# MultiRankIt: ランキング学習と大規模言語モデルによる物理世界検索

○兼田寛大，長嶋隼矢，是方諒介，杉浦孔明（慶應義塾大学）

少子高齢化に伴い在宅介助者の不足が問題となっており, 生活支援ロボットはその解決策として期待されている．本研究では，human-in-the-loop 設定においてロボットが家庭環境内の物体を操作する自然言語指示から対象物体を検索する learning-to-rank physical objects タスクを扱う．例として，“Go to the bedroom and open the window to the left of the box” という指示文が与えられたときに，対象物体である “the window to the left of the box” をより高くランク付けしたリストを出力することが望ましい．しかし，複雑な参照表現を含む指示文に対する検索性能はまだ十分ではない．そこで本論文では，大規模言語モデルを用いて指示文の分類および標準形への言い換えを行う Task Paraphraser モジュールと，セグメンテーションマスク重畳画像の特徴量を扱い，対象物体およびその周辺物体の関係をモデル化する Crossmodal Segmented Objects Encoder を提案する．

## 1. はじめに

少子高齢化に伴い在宅介助者の不足が社会問題となっており，生活支援ロボットはその解決策として期待されている．生活支援ロボットの応用におけ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．この設定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．

そこで本研究では，家庭環境内の物体を操作する自然言語指示から対象物体を検索する LTRPO タスク [1] を扱う．図 1 に LTRPO タスクの典型的な場面を示す．指示文として “Go to the bedroom and open the window to the left of the tallboy as wide as possible” が与えられた場合に，本タスクにおいてモデルは対象物体を上位にランク付けしたリストを出力することが望ましい．ここで，human-in-the-loop 設定では人間の注意力は限られているため, 人間が容易に選択できる数の対象物体を提示することが重要である．

近年の CLIP [2] 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, 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．例えば，[1] では指示文中の対象物体以外の句に関連する画像を上位に検索する誤りが多いことが報告されている．そこで本研究では，大規模言語モデルとマルチモーダル基盤モデルを用いて LTRPO タスクを扱う手法を提案する．

本研究の新規性は以下の通りである．

- 大規模言語モデルを用いて冗長な情報を含む指示文からタスクを分類し，さらに各タスクの標準形に言い換える Task Paraphraser モジュールを提案する．
- SAM [3] を用いて獲得したセグメンテーションマスク重畳画像の特徴量を扱い，対象物体およびその周辺物体の関係をモデル化する Crossmodal Segmented Objects Encoder を導入する．

## 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おり [4–8]，[4] はマルチモーダル言語処理分野に関するタスクと手法について包括的にまとめている．

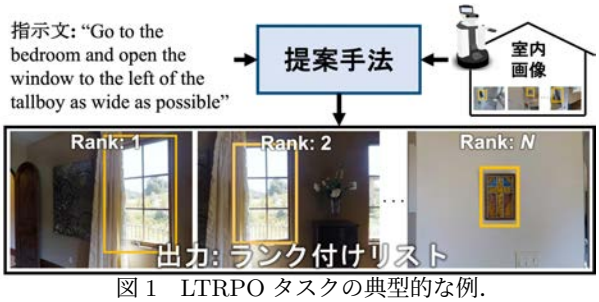


図 1　LTRPO タスクの典型的な例.

本論文で扱う LTRPO タスクに関連する分野として，クロスモーダル検索とマルチモーダル言語処理に関連したロボティクス分野が挙げられる．クロスモーダル検索は，特定の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る．[5] は，画像およびテキストを用いた画像検索において， 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を結合する Text Image Residual Gating を提案している．[6] は画像に対するテキストを補正し，そのテキストと画像中のグローバルおよびローカルな特徴量との関係をモデル化している．

マルチモーダル言語処理に関連したロボティクス分野においては，物体操作指示文理解を扱う手法 [7,8] や，LTRPO タスクを扱う手法 [1] が提案されている．[7] は，対象物体および配置目標を物体検出により抽出した領域から特定する手法を提案し，[8] は，領域単位の対象物体と他の物体との関係を直接学習する Target-dependent UNITER を提案している．

## 3. 問題設定

LTRPO タスクはユーザによる物体操作に関する指示文から，生活支援ロボットが家庭環境内で撮影した物体の領域画像を検索するタスクである．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が上位にランク付けされることが望ましい．本論文で扱う用語は [1] に準拠する．

