VLA における参照表現理解モデルに基づく自由形式の物体操作指示文の理解性能向上

○妹尾幸樹，神原元就，後神美結，杉浦孔明（慶應義塾大学）

本研究では，言語指示に基づく物体操作タスクを扱う．既存手法では，多様な参照表現を含む指示文に基づく物体操作を適切に行うことは困難である．そこで本研究では，多様な参照表現を含む指示文に基づく物体操作タスクにおいて，操作対象物体の予測機構を用いた軌道生成手法を提案する．本手法では，Referring Expression Comprehension モジュールおよび Instruction Interpreter モジュールを導入することにより，事前学習済み VLA モデルを再訓練することなく，その参照表現理解能力を強化することが可能となる．提案手法の有効性を検証するため，シミュレータ環境上で構築したベンチマークを用いて実験を行った．その結果，提案手法は成功率においてベースライン手法を上回った．

1. はじめに

少子高齢化と労働力不足が深刻化する現在の社会情勢において，人間と連携した作業を遂行できるロボットの需要が急速に高まっている．なかでも，家庭内での生活支援や農業現場における労働支援が代表的な導入事例として挙げられる．このような場合に，人間の与えた自然言語指示をロボットが理解し，適切に遂行することが出来れば利便性が高い．しかし，自然言語により物体を特定するためには，複雑な参照表現が用いられる場合が多く，それらを適切に理解することは困難である．実際，言語指示に基づく物体操作手法は多く存在するものの，多様な参照表現に対する理解性能は依然として不十分である．

本研究では，多様な参照表現を含む自然言語指示に基づく物体操作タスクを扱う．たとえば，生活支援ロボットに “コップの隣にあるリモコンを取って” と指示した際に，ロボットが周囲の物体の中から適切な対象を特定し，正しく把持するようなシナリオが考えられる．

実世界の多様な物体操作タスクにおいて，Vision-Language-Action (VLA) モデルにより高い性能が得られたことが報告されており，これらのモデルの訓練にはしばしば大規模なデータセットが用いられている [1–4]．しかしながら，ほとんどのデータセットでは，指示文に含まれる参照表現の多様性に限りがある．結果として，これらのデータセットを用いて訓練された多くの VLA モデルは，多様な参照表現を含む指示文を理解することが困難である．

そこで本研究では，多様な参照表現を含む指示文に基づく物体操作タスクにおいて，操作対象物体の予測機構を用いた軌道生成手法を提案する．提案手法における既存手法との主要な違いは，参照表現によって指定された物体を特定する Referring Expression Comprehension (REC) モジュールの導入である．本モジュールは，物体マスクや画像全体から得られる多様な視覚特徴および指示文の言語特徴を統合し，参照表現が示す物体を推定する．これにより，既存データセットで訓練された VLA モデルの参照表現に対する理解性能が向上し，結果として物体操作における性能も向上することが期待される．

提案手法の主要な貢献は以下の通りである．

- 多様な参照表現を含む指示文に基づく物体操作タスクにおいて，VLA の参照表現への理解性能を強化することを目的としたパイプラインを提案する．
- パイプラインにおいて，参照表現によって指示された物体を特定する REC モジュールを導入する．

”Can you take the red object next to the orange?”

図 1 自然言語指示に基づく物体操作タスクの具体例．画像は左から時刻順である．

- 対照学習において unlabeled positive を考慮するために， IoU を用いた対照損失を導入する．

2. 関連研究

マルチモーダル言語処理に関する研究は広く行われている [5,6]．たとえば，参照表現理解タスク [7] は言語により指定された対象を画像内から特定するタスクである．また，生活支援ロボットのための参照表現理解を目的として，物体操作指示文に基づく参照表現理解タスクに取り組んだ研究も存在する [8,9]．

ロボティクス分野では，自然言語指示文に基づく物体操作タスクがある．VLA モデルは，言語指示と視覚情報に基づいて，ロボットの軌道を生成するモデルであり，本タスクにおいて良好な結果が報告されている [2,10]．一方で，VLA モデルはロボットと実環境を用いて大量の教示データを収集する必要があり，訓練コストが極めて高いという課題を抱えている．

