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

Finding Everyday Objects Using Physical-World Search Engines: a Learning–To–Rank Approach

兼田 寛大 *1 Kanta Kaneda　神原 元就 *1 Motonari Kambara　杉浦 孔明 *1 Komei Sugiura

*1慶應義塾大学 Keio University

In this study, we focus on the learning-to-rank physical objects task, which involves retrieving target objects from open-vocabulary user instructions in a human-in-the-loop setting. We propose MultiRankIt, which introduces the Crossmodal Noun Phrase Encoder to model the relationship between referring expressions and target bounding box, and the Crossmodal Region Feature Encoder to model the relationship between the target object and its surrounding contextual environment. Our model outperforms the baseline method in terms of mean reciprocal rank and recall@K.

1. はじめに

少子高齢化に伴い在宅介護者の不足は喫緊の社会問題となっており，生活支援ロボット（DSR）はその解決策として期待されている．DSR の応用に対する現実的なアプローチとして，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 設定が考えられる．そこで本研究では，DSR がユーザのためのサイバネティックアバター [Ishiguro 21] として機能することを考える．また，このような環境ではオペレータに適切な選択肢を提供することが重要である．

本論文では，human-in-the-loop 設定において，家庭環境内の物体を操作するようなオープンボキャブラリのユーザ指示から対象物体を検索するタスクに着目し，本タスクを learning–to–rank physical objects (LTRPO) タスクと定義する．図 1 に本タスクの典型的な場面を示す．“Go to the bathroom with a picture of a wagon. Bring me the towel under the picture directly across from the sink” という指示文が与えられた場合，本タスクにおいてモデルは対象物体のランク付けリストを出力することが期待される．Human-in-the-loop 設定では人間の注意力は限られているため，ユーザやオペレーターが容易に選択できる数の対象物体を表示することが重要である．

しかし，複雑な指示文から対象物体を正確に特定することは困難であるため，LTRPO タスクは容易ではない．具体的には，単純な参照表現を用いた参照表現理解タスクにおける人間の成功率は 90.76%であるのに対し，SoTA モデル [Yu 18] では 48.98%にとどまっている．さらに，LTRPO タスクは複数の参照表現を含む指示から対象物体を特定するため，参照表現理解タスクよりも困難である．(例: “Go to the dining room with the round table. Pick up the bottle on it ’’)

近年の CLIP [Radford 21] などのマルチモーダル表現学習の発展により，クロスモーダル検索の性能は向上している．しかし，上述のように，複雑な参照表現を含むテキストに対するクロスモーダル検索の性能はまだ十分ではない．これは，既存の手法が単純な表現しか用いないタスクを扱うことが多いためである．(例: “A photo of {label}” [Radford 21])

このような背景から，本論文では human-in-the-loop 設定においてオープンボキャブラリの指示文から対象物体を検索する MultiRankIt を提案する．MultiRankIt では，指示文から抽出された句を扱い，抽出された句と対象物体領域の関係をモデル化する Crossmodal Noun Phrase Encoder (CNPE) を導入する．これにより，複雑な参照表現を含む指示文から適切な対象物体を検索することが期待される．さらに，複数の周辺画像を扱う Crossmodal Region Feature Encoder (CRFE) を導入する．対象物体領域に近接した画像だけでなく，より広い範囲を捉えた周辺画像を扱うことで，対象物体とその周囲の環境との関係を効果的にモデル化することが期待される．

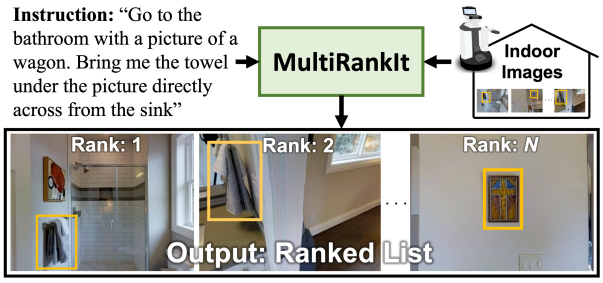

図 1: LTRPO タスクの典型的な例.