本タスクにおける入力は指示文 $\boldsymbol{x}_{\mathrm{I}}$，環境中の対象物体領域を含む $N_{\mathrm{targ}}$ 枚の画像で構成されたリスト $T = \{\boldsymbol{x}_{\mathrm{t}}^{(n)} | n = 1, 2, ..., N_{\mathrm{targ}}\}$ ，および $\boldsymbol{x}_{\mathrm{t}}^{(n)}$ の周辺を写した

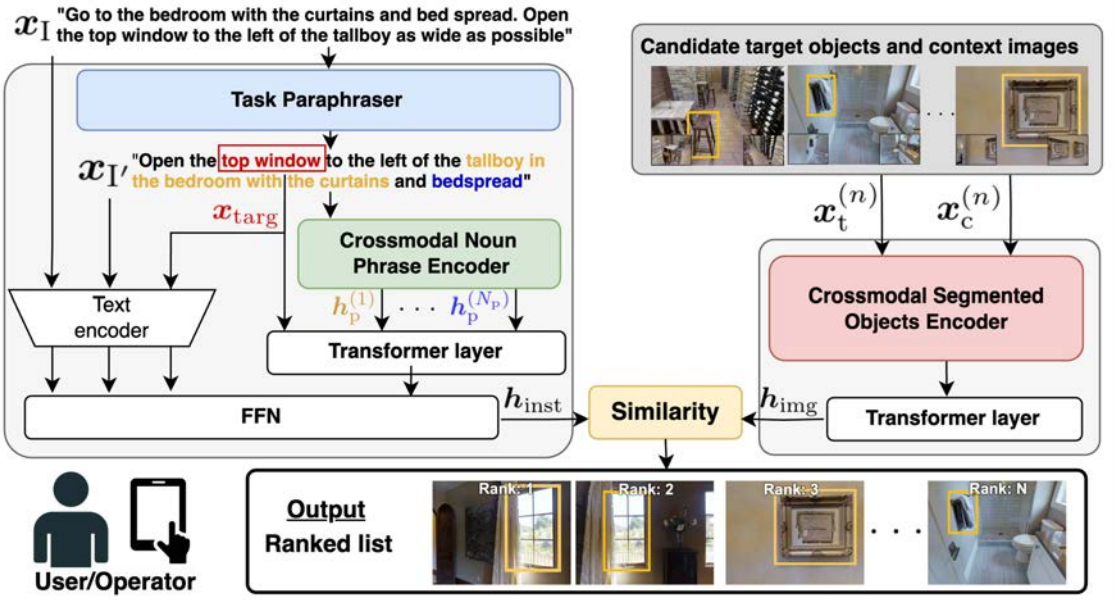


図 2　提案手法の構造．“FFN” は順伝播型ネットワーク層を示す．

画像で構成されたリスト $C = \{\boldsymbol{x}_{\mathrm{c}}^{(n)} | n = 1, 2, ..., N_{\mathrm{targ}}\}$ である．注意として，$T$ には同一の対象物体を写した画像も含まれる (例: 別視点から撮影された画像)．出力は対象物体領域について $T$ をランク付けしたリストである．

## 4. 提案手法

図 2 にモデルの構造を示す．提案手法は主に Task Paraphraser (TP), Crossmodal Segmented Objects Encoder (CSOE) および Crossmodal Noun Phrase Encoder (CNPE) の 3 つのモジュールから構成される．

入力 $\boldsymbol{x}$ を以下のように定義する．

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

ここで $\boldsymbol{x}_{\mathrm{I}} \in \mathbb{R}^{V \times L}$，$\boldsymbol{x}_{\mathrm{t}}^{(n)} \in \mathbb{R}^{3 \times W \times H}$，および $\boldsymbol{x}_{\mathrm{c}}^{(n)} \in \mathbb{R}^{N_{\mathrm{c}} \times 3 \times 256 \times 256}$ はそれぞれ語彙数 $V$ 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，幅 $W$ と高さ $H$ の対象物体領域，および $\boldsymbol{x}_{\mathrm{t}}^{(n)}$ の周辺を写した画像を $N_c$ 枚連結して得られた周辺画像を示す．また，$N_{\mathrm{targ}}$ は環境中の対象物体領域の数を示す．