3. 問題設定

本研究では，自然言語指示に基づき物体操作を行うタスクを扱う．入力は，自然言語指示文および各時刻におけるロボットの一人称視点画像であり，出力はエンドエフェクタの軌道である．

本タスクでは，指示文によって指定された物体操作を適切に行うことが望ましい．ここで，図 1 に本タスクの具体例を示す．図において，画像は左から時刻順である．この例は，指示文として “Can you take the red object next to the orange?” が与えられた際に，マニピュレータが指示文に基づいてリンゴを持ち上げている様子を示す．なお，本研究では，指示文により画像内に映る物体の中から操作対象を一意に特定できることを前提とする．

4. 提案手法

本研究では，RT-1 を拡張し，多様な参照表現を含む指示文に対する理解性能の向上を可能とする VLA モデルのためのパイプラインを提案する．図 2 に提案手法のモデル構造を示す．本パイプラインの主要モジュール

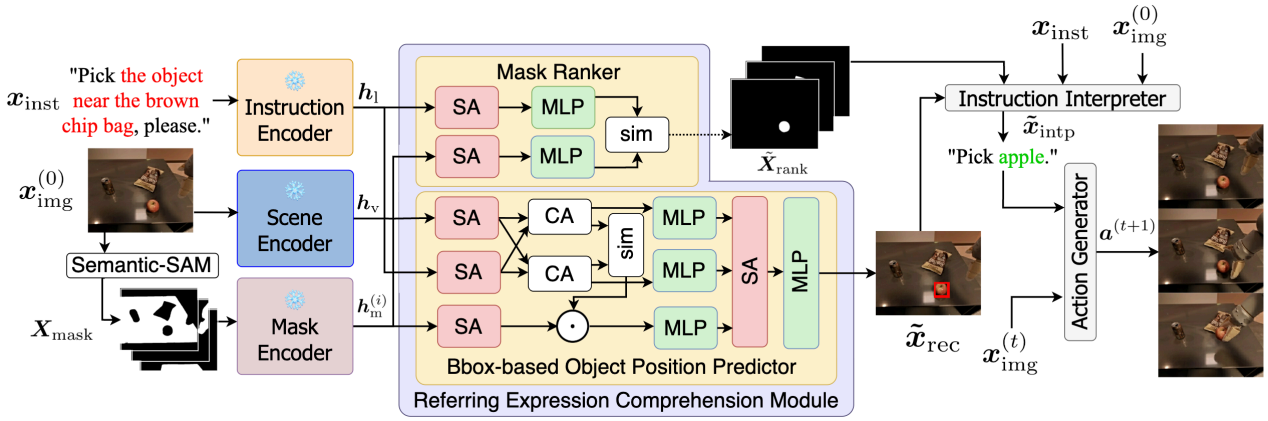

図 2 提案手法のモデル構造．“MLP”，“SA”，“CA”，“sim” および “⊙” は，それぞれ多層パーセプトロン，self-attention，cross-attention，cos 類似度およびアダマール積を表す．

は，Referring Expression Comprehension (REC) モジュール，Instruction Interpreter (II) モジュールおよび Action Generator (AG) モジュールの 3 つである．モデルの入力は $\boldsymbol{x} = \{\boldsymbol{X}_{\text{img}}, \boldsymbol{x}_{\text{inst}}\}$，出力は軌道 $\{\boldsymbol{a}^{(1)}, \ldots, \boldsymbol{a}^{(T)}\}$ である．ここで，$\boldsymbol{X}_{\text{img}} = \{\boldsymbol{x}_{\text{img}}^{(t)} \mid t = 0, \ldots, T\}$ であり，$\boldsymbol{x}_{\text{inst}}$，$\boldsymbol{x}_{\text{img}}^{(t)}$，$t$ および T はそれぞれ指示文，画像，時刻のインデックスおよび最大時刻を表す．

