マルチモーダルLLMおよび視覚言語基盤モデルに基づく大規模物体操作データセットにおけるタスク成功判定

Task Success Prediction on Large-Scale Object Manipulation Datasets Based on Multimodal LLMs and Vision-Language Foundation Models

齋藤 大地 Daichi Saito　神原 元就 Motonari Kambara　九曜 克之 Katsuyuki Kuyo　杉浦 孔明 Komei Sugiura

慶應義塾大学
Keio University

For enhancing model performance in object manipulation tasks, high-performance prediction mechanisms for task success are crucial. However, existing methods are still insufficient in performance. Moreover, existing prediction mechanisms are designed to address only specific tasks, making it challenging to accommodate a diverse range of tasks. Therefore, our study aims to develop a task success prediction mechanism that can handle multiple object manipulation tasks. A key novelty of the proposed method is the introduction of λ-Representation, which preserves all types of visual features: visual charactaristics such as colors and shapes; features aligned with natural language; features structured through natural language. For the experiments, we newly built datasets for task success prediction in object manipulation tasks based on the RT-1 dataset and VLMbench. The results show that the proposed method outperforms all baseline methods in accuracy.

1. はじめに

高齢化が進む現代社会において，在宅介助者不足が深刻な社会問題となっているおり，生活支援ロボットはその解決策として期待されている．生活支援ロボットにとって，物体操作タスクは必要不可欠である．また，物体操作におけるモデル性能向上のためには，高性能なタスク成功判定機構が重要であるが，その性能は現状不十分である．

本研究では，指示文および物体操作前後の画像に基づいて，物体操作の成功および失敗を正しく予測することを目的とする．また，本タスクを Success Prediction for Object Manipulation (SPOM) タスクと定義する．

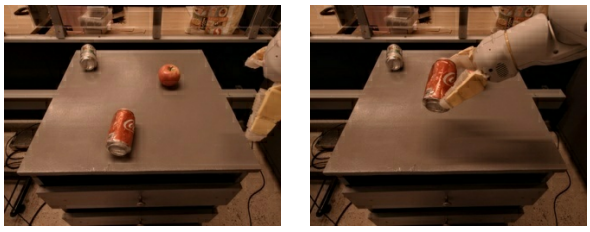
本研究では，物体操作に対するタスク成功判定を行う手法を提案する．提案手法は，[Obinata 23] をはじめとするタスク成功判定を扱う既存手法と関連が深い．我々の手法は，λ-Representation を導入することにより指示文に含まれる複雑な参照表現に対する理解を強化する点で，それらの既存手法と異なる．

本研究の新規性は以下のとおりである．

- 色や形状などの視覚的な特徴量，自然言語にアラインされた特徴量，および言語を媒介として構造化された特徴量の 3 種類の潜在表現を保持した λ-Representation を提案する．
- 指示文および λ-Representation に対して cross-attention を計算する λ-Representation Encoder を導入する．

2. 問題設定

本研究で扱う SPOM タスクは，指示文およびロボットによる物体操作前後の画像が与えられたとき，物体操作の成功および失敗に関する 2 値分類を行うものである．本タスクでは，ロボットによる物体操作の成否を正しく予測することが望ましい．図 1 に本タスクの具体例を示す．この例では，“pick coke can”という指示文とともに，図 1 に示す物体操作前後の画像が与えられている．この例の場合，ロボットが指示文通りに物体操作を実行しているため，成功と予測することが期待される．入力は，指示文およびロボットによる物体操作前後の画像である．SPOM タスクにおいて求められる出力は，ロボットが指示文に従った物体操作に成功した確率の予測値 $P(\hat{y}=1)$ である．本研究では，一人称視点の画像を入力として扱うことを前提とする．そのため，本タスクでは，視界の一部がアームによって遮られる可能性がある．また，本タスクの評価尺度には，分類精度を使用する．



(a) “pick coke can” (b)

図 1: SPOM タスクの具体例.