本研究は learning-to-rank タスク [Liu 09] として物理世界検索を扱い，DSR がユーザやオペレータに対して対象物体候補のランク付けリストを出力する点で既存手法と異なる．提案手法は自動化と人為的介入のバランスを取ることを目的としている．本研究の新規性であるマルチモーダル物体検索への learning–to–rank アプローチの導入は，マルチモーダル画像検索（例: [Wu 21]）などの他タスクにも応用可能であると考えられる．本研究の独自性は以下の通りである．

- Human-in-the-loop 設定において，ユーザによるオープンボキャブラリの指示文から対象物体を特定する新しいアプローチである MultiRankIt を提案する．
- 参照表現を含む句と対象物体領域との関係をモデル化する CNPE を導入する．
- 対象物体と複数の周囲画像との関係をモデル化する CRFE を導入する．

2. 関連研究

Vision and Language (V&L) の研究分野では VQA [Zhou 20] や画像キャプション生成 [Yang 20]，V&L Navigation [Anderson 21] など，多くの研究が行われている．[Uppal 22] は V&L タスクに関する様々なタスクと最新の手法について包括的な概観を提示している．以下では，V&L 分野における関連研究，特に本研究と密接に関連するクロスモーダル検索とロボティクス分野における V&L に焦点を当てる．

連絡先: 兼田寛大，慶應義塾大学，神奈川県横浜市港北区日吉 3-14-1，k.kaneda@keio.jp

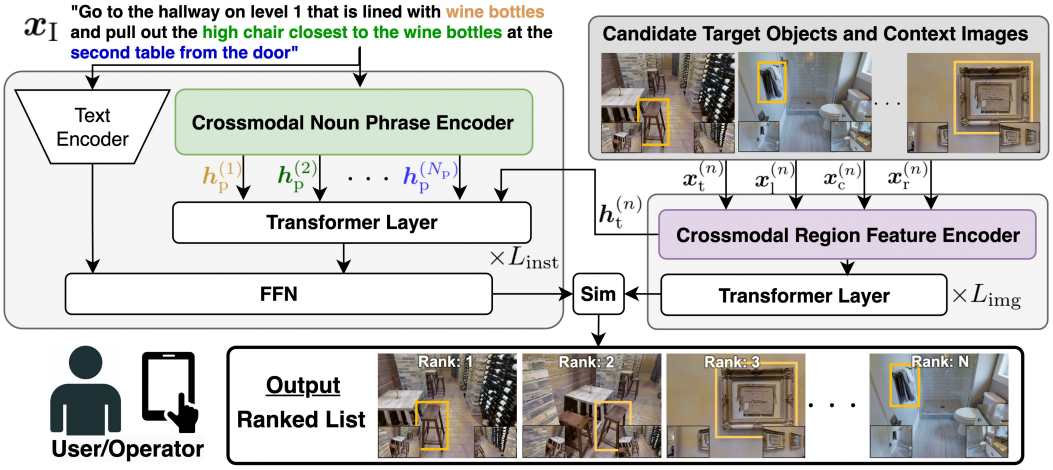

図 2: MultiRankIt の構造．“FFN” は順伝播型ネットワーク層を示す．

クロスモーダル検索は，あるモダリティで表現されたクエリに基づき，別のモダリティからサンプルを検索するタスクであり，幅広い応用がなされている (例: ファッション検索 [Wu 21])．[Vo 19] は，画像とそれに付随するテキストからなる入力クエリを用いた画像検索手法を提案している．この論文では，ゲーティング関数と残差接続を用いて参照画像とテキストクエリを結合する TIRG (Text Image ResidualGating) を紹介している．DCNet [Kim 21] は TIRG を拡張し，さらにターゲット画像と参照画像の差分を用いてロバストなマルチモーダル表現を学習する Correction Network を導入している．

