自動評価尺度を用いた強化学習および マルチモーダル基盤モデルに基づく物体操作指示文生成

○勝又圭，神原元就，杉浦孔明（慶應義塾大学）

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．多くの画像キャプショニング手法は複数画像を入力として扱う機構を持ち合わせていない．そこで本研究では，対象物体画像と配置目標画像を適切に扱うことを可能にする Dual Image Caption Generator モジュールを導入する．また学習ベースの自動評価尺度に基づく最適化を行う訓練手法 Human Centric Calibration Phase を導入する．結果として，提案手法は全ての自動評価尺度において，マルチモーダル大規模言語モデルを含む全てのベースライン手法を上回った．

1. はじめに

高齢化社会では被介助者の増加に伴い介助者不足が社会問題となり得る．ロボットによる日常生活や家事における支援は、介助負担の軽減や被介助者の生活の質向上に繋がる．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 [1,2] 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．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．また，SCST [3] を拡張した学習ベースの評価尺度に基づく訓練手法 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．

- 対象物体画像及び配置目標画像を適切に扱う Dual Image Caption Generator モジュールを導入する．
- 画像の複数モダリティの特徴量を統合した新たな Grid，Region 特徴量を抽出するモジュール Holo Grid Vision Encoder 及び Multimodal Region Encoder を導入する．
- 学習ベースの自動評価尺度に基づく訓練手法 HCCP を導入する．

2. 関連研究

Vison & Language とロボティクスの複合研究は広く行われており，これらについていくつかのサーベイ論文が存在する [4,5]．また，本タスクで扱う物体操作指示文生成に関連する分野として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あげられる．例えば，ユーザーの指示に基づき物体操作タスクを実行するための生活支援ロボットのベンチマーク競技が広く行われている [6–8]．また VLN [9] は，ロボットが画像と自然言語指示文により指定された物体あるいは領域へのナビゲーションを行うタスクである．本論文に関連が深い分野として画像キャプショニング分野があげられる．画像キャプショニングは与えられた画像に基づき文を生成するタスクである．2 枚の画像を入力として扱う画像キャプショニングタスクとして，2 枚の画像間の違いについての説明文を生成する Scene change captioning タスク [10,11] が挙げられる．

(a) (b)

図 1: 物体操作指示文生成タスクの例． (a),(b) はそれぞれ対象物体画像, 配置目標画像を示す．

3. 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 orange cushion on the sofa to the wooden table in the living room.”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．

4. 提案手法

本研究では，物体操作指示文生成モデルを提案する．図 2 に提案手法の概要図を示す．提案手法は主に Multimodal Region Encoder，Holo Grid Vision Encoder，Dual Image Caption Generator の 3 つのモジュールから構成される. モデルの入力を $\boldsymbol{x} = \{\boldsymbol{X}_{\mathrm{tar}}, \boldsymbol{X}_{\mathrm{dst}}\}$ と定義する．ここで $\boldsymbol{X}_{\mathrm{tar}}$ 及び $\boldsymbol{X}_{\mathrm{dst}}$ はそれぞれ対象物体画像及び配置目標画像を示す．

