Multimodal LLM と二重緩和損失に基づく実世界検索エンジン

○八島大地，是方諒介，杉浦孔明（慶應義塾大学）

本研究では，open-vocabulary な指示文に基づき日常物体を指定された家具に運ぶ，生活支援ロボットの構築を目的とする．特に，ロボットの事前探索を通じて屋内環境で収集した画像から，対象物体および配置目標の画像を検索するタスクに焦点を当てる．類似物体が多数存在する環境では，対象物体と配置目標の両方を含む指示文に基づき正しく画像を検索することは困難である．本問題に対処するため，正解画像との類似性が高く部分的に正しいとみなされる画像に対して soft positive ラベルを付与し，それらの埋め込み関係を最適化する損失関数を提案する．標準的な屋内環境で撮影された実画像および参照表現を含む指示文から成るデータセットを用いて提案手法を評価した結果，画像検索設定における標準的な評価指標においてベースライン手法を上回った．

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．ロボットに対して自然言語で日常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，open-vocabulary 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．図 1 に本タスクの具体例を示す．まず，ロボットは事前探索を通じて屋内環境の画像を収集する．“Please get the right red towel hanging on the metal towel rack and put it in the white washing machine on the left.” という指示文が与えられた場合，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「金属製のタオルラックの右に掛かっている赤いタオル」および「左側の白い洗濯機」を上位にランク付けすることが望ましい．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 open-vocabulary 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 5.1 節で示すように，CLIP [1] 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 [2–4]．[3, 4] 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 [5] 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 [1, 3, 5] は主に InfoNCE [6] 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 soft 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，soft 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．本研究の新規性は以下である．

- Soft Labeler を用いて ground truth (GT) 画像に類似した画像に soft positive ラベルを付与し，Soft Positive Contrastive (SPC) 損失関数を用いて positive, soft positive, および negative ペア間の関係を最適化する Soft Positive Augmented Contrastive (SPAC) モジュールを提案する．
- 領域分割された画像にマークを付けるプロンプト手法を使用し，領域ごとに詳細な特徴抽出を行うマルチモーダル大規模言語モデル (MLLM) を用いた Spatial Overlay Grounding (SOG) モジュールを提案する．
- 色，テクスチャ，および形状などの特徴を捉えるための visual encoder ，言語特徴とアラインされた特徴量を抽出する multimodal encoder，言語特徴および視覚特徴の両方をもつ潜在特徴を抽出する MLLM から埋め込みを取得し，SOG モジュールからの出力と統合して包括的な視覚特徴を取得する X-Fusion (XF) モジュールを導入する．

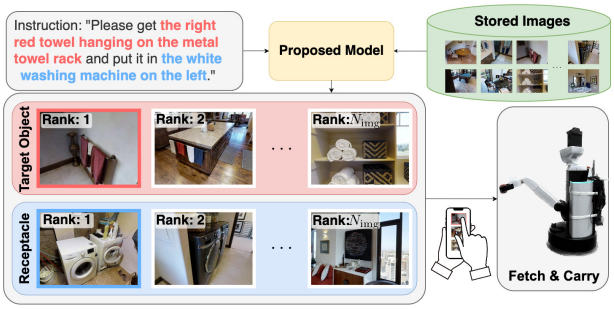

図 1 IROV-FC タスク [5] の具体例

2. 問題設定

本研究では，Image Retrieval-based Open-Vocabulary Fetch-and-Carry (IROV-FC) タスク [5] を扱う．本タスクではロボットが open-vocabulary 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．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．

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 SOG，XF，および SPAC の主に 3 つのモジュールから構成される．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，