また，ロボット分野での V&L に関連する研究として，以下のような既存研究がある．[Magassouba 19] は，制約のない文章からユーザが指示した対象物体を予測することを目的とした，multimodal language understanding task for fetching instructions タスクを扱っている．Target-Dependent UNITER [Ishikawa 21] は，テキストと視覚特徴の関係をモデル化するために，UNITER ベースの変換器アーキテクチャを導入している．HLSM-MAT [Ishikawa 22] は，摂動更新に 2 種類のモーメントを用いる Moment-based Adversarial Training アルゴリズムにより，指示文に基づく家事タスクの実行における V&L タスクを扱っている．

前述のように，本手法は，DSR がユーザやオペレータに対して対象物体の候補のランク付けを行う learning–to–rank タスク [Liu 09] として扱う点で，既存手法（例: [Vo 19]）と異なっている．また，[Hatori 18] が固定視点による 1 枚の画像から対象物体を特定するのに対し，本手法は家庭内環境を示す複数の画像から対象物体を検索する点でも [Hatori 18] と異なり，また本手法では対象物体とその周囲の環境画像との関係をモデル化するモジュールを導入している．加えて名詞句を扱う既存の手法（例: [Subramanian 22]）とは異なる点として，参照表現を含む句と対象物体の領域画像との関係をモデル化するモジュールを導入していることが挙げられる．

3. 問題設定

本論文では，オープンボキャブラリの指示文から，DSR が家庭環境内で撮影した対象物体画像を検索するタスクである LTRPO タスクを扱う．本タスクでは，出力されるリストにおいて適切な対象物体領域を含む画像が上位にランク付けされることが望ましい．本論文で使用する用語を以下のように定義する．

- **指示文:** DSR が家事を行うためのオープンボキャブラリの指示文．
- **対象物体:** 指示の対象である物体．
- **対象物体領域:** 対象物体のバウンディングボックス．
- **周辺画像:** 対象物体とその周辺を撮影した画像．

本モデルへの入力は指示文，対象物体領域，周辺画像であり，出力はランク付けされたリストである．画像中の対象物体の座標は与えられていると仮定し，また MRR（Mean reciprocal rank）と Recall@K でモデルを評価する．

4. 提案手法

入力 $\boldsymbol{x}$ を以下のように定義する．

$$\boldsymbol{x} = (\boldsymbol{x}_{\mathrm{I}}, T, C) \tag{1}$$

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

$$C = \{X_{\mathrm{c}}^{(n)} | n = 1, 2, ..., N_{\mathrm{targ}}\} \tag{3}$$

$$X_{\mathrm{c}}^{(n)} = (\boldsymbol{x}_{\mathrm{c}}^{(n)}, \boldsymbol{x}_{\mathrm{l}}^{(n)}, \boldsymbol{x}_{\mathrm{r}}^{(n)}) \tag{4}$$

ここで $\boldsymbol{x}_{\mathrm{I}} \in \mathbb{R}^{V \times L}$ は語彙数 V と最大トークン長 L で 1–of–K ベクトルとしてトークン化された指示文，$\boldsymbol{x}_{\mathrm{t}}^{(n)} \in \mathbb{R}^{3 \times W \times H}$ は幅 W と高さ H の対象物体領域を示す．また，$\boldsymbol{x}_{\mathrm{t}}^{(n)} \in \mathbb{R}^{3 \times 256 \times 256}$, $\boldsymbol{x}_{\mathrm{t}}^{(n)} \in \mathbb{R}^{3 \times 256 \times 256}$, $\boldsymbol{x}_{\mathrm{t}}^{(n)} \in \mathbb{R}^{3 \times 256 \times 256}$ は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，そして $\boldsymbol{x}_{\mathrm{c}}^{(n)}$ の右側の画像を示す．

まず，$\boldsymbol{x}_{\mathrm{I}}$ と $(\boldsymbol{x}_{\mathrm{t}}^{(n)}, X_{\mathrm{c}}^{(n)})$ の間の類似度を計算するため，クエリ $(\boldsymbol{x}_{\mathrm{I}}, \boldsymbol{x}_{\mathrm{t}}^{(n)}, X_{\mathrm{c}}^{(n)})$ を作成する．ここで，1 章でも述べたように，LTRPO タスクの指示文には対象物体やランドマーク，部屋などに関する複雑な参照表現が多く含まれている．そこで，$\boldsymbol{x}_{\mathrm{I}}$ から抽出した名詞句と前置詞句を扱う CNPE を導入し，抽出した句と $\boldsymbol{x}_{\mathrm{t}}$ の関係をモデル化する．

