マルチモーダル基盤モデルによる対象物体抽出に基づく日常物体検索および物体操作

◯長嶋隼矢，是方諒介，兼田寛大，杉浦孔明（慶應義塾大学）

本研究では，human-in-the-loop 設定において，ロボットが自然言語指示文に基づいて日常物体の検索および物体操作を行う手法の構築を目的とする．既存のクロスモーダル検索手法では，複雑な参照表現を含む指示文から対象物体を正確に特定することが困難である．本論文では，マルチモーダル基盤モデルにより指示文から対象物体表現を抽出する Target Phrase Extractor およびセグメンテーションマスクを獲得し物体の輪郭および形状に関する情報を扱う Object Segmentation Region Extractor を提案する．結果として，物体操作指示文およびシミュレータで収集された屋内の実画像で構成されるデータセットにおいて，提案手法がベースライン手法を検索タスクにおける標準的な評価指標で上回った．さらに，日常物体を用いた実機実験を標準化された家庭内環境で行い，提案手法によって得られた検索結果に対するユーザの選択に従った物体操作が実行可能であることを示した．

1. はじめに

少子高齢化が進行する現代社会において，在宅介助者不足が社会問題となっている．その一つの解決策として，被介助者を物理的に支援することが可能な生活支援ロボットに注目が集まっている．生活支援ロボットの応用に対する現実的なアプローチとして，自動化とユーザによる介入を組み合わせた human-in-the-loop 設定が考えられる．このような設定では，ユーザに適切な選択肢を提供することが重要である．

本研究では，図 1 に示すように家庭環境内の物体を操作するような自然言語指示から対象物体を検索し，検索結果の中からユーザによって選択された物体の把持を行うタスクに着目する．本タスクを learning–to–rank physical objects for fetching (LTRPO-fetch) タスクと定義する．具体的には，指示文として“ Go into the living room and pick up the yellow cup on the square table. ”が与えられるとする．このとき，ロボットが環境中の物体の中から，対象物体である黄色いカップをより高くランク付けし，ユーザによって選択された物体の把持に成功しユーザの元に届けることが望ましい．

LTRPO-fetch タスクは，複数の参照表現を含む複雑な指示文から指定された対象物体を正確に特定する必要がある点で困難である．具体的には，REVERIE データセットにおける参照表現理解タスクでの人間の成功率は 90.76%であるのに対し，一般的な手法 [1] では 48.98%に過ぎないと報告されている [2]．[3] に代表されるようにクロスモーダル検索の性能は向上している一方で，複雑な参照表現を含む指示文における性能は依然として不十分である．これは既存の手法が，複雑な指示文の中から対象物体の表現を正しく抽出することが困難であるためだと考えられる．

本研究では，物体検索タスクにおいてマルチモーダル基盤モデルを用い，対象物体に関する表現および物体の輪郭や形状を扱う手法を提案する．既存手法と異なる点の 1 つ目として，大規模言語モデルを用いて指示文から対象物体表現を抽出する点がある．抽出された対象物体表現を学習に用いることで，対象となる物体における画像と言語の類似度を高めるように学習され，対象とする物体領域を高くランク付けすることが期待される．既存手法と異なる点の 2 つ目として，SAM [4] を用いて環境中の画像から物体の形状および輪郭を抽出する点がある．抽出された物体ごとの細かい領域を学習に用いることで，参照表現理解における精度を向上させることが期待される．

