物体操作指示文生成モデルに基づく
モバイルマニピュレーションのためのデータセット拡張
Dataset Augmentation Based on Instruction Generation Model for Open-Vocabulary Mobile Manipulation

勝又 圭 Kei Katsumata　神原 元就 Motonari Kambara　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

We consider the problem of generating free-form mobile manipulation instructions based on a target object image and receptacle image. In this study, we propose a model that handles both the target object and receptacle to generate free-form instruction sentences for mobile manipulation tasks. Moreover, we introduce a novel training method that effectively incorporates the scores from both learning-based and n-gram based automatic evaluation metrics as rewards. This method enables the model to learn the co-occurrence relationships between words and appropriate paraphrases. Results demonstrate that our proposed method outperforms baseline methods including representative multimodal large language models on standard automatic evaluation metrics. Moreover, physical experiments reveal that using our method to augment data on language instructions improves the performance of an existing multimodal language understanding model for mobile manipulation.

1. はじめに

生活支援ロボットの発展は、高齢者介護施設や障がい者の日常支援など、様々な場面で有用である。特に、高齢者介護施設への生活支援ロボットの導入は、介護者の負担を大幅に軽減し、高齢者人口の増加に伴う介助者不足に対応する上で重要である。特に，自然言語によりロボットへ日常タスクを指示することができれば利便性が高い．一方，自然言語指示文に基づき日常タスクを実行するためのマルチモーダル言語理解モデルは未だ性能が不十分である．理解能力を高めるためには，高品質な自然言語指示文を含むデータセットを用いた訓練が必要である．データセット構築には人間によるアノテーションが不可欠だが，コストが高いという課題がある．よって，高品質な指示文を自動生成できれば利便性が高い． 本研究では，対象物体画像及び配置目標画像に基づき物体操作指示文の生成を行う物体操作指示文生成タスクを扱う．本研究の概要を 図 1 に示す．本タスクでは対象物体画像と配置目標画像の 2 枚画像それぞれに存在する対象物体と配置目標の双方を考慮した指示文を生成する必要がある．既存の画像キャプション生成モデル [Nguyen 22, Li 23] は対象物体画像及び配置目標画像の複数画像を入力として扱う構造を持たず，適切に扱うことができない．そこで本研究では，対象物体及び配置目標を適切に含む物体操作指示文を生成するモデルを提案する．我々の提案手法では，複数の視覚的特徴量をテキスト特徴量に対して適切なアライメントを行う Triplet Qformer を導入する。また，学習ベース及び n-gram ベースの自動評価指標に基づく訓練手法 Human Centric Calibration Phase (HCCP) を提案する．HCCP を導入することで単語の共起関係やパラフレーズを学習し、人間による付与文の品質に近い指示文の生成が期待される．提案手法の新規性は以下である．

- 複数の視覚的特徴量をそれぞれにテキスト特徴とアライメントを行う Triplet Qformer を導入する．
- 学習ベース及び n-gram ベースの自動評価尺度に基づく訓練手法 HCCP を導入する．

連絡先: 勝又圭，慶應義塾大学，神奈川県横浜市港北区日吉 3-14-1，ke59ka77@keio.jp

2. 問題設定

本研究では, 物体操作指示文生成タスクを扱う．本タスクでは対象物体画像, 及び配置目標画像に基づき物体操作指示文の生成を行うことを目的とする．本タスクでは，入力画像群に基づき， 対象物体と配置目標を含んだ適切な指示文が生成されることが望ましい．図 1 に本タスクの具体例を示す．図 1 のような対象物体画像と配置目標画像が与えられた時，それぞれに含まれる対象物体と配置目標に基づき，“Move the lamp on the table to the dining table in the kitchen.”のような指示文を生成することを目的とする．