入力特徴量として，$\boldsymbol{x}_{\text{img}}^{(0)}$ の画像特徴量 $\boldsymbol{h}_{\text{v}}$，マスクの特徴量 $\boldsymbol{h}_{\text{m}}^{(i)}$ および $\boldsymbol{x}_{\text{inst}}$ の言語特徴量 $\boldsymbol{h}_{\text{l}}$ を抽出する．それぞれの特徴量についての抽出方法は以下の通りである．まず，$\boldsymbol{x}_{\text{img}}^{(0)}$ を入力とし，SigLIP2 画像エンコーダ [11] および DINOv2 [12] を用いて画像特徴量 $\boldsymbol{h}_{\text{vsi}}$ および $\boldsymbol{h}_{\text{vdi}}$ を得る．次に，マルチモーダル大規模言語モデル (Multimodal Large Language Model, MLLM) を用いて $\boldsymbol{x}_{\text{img}}^{(0)}$ についてのキャプションを生成し，Stella [13] を用いて言語特徴量 $\boldsymbol{h}_{\text{vml}}$ を得る．これらより $\boldsymbol{h}_{\text{v}}$ を以下のように得る．

$$\boldsymbol{h}_{\text{v}} = [\boldsymbol{h}_{\text{vsi}}, \boldsymbol{h}_{\text{vdi}}, \boldsymbol{h}_{\text{vml}}]^{\top}$$

続いて，マスク特徴量 $\boldsymbol{h}_{\text{m}}^{(i)}$ の抽出方法について以下で説明する．まず，Semantic-SAM [14] を用いて $\boldsymbol{x}_{\text{img}}^{(0)}$ からマスク群 $\boldsymbol{X}_{\text{mask}} = \{\boldsymbol{x}_{\text{mask}}^{(i)} \mid i = 1, \ldots, N_{\text{mask}}\}$ を生成する．ここで，$\boldsymbol{x}_{\text{mask}}^{(i)}$，$i$ および N_{mask} はそれぞれマスク，マスクのインデックスおよびマスクの数を表す．次に，$\boldsymbol{x}_{\text{img}}^{(0)}$ と $\boldsymbol{x}_{\text{mask}}^{(i)}$ を入力とし，AlphaCLIP [15] を用いて画像特徴量 $\boldsymbol{h}_{\text{al}}^{(i)}$ を得る．また，$\boldsymbol{x}_{\text{mask}}^{(i)}$ を囲むバウンディングボックスのクロップ画像を入力とし，SigLIP2 画像エンコーダおよび DINOv2 を用いて画像特徴量 $\boldsymbol{h}_{\text{msi}}^{(i)}$ および $\boldsymbol{h}_{\text{mdi}}^{(i)}$ を得る．続いて，$\boldsymbol{x}_{\text{mask}}^{(i)}$ を囲むバウンディングボックスを $\boldsymbol{x}_{\text{img}}^{(0)}$ に重畳した画像を生成する．この画像についてのキャプションを MLLM を用いて生成し，Stella を用いて言語特徴量 $\boldsymbol{h}_{\text{mml}}^{(i)}$ を得る．これらより，$\boldsymbol{h}_{\text{m}}^{(i)}$ を以下のように得る．

$$\boldsymbol{h}_{\text{m}}^{(i)} = [\boldsymbol{h}_{\text{al}}^{(i)}, \boldsymbol{h}_{\text{msi}}^{(i)}, \boldsymbol{h}_{\text{mdi}}^{(i)}, \boldsymbol{h}_{\text{mml}}^{(i)}]^{\top}$$

また，$\boldsymbol{h}_{\text{l}}$ を以下のように得る．

$$\boldsymbol{h}_{\text{l}} = [\boldsymbol{h}_{\text{st}}, \boldsymbol{h}_{\text{ns}}^{\top}]^{\top}$$

ここで，$\boldsymbol{h}_{\text{st}}$ は Stella に $\boldsymbol{x}_{\text{inst}}$ を入力して取得した言語特徴量であり，$\boldsymbol{h}_{\text{nt}}$ は MLLM を用いて抽出した $\boldsymbol{x}_{\text{inst}}$ の名詞節を入力とし，SigLIP2 言語エンコーダを用いて取得した言語特徴量である．

4.1 REC モジュール

REC モジュールは，VLA における多様な参照表現を含む指示文への理解性能向上を目的として，操作対象物体を特定するモジュールである．VLA モデルの訓練に用いられるデータセットに多様な参照表現が含まれていない場合，VLA モデルがそれらを適切に理解することは困難である．本モジュールを用いることで，VLA モデルを再訓練することなく，参照表現の理解性能の向上が期待される．