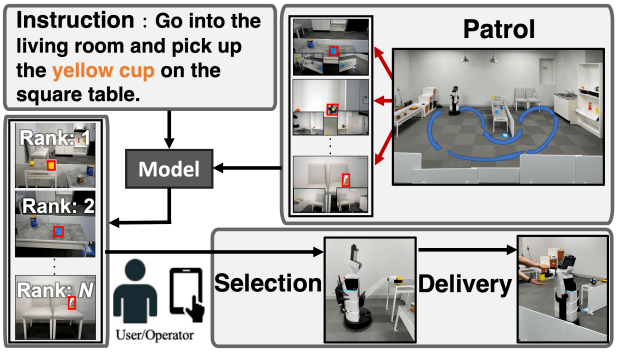

図 1 LTRPO-fetch タスクの例

2. 関連研究

マルチモーダル言語処理分野の研究は広く行われている [5–8]．[5] は，画像と言語の複合分野に関する様々なタスクおよび最新の手法について包括的な概観を提示している．[6] は，画像およびテキストからなるクエリを用いた検索タスクにおいて， ゲーティング関数と残差接続を用いて画像およびテキストを結合する Text Image Residual Gating を提案している．ロボティクス分野においては，物体操作指示文理解に関する手法 [7,8] が提案されている．[7] は，物体検出により抽出した各領域から対象物体および配置目標を特定するタスクを扱う．[8] は，画像内の関連領域に着目し，対象物体と他の物体との関係を学習する手法を提案している．

3. 問題設定

LTRPO-fetch タスクは，物体検索と物体操作という 2 つのサブタスクにより構成される．物体検索においては，指示文の対象物体領域がより上位にランク付けされることが望ましい．また，物体操作においては，ユーザが選択した対象物体領域の物体を把持し，ユーザの元に届けることが望ましい．

本論文で使用する用語は [3] に準拠する．物体検索において，画像中の対象物体の座標は与えられることを前提とする．また，物体操作において，ロボットの移動，把持，および配置に関する軌道生成はヒューリスティックに行われるものとする．

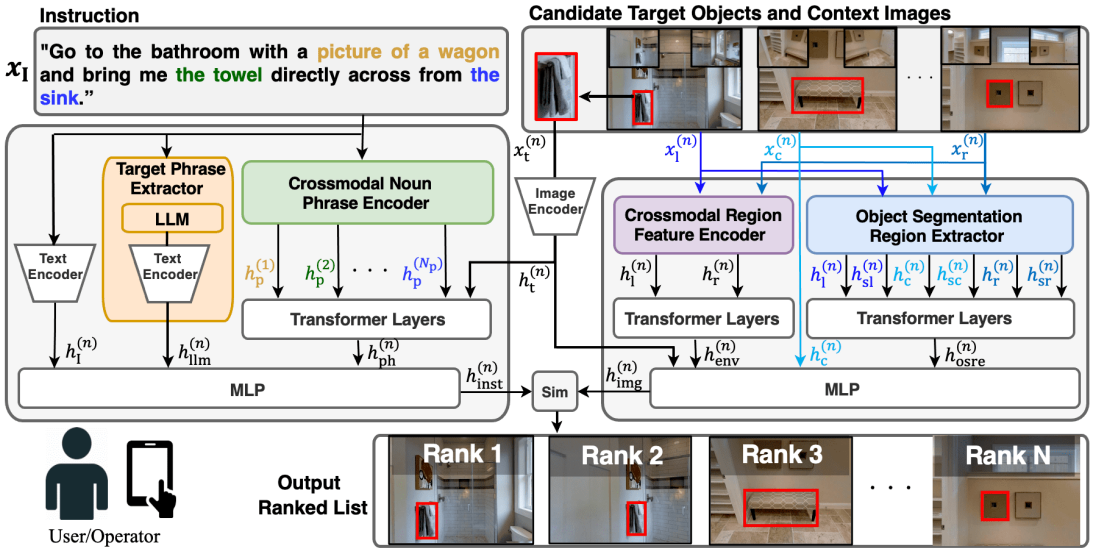

図 2 提案手法のモデル構造．Instruction，Candidate Target Objects，および Context Images はそれぞれ指示文，対象物体領域の候補，および周辺画像を示す．