ここで，実世界に適用できることは，実用上重要である．そこで本研究では，RT-1 データセット [Brohan 22] をもとに実環境データセットを構築し利用した．一方，実環境においては，環境の完全な再現が困難である．したがって，本研究において利用した実環境データセットにおける性能を実環境における物体操作実行時に再現することは困難である．しかるに，シミュレーション環境では環境の再現が容易であるため，モデルの再現性を担保可能である．以上より，本研究では，実環境データセットおよびシミュレーションデータセットを両方使用する．

3. 提案手法

提案手法における主要モジュールは，λ-Representation Encoder である．図 2 に，提案手法のネットワーク構造を示す．

ここで，モデルの入力を $\boldsymbol{x} = \{\boldsymbol{x}_{\text{inst}}, \boldsymbol{x}_{\text{before}}, \boldsymbol{x}_{\text{after}}\}$ と定義する．この式において，$\boldsymbol{x}_{\text{inst}} \in 1, 0^{V \times L}$ は指示文，$\boldsymbol{x}_{\text{before}} \in \mathbb{R}^{3 \times H \times W}$ および $\boldsymbol{x}_{\text{after}} \in \mathbb{R}^{3 \times H \times W}$ はそれぞれ物体操作前および物体操作後の画像を表す．また，V は語彙サイズ，L は最大トークン長，H, W はそれぞれ画像の高さと幅を表す．

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

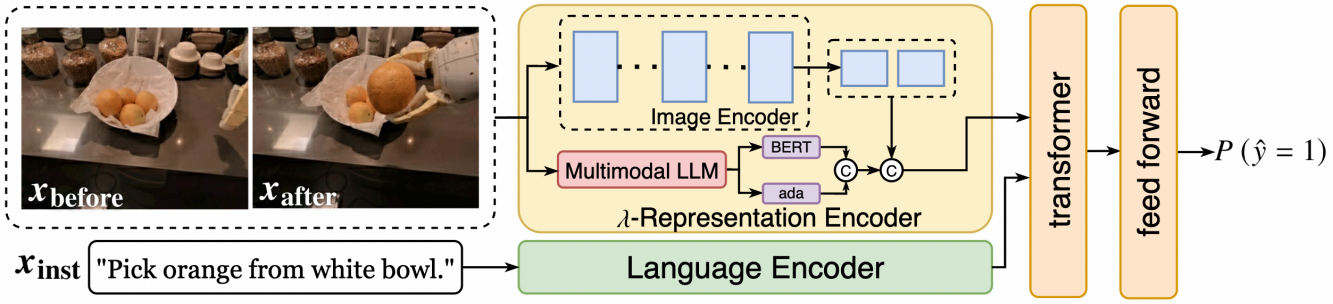


図 2: 提案手法のネットワーク構造.

3.1 Language Encoder

$\boldsymbol{x}_{\text{inst}}$ について，Language Encoder を用いて言語特徴量を抽出する．Language Encoder で行われる処理は以下である．まず，BERT [Devlin 18] を使用して $\boldsymbol{x}_{inst}$ から言語特徴量を抽出し，その CLS トークンにあたる特徴量 $I_{\text{BERT}} \in \mathbb{R}^{d_{\text{BERT}}}$ を得る．また，CLIP テキストエンコーダ [Radford 21] を使用して $\boldsymbol{x}_{\text{inst}}$ から言語特徴量 $\boldsymbol{l}_{\text{CLIP}} \in \mathbb{R}^{d_{\text{CLIP}}}$ を得る．同様に，text-embedding-ada-002 [ada] を用いて $\boldsymbol{x}_{\text{inst}}$ から $\boldsymbol{l}_{\text{ada}} \in \mathbb{R}^{d_{\text{ada}}}$ を得る．ここで，$d_{\text{BERT}}, d_{\text{CLIP}}, d_{\text{ada}}$ はそれぞれ BERT, CLIP テキストエンコーダ，text-embedding-ada-002 の出力次元数である．最後に，それぞれの特徴量を結合し，言語特徴量 $\boldsymbol{h}_l = \{\boldsymbol{l}_{\text{BERT}}, \boldsymbol{l}_{\text{CLIP}}, \boldsymbol{l}_{\text{ada}}\}$ を得る．