本モジュールは，$\boldsymbol{h}_{\text{v}}$，$\boldsymbol{h}_{\text{m}}^{(i)}$ および $\boldsymbol{h}_{\text{l}}$ を入力とし，順位づけされたマスク系列 $\tilde{\boldsymbol{X}}_{\text{rank}}$ および操作対象物体を表すバウンディングボックス $\tilde{\boldsymbol{x}}_{\text{rec}}$ を予測する．また，本モジュールは Mask Ranker (MR) モジュールと Bbox-based Object Position Predictor (BOPP) モジュールの 2 つのサブモジュールから構成される．それぞれのサブモジュールについて以下で説明する．

MR モジュールは，操作対象物体を特定するために，マスクの順位付けを行うモジュールである．まず，$\boldsymbol{h}_{\text{mask}}^{(i)} = \text{MLP}(\text{SA}(\boldsymbol{h}_{\text{m}}^{(i)}))$ を得る．ここで，$\text{SA}(\cdot)$ および $\text{MLP}(\cdot)$ は self-attention および多層パーセプトロンを表す．また，言語特徴量 $\boldsymbol{h}_{\text{txt}} = \text{MLP}(\text{SA}(\boldsymbol{h}_{\text{l}}))$ を得る．そして，$\boldsymbol{h}_{\text{mask}}^{(i)}$ と $\boldsymbol{h}_{\text{txt}}$ の cos 類似度に基づく順位づけにより $\boldsymbol{X}_{\text{mask}}$ を並び替えたマスク系列 $\tilde{\boldsymbol{X}}_{\text{rank}}$ を得る．

次に，BOPP モジュールについて以下で説明する．このモジュールは，操作対象物体を特定することを目的として，バウンディングボックスを生成するモジュールである．MR モジュールは，生成されたマスクを順位づけするため，物体の形状を考慮した出力を得ることが可能である．一方で，生成されたマスクが捉えられない詳細な物体位置を示すことが困難である．このモジュールを用いることにより，マスク生成モデルが考慮できなかった物体情報を補完することで，より詳細な物体位置の特定が可能になることが期待される．

まず，$\boldsymbol{x}_{\text{img}}^{(0)}$ についての画像特徴量 $\boldsymbol{h}_{\text{scene}} = \text{MLP}(\text{SA}(\boldsymbol{h}_{\text{v}}))$ を得る．次に，$\boldsymbol{h}_{\text{txt}}$ と $\boldsymbol{h}_{\text{scene}}$ を用いてマルチモーダル特徴量 $\boldsymbol{h}_{\text{cs}} = \text{CA}(\boldsymbol{h}_{\text{txt}}, \boldsymbol{h}_{\text{scene}})$ および $\boldsymbol{h}_{\text{ct}} = \text{CA}(\boldsymbol{h}_{\text{scene}}, \boldsymbol{h}_{\text{txt}})$ を得る．ここで，$\text{CA}(\cdot, \cdot)$ は，cross-attention [16] を表す．続いて，マスク特徴量 $\boldsymbol{h}_{\text{cm}}$ を以下のように計算する．

$$\boldsymbol{h}_{\text{cm}} = \sum_{i=1}^{N_{\text{mask}}} \text{sim}(\boldsymbol{h}_{\text{mask}}^{(i)}, \boldsymbol{h}_{\text{ct}}) \cdot \boldsymbol{h}_{\text{mask}}^{(i)}$$

ただし，$\text{sim}(\cdot, \cdot)$ は cos 類似度を表す．最終的に $\tilde{\boldsymbol{x}}_{\text{rec}}$

を以下のように計算する．

$$\tilde{\boldsymbol{x}}_{\mathrm{rec}} = \mathrm{MLP}(\mathrm{SA}([\boldsymbol{h}'_{\mathrm{cm}}, \boldsymbol{h}'_{\mathrm{cs}}, \boldsymbol{h}'_{\mathrm{ct}}]^\top))$$

ここで，$\boldsymbol{h}'_{\mathrm{cm}} = \mathrm{MLP}(\boldsymbol{h}_{\mathrm{cm}})$，$\boldsymbol{h}'_{\mathrm{cs}} = \mathrm{MLP}(\boldsymbol{h}_{\mathrm{cs}})$ および $\boldsymbol{h}'_{\mathrm{ct}} = \mathrm{MLP}(\boldsymbol{h}_{\mathrm{ct}})$ である．