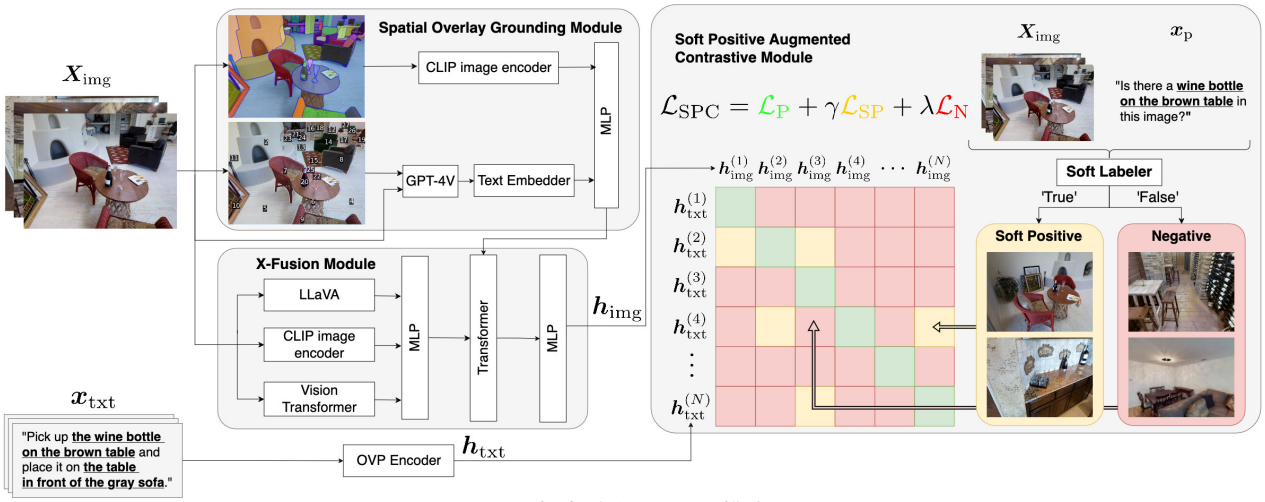

図 2　提案手法のモデル構造

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．

3.1　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： [7,8]) 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．

本モジュールは $\boldsymbol{X}_{\text{img}}$ を入力とする．まず，領域分割のための基盤モデルである SAM [7] および SEEM [8] を用いて得られた領域分割マスクを $\boldsymbol{x}_{\text{img}}^{(i)}$ に重畳することで，それぞれ $\boldsymbol{x}_{\text{GS}}^{(i)} \in \mathbb{R}^{3\times W\times H}$ および $\boldsymbol{x}_{\text{SS}}^{(i)} \in \mathbb{R}^{3\times W\times H}$ を得る．次に，事前学習された CLIP image encoder [1] を用いて $\boldsymbol{x}_{\text{GS}}^{(i)}$ から視覚特徴量 $\boldsymbol{v}_{\text{GS}}^{(i)} \in \mathbb{R}^{d_{\text{GS}}}$ を得る．ここで d_{GS} は出力次元数である．その後，$\boldsymbol{x}_{\text{SS}}^{(i)}$ の各領域分割マスクに対して，数字のついたマークを付与することで $\boldsymbol{x}_{\text{SGM}}^{(i)} \in \mathbb{R}^{3\times W\times H}$ を得る．各領域分割マスクの中心点に対してマークが割り当てられ，配置される．ただし，マークは最小面積の領域から順に番号が振られ，他のマスクと領域が重複しているものは除かれる．これは [9] で議論されているように，マークとマスク領域の重なりを回避するために行う．また，画像内の家具などのレイアウトの説明を指示するプロンプト，$\boldsymbol{x}_{\text{img}}^{(i)}$，および $\boldsymbol{x}_{\text{SS}}^{(i)}$ を GPT-4V [10] に入力し，画像に対する説明文を取得する．説明文は text-embedding-large-3 で埋め込まれたのち，多層パーセプトロンに入力され，出力として視覚特徴量 $\boldsymbol{v}_{\text{SGM}}^{(i)} \in \mathbb{R}^{d_{\text{SGM}}}$ を得る．ここで d_{SGM} は出力次元数を表す．GPT-4V による画像説明能力を強化するために，入力として $\boldsymbol{x}_{\text{img}}^{(i)}$ および $\boldsymbol{x}_{\text{SGM}}^{(i)}$ の 2 枚を用いることで，マークが物体の重要な領域と重なることや，マスクされた色を実際の色と誤認識することを防ぐ．最後に $\boldsymbol{v}_{\text{GS}}^{(i)}$ および $\boldsymbol{v}_{\text{SGM}}^{(i)}$ を連結して多層パーセプトロンに入力し，本モジュールの出力 $\boldsymbol{h}_{\text{SOG}} \in \mathbb{R}^{d_{\text{SOG}}}$ を得る．ここで d_{SOG} は出力次元数を示す．

3.2　XF

XF モジュールでは，visual encoder (例：ViT [11])，multimodal encoder (例：CLIP)，および潜在特徴量が得られる MLLM (例：LLaVA [12]) 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する．