物体操作指示文生成タスクにおいて，入力は対象物体画像および配置目標画像，出力は対象物体及び配置目標を含む，モバイルマニピュレータのための物体操作指示文である．本論文で用いる用語を以下のように定義する．対象物体は，指示に基づき把持される日常物体であり，対象物体画像は対象物体が写る画像である．配置目標は対象物体が置かれる家具を指し，配置目標画像は配置目標が写る画像である．

3. 提案手法

図1に，提案手法のモデル構造を示す．提案手法は主に Multi Image Feature Generator，Triplet Qformer 及び LLM Decoder の 3 つのモジュールから構成される．モデルの入力を $\boldsymbol{x} = \{\boldsymbol{X}_{\mathrm{tar}}, \boldsymbol{X}_{\mathrm{rec}}\}$ と定義する．ここで $\boldsymbol{X}_{\mathrm{tar}}$ 及び $\boldsymbol{X}_{\mathrm{rec}}$ はそれぞれ対象物体画像及び配置目標画像を示す．

3.1 Multi Image Feature Generator

Multi Image Feature Generator は物体に関する情報を含む特徴量である Region 特徴量 $\boldsymbol{H}_{\mathrm{R}}$ 及び画像全体から得られる Grid 特徴量 $\boldsymbol{H}_{\mathrm{G}}$ を抽出するモジュールである．本モジュールは Region encoder branch 及び Grid encoder branch の 2 つのブランチから構成される．

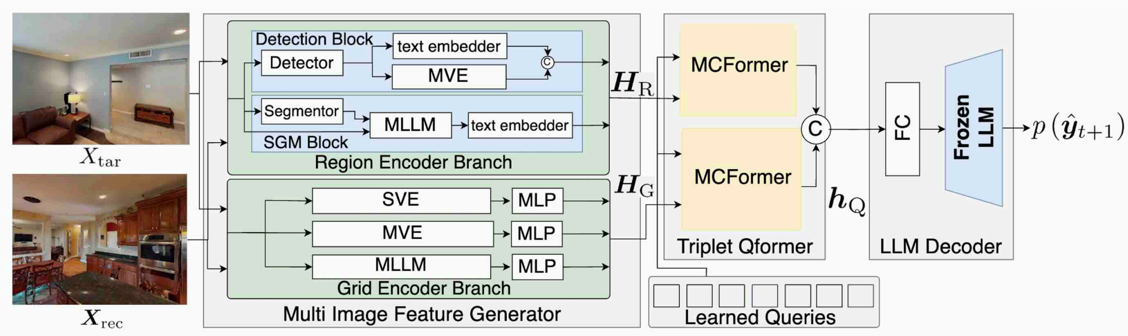

図 1: 提案手法のモデル構造．SVE, MVE, MLLM, 及び FC は，シングルモーダル画像エンコーダ, マルチモーダル画像エンコーダ, 大規模言語モデル, 及び 全結合層を示す．

3.1.1 Region encoder branch

Region encoder branch は物体に関する情報を含むマルチモーダルな Region 特徴量を抽出する．本ブランチは detection block 及び spatial grounded marks (SGM) block から構成される．本モジュールでは入力 $\mathbf{X}_{\mathrm{tar}}$，$\mathbf{X}_{\mathrm{rec}}$ を個別に扱う．以下は簡略化のため $\mathbf{X}_{\mathrm{tar}}$ に対しての処理についてのみ説明を行う．本モジュールでは 2 種類の Region 特徴量 detection image feature $\boldsymbol{h}_{\mathrm{R,D}}$ 及び SGM 特徴量 $\boldsymbol{h}_{\mathrm{R,S}}$ を $\boldsymbol{X}_{\mathrm{tar}}$ から取得する．