4.2 II モジュール

II モジュールは，VLA が理解可能な指示文を生成することを目的として，参照表現を任意の粒度の表現に変換するモジュールである．本モジュールを用いない場合，VLA モデルは，REC モジュールの出力を直接処理することが求められる．これは，VLA モデルの訓練において，一般的にバウンディングボックスによるビジュアルプロンプトを含む画像が用いられないため，困難である．本モジュールは $\boldsymbol{x}_{\mathrm{inst}}$，$\boldsymbol{x}^{(0)}_{\mathrm{img}}$，$\tilde{\boldsymbol{X}}_{\mathrm{rank}}$ の上位 K 個および $\tilde{\boldsymbol{x}}_{\mathrm{rec}}$ を入力として，MLLM を用いて新しい指示文 $\tilde{\boldsymbol{x}}_{\mathrm{intp}}$ を生成する．

AG モジュールは，物体操作を行うために，事前学習済み VLA モデルを用いて，エンドエフェクタの軌道を生成するモジュールである．本モジュールは，各時刻で $\boldsymbol{x}^{(t)}_{\mathrm{img}}$ および $\tilde{\boldsymbol{x}}_{\mathrm{intp}}$ を入力とし，次時刻におけるエンドエフェクタの行動 $\boldsymbol{a}^{(t+1)}$ を計算する．本研究では，VLA モデルとして，RT-1 を用いた．

4.3 損失関数

MR モジュールでは，マスク-指示文間およびマスク-マスク間の positive ペアの cos 類似度を最大化する．類似性を考慮する際，多くの対照学習手法は InfoNCE を損失関数として利用している [17]．InfoNCE 損失を用いる場合，$\boldsymbol{y}_{\mathrm{mask}}$ と，それと領域を一部共有するマスクは negative ペアとみなされる．しかしながら，それらは類似性の観点から，negative ではなく unlabeled positive として対称性を緩和することが望ましい．さらに，画像-言語間の対照学習において，画像-画像間の対照損失を併用することで，より高品質な特徴量が得られることが報告されている [18]．これらのことから，本手法で用いる損失関数を $\mathcal{L}_{\mathrm{rank}} = \mathcal{L}_{\mathrm{mt}} + \lambda_{\mathrm{mm}}\mathcal{L}_{\mathrm{mm}}$ とする．この損失関数を構成する要素は，それぞれマスク-言語間およびマスク-マスク間の対照損失であり，以下のように定義される．

$$\mathcal{L}_{\mathrm{mt}} = \frac{1}{N_{\mathrm{mask}}} \sum_{i}^{N_{\mathrm{mask}}} (\alpha_{\mathrm{iou}}(\boldsymbol{x}^{(i)}_{\mathrm{mask}}, \boldsymbol{y}) - s_{\mathrm{mt}})^2$$

$$\mathcal{L}_{\mathrm{mm}} = \frac{1}{N^2_{\mathrm{mask}}} \sum_{i_1,i_2}^{N_{\mathrm{mask}}} (\alpha_{\mathrm{iou}}(\boldsymbol{x}^{(i_1)}_{\mathrm{mask}}, \boldsymbol{x}^{(i_2)}_{\mathrm{mask}}) - s_{\mathrm{mm}})^2$$

ここで，$\alpha_{\mathrm{iou}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = 2\sqrt{\mathrm{IoU}(f_{\mathrm{bbox}}(\boldsymbol{x}_1), f_{\mathrm{bbox}}(\boldsymbol{x}_2))}$，$s_{\mathrm{mt}} = 1 + \mathrm{sim}(\boldsymbol{h}^{(i)}_{\mathrm{mask}}, \boldsymbol{h}_{\mathrm{txt}})$ および $s_{\mathrm{mm}} = 1 + \mathrm{sim}(\boldsymbol{h}^{(i_1)}_{\mathrm{mask}}, \boldsymbol{h}^{(i_2)}_{\mathrm{mask}})$ である．ただし，$f_{\mathrm{bbox}}(\cdot)$，$\mathrm{IoU}(\cdot,\cdot)$ および λ_{mm} は，それぞれマスクを内包する最小のバウンディングボックス，IoU および重み係数を表す．

