二重緩和損失を用いたマルチモーダル検索に基づく生活支援ロボットによる物体操作

Open-Vocabulary Mobile Manipulation Based on Double Relaxed Contrastive Loss

八島 大地 Daichi Yashima　是方 諒介 Ryosuke Korekata　杉浦 孔明 Komei Sugiura

慶應義塾大学
Keio University

In this study, we develop a DSR that transports everyday objects to specified pieces of furniture based on open-vocabulary instructions. Our approach focuses on retrieving images of target objects and receptacles from pre-collected images of indoor environments. We propose a multimodal model, which learns diverse and robust representations from among positive, unlabeled positive, and negative samples. The experimental results demonstrate that our model outperformed existing baseline models across standard image retrieval metrics. Moreover, we performed physical experiments using a DSR to evaluate the performance of our approach in a zero-shot transfer setting. The experiments involved the DSR to carry objects to specific receptacles based on open-vocabulary instructions, achieving an overall success rate of 75%.

1. はじめに

労働力不足および少子高齢化が進行する現代社会において，物を運ぶことができ，人間の代わりに働く移動ロボットは，様々な場面で重要性が高まっている．こうした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より利便性が向上する．

本研究では，指定された家具に日常物体を運搬する生活支援ロボットを扱う．ロボットは，自由形式な指示文を使用して環境画像群の中から対象物体や配置目標の画像を検索する．例えば，“Please carry the apple on the chair to the table next to the scissors.” という指示文が与えられた場合を想定する．図 1 に示された環境において事前に収集された環境画像群の中から対象物体および配置目標としてそれぞれ「椅子の上にあるりんご」および「横にハサミが置かれた机」を上位にランク付けすることが望ましい．そして検索された画像リストの中からユーザが選択した画像を基に，ロボットが対象物体を配置目標に運搬することが期待される．

本タスクは，多数の類似物体が存在する環境画像群から，複雑な参照表現を含む自由形式な指示文に基づき対象物体や配置目標を特定する点が困難である．実際に 4.1 節で示すように，CLIP [Radford 21] などの代表的な基盤モデルを本タスクに直接適用するだけでは不十分である．

物体操作指示文に基づき，環境内の物体から対象物体を識別する研究は広く行われている [Korekata 23, Kaneda 24, Sigurdsson 23]．MultiRankIt [Kaneda 24] および RREx-BoT [Sigurdsson 23] は，屋内環境でのマルチモーダル検索に取り組んでいる．本研究は [Korekata 25] と同様に画像検索設定において，単一の指示文で対象物体および配置目標の両方を扱う．近年，マルチモーダル表現モデルはクロスモーダル検索の性能を向上させている．これらの既存手法 [Korekata 25, Kaneda 24, Radford 21] は主に InfoNCE [Oord 18] を損失関数として利用している．これらの既存手法の多くは，positive 画像との類似性が高く部分的に正しいとみなされる unlabeled positive がバッチ内に存在するとき，これらを negative としてみなすため不適切な場合がある．このように，類似画像に厳密なアノテーションが付与されない背景として，アノテーションに伴う労働力的，時間的コストなどの制約がある．本研究では，unlabeled positive ラベルを活用し，新たな損失関数を組み込んだ手法を提案する．本研究の貢献は以下である．

連絡先: 八島大地，慶應義塾大学，神奈川県横浜市港北区日吉 3-14-1，ydaichi1207@keio.jp

図 1: 実機実験環境

- Dense Representation Learning (DRL) モジュールを提案する．本モジュールは，Dense Labeler を用いて positive 画像に類似した画像に unlabeled positive ラベルを付与し，Double Relaxed Contrastive (DRC) 損失関数を用いて positive, unlabeled positive, および negative ペア間の関係を最適化する．
- 提案手法をゼロショット転移設定における性能を検証するため，ロボットを用いた実機実験を実施し，既存手法を上回る結果が得られた．

2. 問題設定

本研究では，Image Retrieval-based Open-Vocabulary Fetch-and-Carry (IROV-FC) タスク [Korekata 25] を扱う．本タスクではロボットが自由形式な指示文に基づき，対象物体および配置目標の画像を検索し，対象物体を配置目標へ運搬する．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，対象物体および配置目標の画像が，それぞれの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．対象物体および配置目標は検索された画像リストの中からユーザによって選択される．