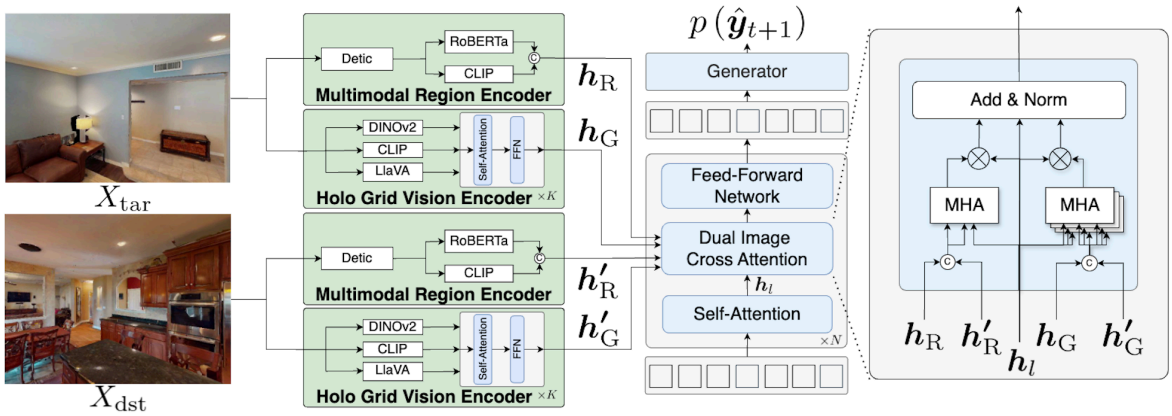


図 2: 提案手法のネットワーク図．⊗，ⓒ及び MHA はゲート機構，連結及び Multi Head Attention を表す．

4.1 Multimodal Region Encoder

Multimodal Region Encoder は物体に関する情報を含む特徴量である，Region 特徴量を取得するモジュールである．Region 特徴量が用いられる既存手法では物体に関する画像特徴量のみを用いる場合が多い．物体検出器の予測ラベルに基づくテキスト特徴量は，多くの場合有効に用いられていなかったものの，物体の存在有無が重要な指示文生成において有効な特徴量であると考えられる．よって本モジュールでは，物体検出で得られた物体領域特徴量と予測ラベルに基づくテキスト特徴量の 2 つのモダリティから得られる Region 特徴量を統合した Multimodal Region 特徴量を獲得する．

本モジュールの入力 $\mathbf{X}_{\text{tar}}$，$\mathbf{X}_{\text{dst}}$ から予測ラベルつき物体領域群 $\boldsymbol{D}$，$\boldsymbol{D'}$ を得る．獲得した $\boldsymbol{D}$，$\boldsymbol{D'}$ に対して画像特徴量 $\boldsymbol{v}_{\text{De}}$，$\boldsymbol{v}'_{\text{De}}$ 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 $\boldsymbol{s}$ 及び $\boldsymbol{s}'$ を抽出する．そしてそれぞれ画像 $\mathbf{X}_{\text{tar}}$ 及び $\mathbf{X}_{\text{dst}}$ に対応する本モジュールの出力である Region 特徴量 $\boldsymbol{h}_R = [\boldsymbol{v}_{\text{De}}; \boldsymbol{s}]$ 及び $\boldsymbol{h}'_R = [\boldsymbol{v}'_{\text{De}}; \boldsymbol{s}']$ が獲得される．

4.2 Holo Grid Vision Encoder

Holo Grid Vision Encoder 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．Holo Grid Vision Encoder は 3 種類の潜在表現を組み合わせた λ-Representation [12] を拡張した Holo Grid 特徴量を抽出する．

本モジュールの入力 $\mathbf{X}_{\text{tar}}$，$\mathbf{X}_{\text{dst}}$ に対応する Holo Grid 特徴量 $\boldsymbol{h}_{\text{G}}, \boldsymbol{h}'_{\text{G}}$ を出力する．まず，画像 $\boldsymbol{X}_{\text{tar}}$ について，画像特徴量 $\boldsymbol{v}_{\text{D}}$，$\boldsymbol{v}_{\text{C}}$ ，$\boldsymbol{v}_{\text{L}}$ をそれぞれ抽出する．同様に $\boldsymbol{X}_{\text{dst}}$ についても $\boldsymbol{v}'_{\text{D}}$，$\boldsymbol{v}'_{\text{C}}$，$\boldsymbol{v}'_{\text{L}}$ を抽出する．本モジュールの出力である画像 $\boldsymbol{X}_{\text{tar}}$，及び $\boldsymbol{X}_{\text{dst}}$ に対応する Holo Grid 特徴量を $\boldsymbol{h}_{\text{G}} = \{\boldsymbol{h}_{\text{G,D}}, \boldsymbol{h}_{\text{G,C}}, \boldsymbol{h}_{\text{G,L}}\}$ 及び $\boldsymbol{h}'_{\text{G}} = \{\boldsymbol{h}'_{\text{G,D}}, \boldsymbol{h}'_{\text{G,C}}, \boldsymbol{h}'_{\text{G,L}}\}$ と表す．ここで，$\boldsymbol{h}_{\text{G,D}}, \boldsymbol{h}_{\text{G,C}}, \boldsymbol{h}_{\text{G,L}}$, $\boldsymbol{h}'_{\text{G,D}}, \boldsymbol{h}'_{\text{G,C}}$及び $\boldsymbol{h}'_{\text{G,L}}$ は $\boldsymbol{v}_{\text{D}}, \boldsymbol{v}_{\text{C}}, \boldsymbol{v}_{\text{L}}, \boldsymbol{v}'_{\text{D}}, \boldsymbol{v}'_{\text{C}}$及び $\boldsymbol{v}'_{\text{L}}$ を K 層の Self-Attention 層で計算を行い獲得する．