3.2 λ-Representation Encoder

既存の Vision and Language Understanding(VLU) モデルには，画像特徴量抽出において主に 3 つの手法がある．一つ目は，ResNet や ViT をはじめとした画像エンコーダを用いて，テクスチャやエッジなどの視覚的な特徴量を抽出する方法である [Korekata 23]．ここで，それらの視覚表現を Scene Representation と定義する．二つ目は，マルチモーダル LLM やキャプショニングモデルを用いることで，言語を媒介として複雑な参照関係や位置関係を直接的に埋め込んだ構造的な特徴量を抽出する方法である．ここで，これらの視覚表現を Narrative Representation と定義する．三つ目は，CLIP 画像エンコーダなどのマルチモーダル画像エンコーダを用いて，自然言語にアラインされた特徴量を抽出する方法である [Kaneda 24]．ここで，それらを Aligned Representation と定義する．

しかし，多くの既存モデルでは，これらの全てを並列に扱っていないため，画像特徴量の表現力が不十分である．例えば，Scene Representation のみを用いた場合，画像に含まれる形状や色などの視覚的な情報は抽出可能である．しかし，位置関係等の複雑な参照関係を抽出し利用することは困難である．また，Narrative Representation のみを用いた場合，言語を媒介とすることで，構造的な特徴量を抽出可能である．しかし，テクスチャを含む詳細な視覚的な特徴を全て記述することは同様に困難である．これらの特徴量と異なり，Aligned Representation は，自然言語にアラインされた画像特徴量であり，その点で Scene Representation および Narrative Representation とそれぞれと類似した性質を持つ．しかし，自然言語を媒介とした特徴抽出を行っておらず，ゆえに指示文中の複雑な参照表現を構造的に理解する性能は不十分である．

以上より，これらの特徴量は，全てを並列的に用いることで，十分な視覚表現が獲得できることが期待される．そこで，提案手法では，Scene Representation, Aligned Representation, Narrative Representation の 3 つを全て並列に用いた λ-Representation を提案する．また，λ-Representation を抽出する λ-Representation Encoder を導入する．本モジュールでは，3 種類の視覚表現を抽出し重要度に応じて重み付けを行い，最終的に λ-Representation として結合する．λ-Representation Encoder では，CLIP の中間出力及び最終出力を用いることで，それぞれ Scene Representation 及び Aligned Representation を獲得する．また，Multimodal LLM による長文キャプションを用いて，Narrative Representation を獲得する．

本モジュールにおける入力は，$\boldsymbol{x}_{\text{before}}$ および $\boldsymbol{x}_{\text{after}}$ である．以降，$\boldsymbol{x}_{\text{after}}$ は $\boldsymbol{x}_{\text{before}}$ と同様の処理が行われるため，$\boldsymbol{x}_{\text{before}}$ についてのみ説明する．本モジュールでは，まず，Multi-resolution 特徴量 $\boldsymbol{h}_m = \{\boldsymbol{h}_a, \boldsymbol{h}_s\}$ を抽出する．ここで，$\boldsymbol{h}_a \in \mathrm{R}^{\mathrm{d_{CLIP}}}$ は CLIP 画像エンコーダの最終出力を表し，$\boldsymbol{h}_s \in \mathbb{R}^{C_s \times H_s \times W_s}$ は CLIP 画像エンコーダの中間出力に対して畳み込みを適用することで得られる．ただし，C_s, H_s, W_s はそれぞれ CLIP 画像エンコーダの中間出力におけるチャネル数，高さ，幅を表す．Multi-resolution 特徴量は，複数の解像度で得られた画像特徴量である．従って，CLIP 画像エンコーダを基に獲得した Multi-resolution 特徴量を導入することで，自然言語によくアラインされた特徴量および色や形などの視覚特徴を保持した特徴量を同時に扱うことが可能となる．

また，上述の定義から，$\boldsymbol{h}_a$ および $\boldsymbol{h}_s$ はそれぞれ Aligned Representation と Scene Representation を表す．次に，InstructBLIP [Dai 23] を使用して，$\boldsymbol{x}_{\text{before}}$ から Narrative Representation を抽出する．その後，BERT および text-embedding-ada-002 を並列に使用して得られた特徴量を結合し，$\boldsymbol{h}_n$ を得る．続いて，物体操作前の画像に対する λ-Representation である $\boldsymbol{\lambda}_{\text{before}} = \{\boldsymbol{h}_s, \boldsymbol{h}_a, \boldsymbol{h}_n\}$ を得る．本モジュールにおける最終出力 $P(\hat{y} = 1)$ は，マニピュレータが物体操作を適切に実行した確率の予測値を示す．ただし，$\hat{y}$ は予測ラベルであり，成功を 1，失敗を 0 とする．まず，n_{enc} 層の transformer encoder $f_e(\cdot)$ を用いて，特徴量 $\boldsymbol{h}_v = f_e(\boldsymbol{\lambda}_{\text{before}}, \boldsymbol{\lambda}_{\text{after}})$ を得る．その後，n_{dec} 層の transformer decoder $f_d(\cdot)$ と feed forward network $f_f(\cdot)$ を利用して，$P(y=1) = f_f(f_d(\boldsymbol{h}_v, \boldsymbol{h}_l))$ を得る．なお，損失関数には交差エントロピー誤差を使用する．