TP は冗長な情報を含む $\boldsymbol{x}_{\mathrm{I}}$ からタスクを分類し，さらに各タスクの標準形に言い換える．まず，大規模言語モデルである ChatGPT を用いて $\boldsymbol{x}_{\mathrm{I}}$ を以下の 5 種類のタスクに分類する．(i) Carry：対象物体をある地点から目標領域へ移動させるタスク，(ii) Retrieval：対象物体を把持するタスク（目標領域が明示されていない場合やユーザへ物体移動を行う場合を含む），(iii) Manipulation：物体操作を扱うタスク，(iv) Open/Close：ドアや棚などの開閉を扱うタスク，(v) その他：上記以外のタスク．これらのタスクは，国際補助犬パートナー協会が定義した介助犬タスク [1] のうち，生活支援ロボットが実行可能なタスクを抽出したものである．介助犬は人間の支援を目的としているため，上記タスクは生活支援ロボットにおいても重要なタスクであると考えられる．上記を行うプロンプトの抜粋は以下の通りである．“Robots can perform various tasks and here are the definitions for each task: <Object Transportation>: This task involves a robot moving an object from one location to another ...” 例として，$\boldsymbol{x}_{\mathrm{I}}$ が “Go to the bedroom with french doors and a white chair and open the window farthest from that chair” の場合，open/close タスクに分類される．また上記の 5 種類のタスクに対し，介助犬タスクをもとに標準形を作成する．

次に，同様に大規模言語モデルを用いて $\boldsymbol{x}_{\mathrm{I}}$ から 対象物体，対象物体の位置および目標領域に関する句を抽出し，各タスクの標準形への言い換えを行うプロンプトから $\boldsymbol{x}_{\mathrm{I}'}$ を獲得する．プロンプトの抜粋は以下の通りである．“If the instruction is classified as <Object Transportation>, paraphrase the instruction in the following format: <Object Transportation>: Carry <object> from <original place> to <destination> ...” 例として，$\boldsymbol{x}_{\mathrm{I}}$ が “Go to the hallway on the third floor with a silver and black corded phone and white vase in it. Take the photo in the dark frame off of the wall” であるとき，$\boldsymbol{x}_{\mathrm{I}'}$ として “Retrieve the photo in the dark frame from the wall in the hallway with a corded phone and a white vase,” $\boldsymbol{x}_{\mathrm{targ}}$ として “photo in the dark frame” が得られる．最後に，事前学習済みの CLIP テキストエンコーダを用いることで，$\boldsymbol{x}_{\mathrm{I}'}$ から $\boldsymbol{h}_{\mathrm{I}'}$ を得る．

CSOE は SAM [3] を用いて獲得したセグメンテーションマスク重畳画像の特徴量を扱い，対象物体およびその周辺物体の関係をモデル化する．まず $N_c$ 枚の周辺画像に SAM を適用することで $\boldsymbol{x}_{\mathrm{c}}$ に写っている物

[1] http://www.iaadp.org/tasks.html

表 1　定量的結果．“R@K” は Recall@K を示す．

| | | 条件 | | テスト集合 | | | | |
|---|---|---|---|---|---|---|---|---|
| | | w/TP | w/CSOE | MRR | R@1[%] | R@5[%] | R@10[%] | R@20[%] |
| (a) | CLIP-extended [2] | | | 0.415 ± 0.009 | 14.0 ± 1.0 | 45.3 ± 1.7 | 63.8 ± 2.5 | 80.8 ± 2.0 |
| (b) | MultiRankIt [1] | | | 0.501 ± 0.008 | 18.3 ± 1.0 | 52.2 ± 1.4 | 69.8 ± 1.5 | 83.8 ± 0.6 |
| (c) | 提案手法 | | ✓ | 0.527 ± 0.013 | 18.3 ± 0.7 | 56.1 ± 2.8 | 74.8 ± 1.1 | 89.7 ± 1.0 |
| (d) | | ✓ | | 0.546 ± 0.008 | 19.7 ± 0.3 | 57.8 ± 1.0 | 76.2 ± 0.9 | 90.5 ± 0.8 |
| (e) | | ✓ | ✓ | **0.579 ± 0.010** | **22.3 ± 0.9** | **59.4 ± 0.8** | **77.5 ± 1.2** | **90.8 ± 0.8** |