4.3 Dual Image Caption Generator

Dual Image Caption Generator は $\boldsymbol{h}_{\text{G}}$，$\boldsymbol{h}'_{\text{G}}$，$\boldsymbol{h}_{\text{R}}$，$\boldsymbol{h}'_{\text{R}}$ を基に，指示文を生成するモジュールである．本モジュールは L 層の Transformer 層で構成されている．各層は Self-Attention 層，Dual Image Cross Attention 層，及び Feed Forward Network (FFN) 層からなる．

まず，Self-attention 層では，時刻 0 から t の予測単語系列の埋め込み特徴量 $\hat{\boldsymbol{y}}_{0:t}$ について，埋め込み行列を用いて言語特徴量 $\boldsymbol{h}_l$ を抽出する．$\boldsymbol{h}_l$ について，Self-attention を計算し，特徴量 $\boldsymbol{h}'_l$ を得る．図 2 に示すように，Dual Image Cross Attention 層は 4 つの Multi Head Attention (MHA) ブロック及び正規化層を持つ．4 つの MHA ブロックでは図のように，それぞれの視覚特徴量 $\boldsymbol{H}_{\text{G,D}} = [\boldsymbol{h}_{\text{G,D}}; \boldsymbol{h}'_{\text{G,D}}]$，$\boldsymbol{H}_{\text{G,C}} = [\boldsymbol{h}_{\text{G,C}}; \boldsymbol{h}'_{\text{G,C}}]$，$\boldsymbol{H}_{\text{G,L}} = [\boldsymbol{h}_{\text{G,L}}; \boldsymbol{h}'_{\text{G,L}}]$，$\boldsymbol{H}_{\text{R}} = [\boldsymbol{h}_{\text{R}}; \boldsymbol{h}'_{\text{R}}]$ をキー及びバリュー，言語特徴量 $\boldsymbol{h}'_l$ をクエリとして Cross-attention を計算する．ここで，各ブロックのパラメータは共有しない．各 MHA の出力として $\boldsymbol{a}_m$ $(m = 1, ..., 4)$ を得る．最終的に特徴量 $\boldsymbol{h}_a = \text{LN}\left(\sum_{m=1}^{4} \boldsymbol{c}_m \otimes \boldsymbol{a}_m + \boldsymbol{h}_l\right)$ を獲得する．ここで $\boldsymbol{c}_m = \text{sigmoid}\left(W[\boldsymbol{a}_m; \boldsymbol{h}_l] + \boldsymbol{b}_m\right)$ であり，$\text{sigmoid}\left(\cdot\right)$, W, b, $\text{LN}\left(\cdot\right)$ はそれぞれシグモイド関数，訓練可能な重み行列，バイアス項及びレイヤー正規化を表す．最終的に $\boldsymbol{h}_a$ について，FFN 層及びソフトマックス関数を適用することで，トークン $\hat{\boldsymbol{y}}_{t+1}$ についての予測確率 $p(\hat{\boldsymbol{y}}_{t+1}|\boldsymbol{x}, \hat{\boldsymbol{y}}_{1:t})$ を得る．

4.4 訓練方法

本モデルは、異なる損失関数を用いる 2 つのステージ，Probability Distribution Matching Phase (PDMP) と HCCP によって訓練する．PDMP ではクロスエントロピー関数を用いた訓練を行う．本ステージにおいて，検証集合で対象物体画像および配置目標画像に対応する Polos [15] の平均値が一番高いモデルを選択する．選択したモデルについて，HCCP において追加の訓練を行う．HCCP では SCST [3] を拡張した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 $\mathcal{L}_{\text{HCCT}}$ を以下のように定義する．

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

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

表 1: ベースライン手法との定量的比較結果.

手法	Polos		SPICE	CIDEr	BLEU4
	対象物体画像	配置目標画像			
GRIT [1]	41.02 ±1.80	38.82 ±1.81	20.20 ±0.88	61.35 ±5.85	11.32 ±0.71
BLIP-2 [2]	41.05 ±1.33	43.39 ±1.26	17.13 ±0.53	38.03 ±5.04	8.40 ±1.18
GPT-4V [13]	38.91 ±0.40	39.28 ±0.36	14.79 ±0.41	23.07 ±0.91	5.88 ±0.30
Gemini [14]	29.16 ±0.50	29.26 ±0.53	10.97 ±0.51	25.56 ±1.55	5.04 ±0.40
提案手法	**49.16** ±0.31	**49.11** ±0.34	**24.48** ±0.55	**79.46** ±1.28	**14.21** ±0.52