4. 実験設定

本研究では，SPOM タスクのために新たに SP-RT-1 データセットと SP-VLMbench データセットを構築した．SP-RT-1 データセットは，RT-1 データセットをもとに構築され，指示文，物体操作前後の画像，およびラベルを含む．RT-1 データセットは, 実世界の物体操作のための標準的な大規模データセットである.RT-1 データセットには，指示文, 操作中に撮影された画像, そして人間によってラベル付けされた二値の reward が含まれている．SPOM タスクのために, 各エピソードの最初と最後の画像を収集した. さらに，VLMbench シミュレーション環境 [Zheng 22] を使用して SP-VLMbench データセットを構築した．SP-RT-1 データセットと同様に，SP-VLMbench データセットには，指示文と操作前後の画像が含まれている．VLMbench は，物体操作タスクのための標準的なベンチマー

表 1: ベースライン手法との比較および Ablation Study の定量的結果.

Model	SR	NR	Accuracy [%]	
			SP-RT-1	SP-VLMbench
InstructBLIP [Dai 23]			52.30 ± 0.74	41.30 ± 0.87
Gemini [GeminiTeam 23]			64.12 ± 1.21	56.15 ± 2.69
GPT-4V [Achiam 23]			69.12 ± 0.78	57.81 ± 0.79
UNITER [Chen 20]			69.08 ± 1.77	68.22 ± 0.90
Ours (i)		✓	73.30 ± 1.27	75.80 ± 0.62
Ours (ii)	✓		72.02 ± 1.55	73.74 ± 1.84
Ours (iii)	✓	✓	**74.50 ± 1.44**	**78.92 ± 0.68**

クであり、自然言語の指示、各操作の成功または失敗を示すラベル、および 5 つのカメラ視点 (前方，頭部，右肩，左肩，手首) から撮影された画像を取得することができる．VLMbench 環境では，物体がテーブル上にランダムに配置され，それに合わせてテンプレートから指示文を自動的に生成する．本研究では実機への応用を考慮するため，ロボットに取り付けられたカメラのみを利用した．すなわち，前方に設置されたカメラを除く 4 視点のカメラを用いて物体操作前後の画像を収集した．

本研究では，正例における指示文を変更することで負例を追加した．VLMbench の指示文において，対象物を特定するために物体の色や形，大きさ，位置などに関する表現が用いられている．そこで，正例の指示文において，対象物を特定する表現をその他のサンプルで使用されている表現にランダムに変更することで，データ拡張を行った．例えば，ある正例の指示文が “Pick up the cube and place it into the black container.” である場合に，“black” を “red” に変更することで新たに負例を作成した．SP-RT-1 データセットは、合計 13,915 サンプルで構成され、語彙サイズは 49、総単語数は 78,790、平均文長は 5.66 である．さらに、SP-RT-1 データセットは、合計 8,326 サンプルを含み、語彙サイズは 54、総単語数は 67,233、平均文長は 8.79 である．本研究では，訓練集合および検証集合をそれぞれパラメータの推定およびハイパーパラメータの選択に使用した．また，テスト集合を性能の評価に使用した．提案手法における訓練可能パラメータ数は約 48M であり，総積和演算数は約 7.1G であった．SP-RT-1 データセットにおけるモデルの訓練時間および 1 サンプルあたりの推論時間は，それぞれ約 34 分および約 4.8ms であった．また，SP-VLMbench データセットでは，それぞれ約 36 分および約 4.8ms であった．なお，各エポックごとに検証集合で精度を計算し，最も高い精度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. 実験結果