体のセグメンテーションマスクを獲得する．次に，セグメンテーションマスクから $N_s$ 個の周辺物体の領域画像 $\boldsymbol{x}_{\mathrm{s}}^{(k)}(k=1,2,...,N_{\mathrm{s}})$ を得る．それらの領域画像に事前学習済みの CLIP 画像エンコーダを用いることで画像特徴量 $\boldsymbol{h}_{\mathrm{s}}^{(k)}(k=1,2,...,N_{\mathrm{s}})$ を得る．同様に，$\boldsymbol{x}_{\mathrm{t}}^{(n)}$ および $\boldsymbol{x}_{\mathrm{c}}^{(n)}$ から画像特徴量 $\boldsymbol{h}_{\mathrm{t}}^{(n)} \in \mathbb{R}^{768}$ および $\boldsymbol{h}_{\mathrm{c}}^{(n)} \in \mathbb{R}^{768}$ も得る．最後に $L_{\mathrm{csoe}}$ 層の transformer 層 を用いて $\boldsymbol{h}_{\mathrm{t}}^{(n)} \in \mathbb{R}^{768}$, $\boldsymbol{h}_{\mathrm{c}}^{(n)} \in \mathbb{R}^{768}$ の関係をモデル化し，画像に関する特徴量 $\boldsymbol{h}_{\mathrm{img}}$ を得る．ここで，各 transformer 層は multi-head self attention 層 [9] と順伝播型ネットワーク層によって構成される．

上記で説明したように，CSOE では対象物体の矩形領域を CLIP 画像エンコーダに入力し $\boldsymbol{h}_{\mathrm{img}}$ を獲得する．そのため，$\boldsymbol{h}_{\mathrm{img}}$ との類似度を適切に計算するためには CLIP のマルチモーダル埋め込み空間に適した入力を扱うことが重要である．ここで CLIP では “A photo of {label}” のような文を扱っているため，CNPE では指示文全体から名詞句および前置詞句を抽出し，それらと対象物体の句の関係をモデル化する．まず，Stanford Parser [10] を用いて $\boldsymbol{x}_{\mathrm{I}'}$ から名詞句および前置詞句を抽出する．次に，隣接する名詞句および前置詞句をグループ化し，$N_{\mathrm{p}}$ 個の句 $\boldsymbol{x}_{\mathrm{p}}^{(k)}(k=1,2,...,N_{\mathrm{p}})$ を得る．その後，CLIP のテキストエンコーダを用いて $\boldsymbol{x}_{\mathrm{I}'}$ および $\boldsymbol{x}_{\mathrm{p}}^{(k)}$ からそれぞれ言語特徴量 $\boldsymbol{h}_{\mathrm{I}'} \in \mathbb{R}^{768}$ および $\boldsymbol{h}_{\mathrm{p}}^{(k)} \in \mathbb{R}^{768}$ を獲得する．また，TP で得られた $\boldsymbol{x}_{\mathrm{targ}}$ から同様に言語特徴量 $\boldsymbol{h}_{\mathrm{targ}} \in \mathbb{R}^{768}$ を得るその後，指示文に対する特徴量 $\boldsymbol{h}_{\mathrm{inst}}$ を以下のように計算する．

$$\boldsymbol{h}_{\mathrm{inst}} = f_{\mathrm{FFN}}\left(\left[f_{\mathrm{trm}}\left(\left[\boldsymbol{h}_{\mathrm{targ}}; \boldsymbol{h}_{\mathrm{p}}^{(1)}; ...\boldsymbol{h}_{\mathrm{p}}^{(k)}\right]\right); \boldsymbol{h}_{\mathrm{I}'}; \boldsymbol{h}_{\mathrm{I}}\right]\right)$$

ここで，$f_{\mathrm{trm}}$ と $f_{\mathrm{FFN}}$ はそれぞれ $L_{\mathrm{trm}}$ 層の transformer 層 および順伝播型ネットワーク層を示す．

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

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

最後に，本モデルの出力は $s(\boldsymbol{x}_{\mathrm{I}}, \boldsymbol{x}_{\mathrm{t}}^{(n)}, \boldsymbol{x}_{\mathrm{c}}^{(n)})$ に基づく $T$ のランク付きリストを出力する．