Detection block では $\boldsymbol{X}_{\mathrm{tar}}$ から予測ラベルつき物体領域群 $\boldsymbol{D}_k$ を得る．ここで k は検出された物体数を表す．獲得した $\boldsymbol{D}_k$ に対して画像特徴量 $\boldsymbol{v}_{\mathrm{De},k}$ を抽出する．また物体領域群に付与された予測ラベルを用いて，テキスト特徴量 $\boldsymbol{s}_k$ を抽出する．そして本ブロックの出力 $\boldsymbol{h}_{\mathrm{R,D}} = [\boldsymbol{v}_{\mathrm{De},k}; \boldsymbol{s}_k]$ を得る．

SGM block では segmentation モデルを用いて記号を重畳した画像を生成する．これに対し、MLLM を用いて詳細情報を記述し，SGM 特徴量 $\boldsymbol{h}_{\mathrm{R,S}}$ を取得する．詳細に関しては [Yashima 25] を参照されたい．同様の処理を $\boldsymbol{X}_{\mathrm{rec}}$ に行い，Region 特徴量 $\boldsymbol{H}_{\mathrm{R}} = \left\{[\boldsymbol{h}_{\mathrm{R,D}}; \boldsymbol{h}'_{\mathrm{R,D}}], [\boldsymbol{h}_{\mathrm{R,S}}; \boldsymbol{h}'_{\mathrm{R,S}}]\right\}$ を得る．

3.1.2 Grid encoder branch

Grid encoder branch は画像全体から得られる Grid 特徴量を抽出するモジュールである．先述した Region 特徴量は物体検出器の性能に依存するため，重要な物体が見落とされる可能性がある．また，Region 特徴量は物体領域間の特徴量を含んでおらず，物体間の位置関係を抽出することが困難である．一方で，前述したような位置関係等の複雑な参照関係や視覚的な特徴は物体操作指示文生成において重要である．よって，画像全体から Grid 特徴量を抽出し利用することが効果的であると考えられる．Grid encoder branch は 3 種類の潜在表現を組み合わせた λ-Representation [Goko 24] を拡張した Grid 特徴量を抽出する．Grid 特徴量 $\boldsymbol{h}_{\mathrm{G,D}}$, $\boldsymbol{h}_{\mathrm{G,C}}$, 及び $\boldsymbol{h}_{\mathrm{G,L}}$ を $\boldsymbol{X}_{\mathrm{tar}}$ から以下のように取得する．$(\boldsymbol{h}_{\mathrm{G,D}}, \boldsymbol{h}_{\mathrm{G,C}}, \boldsymbol{h}_{\mathrm{G,L}}) = (\mathrm{MLP}(\boldsymbol{v}_{\mathrm{D}}), \mathrm{MLP}(\boldsymbol{v}_{\mathrm{C}}), \mathrm{MLP}(\boldsymbol{v}_{\mathrm{L}}))$．ここで $\boldsymbol{v}_{\mathrm{D}}$, $\boldsymbol{v}_{\mathrm{C}}$ 及び $\boldsymbol{v}_{\mathrm{L}}$ はそれぞれシングルモーダル画像エンコーダ, マルチモーダル画像エンコーダ及びマルチモーダル大規模言語モデル (MLLM) から得られた視覚特徴量を表す．また，MLP $(\cdot)$ は多層パーセプトロンを表す．同様に X_{rec} からも視覚特徴量 $\boldsymbol{h}'_{\mathrm{G,D}}$, $\boldsymbol{h}'_{\mathrm{G,C}}$ 及び $\boldsymbol{h}'_{\mathrm{G,L}}$ を得る．これによって本モジュールの出力 $\boldsymbol{H}_{\mathrm{G}} = \left\{[\boldsymbol{h}_{\mathrm{G,D}}; \boldsymbol{h}'_{\mathrm{G,D}}], [\boldsymbol{h}_{\mathrm{G,C}}; \boldsymbol{h}'_{\mathrm{G,C}}], [\boldsymbol{h}_{\mathrm{G,L}}; \boldsymbol{h}'_{\mathrm{G,L}}]\right\}$ を得る．

