Dense Text を用いたマルチモーダル LLM に基づく大規模屋内環境における物体検索

○今井悠人，是方諒介，杉浦孔明（慶應義塾大学）

本研究では，open-vocabulary かつ多様な自然言語文に基づき，移動ロボットが屋内環境で撮影した対象物体の画像を検索するタスクを扱う．既存手法では，本タスクで扱う多様な屋内環境の画像および複雑な入力文の処理において，画像中のテキスト情報 (dense text) を考慮した参照表現理解に問題がある．そこで，本研究では dense text を用いてマルチモーダル大規模言語モデルから獲得される構造的な画像特徴量および入力文の複雑さおよび曖昧性に頑健な言語特徴量を提案する．約 3,000 平方メートルにわたる大規模な屋内環境から収集されたデータセットにおける実験の結果，標準的な評価尺度において提案手法がベースライン手法を上回った．

1. はじめに

医療，物流，および農業などの労働集約型産業から日常生活に至るまで，任意の物体の位置を検索し特定できるシステムは常時および非常時を問わず利便性が高い．例えば，大規模な病院や倉庫において，ユーザが入力した文に基づき移動ロボットに物体操作を自動で実行させることができれば，人間の労力を削減することが可能である．しかし，ユーザの入力した文はしばしば open-vocabulary であり，長く複雑で曖昧さを含むため，このようなシステムの実現は困難である．

本研究では，ユーザによって与えられる open-vocabulary かつ多様な自然言語文に基づき，移動ロボットが屋内環境で撮影した物体の画像を検索する learning-to-rank physical objects (LTRPO) タスク [1] を扱う．本タスクは，open-vocabulary で参照表現を含むユーザの入力文に基づき，屋内環境の様々な視覚的コンテクストにおいて撮影された物体を正確にランク付けする必要がある点が困難である．実際に，6.1 節で示すように，視覚言語基盤モデルとして代表的な CLIP [2] を本タスクへ直接的に適用するだけでは性能が不十分である．物体検索に関する研究は広く行われている [1,3,4]．しかし，既存手法では LTRPO タスクで扱う多様な屋内環境の画像および複雑な入力文の処理において，テキスト情報を考慮した参照表現理解に問題がある．

本研究では，上述の問題に対し，画像中のテキスト情報を考慮したランキング学習手法を提案する．提案手法の新規性は以下である．

- dense text に基づき画像から言語を媒介としてマルチモーダル大規模言語モデル (MLLM) から得られる高次の視覚特徴量および複数の観点に基づく視覚特徴量を組み合わせた，Dense Structural Multimodal Encoder(DSME) モジュールを導入する．
- クエリとして与えられる文からマルチモーダルな特徴量を効果的に獲得するため，PromCSE [5] および CLIP に基づく Universal Query Encoder(UQE) モジュールを導入する．

2. 関連研究

マルチモーダル言語理解に関する研究は，クロスモーダル検索および参照表現理解などの分野において広く行われている [6, 7] ．ロボットへの指示文から対象物体を特定するタスクに関して，物体の矩形領域を分類問題として扱う手法 [8, 9] やクロスモーダル検索として扱う手法 [4, 10] などが研究されている．

本研究では，クロスモーダル検索に基づく LTRPO タスク [1] に取り組む．本タスクに適用可能な既存手法において，物体やランドマークの特定に有用であるとされる dense text [11, 12] に着目した手法は少ない．そこで本研究では，dense text と MLLM を組み合わせることで得られる構造的な特徴量および画素・空間関係・画像・対象物体に関する複数の粒度から抽出される特徴量を統合した視覚特徴量を導入する．