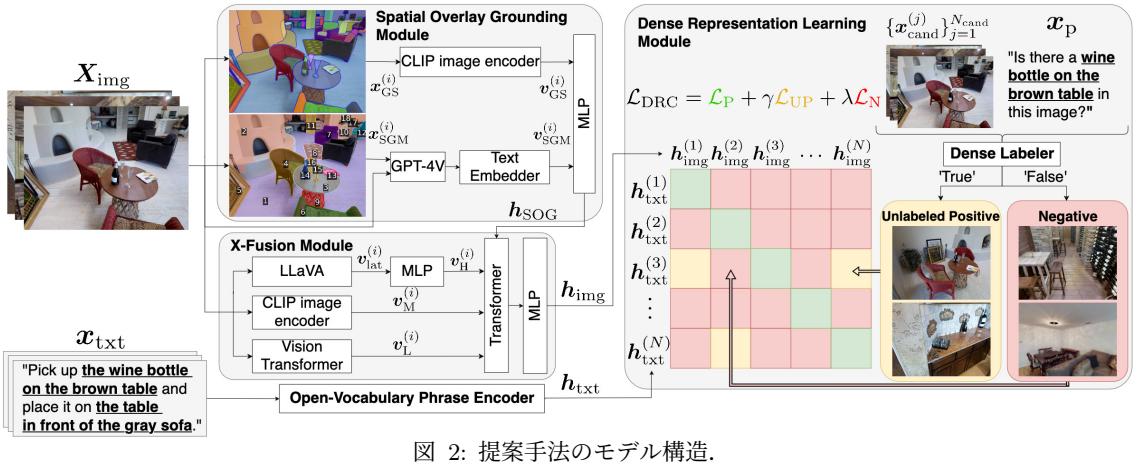

図 2: 提案手法のモデル構造.

動作実行においては，ロボットは対象物体を把持し，それを配置目標まで運搬することが求められる．

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，Spatial Overlay Grounding (SOG)，X-Fusion (XF)，および Dense Representation Learning (DRC) という主に 3 つのモジュールから構成される．

3.1 入力

本モデルへの入力を，$\boldsymbol{x} = \{\boldsymbol{x}_{\text{txt}}, X_{\text{img}}\}, X_{\text{img}} = \{\boldsymbol{x}_{\text{img}}^{(i)}\}_{i=1}^{N_{\text{img}}}$ と定義する．ここで，$\boldsymbol{x}_{\text{txt}} \in \{0,1\}^{V\times L}$，$V$，$L$，$\boldsymbol{x}_{\text{img}} \in \mathbb{R}^{3\times W\times H}$，$W$，$H$，および N_{img} は，それぞれトークナイズされた指示文，語彙サイズ，最大トークン長，画像，画像の幅，画像の高さ，およびランク付けされる画像数である．

3.2 SOG

SOG モジュールでは，2 つの並列入力を通じて視覚特徴量を取得する．一方の入力では，領域分割マスクを重畳した画像に multimodal encoder を用いる．他方では，領域分割マスクを重畳した画像の領域ごとにマーキングを付け，MLLM を用いる．既存手法の多くは，視覚的特徴を $\boldsymbol{x}_{\text{img}}^{(i)}$ から全体的に，もしくは単一の物体に焦点を当てて取得することが多い．その結果，物体の誤認識などが生じる可能性がある．一方，提案手法では，領域分割のための基盤モデル (例： [Kirillov 23, Zou 23]) を用いて，特徴抽出を行う．輪郭，色，および物体間の位置関係に関する補助的な情報を与えることで，視覚的な誤りを減少させる．詳細については [Yashima 25] を参照されたい．