4. 提案手法

モデル全体は 4 つの主要モジュールから構成され，それぞれ Target Phrase Extractor (TPE)，Object Segmentation Region Extractor (OSRE)，Crossmodal Noun Phrase Encoder (CNPE)，および Crossmodal Regional Feature Encoder (CRFE) である．図 2 にモデルの構造を示す．入力 $\boldsymbol{x}$ を以下のように定義する．

$$\boldsymbol{x} = (\boldsymbol{x}_{\mathrm{I}}, T, C)$$

$$T = \{\boldsymbol{x}_{\mathrm{t}}^{(n)} | n = 1, 2, ..., N_{\mathrm{targ}}\}$$

$$C = \{X_{\mathrm{c}}^{(n)} | n = 1, 2, ..., N_{\mathrm{targ}}\}$$

$$X_{\mathrm{c}}^{(n)} = (\boldsymbol{x}_{\mathrm{c}}^{(n)}, \boldsymbol{x}_{\mathrm{l}}^{(n)}, \boldsymbol{x}_{\mathrm{r}}^{(n)})$$

ここで，$\boldsymbol{x}_{\mathrm{I}}^{(n)} \in \{0,1\}^{V \times L}$ は語彙数 V と最大トークン長 L で one-hot ベクトルとしてトークン化された指示文であり，$\boldsymbol{x}_{\mathrm{t}}^{(n)} \in \mathbb{R}^{3 \times W \times H}$ は幅 W と高さ H の対象物体領域を示す．また，$\boldsymbol{x}_{\mathrm{c}}^{(n)} \in \mathbb{R}^{3 \times 256 \times 256}$，$\boldsymbol{x}_{\mathrm{l}}^{(n)} \in \mathbb{R}^{3 \times 256 \times 256}$，および $\boldsymbol{x}_{\mathrm{r}}^{(n)} \in \mathbb{R}^{3 \times 256 \times 256}$ はすべて周辺画像であり，それぞれ対象物体領域が含まれる画像，$\boldsymbol{x}_{\mathrm{c}}^{(n)}$ の左側の画像，および $\boldsymbol{x}_{\mathrm{c}}^{(n)}$ の右側の画像を示す．なお，N_{targ} は環境中の対象物体候補領域の数である．

事前学習済みの CLIP 画像エンコーダ（ViT/L-14）[9] を用い，$\boldsymbol{x}_{\mathrm{t}}^{(n)}$，$\boldsymbol{x}_{\mathrm{l}}^{(n)}$，および $\boldsymbol{x}_{\mathrm{r}}^{(n)}$ からそれぞれ視覚特徴 $\boldsymbol{h}_{\mathrm{t}}^{(n)} \in \mathbb{R}^{768}$，$\boldsymbol{h}_{\mathrm{c}}^{(n)} \in \mathbb{R}^{768}$，$\boldsymbol{h}_{\mathrm{l}}^{(n)} \in \mathbb{R}^{768}$，および $\boldsymbol{h}_{\mathrm{r}}^{(n)} \in \mathbb{R}^{768}$ を求める．さらに SAM [4] を用い，$\boldsymbol{x}_{\mathrm{c}}^{(n)} \in \mathbb{R}^{768}$，$\boldsymbol{x}_{\mathrm{l}}^{(n)} \in \mathbb{R}^{768}$，および $\boldsymbol{x}_{\mathrm{r}}^{(n)} \in \mathbb{R}^{768}$ から，それぞれセグメンテーションマスクを重畳した画像 $\boldsymbol{x}_{\mathrm{sc}}^{(n)} \in \mathbb{R}^{3 \times 256 \times 256}$，$\boldsymbol{x}_{\mathrm{sl}}^{(n)} \in \mathbb{R}^{3 \times 256 \times 256}$ および $\boldsymbol{x}_{\mathrm{sr}}^{(n)} \in \mathbb{R}^{3 \times 256 \times 256}$ を求める．それらに同様に CLIP 画像エンコーダを用い，$\boldsymbol{h}_{\mathrm{sc}}^{(n)} \in \mathbb{R}^{768}$，$\boldsymbol{h}_{\mathrm{sl}}^{(n)} \in \mathbb{R}^{768}$，および $\boldsymbol{h}_{\mathrm{sr}}^{(n)} \in \mathbb{R}^{768}$ を求める．