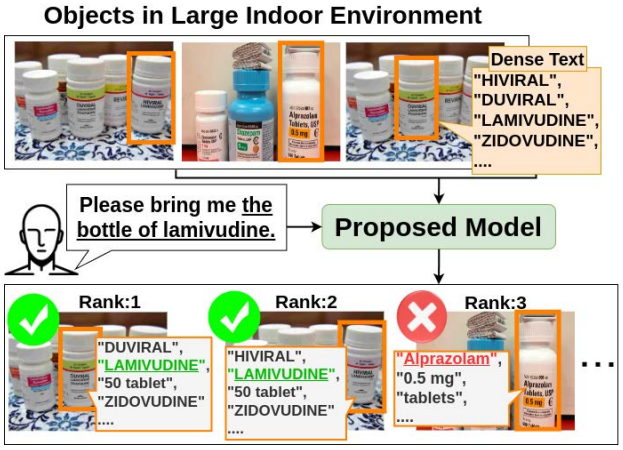

図 1　LTRPO タスクの具体例

3. 問題設定

本研究では，LTRPO タスク [1] を扱う．本タスクは，ユーザによるクエリに基づき，移動ロボットが屋内環境で撮影した物体の画像を検索するタスクである．ここで，クエリは物体操作に関する指示文や物体に関する疑問文など，open-vocabulary かつ多様な自然言語文によって構成される．

本タスクの入出力は以下である．

- 入力：open-vocabulary なクエリおよび屋内環境において事前に収集された画像群
- 出力：ランク付けされた画像群

本タスクでは，クエリの対象となる物体に関する画像が上位にランク付けされたリストを出力することが望ましい．図 1 に LTRPO タスクの具体例を示す．例えば，クエリとして “Please bring me the bottle of lamivudine.” が与えられた際，“lamivudine” というラベルのある瓶を上位にランク付けすることが望ましい．

本論文で扱う用語を以下のように定義する．

- クエリ：物体操作に関する指示文や物体に関する疑問文など，open-vocabulary な自然言語文
- 対象物体：移動ロボットの動作対象となる物体
- 対象物体領域：画像に含まれる対象物体の矩形領域

本研究では，事前の画像撮影の際に記録する座標に基づきロボットが撮影地点まで移動可能であることを前提とし，Mean Reciprocal Rank (MRR) および Recall@K (K=1,5) でモデルを評価する．

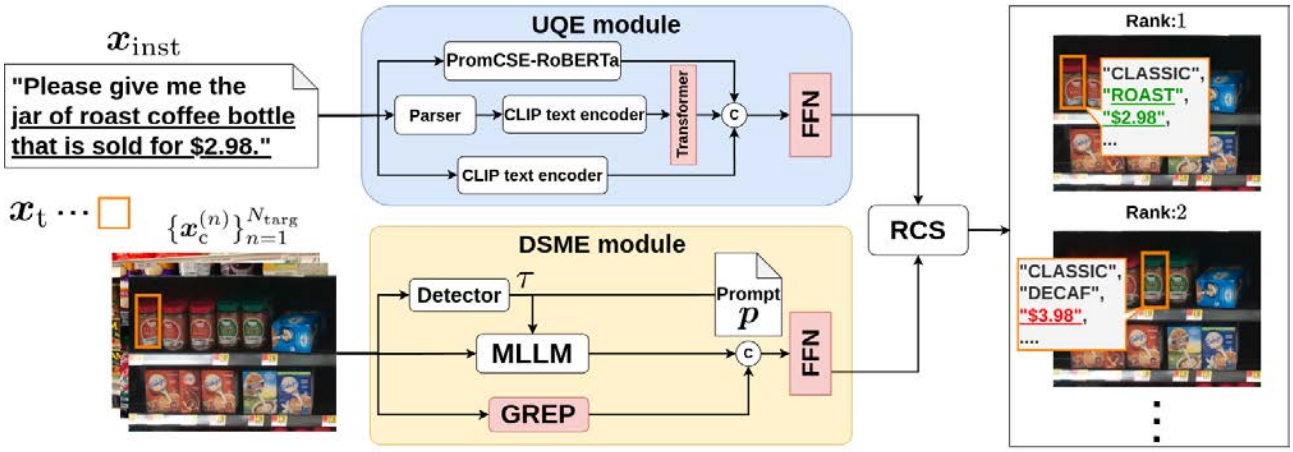

図 2　提案手法のモデル構造

4.　提案手法

本研究では，MLLM が内部的に持つ常識的知識を活用した検索手法を提案する．具体的には，大規模屋内空間環境中の物体をクエリに基づき検索するタスクにおいて，MLLM および dense text を用いて得られる言語を媒介とした視覚特徴量を導入する．本拡張は視覚特徴を強化するアプローチであるため，テキストを含む画像を入力とする既存のタスクに幅広く適用可能である．

図 2 に，提案手法のモデル構造を示す．提案手法は Dense Structural Multimodal Encoder(DSME) モジュール，Universal Query Encoder(UQE) モジュール，および RCS モジュール [4] の 3 つから構成される．