CNPE は以下の手順で $\boldsymbol{x}_{\mathrm{I}}$ から N_p 個の句を抽出する．まず，Stanford Parser [Schuster 16] を用いて名詞句と前置詞句を抽出する．次に，隣接する名詞句と前置詞句をグループ化し，N_{p} 個の句 $\boldsymbol{x}_{\mathrm{p}}^{(k)} (k = 1, 2, ..., N_{\mathrm{p}})$ を得る．その後，事前学習済みの CLIP text encoder [Radford 21] を用いて $\boldsymbol{x}_{\mathrm{I}}$ と $\boldsymbol{x}_{\mathrm{p}}^{(k)}$ からそれぞれ言語特徴量 $\boldsymbol{h}_{\mathrm{I}} \in \mathbb{R}^{768}$ と $\boldsymbol{h}_{\mathrm{p}}^{(k)} \in \mathbb{R}^{768}$ を得る．

次に，$\boldsymbol{h}_{\mathrm{p}}^{(k)}$ と $\boldsymbol{h}_{\mathrm{t}}^{(n)}$ の関係を以下のように計算する．

$$\boldsymbol{h}_{\mathrm{p}} = f_{\mathrm{inst}}(\boldsymbol{h}_{\mathrm{p}}^{(1)}; \boldsymbol{h}_{\mathrm{p}}^{(2)}; ...\boldsymbol{h}_{\mathrm{p}}^{(k)}; \boldsymbol{h}_{\mathrm{t}}^{(n)}) \tag{5}$$

ここで，f_{inst} は L_{inst} 層の transformer 層，$\boldsymbol{h}_{\mathrm{t}}^{(n)}$ は $\boldsymbol{x}_{\mathrm{t}}^{(n)}$ から得られた画像特徴量を示す．また，各 transformer 層は multi-head self attention 層 [Vaswani 17] と順伝播型ネットワーク層によって構成される．最後に，指示文に対する埋め込み表現 $\boldsymbol{h}_{\mathrm{inst}} = f_{\mathrm{FFN}}(\boldsymbol{h}_{\mathrm{I}}; \boldsymbol{h}_{\mathrm{p}})$ を得る．ここで，f_{FFN} は順伝播型ネットワーク層を示す．

CRFE は，対象物体領域に近接した画像だけでなく，より広い範囲を捉えた画像を扱うことで，対象物体とその周囲の文脈環境との関係を計算する．まず，事前学習済みの CLIP image

表 1: 定量的結果.

		条件		テスト集合				
		w/ CNPE	w/ $X_{\mathrm{c}}^{(n)}$	MRR	R@1[%]	R@5[%]	R@10[%]	R@20[%]
(a)	ベースライン		✓	0.415 + 0.009	14.0 + 1.0	45.3 + 1.7	63.8 + 2.5	80.8 + 2.0
(b)	提案手法	✓		0.373 + 0.015	12.1 + 0.5	39.6 + 1.4	56.1 + 1.1	70.2 + 0.7
(c)			✓	0.426 + 0.004	14.6 + 0.4	45.3 + 0.5	66.1 + 1.7	80.6 + 0.9
(d)		✓	✓	**0.501 + 0.008**	**18.3 + 1.0**	**52.2 + 1.4**	**69.8 + 1.5**	**83.8 + 0.6**