3.2 Triplet Qformer

Triplet Qformer はテキスト特徴量を軸として 2 種類の視覚特徴と付与文から得られるテキスト特徴量のアライメントを行うモジュールである．これにより，3 種類の特徴を間接的にアライメントを行うことが可能となる．本モジュールは $\boldsymbol{H}_{\mathrm{G}}$, $\boldsymbol{H}_{\mathrm{R}}$, 及び学習クエリを入力として受け取り，マルチモーダル特徴量 $\boldsymbol{h}_{\mathrm{Q}}$ を出力とする．また，事前学習時には付与文 $\boldsymbol{y}$ も入力とする．Triplet Qformer は二つの Multi Image Cross Attention Transformers (MCFormers) から構成される．MCFormer は複数の視覚的特徴量をテキスト特徴量とアライメントを行う Qformer [Li 23] を拡張した構造である．

3.3 LLM Decoder

既存の画像キャプションモデルは，新規の場面や物体を含むドメイン外の画像対しての汎用性が低い．そこで、本研究では事前学習済み大規模言語モデル (LLM) をデコーダーとして使用する．LLM Decoder は，$\boldsymbol{h}_{\mathrm{Q}}$ に基づいて指示文を生成する．本モジュールでは，まず全結合層を用いて $\boldsymbol{h}_{\mathrm{Q}}$ を線形変換する．次に，LLM を用いて，次のトークンの確率 $p(\boldsymbol{y}_{t+1}|\boldsymbol{x}, \boldsymbol{y}_{1:t})$ を取得する．ここで，$\boldsymbol{y}_{t+1}$ は時刻 $t+1$ で予測されたトークン，$\boldsymbol{y}_{1:t}$ は時刻 t までの予測されたトークン列を表す．

3.4 訓練方法

本手法の訓練は Triplet Qformer pre-training phase (TQPP), probability distribution matching phase (PDMP), 及び HCCP の 3 つのステージから構成される．TQPP では Triplet Qformer を Qformer [Li 23] と同様の損失関数により事前学習を行う．また，PDMP ではクロスエントロピー損失関数によりモデル全体を学習する．HCCP では学習ベースの評価尺度を用いる損失関数 HCCT を用いて訓練を行う．損失関数 $\mathcal{L}_{\mathrm{HCCT}}$ を以下のように定義する．

$$\mathcal{L}_{\mathrm{HCCT}} = -\frac{1}{k}\sum_{i=1}^{k}(r(\mathbf{w}_i) - b)\log \mathrm{p}(\mathbf{w}_i)$$

ここで $\mathbf{w}_i$，$r(\mathbf{w}_i)$，b，k はそれぞれビーム内の i 番目の生成文，報酬関数，報酬基準，バッチ内のサンプルのインデックスを表す．また，$r(\mathbf{w}_i)$ と b をそれぞれ $r(\mathbf{w}_i) = \frac{1}{2}\left(\mathrm{mean}\left(\mathrm{P}_{\mathrm{tar}}(\mathbf{w}_i), \mathrm{P}_{\mathrm{rec}}(\mathbf{w}_i)\right), \mathrm{C}(\mathbf{w}_i)\right)$ 及び $b = \frac{1}{k}\sum_{i=1}^{k} r(\mathbf{w}_i)$ と定義する．ここで $\mathrm{P}_{\mathrm{tar}}(\cdot)$，$\mathrm{P}_{\mathrm{rec}}(\cdot)$ 及び $\mathrm{C}(\cdot)$ は対象物体画像，配置目標画像に対する Polos 及び CIDEr [Vedantam 15] の値を表す．Polos は既存の n-gram に基づいた自動評価尺度と比較して人間による評価との相関係数が高い．そのため，組み合わせて使用することで，より人間による付与文の品質に近い指示文の生成が可能であると考えられる．