TPE では，大規模言語モデルである ChatGPT [10] により，$\boldsymbol{x}_{\mathrm{I}}^{(n)}$ から対象物体表現 $\boldsymbol{x}_{\mathrm{llm}}^{(n)}$ を抽出する．$\boldsymbol{x}_{\mathrm{llm}}^{(n)}$ の抽出には以下のプロンプトを用いる．“[instruction]. Extract the portion of the above instruction that indicates the target object. Please enclose the information with #. Output the information only.” ここで，[instruction] に $\boldsymbol{x}_{\mathrm{I}}^{(n)}$ が入る．例として，$\boldsymbol{x}_{\mathrm{I}}^{(n)}$ が”Go to the laundry room and bring me the plant on the shelf. ” であるとき，$\boldsymbol{x}_{\mathrm{llm}}^{(n)}$ として “the plant on the shelf” が得られる．$\boldsymbol{x}_{\mathrm{llm}}^{(n)}$ から，CLIP テキストエンコーダを用いることで，対象物体表現の特徴量 $\boldsymbol{h}_{\mathrm{llm}}^{(n)}$ を得る．

OSRE では，周辺画像およびセグメンテーションマスク重畳画像の特徴量，$\boldsymbol{h}_{\mathrm{c}}^{(n)}$，$\boldsymbol{h}_{\mathrm{l}}^{(n)}$，$\boldsymbol{h}_{\mathrm{r}}^{(n)}$，および $\boldsymbol{h}_{\mathrm{sc}}^{(n)}$，$\boldsymbol{h}_{\mathrm{sl}}^{(n)}$，$\boldsymbol{h}_{\mathrm{sr}}^{(n)}$ を結合し，L_{osre} 層の transformer 層 f_{osre} に入力することによって物体の輪郭および形状に関する特徴量 $\boldsymbol{h}_{\mathrm{osre}}^{(n)} \in \mathbb{R}^{768}$ を抽出する．なお，各 transformer 層は [11] に従う．

CNPE では，$\boldsymbol{x}_{\mathrm{I}}^{(n)}$ から抽出した名詞句ならびに前置詞句，$\boldsymbol{x}_{\mathrm{llm}}^{(n)}$，および $\boldsymbol{x}_{\mathrm{t}}^{(n)}$ の関係をモデル化する．まず [3] と同様の手法を用いて，$\boldsymbol{x}_{\mathrm{I}}^{(n)}$ から $\boldsymbol{N}_{\mathrm{p}}$ 個の名詞句および前置詞句を含む句 $\boldsymbol{x}_{\mathrm{p}}^{(k)} (k = 1, 2, ..., \boldsymbol{N}_{\mathrm{p}})$ を得る．そして，CLIP のテキストエンコーダを用い，$\boldsymbol{x}_{\mathrm{I}}^{(n)}$ および $\boldsymbol{x}_{\mathrm{p}}^{(k)}$ からそれぞれ言語特徴量 $\boldsymbol{h}_{\mathrm{I}}^{(k)} \in \mathbb{R}^{768}$ および $\boldsymbol{h}_{\mathrm{p}}^{(k)} \in \mathbb{R}^{768}$ を得る．次に，$\boldsymbol{h}_{\mathrm{p}}^{(k)}$ および $\boldsymbol{h}_{\mathrm{t}}^{(n)}$ を結合し，L_{ph} 層の transformer 層に入力することによって指示文の句に対する埋め込み表現 $\boldsymbol{h}_{\mathrm{ph}}^{(n)}$ を得る．ここで，各 transformer 層の構造は f_{osre} と同様である．最後に，$\boldsymbol{h}_{\mathrm{I}}^{(n)}$, $\boldsymbol{h}_{\mathrm{llm}}^{(n)}$, $\boldsymbol{h}_{\mathrm{ph}}^{(n)}$ を順伝播型ネットワーク層に入力することによって指示文に対する埋め込み表現 $\boldsymbol{h}_{\mathrm{inst}}^{(n)}$ を得る．