encoder を用いて $\boldsymbol{x}_{\mathrm{t}}^{(n)}, \boldsymbol{x}_{\mathrm{c}}^{(n)}, \boldsymbol{x}_{\mathrm{l}}^{(n)}, \boldsymbol{x}_{\mathrm{r}}^{(n)}$ から画像特徴量 $\boldsymbol{h}_{\mathrm{t}}^{(n)} \in \mathbb{R}^{768}$, $\boldsymbol{h}_{\mathrm{c}}^{(n)} \in \mathbb{R}^{768}$, $\boldsymbol{h}_{\mathrm{l}}^{(n)} \in \mathbb{R}^{768}$, $\boldsymbol{h}_{\mathrm{r}}^{(n)} \in \mathbb{R}^{768}$ を得る．その後，画像に対する埋め込み表現 $\boldsymbol{h}_{\mathrm{targ}} = f_{\mathrm{img}}(\boldsymbol{h}_{\mathrm{t}}^{(n)}; \boldsymbol{h}_{\mathrm{c}}^{(n)}; \boldsymbol{h}_{\mathrm{l}}^{(n)}; \boldsymbol{h}_{\mathrm{r}}^{(n)})$ を計算する．ここで，f_{img} は L_{img} 層の transformer 層を示す．

最後に，$\boldsymbol{x}_{\mathrm{I}}$ と $(\boldsymbol{x}_{\mathrm{t}}^{(n)}, X_{\mathrm{c}}^{(n)})$ の類似度をコサイン類似度を用いて以下のように計算する．

$$s(\boldsymbol{x}_{\mathrm{I}}, \boldsymbol{x}_{\mathrm{t}}^{(n)}, X_{\mathrm{c}}^{(n)}) = \frac{\boldsymbol{h}_{\mathrm{inst}} \cdot \boldsymbol{h}_{\mathrm{targ}}}{\|\boldsymbol{h}_{\mathrm{inst}}\| \|\boldsymbol{h}_{\mathrm{targ}}\|} \tag{6}$$

本モデルの出力は $s(\boldsymbol{x}_{\mathrm{I}}, \boldsymbol{x}_{\mathrm{t}}^{(n)}, X_{\mathrm{c}}^{(n)})$ に基づいて T をランク付けしたリストである．各バッチにおける損失は以下のように計算される．

$$L_{\mathcal{B}} = -\frac{1}{|\mathcal{B}|} \sum_{\boldsymbol{x}_{\mathrm{t}}^{(n)} \in \mathcal{B}} \log \frac{\exp\left(s(\boldsymbol{x}_{\mathrm{I}}, \boldsymbol{x}_{\mathrm{t}}^{(n)}, X_{\mathrm{c}}^{(n)})\right)}{\sum_{j=1}^{|\mathcal{B}|} \exp\left(s(\boldsymbol{x}_{\mathrm{I}}, \boldsymbol{x}_{\mathrm{t}}^{(n)}, X_{\mathrm{c}}^{(n)})\right)} \tag{7}$$

ここで $|\mathcal{B}|$ はバッチサイズを示し，この損失関数は InfoNCE [Radford 21] において $\boldsymbol{x}_{\mathrm{I}}$ のみを考慮した場合と等しい．

5. 実験

5.1 データセットと実験設定

我々の知る限り，LTRPO タスクのための標準的なデータセットは存在しない．Object-goal navigation タスクの標準データセットである REVERIE データセット [Qi 20] は，実環境の室内画像と物体操作指示を扱っているという点で，物体位置特定タスクにとって重要な要素を含むが，REVERIE データセット単体では LTRTO タスクには適さない．そこで，我々は LTRPO タスクのための新しいデータセットである LTRRIE (Learning–to–Rank in Real Indoor Environments) データセットを構築した．

LTRRIE データセットは以下の手順で得られた．まず，REVERIE データセットから指示文を，Matterport3D Simulator [Chang 17] からパノラマ画像を収集した．次に，REVERIE データセットで提供された対象物体の座標を用いてパノラマ画像を切り出し，対象物体領域を取得した．なお，元画像の端に対象物体領域があるサンプルは，切り出すと対象物体全体が含まれない可能性があるため除外した．REVERIE データセットに含まれる指示文は，Amazon Mechanical Turk を利用して 1000 人以上のアノテータによって収集された．アノテータには，移動経路の動画とランダムに選択された対象物体が提示され，その対象物体を操作するための指示文を作成するよう指示された．