5.1 定量的結果

表 1 にベースライン手法と提案手法との比較に関する定量的結果を示す．数値は，それぞれ 5 回実験における平均値および標準偏差を表す．表中の太字は最大値を表す．また，SR，NR はそれぞれ Scene Representation，Narrative Representation を表す．ベースライン手法には，InstructBLIP, Gemini [GeminiTeam 23], GPT-4V [Achiam 23]，および UNITER を用いた．なお，Instruct BLIP, Gemini, および GPT-4V は zero-shot で使用した．以下の理由により，各手法をベースライン手法として用いた．UNITER については，VQA を含む多くの Vision-and-Language タスクにおいて良好な結果が得られている手法であるため利用した．また，GPT-4V および Gemini は，非常に大規模なデータセットを用いて事前訓練された代表的な大規模視覚言語モデルであり，さまざまなタスクにおいて良好な結果が示されている [Achiam 23, GeminiTeam 23]．このことから，本研究で扱うタスクについても適用可能であると考え，利用した．また，本研究では，2 値分類のタスクにおいて標準的である分類精度を評価尺度に使用した．

表 1 に示すように、SP-RT-1 データセットにおける提案手法の精度は 74.50% であるのに対し，InstructBLIP、Gemini，GPT-4V，および UNITER はそれぞれ 52.30%，64.12%，69.12%，69.08% でった．また，SP-VLMbench データセットにおける提案手法の精度は 78.92% であり，InstructBLIP，Gemini，GPT-4V，および UNITER はそれぞれ 41.30%，56.15%，57.81%，68.22% であった．分類精度においてベースライン手法と提案手法の性能差は統計有意であった ($p < 0.05$).

5.2 定性的結果

図 3 に提案手法の成功例を示す．図 3 には，指示文および物体操作前後の画像を示した．まず，図 3(i) は SP-RT-1 データセットにおける例を示しており，指示文は “place redbull can into middle drawer” であった．この例において，ロボットは適切に物体操作を実行しているため，正解ラベルは成功である．提案手法はこの例において成功と正しく予測したのに対し，それぞれのベースライン手法は失敗と誤って予測した．また，図 3(ii) も同様に SP-RT-1 データセットにおける例を示しており，指示文は “place water bottle upright” であった．この例において，ロボットは指示文通りに物体操作を実行できていないため，正解ラベルは失敗である．提案手法はこの例において失敗と正しく予測したのに対し，それぞれのベースライン手法は成功と誤って予測した．また，図 3(iii) は SP-VLMbench データセットにおける例を示しており，指示文は “Drop the red pencil into the navy container.” であった．この例において，ロボットは適切に物体操作を実行しているため，正解ラベルは成功である．この例についても同様に，提案手法は成功と正しく予測したものの，それぞれのベースライン手法は誤って予測した．これらのことから，提案手法の色や空間関係に関する参照表現を理解する性能が向上していることが示唆される．

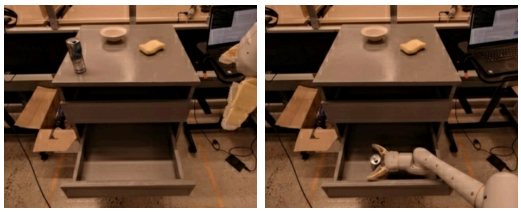
5.3 Ablation Studies

Ablation 条件として以下の 2 つを定めた．

Scene Representation Ablation λ-Representation Encoder から Narrative Representation を取り除くことによる性能への影響を調査した．その結果，Model (i) の SP-RT-1 データセットにおける精度は 73.30%であり，Model (iii) と比較して 1.20 ポイント下回った．また，Model (i) の SP-VLMbench データセットにおける精度は 75.80%であり，Model (iii) と比較して 3.12 ポイント下回った．このことから，Narrative Representation が自然言語を媒介として構造化された特徴量を抽出し視覚表現を強化したことが示唆される．

Narrative Representation Ablation λ-Representation Encoder から Scene Representation を取り除くことによる性能への影響を調査した．その結果，Model (ii) の SP-RT-1 データセットにおける精度は 72.02%であり，Model (iii) と比較し

(i)



(a) (b)

“place redbull can into middle drawer.”