4. 実験設定

本実験では LTRRIE-FC データセットにおける HM3D-FC サブセット [Korekata 25] を用いた．本研究では GeForce RTX 4090, 64GB RAM, Intel Core i9-13900KF でモデルの訓練及び推論を行った．本提案手法の学習には 16 時間，1 サンプルあたりの推論には 92 ms の時間を要した. 各エポック毎に検証集合を用いて各種自動評価尺度によるスコアを計算した．検証集合で Polos [Wada 24] スコアが一番高いモデルを選択し，テ

表 1: ベースライン手法との定量的比較結果．表中の TQ は Triplet Qformer を表す．

手法	Polos		SPICE	CIDEr	BLEU4
	対象物体画像	配置目標画像			
(i) GRIT [Nguyen 22]	41.2 ± 1.8	39.0 ± 1.8	19.8 ± 0.5	59.2 ± 2.8	10.5 ± 0.3
(ii) BLIP-2 [Li 23]	42.4 ± 1.4	40.1 ± 1.5	16.9 ± 0.4	36.6 ± 4.9	8.5 ± 1.2
(iii) Gemini [Reid 24]	29.5 ± 0.2	29.4 ± 0.2	11.2 ± 0.6	26.2 ± 1.6	5.2 ± 0.4
(iv) GPT-4o [Achiam 23]	34.5 ± 0.2	35.5 ± 0.06	15.2 ± 0.5	33.4 ± 1.5	7.5 ± 0.2
(v) 提案手法 wo/HCCP	44.7 ± 0.1	44.4 ± 0.2	20.9 ± 0.5	55.3 ± 1.6	10.2 ± 0.5
(vi) 提案手法 wo/TQ	46.2 ± 0.2	44.7 ± 0.2	18.3 ± 0.3	55.3 ± 1.4	10.2 ± 0.2
(vii) 提案手法	**50.9** ± 0.8	**50.7** ± 0.7	**22.7** ± 0.4	**64.8** ± 4.0	**11.5** ± 1.1

スト集合にて評価した．

5. 実験結果

5.1 定量的結果

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．実験はそれぞれ 5 回ずつ行い，その平均及び標準偏差を示した．また，表中の太字の数値は各指標における最も高い数値を表す．本研究では GRIT [Nguyen 22], BLIP-2 [Li 23]. Gemini [Reid 24], GPT-4o [Achiam 23] をベースライン手法とした．画像キャプショニングにおける代表的な手法のため GRIT 及び BLIP-2 を選定した．また，Gemini，GPT-4o は多くの Vision & Language タスクで良好な結果が報告されている代表的な MLLM であるため用いた．本研究では評価尺度として BLEU4，CIDEr [Vedantam 15]，SPICE [Anderson 16] 及び Polos [Wada 24] を用いた．主要尺度は学習ベースの自動評価尺度である Polos に加え，画像キャプション生成において標準的な自動評価尺度である SPICE 及び CIDEr とした．画像キャプショニングで標準的な自動評価尺度であるため BLEU4,CIDEr，及び SPICE を利用した．また，Polos は人間の評価と相関係数が高い尺度であることから利用した．

表 1 より，提案手法は全ての自動評価尺度においてベースライン手法群を上回った．具体的には，提案手法は，Polos において，最もスコアが高かったベースライン手法である BLIP-2 に比べ対象物体画像及び配置目標画像についてそれぞれ 8.5 ポイント及び 10.6 ポイント向上した．また，SPICE 及び CIDEr においても，最もスコアが高かったベースライン手法である GRIT と比較してそれぞれ 2.9 ポイント及び 5.6 ポイント向上した．使用した全ての評価指標において，ベースライン手法との性能差は統計有意であった（$p<0.05$）．