3.3 XF

XF モジュールでは，visual encoder (例：ViT [Dosovitskiy 21])，multimodal encoder (例：CLIP [Radford 21])，および潜在特徴量が得られる MLLM (例：LLaVA [Liu 23]) から包括的に 3 種類の視覚特徴量を取得する．これらの埋め込み表現には，次のような特徴がある．visual encoder は，物体の色，テクスチャ，および形状などの特徴を取得するが，複雑な参照関係を扱うことができない．また，multimodal encoder は言語と視覚がアラインされた埋め込みを抽出する一方，構造化された情報を取得することが難しい．他方，潜在特徴量が得られる MLLM は，トークナイザを利用する必要がなく，LLM および vision encoder からそれぞれ取得された特徴を組み合わせるため，埋め込みを通じて言語特徴と視覚特徴の両方を持った構造的な特徴を取得することができる．したがって，これらの 3 種類の埋め込み表現および SOG モジュール視覚的特徴を並列して使用することで，それぞれの補完的な強みを活用できると考えられる．詳細については [Yashima 25] を参照されたい．

3.4 SPAC

SPAC モジュールは，DRC 損失関数を用いて，positive ペアのコサイン類似度を最大化させ，unlabeled positive および negative ペアの対照性を緩和しつつモデルを最適化する．近年のマルチモーダル事前学習手法は主に InfoNCE [Oord 18] を損失関数として利用している [Radford 21]．しかしながら，データセット内に類似画像が含まれる状況では，これらは negative ではなく，unlabeled positive として扱われることが望ましい．したがって，unlabeled positive ラベルの付与およびそれらを適切に最適化可能な損失関数の導入が重要である．本問題に対処するため，DRC 損失関数を $\mathcal{L}_{\text{DRC}} = \mathcal{L}_{\text{P}} + \gamma\mathcal{L}_{\text{UP}} + \lambda\mathcal{L}_{\text{N}}$ と定義する．ここで，γ および λ は重み係数である．また，$\mathcal{L}_{\text{DRC}}$ を構成する各項はそれぞれ positive，unlabeled positive，および negative ペアに対する損失であり，以下のように定義する．

$$\mathcal{L}_{\text{P}} = \sum_{i} \left(1 - \text{sim}\left(\boldsymbol{h}_{\text{txt}}^{(i)}, \boldsymbol{h}_{\text{img}}^{(i)}\right)\right)^2$$

$$\mathcal{L}_{\text{UP}} = \sum_{(i,j)\in\mathcal{S}} \max\left(\alpha - \text{sim}\left(\boldsymbol{h}_{\text{txt}}^{(i)}, \boldsymbol{h}_{\text{img}}^{(j)}\right), 0\right)^2$$

$$\mathcal{L}_{\text{N}} = \sum_{(i,j)\notin\mathcal{S}} \max\left(\text{sim}\left(\boldsymbol{h}_{\text{txt}}^{(i)}, \boldsymbol{h}_{\text{img}}^{(j)}\right), 0\right)^2$$

ここで，$\text{sim}(\cdot,\cdot)$，$\mathcal{S}$，および α は，それぞれコサイン類似度，unlabeled positive に対応する添字集合，および unlabeled positive ペアのコサイン類似度の閾値である．

本モジュールの入力は $\boldsymbol{h}_{\text{txt}} \in \mathbb{R}^{d_{\text{txt}}}$ および $\boldsymbol{h}_{\text{img}} \in \mathbb{R}^{d_{\text{img}}}$ である．ここで $\boldsymbol{h}_{\text{txt}}$ および d_{txt} はそれぞれ OVP encoder の出力および出力次元数であり，$\boldsymbol{h}_{\text{img}}$ および d_{img} はそれぞれ XF module の出力および出力次元数である．OVP encoder は [Korekata 25] に準拠する text encoder であり，対象物体および配置目標の両方を含む自由形式な指示文を処理し，予測の対象に応じた言語特徴量 $\boldsymbol{h}_{\text{txt}}$ を出力する．また，$\mathcal{S}$ は $\mathcal{S} = \text{Dense Labeler}(\boldsymbol{x}_{\text{p}}, \{\boldsymbol{x}_{\text{img}}^{(1)}, \ldots, \boldsymbol{x}_{\text{img}}^{(N_{\text{cand}})}\})$ で得られる．ここで $\boldsymbol{x}_{\text{p}}$ は対象物体または配置目標が画像中に存在するか否かを判定するために使用されるプロンプトを表す．まず，画像検索タスクで利用される既存の事前学習モデル (例：[Kaneda 24, Radford 21]) を用いて，すべての指示文と画像間における類似度スコア $-1 \leq \text{sim}(\boldsymbol{h}_{\text{txt}}^{(i)}, \boldsymbol{h}_{\text{img}}^{(j)}) \leq 1$ を取得する．各