れ $r(\mathbf{w}_i) = \text{mean}\left(\text{mean}\left(\mathrm{P}_{\text{tar}}(\mathbf{w}_i), \mathrm{P}_{\text{dst}}(\mathbf{w}_i)\right), \mathrm{C}(\mathbf{w}_i)\right)$ 及び $b = \frac{1}{k}\sum_{i=1}^{k} r(\mathbf{w}_i)$ と定義する．ここで $\mathrm{P}_{\text{tar}}(\cdot)$, $\mathrm{P}_{\text{dst}}(\cdot)$ 及び $\mathrm{C}(\cdot)$ は対象物体画像に対する Polos，配置目標画像に対する Polos 及び CIDEr [16] 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，Polos を組み合わせることで，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．

5. 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 [17] を用いた．本研究では最適化手法として Adam を採用し，β_1 が 0.9、β_2 が 0.99 とした．HCCP で使用されるビームサーチにおいてビームサイズは 5，ビーム長は 20 である．学習率，バッチサイズ及びエポック数は PDMP では 1.0×10^{-4}，32 及び 20，HCCP では 5.0×10^{-6}，16 及び 10 とした．

本提案手法の訓練可能なパラメータ数は約 268M, 積和演算数は 7.54G であった．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 14 時間，1 サンプルあたりの推論には 13.81 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．検証集合で対象物体画像および配置目標画像に対応する Polos [15] が一番高いモデルを選択し，テスト集合にて評価した．

6. 実験結果

6.1 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 [1], BLIP-2 [2]. Gemini [14], GPT-4V [13] をベースラインとした．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-4V 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 [16], SPICE [18] 及び Polos [15] 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 は我々の知る限り画像キャプション生成タスクのための自動評価尺度として最も人間の評価に近い尺度であることから利用した．

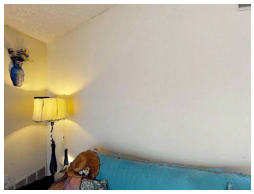
(i)

(a)

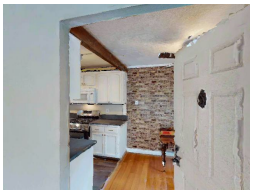
(b)

Ref: “take the white pillow on the bed to the fireplace in the living room”
Baseline: “move the pillow on the bed to the white shelf in the kitchen”
Proposed: “take the white pillow on the bed to the fireplace in the living room”

(ii)



(a)



(b)

Ref: “put the lamp into the upper of the white shelf”
Baseline: “move the pillow on the bed to the white sofa in the living room”
Proposed: “could you move the light on the shelf to the kitchen”

図 3: 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 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 8.16 ポイント及び 5.80 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 4.28 ポイント及び 18.11 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（$p<0.05$）．

6.2 定性的結果

図 3 に提案手法及びベースライン手法 [1] 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白いクッション及び暖炉であった．この例について，参照文は “take the white pillow on the bed to the fireplace in the living room” であった．一方で提案手法は，“take the white pillow on the bed to the fireplace in the living room” と記述した．提案手法は対象物体

表 2: 本提案手法における Ablation Studies の結果.

Ablation Condition	Polos		SPICE	CIDEr	BLEU4
	対象物体画像	配置目標画像			
(i) w/o HCCP	43.89 ±1.16	43.67 ±0.61	24.11 ±0.47	65.76 ±3.94	11.54 ±0.17
(ii) w/o Multimodal Region Encoder	47.46 ±0.43	47.39 ±0.43	22.12 ±0.76	68.28 ±3.97	12.61 ±0.75
(iii) w/o Holo Grid Visual Encoder	48.52 ±0.47	48.54 ±0.45	23.26 ±0.37	72.79 ±4.24	13.31 ±0.35
(iv) Full	**49.16** ±0.31	**49.11** ±0.34	**24.48** ±0.55	**79.46** ±1.28	**14.21** ±0.52

及び配置目標を含む指示文を適切に生成した．一方でベースライン手法は配置目標として白い棚を記述したが存在せず不適切であった．

(ii) の例では対象物体及び配置目標はランプ及び白い棚の上部であった．この例について，参照文は “put the lamp into the upper of the white shelf”であった．提案手法による生成文は，“could you move the light on the shelf to the kitchen”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対象物体及び配置目標にソファの上の枕及びリビングのソファを記述した．対象物体及び配置目標はいずれも存在しない物体であり不適切な指示文と言える．

6.3 Ablation Studies

表 2 に提案手法における Ablation Studies の結果を示す．Ablation 条件は以下の 3 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 2 に示したようにモデル (i) はモデル (iv) と比較して，Polos スコアで 5.27 及び 5.44 ポイント，SPICE 及び CIDEr スコアで 0.37 及び 13.7 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．