本モジュールへの入力は $\boldsymbol{X}_{\text{img}}$ である．まず，事前学習された vision encoder (ViT) を用いて視覚特徴量 $\boldsymbol{v}_{\text{L}}^{(i)} \in \mathbb{R}^{d_{\text{L}}}$ を得る．次に，multimodal encoder (CLIP image encoder) を用いて言語にアラインされた視覚特徴量 $\boldsymbol{v}_{\text{M}}^{(i)} \in \mathbb{R}^{d_{\text{M}}}$ を得る．最後に，$\boldsymbol{x}_{\text{img}}^{(i)}$ とこれを説明させることを指示するプロンプトを MLLM (LLaVA-v1.6-mistral-7b) に入力し，潜在特徴量 $\boldsymbol{v}_{\text{lat}}^{(i)} \in \mathbb{R}^{d_{\text{lat}}}$ を得る．そして，$\boldsymbol{v}_{\text{lat}}^{(i)}$ を多層パーセプトロンに入力し，$\boldsymbol{v}_{\text{H}}^{(i)} \in \mathbb{R}^{d_{\text{H}}}$ を得る．ここで，d_{L}，d_{M}，d_{lat}，および d_{H} は出力次元数である．

本モジュールの出力である画像特徴量 $\boldsymbol{h}_{\text{img}} \in \mathbb{R}^{d_{\text{img}}}$ は次のようにして得られる．まず，$\boldsymbol{v}_{\text{L}}$，$\boldsymbol{v}_{\text{M}}$，$\boldsymbol{v}_{\text{H}}$，および $\boldsymbol{h}_{\text{SOG}}$ が結合されたベクトルを transformer に入力する．得られた出力を多層パーセプトロンに入力し，最終的な画像特徴量 $\boldsymbol{h}_{\text{img}}$ を得る．ここで d_{img} は出力次元数を表す．

3.3　SPAC

SPAC モジュールは，SPC 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，soft 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 [6] を損失関数として利用している [1]．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，soft positive として扱われることが望ましい．したがって，soft positive

表 1　ベースライン手法，提案手法，および Ablation studies における定量的比較結果

	手法	モデル	HM3D-FC			MP3D-FC		
			R@5↑[%]	R@10↑[%]	R@20↑[%]	R@5↑[%]	R@10↑[%]	R@20↑[%]
(i)	CLIP [1]		26.2	44.2	70.8	34.4	54.0	74.7
(ii)	MultiRankIt [3]		28.7 ±3.4	48.3 ±3.4	73.3 ±2.6	35.7 ±9.9	51.7 ±8.9	72.7 ±3.3
(iii)	DM^{2}RM [5]		47.8 ±1.2	67.1 ±2.4	87.0 ±1.1	49.6 ±0.7	64.1 ±3.6	78.5 ±0.5
(iv-a)	提案手法	w/o SPAC	52.5 ±1.4	73.5 ±1.1	91.4 ±0.7	55.4 ±0.7	69.1 ±1.3	80.9 ±1.3
(iv-b)		w/o SOG	52.4 ±2.1	73.6 ±0.7	91.1 ±0.8	54.2 ±1.7	69.8 ±0.7	80.8 ±1.2
(iv-c)		w/o XF	51.9 ±1.4	71.6 ±1.4	86.7 ±1.9	53.2 ±1.3	69.5 ±1.3	79.8 ±1.3
(iv-d)		full	**55.4** ±0.5	**76.3** ±0.9	**91.6** ±0.9	**57.0** ±1.1	**72.4** ±0.7	**82.5** ±0.8

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，SPC 損失関数を $\mathcal{L}_{\mathrm{SPC}} = \mathcal{L}_{\mathrm{P}} + \gamma\mathcal{L}_{\mathrm{SP}} + \lambda\mathcal{L}_{\mathrm{N}}$ と定義する．ここで，γ および λ は重み係数である．また，$\mathcal{L}_{\mathrm{SPC}}$ を構成する各項はそれぞれ positive，soft 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．