BOPP モジュールの損失関数は以下で定義される．

$$\mathcal{L}_{\mathrm{rec}} = \mathcal{L}_{\mathrm{mt}} + \lambda_{\mathrm{l1}}\ell_1(\hat{\boldsymbol{y}}_{\mathrm{rec}}, \boldsymbol{y}_{\mathrm{mask}}) + \lambda_{\mathrm{GIoU}}\mathcal{L}_{\mathrm{giou}}$$

$$\mathcal{L}_{\mathrm{giou}} = (1 - \mathrm{GIoU}(\hat{\boldsymbol{y}}_{\mathrm{rec}}, \boldsymbol{y}_{\mathrm{mask}}))$$

ここで，$\ell_1(\cdot)$ および $\mathrm{GIoU}(\cdot,\cdot)$ はそれぞれ L1 損失および GIoU [19] であり，λ_{l1} および λ_{giou} はいずれも重み係数である．

5. 実験設定

本研究で扱う，指示文に基づく物体操作タスクにおいては，Fractal [10] が標準的データセットである．しかしながら，Fractal にはテンプレートベースの指示文しか含まれておらず，参照表現の多様性が不十分である．そこで本研究では，Fractal からエピソードを抽出して，多様な参照表現を含む指示文に置き換えた Fractal-free データセットを構築した．また，シミュレータ環境で物体操作モデルの評価を行うため，SIMPLER [20] を用いて，100 エピソードから構成されるベンチマークを構築した．

SIMPLER 環境でのベンチマーク構築において，まず，[20] で使用された物体群の中から 10 個を選択した．この上で，ランダムに最大 5 個を配置し，その中の 1 つの物体を持ち上げるよう指示する文を付与させた．

Fractal-free は，1,000 枚の画像，4,849 の指示文およびそれに対応する操作対象の物体のマスクが含まれる実世界データセットである．本データセットに含まれる指示文のうち，422 の指示文は 39 人のアノテータに付与させた．また，残りの指示文については MLLM を用いて収集した．

REC モジュールの実験設定は以下の通りである．MR モジュールは 150 エポック訓練を行った．ここで，$\lambda_{\mathrm{mm}} = 1.0$ とし，学習率およびバッチサイズは 1×10^{-6} および 32 とした．BOPP モジュールは，300 エポック訓練を行った．ここで，$\lambda_{\mathrm{l1}} = 2.0 \times 10^{-3}$，$\lambda_{\mathrm{GIoU}} = 3.3 \times 10^{-2}$ とし，学習率およびバッチサイズは 1×10^{-5} および 16 とした．また，どちらにおいても最適化手法として Adam (β_1=0.9, β_2=0.999) を用いた．これらのモデルの訓練には，GeForce RTX 4090 および Intel Core i9-14900F を搭載した計算機を使用した．REC モジュールの 1 サンプルあたりの推論時間は約 3.9 秒であった．また，RT-1 モジュールの 1 ステップあたりの推論時間は約 1.2×10^{-3} 秒であった．

6. 実験結果

6.1 定量的結果

表 1　ベースライン手法との定量的比較結果

手法	成功率 [%]
RT-1 [10]	17
RT-1 w/ Qwen2.5-VL	66
提案手法	**68**

ベースライン手法と提案手法の定量的比較結果を表 1 に示す．また，表中の太字の数値は指標における最も高い数値を表す．RT-1 [10] および RT-1 w/ Qwen2.5-VL をベースライン手法とした．

RT-1 は，指示文に基づく物体操作タスクにおける標準的な手法であるため選択した [20]．RT-1 w/ Qwen2.5-VL は，Qwen2.5-VL [21] を用いて指示文に含まれる参照表現を単純化し，RT-1 で物体操作を行う手法である．具体的には，Fractal データセット [10] に含まれる指示文と類似したテンプレートベース指示文を生成することを目的として，Qwen2.5-VL に対し，およびを含むプロンプトを入力する．その後，生成された指示文を RT-1 の入力として用い，物体操作を行う．これは，高いマルチモーダル言語処理能力を有する Qwen2.5-VL を用いた単純な変換手法によって，どの程度参照表現の理解能力が向上するのかを評価するためのベースラインになりうると考え採用した．