CRFE では，対象物体領域に近接した画像の左右画像を用いることで，対象物体の周囲の環境を学習する．まず，$\boldsymbol{h}_{\mathrm{l}}^{(n)}$, $\boldsymbol{h}_{\mathrm{r}}^{(n)}$ を L_{env} 層の transformer 層に入力することによって周囲の環境の埋め込み表現 $\boldsymbol{h}_{\mathrm{env}}^{(n)}$ を得る．各 transformer 層の構造は f_{osre} と同様である．次に，$\boldsymbol{h}_{\mathrm{t}}^{(n)}$, $\boldsymbol{h}_{\mathrm{env}}^{(n)}$, $\boldsymbol{h}_{\mathrm{c}}^{(n)}$, $\boldsymbol{h}_{\mathrm{osre}}^{(n)}$ を順伝播型ネットワーク層に入力することで画像に対する埋め込み表現 $\boldsymbol{h}_{\mathrm{img}}^{(n)}$ を得る．

上記で得られた $\boldsymbol{h}_{\mathrm{inst}}^{(n)}$，$\boldsymbol{h}_{\mathrm{img}}^{(n)}$ を用い，$\boldsymbol{x}_{\mathrm{I}}^{(n)}$ と $(\boldsymbol{x}_{\mathrm{t}}^{(n)}, X_{\mathrm{c}}^{(n)})$ の類似度を以下のように定義する．

$$s(\boldsymbol{x}_{\mathrm{I}}, \boldsymbol{x}_{\mathrm{t}}^{(n)}, X_{\mathrm{c}}^{(n)}) = \frac{\boldsymbol{h}_{\mathrm{inst}} \cdot \boldsymbol{h}_{\mathrm{img}}}{\|\boldsymbol{h}_{\mathrm{inst}}\| \|\boldsymbol{h}_{\mathrm{img}}\|}$$

表 1　LTRRIE データセットにおける定量的結果．(“R@K” および太字はそれぞれ Recall@K および最良の値を示す．)

		条件		テスト集合				
		w/TPE	w/OSRE	MRR	R@1[%]	R@5[%]	R@10[%]	R@20[%]
(a)	CLIP extended [9]			0.415 ± 0.009	14.0 ± 1.0	45.3 ± 1.7	63.8 ± 2.5	80.8 ± 2.0
(b)	MultiRankIt [3]			0.501 ± 0.008	18.3 ± 1.0	52.2 ± 1.4	69.8 ± 1.5	83.81 ± 0.6
(c)	提案手法		✓	0.526 ± 0.009	18.3 ± 0.6	55.1 ± 1.1	74.7 ± 1.0	88.9 ± 0.6
(d)		✓		0.557 ± 0.007	20.1 ± 0.4	**60.1 ± 0.7**	**77.9 ± 0.9**	**90.3 ± 0.6**
(e)		✓	✓	**0.563 ± 0.013**	**20.7 ± 0.8**	58.7 ± 1.1	77.7 ± 1.1	90.0 ± 0.5

最後に $s(\boldsymbol{x}_{\mathrm{I}}, \boldsymbol{x}_{\mathrm{t}}^{(n)}, X_{\mathrm{c}}^{(n)})$ に従い，ランク付けされた対象物体領域のリストを出力する．

本モデルの出力は，$s(\boldsymbol{x}_{\mathrm{I}}, \boldsymbol{x}_{\mathrm{t}}^{(n)}, X_{\mathrm{c}}^{(n)})$ に基づく T のランク付きリストである．各バッチにおける損失関数 $L_{\mathcal{B}}$ を以下の式に定義する．

$$L_{\mathcal{B}} = -\frac{1}{|\mathcal{B}|} \sum_{\boldsymbol{x}_{\mathrm{t}}^{(n)} \in \mathcal{B}} \log \frac{\exp\left(s(\boldsymbol{x}_{\mathrm{I}}, \boldsymbol{x}_{\mathrm{t}}^{(n)}, X_{\mathrm{c}}^{(n)})\right)}{\sum_{i=1}^{|\mathcal{B}|} \exp\left(s(\boldsymbol{x}_{\mathrm{I}}, \boldsymbol{x}_{\mathrm{t}}^{(i)}, X_{\mathrm{c}}^{(i)})\right)}$$