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

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

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

ここで，$\mathrm{sim}(\cdot,\cdot)$，$\mathcal{S}$，および α は，それぞれコサイン類似度，soft positive に対応する添字集合，および soft 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 $\boldsymbol{h}_{\mathrm{txt}} \in \mathbb{R}^{d_{\mathrm{txt}}}$ および $\boldsymbol{h}_{\mathrm{img}}$ である．ここで $\boldsymbol{h}_{\mathrm{txt}}$ 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る．OVP encoder は [5] 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 open-vocabulary な指示文を処理し，予測の対象に応じた言語特徴量 $\boldsymbol{h}_{\mathrm{txt}}$ を出力する．詳細は [5] を参照されたい．また，$\mathcal{S}$ は $\mathcal{S} = \mathrm{Soft\ Labeler}(\boldsymbol{x}_{\mathrm{p}}, \{\boldsymbol{x}_{\mathrm{img}}^{(1)}, \ldots, \boldsymbol{x}_{\mathrm{img}}^{(N_{\mathrm{cand}})}\})$ で得られる．ここで $\boldsymbol{x}_{\mathrm{p}}$ 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：[1,3]) 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 $0 \le \mathrm{sim}(\boldsymbol{h}_{\mathrm{txt}}^{(i)}, \boldsymbol{h}_{\mathrm{img}}^{(j)}) \le 1$ を取得する．各 $\boldsymbol{x}_{\mathrm{txt}}$ 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 soft 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，Soft Labeler で得られた $\mathcal{S}$ を用いて，$\mathcal{L}_{\mathrm{SP}}$ は $\mathrm{sim}(\boldsymbol{h}_{\mathrm{txt}}^{(i)}, \boldsymbol{h}_{\mathrm{img}}^{(j)})$ 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 $\mathcal{L}_{\mathrm{N}}$ 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，$\mathcal{L}_{\mathrm{SPC}}$ 損失関数は positive, soft 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 SPC 損失関数および Soft Labeler は，他のクロスモーダル検索タスク (例：[13]) 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 $\mathrm{sim}(\boldsymbol{x}_{\mathrm{txt}}^{(i)}, \boldsymbol{x}_{\mathrm{img}}^{(j)})$ 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 $\hat{Y}_{\mathrm{targ}}$ および $\hat{Y}_{\mathrm{rec}}$ である．$\hat{Y}_{\mathrm{targ}}$ および $\hat{Y}_{\mathrm{rec}}$ 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. 実験設定

本研究では LTRRIE-FC データセット [5] を使用した．本データセットは，HM3D [14] および MP3D [15] 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 [5] 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．

5. 実験結果

5.1 定量的結果

表 1 に，LTRRIE データセット [5] 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，CLIP [1]，MultiRankIt [3]，および DM^{2}RM [5] をベースライン手法とした．CLIP は zero-shot で画像検索タスクに用いることができる代表的な手法であるため選択した．また，MultiRankIt および DM^{2}RM は本タスクと関連が深いタスク [3, 5] 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 (iv-d) 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 (iii) に対してそれぞれ 9.2 および 8.3 ポ

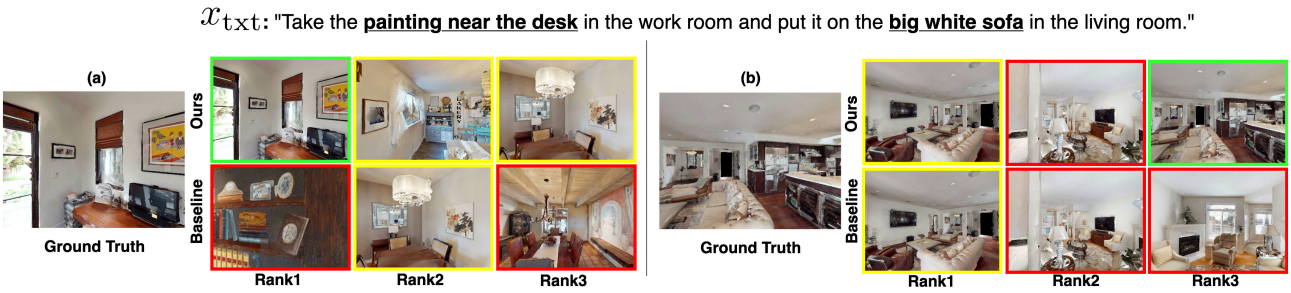

図 3　HM3D-FC テスト集合における提案手法およびベースライン手法 [5] の定性的比較結果

イント上回った．さらに，提案手法 (iv-d) 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 ($p < 0.01$)．

5.2　定性的結果

図 3 に，提案手法とベースライン手法の一つである DM^{2}RM との定性的比較結果を示す．各パネルに，それぞれ対象物体および配置目標に関しての GT 画像および上位 3 件の画像を示す．また，検索された各画像にはそのラベルを示す色が枠線に付けられている．positive ラベル，soft positive ラベル，および negative ラベルは，それぞれ緑，黄色，および赤で色分けされている．