本タスクにおいて標準的であるため，評価尺度として成功率を用いた．表 1 に示す通り，RT-1，RT-1 w/

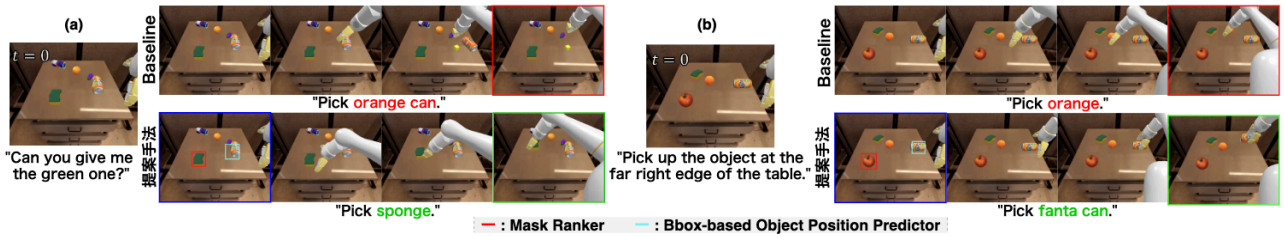

図 3 提案手法およびベースライン手法の定性的結果．各例において，$\boldsymbol{x}_{\text{img}}^{(0)}$，$\boldsymbol{x}_{\text{isnt}}$，各手法が生成した指示文，および物体操作の様子を示す．青枠で囲まれた画像に含まれる赤色と水色のバウンディングボックスは，それぞれ MR モジュールと BOPP モジュールの出力を示す．

Qwen2.5-VL および提案手法の成功率はそれぞれ 17%，66%および 68%であった．このことから，提案手法のスコアは RT-1 より 51.0 ポイント，RT-1 w/ Qwen2.5-VL より 2.0 ポイント上回るスコアであった．

6.2 定性的結果

図 3 に提案手法およびベースライン手法の 1 つである RT-1 w/ Qwen2.5-VL の定性的結果を示す．図 3 の (a) および (b) は提案手法の成功例である．

(a) の $\boldsymbol{x}_{\text{inst}}$ は，“Can you give me the green one?” であり，この指示文が示す操作対象物体はスポンジであった．ベースライン手法は “Pick orange can.” という不適切な指示文を生成し，物体操作に失敗した．一方で，提案手法では MR モジュールの出力に適切な物体が含まれていたため，II モジュールが “Pick sponge.” という指示文を生成し，最終的に物体操作に成功した．

(b) では，“Pick up the object at the far right edge of the table.” が $\boldsymbol{x}_{\text{inst}}$ として与えられ，操作対象物体はオレンジ色の缶であった．ベースライン手法は不適切な指示文 “Pick orange.” を生成し，これを入力としたため物体操作に失敗した．しかし，提案手法の BOPP モジュールは適切な物体を示すバウンディングボックスを出力し，II モジュールは “Pick fanta can.” という指示文を生成したため AG モジュールは適切な物体を把持することに成功した．これらのことから，MR と BOPP モジュールは互いの参照表現理解能力を補完するように機能していることが示唆される．また，II モジュールがそれらの出力を踏まえ，適切な指示文を生成できたことが分かる．

表 2 ablation study の結果

モデル	MR モジュール	BOPP モジュール	成功率 [%]
(i)		✓	65
(ii)	✓		65
(iii)	✓	✓	**68**

6.3 Ablation study

表 2 に Ablation study の結果を示す．また，表中の太字は指標における最も高い数値を表す．Ablation study として，REC モジュールを構成する 2 つのサブモジュールをそれぞれ取り除くことによる性能への寄与を調査した．モデル (i) は MR モジュールを取り除いたものであり，モデル (ii) は BOPP モジュールを取り除いたものである．

モデル (i) とモデル (ii) の成功率はどちらも 65%であり，モデル (iii) と比較すると 3.0 ポイント低下した．この結果から，各サブモジュールはそれぞれ異なる性質の参照表現に対して高い理解性能を有しており，それらが補完的に機能することで性能向上に寄与したことが示唆される．