4.1　入力

提案手法の入力 $\boldsymbol{x}$ を次のように定義する．

$$\boldsymbol{x} = (\boldsymbol{x}_{\text{inst}}, \mathcal{T}, \mathcal{C})$$
$$\mathcal{T} = \left\{ \boldsymbol{x}_t^{(n)} \,\middle|\, n = 1, ..., N_{\text{targ}} \right\}$$
$$\mathcal{C} = \left\{ \boldsymbol{x}_c^{(n)} \,\middle|\, n = 1, ..., N_{\text{targ}} \right\}$$

ここで，$\boldsymbol{x}_{\text{inst}} \in \{0,1\}^{V \times L}$，$\boldsymbol{x}_t^{(n)} \in \mathbb{R}^{3 \times H_t \times W_t}$，および $\boldsymbol{x}_c^{(n)} \in \mathbb{R}^{3 \times H_c \times W_c}$ は，それぞれ one-hot ベクトルとしてトークン化されたクエリ，対象物体領域，および画像を表す．ここで，V，L，N_{targ}，H_t，W_t，H_c，および W_c は，それぞれ語彙サイズ，最大トークン長，対象物体領域の個数，対象物体領域の高さ，幅，画像の高さ，および幅を示す．

4.2　UQE

UQE モジュールでは，$\boldsymbol{x}_{\text{inst}}$ から，マルチモーダル基盤モデルおよび汎用的な文埋め込みモデルに基づき，画像との接地および意味的な類似性を考慮した言語特徴量を得る．LTRPO タスク [1] の入力となる $\boldsymbol{x}_{\text{inst}}$ は，open-vocabulary であり，しばしば長く曖昧性を含む．これに対し，PromCSE によって訓練された RoBERTa [13] を本タスクに適用することで $\boldsymbol{x}_{\text{inst}}$ の曖昧性に頑健な文埋め込みを得られると期待される．

CLIP は比較的短い alt-text で訓練されているため，入力に含まれる複雑な関係性や長い系列に対して脆弱であることが知られている [14]．しかし，既存手法 [1,4] においては，$\boldsymbol{x}_{\text{inst}}$ に対する言語特徴量の抽出は CLIP のみに依存している．そこで，多様な $\boldsymbol{x}_{\text{inst}}$ に対し意味的特徴および対象物体に関する特徴を獲得するために，CLIP，PromCSE，および構文解析を並列に使用した特徴抽出機構を提案する．PromCSE では，同一の文を 2 回エンコーダに入力し，それぞれ異なるドロップアウトマスクを適用して得られる特徴量を正例ペアとして学習する．これにより，多様なクエリに対して意味的な一貫性を保ちながら，微細な違いにも対応可能な特徴量が獲得できると期待される．

本モジュールではまず，$\boldsymbol{x}_{\text{inst}}$ をパーサ [15]，CLIP および PromCSE で学習された RoBERTa に入力し，それぞれ $\{\boldsymbol{x}_{\text{np}}\}_{i=1}^{N_{\text{np}}}$，$\boldsymbol{h}_{\text{cl}} \in \mathbb{R}^{d_{\text{cl}}}$，および $\boldsymbol{h}_{\text{st}} \in \mathbb{R}^{d_{\text{st}}}$ を得る．ただし，N_{np}，d_{cl}，および d_{st} は，それぞれパーサから得られる名詞句の数，CLIP の出力次元数，および RoBERTa の出力次元数を示す．また，$\boldsymbol{h}_{\text{st}}$ は，`[CLS]` トークンの出力に対応する最終層から取得する．さらに，$\{\boldsymbol{x}_{\text{np}}\}_{i=1}^{N_{\text{np}}}$ を CLIP text encoder に入力し，$\boldsymbol{h}_{\text{np}} \in \mathbb{R}^{d_{\text{cl}} \times N_{\text{np}}}$ を得る．これらを用いて，以下の式から本モジュールの出力である言語特徴量 $\boldsymbol{h}_{\text{txt}}$ を得る．

$$\boldsymbol{h}_{\text{txt}} = \text{FFN}\left[\boldsymbol{h}_{\text{cl}}; \text{Transformer}\left(\boldsymbol{h}_{\text{np}}\right); \boldsymbol{h}_{\text{st}}\right]$$