図 3 (a) および (b) は，HM3D-FC テスト集合のサンプルである．$\boldsymbol{x}_{\mathrm{txt}}$ は “Take the painting near the desk in the work room and put it on the big white sofa in the living room.” であった．このとき (a) では，ベースライン手法が GT 画像を 6 位とランク付けしたのに対し，提案手法は GT 画像を 1 位とし，2 位および 3 位も soft positive をランク付けしている．また，(b) においてベースライン手法は同様に GT 画像を 9 位とランク付けし，上位 3 件に入っていない．一方で，提案手法は soft positive を 1 位にランク付けし，GT 画像を 3 位にランク付けしている．これは，$\mathcal{L}_{\mathrm{SPC}}$ を用いて最適化することで positive および soft positive に対する検索性能が向上したためだと考えられる．

5.3　Ablation studies

表 1 に，ablation studies における定量的比較結果を示す．Ablation studies の条件は以下の通りである．

SPAC ablation. SPAC モジュールにおいて，SPC 損失関数の有用性を検証した．モデル (iv-a) および (iv-d) 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 2.8 ポイントおよび 3.3 ポイント減少した．これにより，本タスクにおいて soft positive ラベルを付与して SPC 損失関数を用いることが有用であると示唆される．

SOG ablation. SOG モジュールを取り除き，その有用性を検証した．手法 (iv-b) および (iv-d) 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 2.7 ポイントおよび 2.6 ポイント減少した．本結果から，画像を領域分割することで，visual encoder および MLLM の接地能力が向上することが示唆される．

XF ablation. 同様に XF モジュールを取り除き，その有用性を検証した．手法 (iv-c) および (iv-d) 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 4.7 ポイントおよび 2.9 ポイント減少した．本結果から，XF モジュールを導入することにより，画像内の複雑な物体関係などを捉える能力が向上することが示唆される．

6.　おわりに

本研究では，生活支援ロボットが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 IROV-FC タスク [5] を扱った．標準的な屋内環境で撮影された実画像および参照表現を含む指示文から成るデータセットを用いて提案手法を評価した結果，画像検索設定における標準的な評価指標において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 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

[1] A. Radford, W. Kim, C. Hallacy, A. Ramesh, et al., “Learning Transferable Visual Models From Natural Language Supervision,” ICML, pp.8748–8763, 2021.
[2] R. Korekata, et al., “Switching Head-Tail Funnel UNITER for Dual Referring Expression Comprehension with Fetch-and-Carry Tasks,” IROS, pp.3865–3872, 2023.
[3] K. Kaneda, 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.
[4] G. Sigurdsson, J. Thomason, G. Sukhatme, and R. Piramuthu, “RREx-BoT: Remote Referring Expressions with a Bag of Tricks,” IROS, pp.5203–5210, 2023.
[5] 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，2024．3T5-OS-6b-04.
[6] A. Oord, Y. Li, and O. Vinyals, “Representation Learning with Contrastive Predictive Coding,” arXiv preprint arXiv:1807.03748, 2018.
[7] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. Berg, W. Lo, et al., “Segment Anything,” ICCV, pp.4015–4026, 2023.
[8] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Wang, L. Wang, J. Gao, and Y. Lee, “Segment Everything Everywhere All at Once,” NeurIPS, pp.19769–19782, 2023.
[9] J. Yang, H. Zhang, F. Li, X. Zou, et al., “Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V,” arXiv preprint arXiv:2310.11441, 2023.
[10] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, L. Aleman, D. Almeida, J. Altenschmidt, et al., “GPT-4 Technical Report,” arXiv preprint arXiv:2303.08774, 2023.
[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” ICLR, pp.12888–12900, 2021.
[12] H. Liu, C. Li, Q. Wu, and J. Lee, “Visual Instruction Tuning,” NeurIPS, pp.34892–34916, 2023.
[13] H. Wu, Y. Gao, X. Guo, Z. Al-Halah, et al., “Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback,” CVPR, pp.11307–11317, 2021.
[14] S. Ramakrishnan, A. Gokaslan, E. Wijmans, et al., “Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D Environments for Embodied AI,” NeurIPS, 2021.
[15] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niebner, M. Savva, et al., “Matterport3D: Learning from RGB-D Data in Indoor Environments,” 3DV, pp.667–676, 2017.