5.2 定性的結果

図 2 に提案手法及びベースライン手法 BLIP-2 及び GPT-4o の定性的結果を示す．各 (a)，(b) はそれぞれ対象物体画像，配置目標画像を示す．(i) の例では対象物体及び配置目標は赤い物体及び食器棚であった．この例について，参照文は “Move the red object on the sofa to the cupboard at the corner.” であった．一方で提案手法は，“Move the red object on the sofa to the shelf above the kitchen.” と記述した．提案手法は対象物体及び配置目標を含む指示文を適切な色と参照表現を用いて生成した．一方でベースライン手法は対象物体及び配置目標に関する記述が不適切であった．

(ii) の例では対象物体及び配置目標はクッション及び椅子であった．この例について，参照文は “Pick up the cushion on the sofa and put it on the chair near the window.” であった．提案手法による生成文は，“Could you move the brown cushion on the sofa to the wooden chair in the corner of the room?” であった．提案手法は対象物体及び配置目標を含む指示文を適切に記述した．一方でベースライン手法は色に関する記述を誤るものや存在しない物体を記述するなど不適切な指示文を出力した．

5.3 Ablation Studies

表 1 に提案手法における Ablation Studies の結果を示す．Ablation 条件は以下の 2 条件とした．

HCCP Ablation. 本モデルは異なる損失関数を用いた 2 つのステージ，PDMP と HCCP により訓練を行った．ここで，2 つ目のステージ HCCP を取り除くことで HCCP の有効性を調査した．表 1 に示したようにモデル (v) はモデル (vii) と比較して，Polos スコアで 6.2 及び 6.3 ポイント低くなった．本結果より，HCCT による訓練はより品質の高い指示文が生成されるように適切な学習を行うことができていると言える．

Triplet Qformer Ablation. Triplet Qformer を Qformer [Li 23] に置き換えることで Triplet Qformer の有効性を検証した．表 1 に示したようにモデル (vi) はモデル (vii) と比較して，Polos スコアで 4.7 及び 6.0 ポイント低下した．これらの結果から、Triplet Qformer の導入により 2 種類の視覚特徴がテキスト特徴と適切にアライメントが行われていることが確認できた．

5.4 データセット拡張実験

表 2: データ拡張実験における定量的比較結果

[%]	条件	HM3D-FC (unseen)			
		MRR↑	R@5↑	R@10↑	R@20↑
(i)	full	20.2± 1.7	28.6± 3.3	48.7± 3.7	73.9± 3.0
(ii)	full + Aug	**22.2**± 1.5	**32.0**± 4.1	**51.3**± 2.7	**76.2**± 3.9

提案手法で生成した物体移動指示文の有効性を検証するためにデータセット拡張実験を行った．本実験では IROV-FC task [Korekata 25] を扱う．本実験では評価尺度として画像検索タスク [Kaneda 24,Korekata 25] において標準的な指標である mean reciprocal rank (MRR) 及び recall@K (R@K) を使用した．本実験では IROV-FC タスクに使用できる代表的な手法である MultiRankIt [Kaneda 24] を LTRRIE-FC dataset [Korekata 25] を用いて学習を行った．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．本実験では HM3D-FC dataset の訓練セットに対して提案手法を用いて生成した生成文を用いて Aug-train dataset を作成した．Table 2 は本実験における定量的な結果を示す．モデル (i) 及び (ii) は MRR において 20.2 及び 22.2 ポイントであった．データセット拡張を行なったデータセットにより学習したモデル (ii) が拡張を行なっていないデータセットにより学習したモデル (i) を MRR において

(a) (b) (a) (b)

Ref: "Move the red object on the sofa to the cupboard at the corner."
BLIP-2: "Move the white curtain on the left side of the window to the white shelf on the right side of the window."
GPT-4o: "Take the two glass figurines from the wooden ledge and place them on the kitchen counter next to the sink."
Ours: "Move the red object on the sofa to the shelf above the kitchen."

(i)

Ref: "Pick up the cushion on the sofa and put it on the chair near the window."
BLIP-2: "Could you move the cushion on the sofa and put it on the brown sofa?"
GPT-4o: "Take the brown cushion from the couch and place it on the counter next to the red bottle in the kitchen."
Ours: "Could you move the brown cushion on the sofa to the wooden chair in the corner of the room?"

