計 50 種類の日常物体を用いた．これらの物体は，物体操作に関する研究において標準的である YCB オブジェクト [Calli 15] に準拠する．本実験は，10 種類の物体配置において行われた．各物体配置において，20-30 種類の物体がランダムに選択された家具のランダムな位置に配置された．なお，いくつかの小物体 (e.g., 歯ブラシ) は NICT ケース [Magassouba 18] の中に入れて配置された．

事前動作として，ロボットは事前に定められた 17 個の viewpoint を探索し，Asus Xtion Pro カメラを用いて環境の画像を収集した．なお，経路計画および移動は，事前に作成したマップを用いた標準的な手法に基づいて行われた．このうえで，ユーザは任意の物体を任意の家具へ運搬する指示文を，参照表現を含めてロボットに与えるよう求められる．各物体配置につ

x_{txt}: "Take the white lamp on the desk near the bed, then move it to the white desk near the black chair."

図 3: テスト集合における提案手法とベースライン手法 [Kaneda 24] の定性的比較

表 2: 実機実験における定量的結果

MRR↑ [%]	Recall@10↑ [%]	SR (N_s / N_a) ↑ [%]
39	96	82 (82 / 100)

き 10 回の指示文が与えられ，合計 100 回の試行が行われた．

ユーザから指示文を受け取ったロボットは，以下のような動作を行った．まず，ロボットは保存された画像から対象物体および配置目標を検索し，それぞれの上位 10 枚の画像を WebUI によりユーザに提示した．ここで，提案手法の未知物体に対する頑健性を検証するため，LTRRIE-FC データセットで訓練したモデルを用いたゼロショット転移を行った．次に，提示された画像からユーザが対象物体画像および配置目標画像を選択した．なお，T モードにおいて対象物体画像が上位 10 位に含まれない場合は失敗とみなし，把持動作は行わなかった．その後，ロボットは対象物体画像が撮影された地点へ移動し，物体把持を行った．この際，把持点は，深度画像から得られた点群および SAM [Kirillov 23] により得られた対象物体のセグメンテーションマスクに基づき決定された．最後に，ロボットは R モードにおいて配置目標画像が上位 10 位以内に含まれ，かつ把持動作が成功していた場合にのみ配置動作を行った．

5.2 定量的結果

表 2 に，実機実験における定量的結果を示す．本実験においては，MRR，Recall@10，および SR $= \frac{N_s}{N_a}$ を評価指標として用いた．ここで，N_s および N_a はそれぞれ成功回数および試行回数を示す．結果より，提案手法の MRR，Recall@10，および SR はそれぞれ 39%，96%，および 82%であった．したがって，LTRRIE-FC データセットにおいて訓練されたモデルを用いたゼロショット転移であったにもかかわらず，実環境において未知物体を扱う場合にも提案手法の性能が頑健であったことが示唆される．また，これらの結果は，提案モデルをロボットに統合し，把持および配置動作を含む包括的なシナリオを実現可能であることを示す．

6. おわりに

本研究では，ロボットが open-vocabulary な指示文に基づいて環境中の画像群から対象物体および配置目標を検索して運搬する POR-FC タスクを扱った．本研究の貢献を以下に示す．

- モードトークンおよび LLM を用いた表現特定により予測対象に応じて埋め込み空間を切り替える SPE モジュールを利用し，単一モデルで対象物体および配置目標の両方を個別に検索可能にする手法を提案した．
- 冗長または文法誤りを含む指示文に対して，指示文を標準形に言い換える TP モジュールを導入した．
- SAM [Kirillov 23] により得られたセグメンテーションマスク重畳画像を利用し，物体の形状や輪郭に関する視覚特徴量を扱う SARE モジュールを導入した．
- HM3D [Ramakrishnan 21] に基づき構築したデータセットにおいて，提案手法が既存手法を上回る結果を得た．
- 実機実験の結果，標準環境においてゼロショット転移という条件のもと 80%以上のタスク成功率を達成した．

謝辞

本研究の一部は，JSPS 科研費 23H03478，JST ムーンショット，NEDO の助成を受けて実施されたものである．

参考文献

[Calli 15] Calli, B., Walsman, A., et al.: Benchmarking in Manipulation Research: Using the Yale-CMU-Berkeley Object and Model Set, *IEEE RAM*, Vol. 22, No. 3, pp. 36–52 (2015)

[Chang 18] Chang, A., Dai, A., Funkhouser, T., Halber, M., Niebner, M., et al.: Matterport3D: Learning from RGB-D Data in Indoor Environments, in *3DV*, pp. 667–676 (2018)

[gpt] https://platform.openai.com/docs/models/gpt-3-5

[Iocchi 15] Iocchi, L., Holz, D., et al.: RoboCup@Home: Analysis and results of evolving competitions for domestic and service robots, *AIJ*, Vol. 229, pp. 258–281 (2015)

[Kaneda 24] Kaneda, K., et al.: 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)

[Kirillov 23] Kirillov, A., Mintun, E., Ravi, N., Mao, H., et al.: Segment Anything, in *ICCV*, pp. 4015–4026 (2023)

[Korekata 23] Korekata, R., et al.: Switching Head–Tail Funnel UNITER for Dual Referring Expression Comprehension with Fetch-and-Carry Tasks, in *IROS*, pp. 3865–3872 (2023)

[Magassouba 18] Magassouba, A., Sugiura, K., and Kawai, H.: A Multimodal Classifier Generative Adversarial Network for Carry and Place Tasks From Ambiguous Language Instructions, *IEEE RA-L*, Vol. 3, No. 4, pp. 3113–3120 (2018)

[Qi 20] Qi, Y., Wu, Q., Anderson, P., Wang, X., et al.: REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments, in *CVPR*, pp. 9982–9991 (2020)

[Radford 21] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., et al.: Learning Transferable Visual Models From Natural Language Supervision, in *ICML*, pp. 8748–8763 (2021)

[Ramakrishnan 21] Ramakrishnan, S., Gokaslan, A., et al.: Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D Environments for Embodied AI, in *NeurIPS* (2021)

[Sigurdsson 23] Sigurdsson, G., Thomason, J., Sukhatme, G., and Piramuthu, R.: RREx-BoT: Remote Referring Expressions with a Bag of Tricks, in *IROS*, pp. 5203–5210 (2023)

[Vaswani 17] Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention Is All You Need, *NeurIPS*, Vol. 30, (2017)

[wrs 20] World Robot Summit 2020 Partner Robot Challenge Real Space Rules & Regulations (2020)

[Yamamoto 19] Yamamoto, T., et al.: Development of Human Support Robot as the research platform of a domestic mobile manipulator, *ROBOMECH J.*, Vol. 6, No. 1, pp. 1–15 (2019)

[Zhou 22] Zhou, X., et al.: Detecting Twenty-thousand Classes using Image-level Supervision, in *ECCV*, pp. 350–368 (2022)