表 1: ベースライン手法，提案手法，および Ablation study における定量的結果

[%]	手法	損失関数		HM3D-FC			MP3D-FC		
		InfoNCE [Oord 18]	DRC	R@5↑	R@10↑	R@20↑	R@5↑	R@10↑	R@20↑
(i-a)	MultiRankIt [Kaneda 24]	✓		28.7 ±3.4	48.3 ±3.4	73.3 ±2.6	35.7 ±9.9	51.7 ±8.9	72.7 ±3.3
(i-b)			✓	38.4 ±2.3	58.3 ±1.9	81.3 ±1.1	39.1 ±2.7	57.6 ±0.9	73.5 ±0.8
(ii-a)	DM^{2}RM [Korekata 25]	✓		47.8 ±1.2	67.1 ±2.4	87.0 ±1.1	49.6 ±0.7	64.1 ±3.6	78.5 ±0.5
(ii-b)			✓	50.2 ±1.0	69.0 ±2.1	87.3 ±1.4	53.1 ±1.8	66.8 ±1.1	78.7 ±1.2
(iii-a)	提案手法	✓		48.8 ±0.9	70.9 ±0.5	91.5 ±0.5	54.8 ±0.8	69.5 ±0.8	81.8 ±1.1
(iii-b)			✓	**55.4** ±0.5	**76.3** ±0.9	**91.6** ±0.9	**57.0** ±1.1	**72.4** ±0.7	**82.5** ±0.8

$\boldsymbol{x}_{\text{txt}}$ について，上位 N_{cand} 枚の画像を選択し MLLM に入力する．出力されたテキストで 'True' とされたものに unlabeled positive ラベルを付与し，それらの添字集合 $(i,j) \in \mathcal{S}$ を得る．

次に，Dense Labeler で得られた $\mathcal{S}$ を用いて，$\mathcal{L}_{\text{UP}}$ は $\text{sim}(\boldsymbol{h}_{\text{txt}}^{(i)}, \boldsymbol{h}_{\text{img}}^{(j)})$ が α 未満であるペア $(i,j) \in \mathcal{S}$ に対してペナルティを与える．ここで，max 関数は，類似度が α 未満のペアのみを損失に寄与させ，α を超えるペアを無視することで対照性を緩和する．同様のことが $\mathcal{L}_{\text{N}}$ にも適用され，negative ペアである $(i,j) \notin \mathcal{S}$ に対し，コサイン類似度の二乗和を用いてペナルティを課している．これらの要素を組み込むことで，$\mathcal{L}_{\text{DRC}}$ 損失関数は positive, unlabeled positive, および negative ペアの寄与を調整することにより，多様な表現を学習することが期待される．

提案する DRC 損失関数および Dense Labeler は，他のクロスモーダル検索タスク (例：[Wu 21]) にも適用可能であり，類似した画像がデータセットに含まれる場合に有効である．推論時のモデルの出力は，類似度スコア $\text{sim}(\boldsymbol{x}_{\text{txt}}^{(i)}, \boldsymbol{x}_{\text{img}}^{(j)})$ を基に X_{img} の画像を降順に並び替えた 2 つの画像リスト $\hat{Y}_{\text{targ}}$ および $\hat{Y}_{\text{rec}}$ である．$\hat{Y}_{\text{targ}}$ および $\hat{Y}_{\text{rec}}$ は，それぞれ対象物体および配置目標に関する画像リストである．