(ii)

図 2: 提案手法における成功例．(a)，(b) はそれぞれの対象物体画像，配置目標画像を示す．

表 3: 実機実験における定量的結果

[%]	条件	MRR↑	R@10↑	SR↑
(i)	full	21	51	45 (9/20)
(ii)	full + Aug	**23**	**58**	**55** (11/20)

2.0 ポイント上回った．同様にモデル (ii) はモデル (i) を R@K においても上回った．本実験により提案手法によって生成した指示文により拡張したデータセットが既存の自然言語理解モデルの性能向上に寄与したことを確認した．

5.5 実機実験

本実験ではデータセット拡張実験で学習したモデル (i) 及び (ii) を用いて実機実験を行うことで，提案手法の生成文の有効性を検証する．本実験では，ロボットがユーザーによって入力された指示文に基づき物体操作を行う．4 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 20 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る．実機実験の設定は DM^{2}RM [Korekata 25] に従った． 評価指標として MRR, R@10 及びタスクの成功率 (SR) を用いた．Table 3 に本実験の定量的結果を示す．データ拡張を行なったデータセットで学習したモデル (ii) は 55%の SR を記録し，拡張を行なっていないデータセットで学習したモデル (i) と比較して 10 ポイント上回った．本実験により提案手法によって拡張したデータセットが既存の自然言語理解モデルの実機実験における性能向上に寄与したことが示された．

6. おわりに

本研究では対象物体画像, 及び配置目標画像に基づく物体操作指示文生成タスクを扱った．対象物体画像及び配置目標画像の複数画像を適切に扱うため，複数の視覚的特徴量をそれぞれにテキスト特徴とアライメントを行う Triplet Qformer を導入した．また，より人間の付与する文に近い指示文を生成するため，学習ベース及び n-gram ベースの自動評価尺度に基づく訓練手法 HCCP を導入した．提案手法は，全評価尺度において，MLLM を含むベースライン手法を上回った．また，本モデルを用いて物体操作指示文のデータ拡張を行うことで，既存の自然言語理解モデルの性能が，データセット拡張および実環境での実験の両方において向上することを確認した．

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

[Achiam 23] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., et al.: GPT-4 Technical Report, *arXiv preprint arXiv:2303.08774* (2023)

[Anderson 16] Anderson, P., Fernando, B., et al.: SPICE: Semantic Propositional Image Caption Evaluation, in *ECCV*, pp. 382–398 (2016)

[Goko 24] Goko, M., et al.: Task Success Prediction for Open-Vocabulary Manipulation Based on Multi-Level Aligned Representations, in *CoRL* (2024)

[Kaneda 24] Kaneda, K., 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)

[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)

[Li 23] Li, J., Li, D., et al.: BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models, in *ICML*, pp. 19730–19742 (2023)

[Nguyen 22] Nguyen, V., et al.: GRIT: Faster and Better Image Captioning Transformer Using Dual Visual Features, in *ECCV*, pp. 167–184 (2022)

[Reid 24] Reid, M., Savinov, N., Teplyashin, D., et al.: Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context, *arXiv preprint arXiv:2403.05530* (2024)

[Vedantam 15] Vedantam, R., Zitnick, C., and Parikh, D.: CIDEr: Consensus-based Image Description Evaluation, in *CVPR*, pp. 4566–4575 (2015)

[Wada 24] Wada, Y., et al.: Polos: Multimodal Metric Learning from Human Feedback for Image Captioning, in *CVPR*, pp. 13559–13568 (2024)

[Yashima 25] Yashima, D., et al.: Open-Vocabulary Mobile Manipulation Based on Double Relaxed Contrastive Learning With Dense Labeling, *IEEE RA-L*, pp. 1–8 (2025)