7. おわりに

本研究では，多様な参照表現を含む自然言語指示に基づく物体操作タスクを扱った．本研究の貢献は，多様な参照表現を含む指示文に基づく物体操作タスクにおいて，VLA モデルの参照表現理解性能を強化するためのパイプラインを提案した点にある．また，対照学習において unlabeled positive を考慮するために，IoU に基づく対照損失関数を導入した．結果として，SIMPLER 環境上に構築したベンチマークにおいて，提案手法はベースライン手法を成功率で上回った．

謝辞

本研究の一部は，JSPS 科研費 23K28168，JST ムーンショット，JSPS 特別研究員奨励費 JP23KJ1917 の助成を受けて実施されたものである．

参 考 文 献

[1] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, et al., “RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control,” CoRL, pp.2165–2183, 2023.
[2] K. Black, N. Brown, D. Driess, et al., “π_0: A Vision-Language-Action Flow Model for General Robot Control,” RSS, 2025.
[3] M. Kim, K. Pertsch, S. Karamcheti, et al., “OpenVLA: An Open-Source Vision-Language-Action Model,” CoRL, 2024.
[4] A. O’Neill, A. Rehman, et al., “Open X-Embodiment: Robotic Learning Datasets and RT-X Models: Open X-Embodiment Collaboration,” ICRA, pp.6892–6903, 2024.
[5] S. Uppal, S. Bhagat, D. Hazarika, et al., “Multimodal Research in Vision and Language: A Review of Current and Emerging Trends,” Information Fusion, vol.77, pp.149–171, 2022.
[6] F. Chen, D. Zhang, M. Han, et al., “VLP: A Survey on Vision-language Pre-training,” MIR, vol.20, pp.38–56, 2023.
[7] Y. Qiao, C. Deng, and Q. Wu, “Referring expression comprehension: A survey of methods and datasets,” IEEE TMM, vol.23, pp.4426–4440, 2020.
[8] Y. Iioka, Y. Yoshida, Y. Wada, S. Hatanaka, et al., “Multimodal Diffusion Segmentation Model for Object Segmentation from Manipulation Instructions,” IROS, pp.7590–7597, 2023.
[9] T. Nishimura, K. Kuyo, M. Kambara, and K. Sugiura, “Object Segmentation from Open-Vocabulary Manipulation Instructions Based on Optimal Transport Polygon Matching with Multimodal Foundation Models,” IROS, pp.9549–9556, 2024.
[10] A. Brohan, N. Brown, J. Carbajal, et al., “RT-1: Robotics Transformer for Real-World Control at Scale,” RSS, 2023.
[11] M. Tschannen, A. Gritsenko, X. Wang, M. Naeem, et al., “SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features,” arXiv preprint arXiv:2502.14786, 2025.
[12] M. Oquab, T. Darcet, et al., “DINOv2: Learning Robust Visual Features without Supervision,” TMLR, pp.1–31, 2024.
[13] D. Zhang, et al., “Jasper and Stella: distillation of SOTA embedding models,” arXiv preprint arXiv:2412.19048, 2024.
[14] F. Li, H. Zhang, P. Sun, X. Zou, et al., “Segment and Recognize Anything at Any Granularity,” ECCV, pp.467–484, 2024.
[15] Z. Sun, et al., “Alpha-CLIP: A CLIP Model Focusing on Wherever You Want,” CVPR, pp.13019–13029, 2024.
[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, et al., “Attention Is All You Need,” NIPS, 2017.
[17] A. Radford, et al., “Learning Transferable Visual Models From Natural Language Supervision,” ICML, pp.8748–8763, 2021.
[18] N. Mu, A. Kirillov, et al., “SLIP: Self-supervision meets Language-Image Pre-training,” ECCV, pp.529–544, 2022.
[19] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, et al., “Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression,” CVPR, pp.658–666, 2019.
[20] X. Li, K. Hsu, et al., “Evaluating Real-World Robot Manipulation Policies in Simulation,” CoRL, pp.3705–3728, 2023.
[21] S. Bai, K. Chen, X. Liu, J. Wang, et al., “Qwen2.5-VL Technical Report,” arXiv preprint arXiv:2502.13923, 2025.