LTRRIE データセットは，58 の環境，5501 の命令文，そして 4352 の対象物体領域から構成される．語彙数は 53118 語，総語数は 103118 語，平均文長は 18.78 語である．LTRRIE データセットには，訓練集合に 4210 個，検証集合に 397 個，テスト集合に 501 個のサンプルが含まれる．各集合にはそれぞれ 50, 4, 4 個の環境が含まれ，環境の重複はない．したがって，検証セットに含まれるサンプルは，unseen object とみなすことができる．なお，訓練集合の指示文は REVERIE の train set から，検証集合とテスト集合の指示文は REVERIE データセットの val_unseen set を分割して得られた．訓練集合はモデルのパラメータを更新するため，検証集合はハイパーパラメータを調整するために使用された．そして，テスト集合においてモデルを評価した．ハイパーパラメータは，$L_{\mathrm{inst}} = 4$, $L_{\mathrm{img}} = 4$, $\#A = 4$, $\#H = 756$ とした．ここで，$\#H$ と $\#A$ はそれぞれ f_{img} と f_{inst} の 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 2.0×10^{-5}，バッチサイズは 128 とした．学習には 11GB 搭載の GeForce RTX 2080 および Intel Core i9-9900K を使用した．モデルの学習には約 60 分かかり，推論時間は約 20ms/サンプルであった．

5.2 実験結果

我々の知る限り，LTRPO タスクを扱った既存研究は限られている．そこで，CLIP を拡張し，LTRPO タスクを扱うベースライン手法を開発した．ベースライン手法は，$\boldsymbol{x}_{\mathrm{I}}$, $\boldsymbol{x}_{\mathrm{t}}$, $\boldsymbol{x}_{\mathrm{lcr}}$ を入力とする．ここで，$\boldsymbol{x}_{\mathrm{lcr}}$ は $\boldsymbol{x}_{\mathrm{c}}$， $\boldsymbol{x}_{\mathrm{l}}$, と $\boldsymbol{x}_{\mathrm{r}}$ を横方向に結合し，256×256 ピクセルにリサイズして得られた．$\boldsymbol{x}_{\mathrm{lcr}}$ はその後 CLIP image encoder と MLP に入力され，画像特徴量が得られる．また，$\boldsymbol{x}_{\mathrm{I}}$ を CLIP text encoder に入力し，言語特徴量を得る．これらの特徴量を用いて式 6 に従って類似度を計算し，この類似度に基づいて得られた T のランク付けリストを出力とする．

表 1 にテスト集合におけるベースライン手法，提案手法，および ablation study の定量結果を示す．表中の数値は 5 回の試行における平均値と標準偏差であり，ablation study の詳細については，下記で説明する．タスクの評価尺度には MRR と recall@K（K= 1, 5, 10, 20）を使用し，本論文では MRR を主要評価尺度とした．

MRR は $\mathrm{MRR} = \frac{1}{\mathrm{N_{inst}}} \sum_{\mathrm{j=1}}^{\mathrm{N_{inst}}} \frac{1}{\mathrm{r}_1^{(\mathrm{j})}}$ のように定義される．ここで，N_{inst} と $r_1^{(j)}$ はそれぞれ指示文の数と検索されたサンプルの中で最も高くランク付けされたサンプルの順位を示す．Recall@K は $\mathrm{Recall@K} = \frac{1}{N_{\mathrm{inst}}} \sum_{j=1}^{N_{\mathrm{inst}}} \frac{|A \cap B|}{|A|}$ のように定義される．ここで，A は正解サンプルの集合，B は検索上位 K 個のサンプル集合を示す．MRR と recall@K は learning–to–rank 設定において標準的な評価尺度のため用いた．[Liu 09]

表 1 に示すように，ベースライン手法 (a) と提案手法 (d) の MRR はそれぞれ 0.415 と 0.501 であった．したがって，提案手法 (d) はベースライン手法 (a) よりも 0.086 ポイント MRR が優れていた．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．また，recall@20 を除く全ての結果において，統計的に有意であった ($p < 0.05$)．