ここで，$|\mathcal{B}|$ はバッチサイズを表す．

5. 実験設定

5.1 LTRRIE データセット

本実験では，LTRRIE データセット [3] を用いた．本研究で扱う LTRPO-fetch タスクでは，物体検索において対象物体領域，周辺画像，および指示文から，その指示文の対象物体がより上位にランク付けがされるのが望ましい．そのため，[3] と同様に LTRRIE データセットを用いた．

ハイパーパラメータは，$L_{\mathrm{osre}} = 3$, $L_{\mathrm{ph}} = 5$, $L_{\mathrm{env}} = 5$, $\#A = 4$, $\#H = 768$ とした．ここで，$\#H$ と $\#A$ はそれぞれ各 transformer 層における隠れ層の数と attention head の数である．最適化には Adam を使用し，学習率は 5.0×10^{-5}，バッチサイズは 128 とした．提案手法における訓練可能パラメータ数は約 14.4 億である．また，積和演算数は 3.04×10^{11} である．訓練にはメモリ 48GB 搭載の Quadro RTX 8000 およびメモリ 384GB 搭載の Intel Xeon Gold 6234 を使用した．訓練には約 83 分，1 つの指示文と 1 つの対象物体領域の候補との類似度を計算する推論に約 1.63×10^{-2} 秒を要した．各エポックごとに検証集合で Mean reciprocal rank (MRR) [12] を測定した．このうち最大の MRR を得たモデルを用いて，テスト集合における評価を行った．

5.2 実機実験

実機実験での環境は，家庭内の実環境における片付けタスクのベンチマークである国際的なロボット競技会 World Robot Summit 2020 Partner Robot Challenge/Real Space [13] での標準環境に基づいている．本実験においては，トヨタ自動車製の生活支援ロボット Human Support Robot (HSR) [14] を使用した．本実験で用いた物体は YCB オブジェクト [15] である．

本実験では，5 種類の物体配置を作成した．各物体は無作為に選択された家具の上の無作為な位置に配置された．そのうえで，ロボットに対して “Go to the dining room and give me the spam on the shelf.” などの指示文を与えた．指示文は英語で，合計 50 文与えた．

次に，ロボットの動作について述べる．はじめに，事前に定められた 8 個の waypoint をロボットが初期位置から順に巡回し，環境中の画像収集を行った．各 waypoint は，ロボットが各家具に正対して正面および左右から物体を撮影できるように定めた．ロボットの移動については，事前に与えられた地図を用いて標準的な手法で経路計画を行った．画像の取得には，HSR の頭部に搭載された Asus Xtion Pro カメラを用いた．したがって，1 度の巡回で各家具の正面，左，および右からの視点の画像がそれぞれ 8 枚，すなわち合計 24 枚得られた．なお，各画像に対して事前学習済みの物体検出器を用いて最尤の 5 個の矩形領域を取得した．これらの画像および矩形領域を提案手法への入力とし，推論には LTRRIE データセットで訓練されたモデルを利用することでゼロショット転移を行った．モデルの出力であるランク付きリストの中から，上位 10 個の対象物体領域の候補をユーザに提示し，対象物体領域があれば選択させた．対象物体領域が選択されると，撮影した waypoint，深度画像，および矩形領域を基に把持点を決定し，把持動作を行った．把持動作が終了すると，事前に定められたユーザの位置まで移動し，物体を手渡した．

6. 実験結果

6.1 LTRRIE データセット

表 1 に，ベースライン手法，提案手法，および ablation study の定量的結果を示す．なお，ベースライン手法として，CLIP [9] を本タスク向けに拡張した手法および MultiRankIt [3] を用いた．表中の値は，5 回の試行における平均値と標準偏差を示す．MultiRankIt は，LTRPO-fetch タスクにおける物体検索において良好な結果が報告されているため，ベースラインとした．