各バッチにおける損失関数を以下の式に定義する．

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

ここで，$|\mathcal{B}|$ はバッチサイズを表す．

## 5. 実験

### 5.1 実験設定

本研究では，LTRPO タスクにおいて標準的な LTRRIE データセット [1] を用いた．訓練集合はモデルのパラメータ更新，検証集合はハイパーパラメータ調整のために使用し，テスト集合でモデルを評価した．提案手法における訓練可能パラメータ数は約 4.7 億である．また，積和演算数は $2.45 \times 10^{10}$ である．訓練には，メモリ 48GB 搭載の Quadro RTX 8000 および Intel Xeon Gold 6234 を使用した．モデルの学習には約 90 分かかり，また 1 つの指示文と 1 枚の対象物体領域の類似度の推論には約 25 ms かかった．各エポックごとに検証集合で MRR を計算し，最も高い 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 実験結果

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．ベースライン手法として，CLIP [2] を LTRPO タスクに拡張した手法および MultiRankIt [1] を用いた．また，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．ここで，MultiRankIt は LTRPO において良好な結果が報告されているため，MultiRankIt をベースライン手法とした．

モデルの評価指標として，MRR [11] および Recall@K を用いた．Recall@K は Recall@K $= \frac{1}{N_{\mathrm{inst}}} \sum_{i=1}^{N_{\mathrm{inst}}} \frac{|A_i \cap B_i|}{|A_i|}$ と定義される．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 [12].

表 1 より，ベースライン手法である (a), (b) および提案手法 (e) における MRR はそれぞれ 0.415, 0.501 および 0.579 であった．したがって，提案手法 (e) は MRR において，ベースライン手法 (a) および (b) をそれぞれ 0.164 と 0.078 ポイント上回った．同様に，提案手法は Recall@K (K<=20) においてもベースライン手法を上回った．また，すべての結果において統計的に有意であった．($p < 0.05$)

図 3 に “Go to the bedroom with the where curtains and bed spread Open the window to the left of the tallboy as wide as possible” に対する上位 3 件の検索結果を示す．$\boldsymbol{x}_{\mathrm{I}}$ は open/close タスクに分類され，$\boldsymbol{x}_{\mathrm{I}'}$ は “Open the window to the left of the tallboy in the bedroom with the curtains and bedspread”, $\boldsymbol{x}_{\mathrm{targ}}$ は “window” であった．ここで，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 Reciprocal


図 3　定性的結果．黄色の枠で囲まれた領域は $\boldsymbol{x}_{\mathrm{t}}$ を示す．

Rank (RR) を用いると，RR は 提案手法で 1.0，ベースライン (b) で 0.083 であった．このサンプルにおいて，$\boldsymbol{x}_{\mathrm{I}}$ は open/close タスクに不要な言葉を含む冗長な指示文であり，また文法的にも誤りを含むが，提案手法は適切な $\boldsymbol{x}_{\mathrm{t}}$ を第 1 位として検索することに成功した．この結果より，対象物体の句の抽出ならびに各タスクの標準形への言い換えを行う TP 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．(c) w/o TP：TP 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) w/o CSOE：CSOE 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 studies の定量的結果を示す．条件 (e) と比較して 条件 (c) における MRR は 0.052 ポイント減少した ．この結果より，TP の導入が最も性能に寄与したと考えられる．同様に，条件 (d) の結果より CSOE 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，家庭環境内の物体を操作する自然言語指示から対象物体を検索し，ランク付けリストを出力する LTRPO タスクを扱った．提案手法による貢献は以下の通りである．

- 大規模言語モデルを用いて冗長な情報を含む指示文からタスクを分類し，さらに各タスクの標準形に言い換える TP を提案した．
- セグメンテーションマスク重畳画像の特徴量を扱い，対象物体およびその周辺物体の関係をモデル化する CSOE を導入した．
- LTRRIE データセットにおいて，提案手法はベースライン手法 を MRR および Recall@K において上回った．

**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

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] Y. Chen, Z. Ma, Z. Zhang, Z. Qi, et al., “ViLEM: Visual-Language Error Modeling for Image-Text Retrieval,” CVPR, pp.11018–11027, 2023.

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

[10] S. Schuster, et al., “Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks,” LREC, pp.2371–2378, 2016.

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

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