ここで，FFN，$[\cdot;\cdot]$，および Transformer は，それぞれ順伝播型ネットワーク，ベクトルの連結，および Transformer 層を表す．

4.3　DSME

DSME モジュールでは，MLLM から得られる構造的な特徴量および画素・空間関係・画像・対象物体に関する複数の粒度から得られる視覚特徴量を組み合わせた画像特徴抽出を行う．本研究で扱う大規模な屋内環境では，様々な場所で撮影された $\boldsymbol{x}_c^{(n)}$ が想定される．既存手法の主要なエラー要因の一つは対象物体と完全に異なる物体を上位にランク付けしてしまう失敗であり，既存手法では多様な $\boldsymbol{x}_c^{(n)}$ の扱いに課題がある [1,4]．

これに対し，MLLM の最終層付近から得られる潜在表現はトークナイザによる影響を受けることがなく，視覚的特徴および構造的な特徴の両方を有するため有用である．しかし，MLLM の視覚理解においてはしばしば Object Hallucination [16] が発生するという問題がある．visual prompt に基づく手法 [17] ではこの問題を画像に直接変更を加えることで軽減しているが，視覚的コンテキストが損なわれるため不十分である．

そこで本モジュールでは，$\boldsymbol{x}_c^{(n)}$ に含まれる dense text τ を抽出し，τ をプロンプトに含めることで，MLLM に視覚的コンテキストを損なわない視覚理解を促す．本モジュールではまず，$\boldsymbol{x}_c^{(n)}$ に対し文字認識を実行し，検出器から得られるテキスト集合 τ を得る．続いて，プロンプト $\boldsymbol{p}$ に対し τ を付加して MLLM に入力し，最終層から潜在表現を得る．さらに，GREP モジュール [4] を用いて，画素，空間関係，画像全体，および対象物

表 1　ベースライン手法および提案手法の定量的比較結果

手法	YAGAMI データセット			LTRRIE2.0 データセット		
	MRR [%]	R@1 [%]	R@5 [%]	MRR [%]	R@1 [%]	R@5 [%]
CLIP [2]	19.0	4.9	16.4	31.6	9.2	29.8
MultiRankIt [1]	18.7 ± 0.8	4.4 ± 0.6	17.3 ± 1.3	32.4 ± 2.9	9.6 ± 2.1	31.9 ± 1.8
今井ら [4]	22.9 ± 0.9	6.6 ± 1.1	21.2 ± 1.6	34.5 ± 1.8	9.6 ± 1.1	33.8 ± 2.1
提案手法	**25.6 ± 1.8**	**9.7 ± 0.8**	**23.3 ± 1.8**	**37.9 ± 1.5**	**11.9 ± 2.1**	**37.0 ± 1.5**

表 2　Ablation study における定量的比較結果

モデル	条件		YAGAMI データセット			LTRRIE-2.0 データセット		
	$\boldsymbol{h}_{\mathrm{lt}}$	$\boldsymbol{h}_{\mathrm{st}}$	MRR [%]	R@1 [%]	R@5 [%]	MRR [%]	R@1 [%]	R@5 [%]
(i)		✓	25.1 ± 0.9	8.7 ± 1.9	**24.2 ± 1.9**	36.5 ± 2.1	11.0 ± 1.6	33.3 ± 1.3
(ii)	✓		24.1 ± 1.6	8.8 ± 0.3	**24.2 ± 1.5**	35.1 ± 2.3	9.5 ± 1.6	34.1 ± 3.3
(iii)	✓	✓	**25.6 ± 1.8**	**9.7 ± 0.8**	23.3 ± 1.8	**37.9 ± 1.5**	**11.9 ± 2.1**	**37.0 ± 1.5**

体の 4 つの観点に基づく特徴量 $\boldsymbol{h}_{\mathrm{px}}$，$\boldsymbol{h}_{\mathrm{sp}}$，$\boldsymbol{h}_{\mathrm{en}}$，および $\boldsymbol{h}_{\mathrm{tg}}$ を抽出する．本モジュールの出力である画像特徴量 $\boldsymbol{h}_{\mathrm{img}}$ は，これらに基づき，以下の式より得られる．

$$\boldsymbol{h}_{\mathrm{img}} = \mathrm{FFN}\left[\mathrm{MLLM}\left(\boldsymbol{p}, \tau\right); \boldsymbol{h}_{\mathrm{px}}; \boldsymbol{h}_{\mathrm{sp}}; \boldsymbol{h}_{\mathrm{en}}; \boldsymbol{h}_{\mathrm{tg}}\right]$$