モデルの評価指標として MRR [12] および Recall@K を用いた．なお，MRR を主要な指標とした．Recall@K は $\text{Recall@}K = \frac{1}{N_{\mathrm{inst}}} \sum_{i=1}^{N_{\mathrm{inst}}} \frac{|A_i \cap B_i|}{|A_i|}$ のように定義される．ここで，A_i および B_i はそれぞれ検索対象のサンプル集合および検索上位 K 個のサンプル集合を表す．MRR および Recall@K は，ランキング学習の設定における標準的な指標であるため，使用した [16].

表 1 から，ベースライン手法 (b) および提案手法 (e) の MRR はそれぞれ 0.501 および 0.563 であった．したがって，提案手法はテスト集合での MRR において，ベースライン手法を 0.062 ポイント上回った．同様に，提案手法は Recall@K ($K \leq 20$) においてもベースライン手法を上回った．上記すべての結果において，統計的に有意な性能差があった（$p < 0.01$）．

図 3 に定性的結果を示す．図 3 では，$\boldsymbol{x}_{\mathrm{I}}$ は “Go down the hallway past the long mirrors and open the curtain.” である．また，$\boldsymbol{x}_{\mathrm{llm}}$ は “curtain” であった．ここで，Reciprocal Rank (RR) は検索結果中最上位の適合サンプル文書のランクの逆数で表される評価指標である．このサンプルにおける RR は，提案手法で 1.0，ベースライン手法で 0.05 であった．TPE により対象物体表現 “curtain” を正しく抽出できたことにより，指

表 2　実機実験における定量的結果

MRR@10	Recall@10 [%]	把持成功率 (成功回数 / 試行回数) [%]	タスク成功率 (成功回数 / 試行回数) [%]
0.373	82.0	58.5 (24 / 41)	48.0 (24 / 50)

示文と対象物体領域との類似度がより高くなるように計算されたと考えられる．

Ablation study として以下の 2 つの条件について検証を行った．(c) TPE モジュールを取り除くことで，性能にどの程度の差が生じるかを調査した．(d) OSRE モジュールを取り除くことで，性能にどの程度の差が生じるかを調査した．表 1 に ablation study の定量的な結果を示す．テスト集合の条件 (c) においては，条件 (e) と比較して，MRR が 0.037 ポイント減少した．この結果より，対象物体表現を抽出する TPE の導入が有効であったことが示唆される．

6.2　実機実験

実機実験の評価指標として，MRR@10，Recall@10，把持成功率，およびタスク成功率を用いた．MRR@10 は以下のように定義される．

$$\mathrm{MRR@10} = \frac{1}{N_{\mathrm{inst}}} \sum_{i=1, r_1^{(i)} \leq 10}^{N_{\mathrm{inst}}} \frac{1}{r_1^{(i)}}$$

ここで，N_{inst} および $r_1^{(i)}$ はそれぞれ指示文の数およびランク付きリストにおける対象物体領域のランクを示す．表 2 に MRR@10，Recall@10，把持成功率，およびタスク成功率を示す．表 2 より，MRR@10 は 0.373 であり，Recall@10 は 82.0 であった．つまり，82.0% の割合で対象物体領域が上位 10 個に入った．また，把持成功率が 58.5%であった．物体検索および物体操作を通してのタスク成功率は 48.0%となった．

図 4 に実機実験の定性的結果を示す．赤色の矩形領域は対象物体領域の候補を表す．また，出力のランク付きリストの上位 3 サンプルを示した．なお，各対象物体領域の候補が写る画像に対して，周辺画像である左右の画像を示した．図 4(a) の例では，対象物体であるスパム缶が適切に 1 位としてランク付けされた．ユーザが上位 10 個の中から 1 位の物体領域を選択した．そのうえで，ロボットがスパム缶の把持およびユーザの元へ届けることに成功した．同様に，図 4(b) の例においても対象物体である黄色のカップが適切に 1 位としてランク付けされた．そして，ユーザによって選択された黄色いカップをユーザに届けることに成功した．