図 3 に定性的結果を示す．図 3(a) は $\boldsymbol{x}_{\mathrm{I}}$ “Go to the bathroom with a picture of a wagon and bring me the towel directly across from the sink” に対する上位二件の検索結果を示す．このサンプルでは，提案法の MRR は 1，ベースライン法の MRR は 0.1 であった．ここで，指示文に示された “sink” は $\boldsymbol{x}_{\mathrm{c}}$ に含まれず，一方で $\boldsymbol{x}_{\mathrm{r}}$ に含まれている．この結果より，広い範囲を捉える周辺画像を扱う CRFE の導入が，性能に寄与したと考えられる．

図 3(b) は $\boldsymbol{x}_{\mathrm{I}}$ “Go to the hallway on level 1 that is lined with wine bottles and pull out the high chair closest to the wine bottles at the second table from the door” に対する上位二件の検索結果を示す．このサンプルでは，提案法の MRR は 1，ベースライン手法の MRR は 0.091 であった．このサン

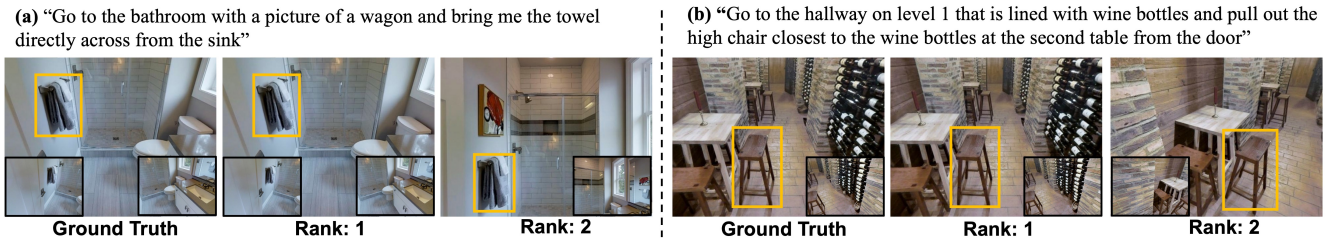

図 3: 定性的結果．黄色の枠で囲まれた領域は $\boldsymbol{x}_{\mathrm{t}}$，緑の点線で囲まれた領域は指示に示されている $\boldsymbol{x}_{\mathrm{t}}$ 以外のオブジェクトを示す．黒枠内の左下と右下の画像は、それぞれ $\boldsymbol{x}_{\mathrm{l}}$ と $\boldsymbol{x}_{\mathrm{r}}$ を示す．一部の $\boldsymbol{x}_{\mathrm{l}}$ と $\boldsymbol{x}_{\mathrm{r}}$ は可視性の都合上省略している．

プルでは $\boldsymbol{x}_{\mathrm{I}}$ に "the high chair closest to the wine bottles at the second table from the door" という複雑な参照表現が含まれているが，提案手法は適切な $\boldsymbol{x}_{\mathrm{t}}$ を "rank 1" として検索することに成功した．この結果より，名詞句と対象物体との関係をモデル化する CNPE モジュールの導入が性能に寄与したと考えられる．

Ablation 条件として，以下を定めた．(i) w/o C: $\boldsymbol{x} = (\boldsymbol{x}_{\mathrm{I}}, T)$ を入力とした場合の性能を調べるため，C を削除した．(ii) w/o CNPE: 性能への影響を調べるため，CNPE を削除した．表 1 は ablation studies の定量的結果を示す．条件 (b) において，テスト集合の MRR は 0.128 ポイント減少した．この結果から，CRFE を導入したことで性能が向上したと考えられる．同様に，条件 (c) の結果より，名詞句を扱う CNPE の導入も性能の向上に寄与していると考えられる．

6. おわりに

本論文では，ユーザからのオープンボキャブラリの指示文から DSR が家庭環境内で撮影した対象物体画像を検索する LTRPO タスクを扱った．提案手法による貢献は以下である．

- Human-in-the-loop 設定において，ユーザによるオープンボキャブラリの指示文から対象物体を特定する新しいアプローチ，MultiRankIt を提案した．
- 参照表現を含む句と対象物体領域との関係をモデル化する CNPE を導入した．
- 対象物体と複数の周囲画像との関係をモデル化する CRFE を導入した．
- すべての評価尺度において提案手法がベースライン手法を上回る性能を確認した．