Multimodal Region Encoder Ablation． Multimodal Region Encoder を取り除くことで，Multimodal Region 特徴量の有効性を調査した．表 2 に示したようにモデル (ii) はモデル (iv) と比較して，Polos スコアで 1.7 及び 1.72 ポイント低かった．Region 特徴量を取り除いたことで，物体レベルの情報がなくなり，物体の存在を適切に認識することが難しかったと考えられる．

Holo Grid Visual Encoder Ablation． Holo Grid Visual Encoder を取り除くことで，Holo Grid Visual Encoder による Holo Grid 特徴量の有効性を調査した．表 2 に示したようにモデル (iii) はモデル (iv) と比較して，Polos スコアで 0.64 及び 0.57 ポイント低下した．これは，Holo Grid 特徴量により補完される物体間の関係などの文脈情報が欠落したことでモデルの性能が低下したためだと考えられる．

7. 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 Dual Image Caption Generator を導入した．また，学習ベースの自動評価尺度を適用した新しい訓練手法 HCCP を導入し，画像の複数次元の特徴量を統合するモジュール，Holo Grid Vision Encoder 及び Multimodal Region Encoder を提案した．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．

謝辞

本研究の一部は，JSPS 科研費 23K28168， JST ムーンショット， NEDO， 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] V. Nguyen, M. Suganuma, and T. Okatani, “GRIT: Faster and Better Image Captioning Transformer Using Dual Visual Features,” ECCV, pp.167–184, 2022.
[2] J. Li, D. Li, et al., “BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models,” ICML, pp.19730–19742, 2023.
[3] S. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, “Self-Critical Sequence Training for Image Captioning,” CVPR, pp.7008–7024, 2017.
[4] R. Firoozi, J. Tucker, S. Tian, A. Majumdar, et al., “Foundation Models in Robotics: Applications, Challenges, and the Future,” arXiv preprint arXiv:2312.07843, 2023.
[5] K. Kawaharazuka, T. Matsushima, A. Gambardella, et al., “Real-World Robot Applications of Foundation Models: A Review,” arXiv preprint arXiv:2402.05741, 2024.
[6] L. Iocchi, et al., “RoboCup@ Home: Analysis and Results of Evolving Competitions for Domestic and Service Robots,” Artificial Intelligence, vol.229, pp.258–281, 2015.
[7] H. Okada, T. Inamura, et al., “What Competitions Were Conducted in the Service Categories of the World Robot Summit?,” AR, vol.33, no.17, pp.900–910, 2019.
[8] S. Yenamandra, A. Ramachandran, K. Yadav, A. Wang, M. Khanna, et al., “HomeRobot: Open Vocabulary Mobile Manipulation,” CoRL, pp.1975–2011, 2023.
[9] P. Anderson, Q. Wu, et al., “Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments,” CVPR, pp.3674–3683, 2018.
[10] Z. Guo, T. Wang, and J. Laaksonen, “CLIP4IDC: CLIP for Image Difference Captioning,” AACL/IJCNLP, pp.33–42, 2022.
[11] L. Yao, W. Wang, and Q. Jin, “Image Difference Captioning with Pre-training and Contrastive Learning,” AAAI, vol.36, pp.3108–3116, June 2022.
[12] 齋藤大地他, “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,” 第 38 回人工知能学会全国大会資料 3O1-OS-16b-02，2024.
[13] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. Aleman, D. Almeida, et al., “GPT-4 Technical Report,” arXiv preprint arXiv:2303.08774, 2023.
[14] M. Reid, N. Savinov, et al., “Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context,” arXiv preprint arXiv:2403.05530, 2024.
[15] Y. Wada, K. Kaneda, D. Saito, and K. Sugiura, “Polos: Multimodal Metric Learning from Human Feedback for Image Captioning,” CVPR, pp.13559–13568, 2024.
[16] R. Vedantam, C. Zitnick, and D. Parikh, “CIDEr: Consensus-based Image Description Evaluation,” CVPR, pp.4566–4575, 2015.
[17] 是方諒介，兼田寛大，長嶋隼矢，今井悠人，杉浦孔明，“大規模言語モデルを用いた Switching 機構付きマルチモーダル検索モデルに基づく生活支援ロボットによる物体操作,” 第 38 回人工知能学会全国大会資料 3T5-OS-6b-04，2024.
[18] P. Anderson, B. Fernando, M. Johnson, and S. Gould, “SPICE: Semantic Propositional Image Caption Evaluation,” ECCV, pp.382–398, 2016.