7.　おわりに

本研究では生活支援ロボットが撮影した物体画像から指示文の対象物体画像を検索し物体操作を行う LTRPO-fetch タスクを扱った．本研究の貢献を以下に示す．

- 大規模言語モデル [10] を用いて複雑な指示文から対象物体表現を抽出する TPE を提案した．
- SAM [4] によるセグメンテーションマスク重畳画像の特徴量を導入することにより，物体の形状および輪郭を扱う OSRE を導入した．
- LTRRIE データセット [3] において，提案手法がベースライン手法 [3] を MRR [12] で上回った．
- 実機実験を行い，LTRPO-fetch タスクにおける提案手法の評価を行った．

図 3　LTRRIE データセットにおける定性的結果

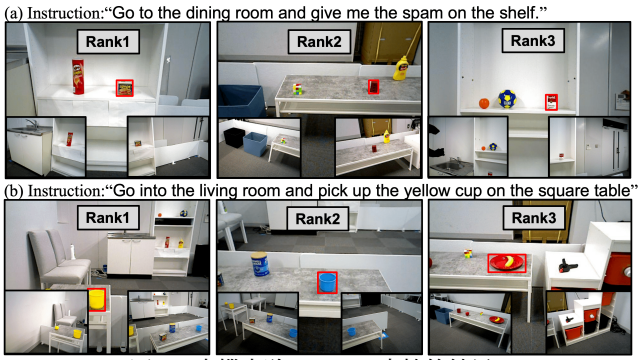

図 4　実機実験における定性的結果

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

[1] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, et al., “MAttNet: Modular Attention Network for Referring Expression Comprehension,” CVPR, pp.1307–1315, 2018.

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

[3] 兼田寛大，神原元就，杉浦孔明，“Learning to Rank Physical Objects: ランキング学習による物理世界検索エンジン，” 2023 年度 人工知能学会全国大会，2023．3G1-OS-24a-01.

[4] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. Berg, W.-Y. Lo, et al., “Segment Anything,” arXiv:2304.02643, 2023.

[5] S. Uppal, S. Bhagat, et al., “Multimodal Research in Vision and Language: A Review of Current and Emerging Trends,” Information Fusion, vol.77, pp.149–171, 2022.

[6] N. Vo, L. Jiang, C. Sun, K. Murphy, J. Li, L. Fei, and J. Hays, “Composing Text and Image for Image Retrieval-An Empirical Odyssey,” CVPR, pp.6439–6448, 2019.

[7] R. Korekata, M. Kambara, Y. Yoshida, S. Ishikawa, Y. Kawasaki, M. Takahashi, et al., “Switching Head–Tail Funnel UNITER for Dual Referring Expression Comprehension with Fetch-and-Carry Tasks,” IROS, 2023. to appear.

[8] S. Ishikawa and K. Sugiura, “Target-Dependent UNITER: A Transformer-Based Multimodal Language Comprehension Model for Domestic Service Robots,” IEEE RA-L, vol.6, no.4, pp.8401–8408, 2021.

[9] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, et al., “Learning Transferable Visual Models from Natural Language Supervision,” ICML, pp.8748–8763, 2021.

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

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, “Attention is All You Need,” NeurIPS, vol.30, pp.5998–6008, 2017.

[12] E. Voorhees and H. Dang, “Overview of the TREC 2001 Question Answering Track.,” Trec, pp.42–51, 2001.

[13] H. Okada, et al., “What competitions were conducted in the service categories of the World Robot Summit?,” Advanced Robotics, vol.33, no.17, pp.900–910, 2019.

[14] 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 Journal, vol.6, no.1, pp.1–15, 2019.

[15] B. Calli, A. Walsman, 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.

[16] T. Liu, “Learning to Rank for Information Retrieval,” Foundations and Trends in Information Retrieval, vol.3, no.3, pp.225–331, 2009.