謝辞

本研究の一部は，JSPS 科研費 20H04269，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

[Anderson 21] Anderson, P., Shrivastava, A., Truong, J., Majumdar, A., Parikh, D., et al.: Sim-to-Real Transfer for Vision-and-Language Navigation, in *CoRL*, pp. 671–681 (2021)

[Chang 17] Chang, A., Dai, A., Funkhouser, T., Halber, M., et al.: Matterport3D: Learning from RGB-d Data in Indoor Environments, *arXiv preprint arXiv:1709.06158* (2017)

[Hatori 18] Hatori, J., Kikuchi, Y., et al.: Interactively Picking Real-World Objects with Unconstrained Spoken Language Instructions, in *IEEE ICRA*, pp. 3774–3781 (2018)

[Ishiguro 21] Ishiguro, H.: The Realisation of an Avatar-Symbiotic Society where Everyone can Perform Active Roles without Constraint, *Advanced Robotics*, Vol. 35, No. 11, pp. 650–656 (2021)

[Ishikawa 21] Ishikawa, S. and Sugiura, K.: Target-Dependent UNITER: A Transformer-based Multimodal Language Comprehension Model for Domestic Service Robots, *IEEE RAL*, Vol. 6, No. 4, pp. 8401–8408 (2021)

[Ishikawa 22] Ishikawa, S. and Sugiura, K.: Moment-based Adversarial Training for Embodied Language Comprehension, in *IEEE ICPR*, pp. 4139–4145 (2022)

[Kim 21] Kim, J., Yu, Y., Kim, H., and Kim, G.: Dual Compositional Learning in Interactive Image Retrieval, in *AAAI*, Vol. 35, pp. 1771–1779 (2021)

[Liu 09] Liu, T.: Learning to Rank for Information Retrieval, *Foundations and Trends in Information Retrieval*, Vol. 3, No. 3, pp. 225–331 (2009)

[Magassouba 19] Magassouba, A., Sugiura, K., Quoc, A., et al.: Understanding Natural Language Instructions for Fetching Daily Objects Using GAN-Based Multimodal Target–Source Classification, *IEEE RAL*, Vol. 4, No. 4, pp. 3884–3891 (2019)

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

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

[Schuster 16] Schuster, S., et al.: Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks, in *LREC*, pp. 2371–2378 (2016)

[Subramanian 22] Subramanian, S., Merrill, W., Darrell, T., Gardner, M., Singh, S., and Rohrbach, A.: ReCLIP: A Strong Zero-shot Baseline for Referring Expression Comprehension, *arXiv preprint arXiv:2204.05991* (2022)

[Uppal 22] Uppal, S., et al.: Multimodal Research in Vision and Language: A Review of Current and Emerging Trends, *Information Fusion*, Vol. 77, pp. 149–171 (2022)

[Vaswani 17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., et al.: Attention is All You Need, in *NeurIPS*, Vol. 30, pp. 5998–6008 (2017)

[Vo 19] Vo, N., Jiang, L., Sun, C., Murphy, K., Li, J., Fei, L., and Hays, J.: Composing Text and Image for Image Retrieval-An Empirical Odyssey, in *IEEE CVPR*, pp. 6439–6448 (2019)

[Wu 21] Wu, H., Gao, Y., Guo, X., Al, Z., Rennie, S., Grauman, K., and Feris, R.: Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback, in *IEEE CVPR*, pp. 11307–11317 (2021)

[Yang 20] Yang, Z., Garcia, N., Chu, C., Otani, M., Nakashima, Y., and Takemura, H.: BERT Representations for Video Question Answering, in *WACV*, pp. 1556–1565 (2020)

[Yu 18] Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., and Berg, T.: MattNet: Modular Attention Network for Referring Expression Comprehension, in *IEEE CVPR*, pp. 1307–1315 (2018)

[Zhou 20] Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., and Gao, J.: Unified Vision-Language Pre-training for Image Captioning and VQA, in *AAAI*, Vol. 34, pp. 13041–13049 (2020)