4. シミュレーション実験

本研究では LTRRIE-FC データセット [Korekata 25] を使用した．本データセットは，HM3D [Ramakrishnan 21] および MP3D [Chang 17] における様々な屋内環境から収集された画像から構築され，人間がアノテーションした自然言語指示文を含む．データセットの詳細は [Korekata 25] を参照されたい．訓練時における最適化手法，バッチサイズ，およびエポック数は AdamW，128，および 20 であった．提案手法の訓練可能なパラメータ数および積和演算数は 201M および 329G であった．訓練は，NVIDIA GeForce RTX4090 および Intel Corei9-13900FK，64 GB の RAM を搭載した計算機上で行った．訓練には約 3 時間，推論時における 1 つの指示文と 100 枚の画像間の計算には約 79 ms を要した．各エポック終了時に検証集合に対して Recall@10 を計算し，その値が最大となったモデルで，テスト集合における評価を行った．

4.1 定量的結果

表 1 に，LTRRIE データセット [Korekata 25] に含まれる HM3D-FC および MP3D-FC のテスト集合における定量的結果を示す．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，最良のスコアを太字で示す．本研究では，MultiRankIt [Kaneda 24]，および DM^{2}RM [Korekata 25] をベースライン手法とした．MultiRankIt および DM^{2}RM は本タスクと関連が深いタスク [Kaneda 24,Korekata 25] において良好な結果を得ているため選択した．ここで，MultiRankIt は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，それらの平均を結果として用いた．評価指標は Recall@K $(K = 5, 10, 20)$ とした．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採用した．本研究では，Recall@10 を主要評価指標とした．表 1 より，HM3D-FC および MP3D-FC テスト集合において提案手法 (iii-b) の Recall@10 がそれぞれ 76.3%および 72.4%であり，ベースライン手法のうち最良の (ii-a) に対してそれぞれ 9.2 および 8.3 ポイント上回った．さらに，提案手法 (iii-b) はすべての評価指標においてベースライン手法群を上回った．これらの性能差はすべて統計有意であった ($p < 0.01$)．

4.2 ablation study

表 1 に ablation study における結果を示す．Ablation study にあたっては，DRC 損失関数を InfoNCE 損失関数に変更する条件を用意し，DRL モジュールにおける DRC 損失関数の有用性を検証した．モデル (iii-a) およびモデル (iii-b) を比較した結果，HM3D-FC および MP3D-FC テスト集合において，それぞれ recall@10 が 5.4 および 2.9 ポイント減少した．これにより，本手法において unlabeled positive を付与して，DRC 損失関数を用いることが有用であると示唆される．

さらに，DRC 損失関数および unlabeled positive の有用性を検証するために，ベースライン手法である MultiRankIt および DM^{2}RM に DRC 損失関数を適用した．表 1 より，どちらの場合においても，ベースライン手法と比較して，HM3D-FC および MP3D-FC テスト集合における全ての評価指標において上回った．このことから DRC 損失関数は他の手法に対しても広く適用可能であることが示唆される．

5. 実機実験

5.1 環境設定および実装

実機実験の設定は DM^{2}RM [Korekata 25] に従う．図 1 に，実機環境を示す．実験環境は $4.0 \times 6.0\text{m}^2$ で，特定のレイアウトに配置された 9 つの家具から構成された環境である [wrs 20]．トヨタ自動車製の Human Support Robot [Yamamoto 19] を使用した．また，YCB Object Set [Calli 15] から日常的な 30 種類の物体を使用した．

本実験は，20 種類の異なる環境設定で，各環境設定ごとに 5 回ずつ実施し，合計で 100 回の試行を行った．各環境において，15-20 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．

環境の画像を事前収集するため，ロボットは事前に定められた 17 点の地点に移動し，環境の RGB-D 画像を Asus Xtion Pro カメラを用いて収集した．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た．

ユーザから指示文を受け取ったロボットは，事前動作にて収集された画像から対象物体および配置目標を検索した．ここで，モデルは LTRRIE-FC データセットで訓練されたゼロショット転移設定において検証された．対象物体および配置目標それぞれの上位 K(=10) 枚の画像を WebUI によりユーザに提示し，ユーザは適切な画像を選択した．対象物体および配置目標として適切な画像が上位 K 位に含まれない場合，その試行は失敗とみなし，ロボットは物体操作を行わないものとし，次に，

表 2: 実機実験における定量的結果

	手法	SR↑ [%]
(i)	MultiRankIt [Kaneda 24]	53 (53 / 100)
(ii)	DM^{2}RM [Korekata 25]	68 (68 / 100)
(iii)	提案手法	**75** (75 / 100)