Status: success

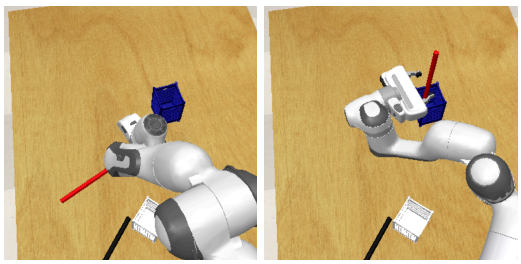
(ii)

(a) (b)

“place water bottle upright”

Status: failure

(iii)



(a) (b)

“Drop the red pencil into the navy container.”

Status: success

図 3: 提案手法の成功例.

て 2.48 ポイント下回った．また，Model (ii) の SP-VLMbench データセットにおける精度は 73.74%であり，Model (iii) と比較して 5.18 ポイント下回った．このことから，Scene Representation が色や形状などの視覚的な情報を捉えることにより視覚表現を強化したことが示唆される．

5.4 エラー分析

表 2 にエラー分析の結果を示す．ここで，失敗例において無作為に選ばれた 100 サンプルについて，以下の 6 つのカテゴリに分類した．

- Multimodal Language Comprehension Error: 視覚情報や言語情報の処理に失敗した例を表す．具体的には，参照表現理解に失敗した場合や，言語情報から関連する物体を正しく特定できなかった場合を含む．
- Occlusion：対象物がアームや他の物体によって半分以上隠れている例を示す．
- Ambiguous Situation: 成功または失敗の解釈が基準によって分かれる場合を表す．
- Narrative Hallucination：マルチモーダル LLM が物体の特徴や位置を誤って説明したり，存在しない物体について記述した例を表す．
- Out-of-Frame：対象物がカメラの画角外に存在する例を指す．
- Ambiguous Instruction：指示文が曖昧な表現を含み，対象物を一意に定めることが困難な場合を示す．

表 2 に示すように，主なエラー原因は Multimodal Language Comprehension Error であった．このエラーは，モデルが指示文に含まれる物体を認識できなかったことに起因する．したがって，モデルが物体を認識できるようにするためには，Narrative Representation のためのテキストプロンプトにおいて，OCR を含む物体の特徴について詳細に記述させることが考えられる．さらに，Open-vocablary の物体検出器を使用して，指示文に含まれる物体の名称を認識させることが考えられる．

表 2: エラー分析の結果.

エラーカテゴリ	サンプル数
Multimodal Language Comprehension Error	45
Occlusion	26
Ambiguous Situation	9
Narrative Hallucination	9
Out-of-Frame	6
Ambiguous Instruction	5
計	100

6. おわりに

本研究では，指示文およびロボットによる物体操作前後の画像が与えられたとき，物体操作の成功および失敗に関する 2 値分類を行う SPOM タスクを扱った．将来研究では，実機ロボットにおける物体操作の成功判定が考えられる．

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

[Achiam 23] Achiam, J., Adler, S., Agarwal, S., et al.: Gpt-4 technical report, *arXiv preprint arXiv:2303.08774* (2023)

[ada] `https://platform.openai.com/docs/models/embeddings`

[Brohan 22] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., et al.: Rt-1: Robotics transformer for real-world control at scale, *arXiv preprint arXiv:2212.06817* (2022)

[Chen 20] Chen, Y.-C., Li, L., Yu, L., Kholy, E., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J.: UNITER: UNiversal Image-TExt Representation Learning, in *ECCV* (2020)

[Dai 23] Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., et al.: InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning (2023)

[Devlin 18] Devlin, J., Chang, M.-W., Lee, K., et al.: Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018)

[GeminiTeam 23] GeminiTeam, G., Anil, R., Borgeaud, S., et al.: Gemini: a family of highly capable multimodal models, *arXiv preprint arXiv:2312.11805* (2023)

[Kaneda 24] Kaneda, K., Nagashima, S., Korekata, R., Kambara, M., and Sugiura, K.: Learning-To-Rank Approach for Identifying Everyday Objects Using a Physical-World Search Engine, *IEEE RAL* (2024)

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

[Obinata 23] Obinata, Y., et al.: Semantic Scene Difference Detection in Daily Life Patroling by Mobile Robots using Pre-Trained Large-Scale Vision-Language Model, in *IROS*, pp. 3228–3233 (2023)

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

[Zheng 22] Zheng, K., Chen, X., Jenkins, C., and Wang, X.: Vlmbench: A compositional benchmark for vision-and-language manipulation, *NeurIPS*, Vol. 35, pp. 665–678 (2022)