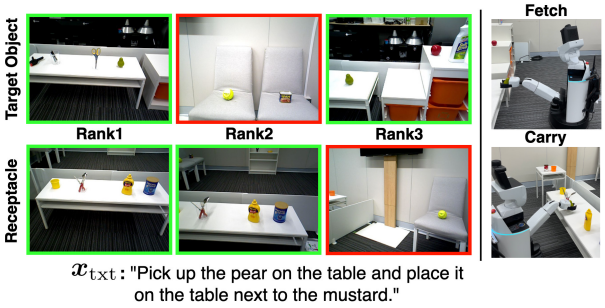

図 3: 実機実験における提案手法の定性的結果

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，把持点はユーザが与えた point prompt に SAM [Kirillov 23] を適用して得られたセグメンテーションマスクおよび，Depth 画像から得られた点群に基づき決定された．最後に，ロボットは配置目標画像が上位 K 位に含まれ，かつ把持動作が成功していた場合のみ配置動作を行った．

5.2 実験結果

表 2 に，実機実験におけるベースライン手法 [Kaneda 24, Korekata 25] および提案手法の定量的結果を示す．本実験においては，SR $= N_s/N_a$ 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．表 2 より，提案手法 (iii) の SR は 75%であり，手法 (ii) を 7 ポイント上回った．図 3 は，実機実験における提案手法の成功例を示す．$\boldsymbol{x}_{txt}$ は “Pick up the pear on the table and place it on the table next to the mustard.” であった．この時，提案手法は対象物体として GT 画像を 1 位にランク付けしている．同様に，配置目標として 1 位および 2 位に GT 画像をランク付けしている．これらの画像は同一の配置目標を異なる角度から撮影されたものであるため，どちらも正解とみなされる．その後，ロボットは対象物体である梨を正しく把持し，適切なテーブルへと配置した．

6. おわりに

本研究では，ロボットが自由形式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 IROV-FC タスク [Korekata 25] を扱った．本研究では，Dense Labeler を用いて unlabeled positive を付与し，positive, unlabeled positive, および negative ペア間の関係を最適化する二重緩和損失を導入した．提案手法は，ロボットを用いた実機実験において，ゼロショット転移設定のもとで頑健な性能を示し，既存手法を上回る結果を得た．

謝辞

本研究の一部は，JSPS 科研費 23K03478，JST ムーンショット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., Singh, A., Srinivasa, S., 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 17] Chang, A., Dai, A., Funkhouser, T., et al.: Matterport3D: Learning from RGB-D Data in Indoor Environments, in *3DV*, pp. 667–676 (2017)

[Dosovitskiy 21] Dosovitskiy, A., et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in *ICLR*, pp. 12888–12900 (2021)

[Kaneda 24] Kaneda, K., Nagashima, S., Korekata, R., 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., Rolland, C., Gustafson, L., Xiao, T., et al.: Segment Anything, in *ICCV*, pp. 4015–4026 (2023)

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

[Korekata 25] Korekata, R., Kaneda, K., et al.: DM2RM: Dual-Mode Multimodal Ranking for Target Objects and Receptacles Based on Open-Vocabulary Instructions, *AR* (2025)

[Liu 23] Liu, H., Li, C., Wu, Q., and Lee, J.: Visual Instruction Tuning, in *NeurIPS*, pp. 34892–34916 (2023)

[Oord 18] Oord, A., Li, Y., and Vinyals, O.: Representation Learning with Contrastive Predictive Coding, *arXiv preprint arXiv:1807.03748* (2018)

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

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

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

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

[Wu 21] Wu, H., et al.: Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback, in *CVPR*, pp. 11307–11317 (2021)

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

[Yashima 25] Yashima, D., Korekata, R., and Sugiura, K.: Open-Vocabulary Mobile Manipulation Based on Double Relaxed Contrastive Learning With Dense Labeling, *IEEE RA-L*, Vol. 10, No. 2, pp. 1728–1735 (2025)

[Zou 23] Zou, X., Yang, J., Zhang, H., Li, F., Li, L., et al.: Segment Everything Everywhere All at Once, in *NeurIPS*, pp. 19769–19782 (2023)