ただし，MLLM はマルチモーダル大規模言語モデルを表し，本研究では LLaVA-NeXT [18] を用いる．

4.4　RCS

上記で得られた $\boldsymbol{h}_{\mathrm{txt}}$ および $\boldsymbol{h}_{\mathrm{img}}$ によって得られる，$\boldsymbol{x}_{\mathrm{inst}}$ および $\boldsymbol{x}_{\mathrm{t}}^{(n)}$ とのコサイン類似度を s とする．モデルの出力は s に基づきランク付けされた対象物体領域集合 $\mathcal{T}'$ である．

本研究では，類似物体との対照性を緩和しつつ，学習効率をバランスする損失 [4] を用いる．InfoNCE 損失 [19] は，$\boldsymbol{x}_{\mathrm{inst}}$ と正例の類似度を最大化し，負例集合 $\mathcal{N}$ のすべての要素との類似度を最小化するよう設計されている．しかし，$\mathcal{N}$ には異なる視点から撮影された同一の対象物体が含まれることがあり，このような場合に InfoNCE 損失は不適切である．ReCo 損失 [20] は，この問題を $\boldsymbol{x}_{\mathrm{inst}}$ と $\mathcal{N}$ のすべての要素との類似度を 0 に近づけることで緩和するが，最適化されない領域が含まれ，学習効率に問題がある．そこで，これらを組合せた，以下の式で定義される損失関数を用いる．

$$\mathcal{L} = \lambda_{\mathrm{InfoNCE}}\mathcal{L}_{\mathrm{InfoNCE}} + \lambda_{\mathrm{ReCo}}\mathcal{L}_{\mathrm{ReCo}}$$

ここで，$\mathcal{L}_{\mathrm{InfoNCE}}$，$\mathcal{L}_{\mathrm{ReCo}}$，$\lambda_{\mathrm{InfoNCE}}$，および λ_{ReCo} は，それぞれ InfoNCE 損失，ReCo 損失，および重み係数である．

5.　実験設定

本研究では，LTRRIE-2.0 データセットおよび YAGAMI データセット [4] を用いた．これらのデータセットは，クエリ，画像，および対象物体領域から構成される．両データセットにおける訓練集合，検証集合，およびテスト集合のサンプル数と得られた環境に関しては，[4] と同一の条件下で実験した．

ハイパーパラメータとして，$\lambda_{\mathrm{InfoNCE}}$，$\lambda_{\mathrm{ReCo}}$，UQE モジュールにおける Transformer 層，隠れ層，および attention head の数をそれぞれ，0.5，0.5，4，768，および 4 とした．最適化手法として Adam を採用し，学習率およびバッチサイズをそれぞれ 3×10^{-5} および 512 に設定して，30 エポックの訓練を実施した．

提案手法の学習可能なパラメータ数は 152M であった．モデルの訓練はメモリ容量 24GB の NVIDIA GeForce RTX 3090 および Intel Core i9 12900K，64GB の RAM を搭載した計算機上で行った．提案手法の訓練には約 90 分を要した．また，推論時における 1 つのクエリと 100 枚の画像間の計算には約 84ms を要した．各エポック終了時に検証集合に対し Mean Reciprocal Rank (MRR) を計算した．このときの MRR が最大となったモデルを用いて，テスト集合における評価を行った．

6.　実験結果

6.1　定量的結果

表 1 にベースライン手法および手法の定量的比較結果を示す．本研究では，CLIP [2]，MultiRankIt [1]，および [4] で提案された手法をベースラインとした．表中の値は 5 回の試行における平均及および標準偏差を表す．ただし，CLIP に関しては，重みを固定した事前学習済みモデルを適用したことから，複数回の試行により同一の結果が得られるため 1 回の試行における値を示す．また，表中の太字は各評価尺度における最も高い数値を表す．

CLIP はマルチモーダル基盤モデルとして広く知られており，fine-tuning を行わずとも text-image retrieval タスクに効果的に適用されているため使用した．また，MultiRankIt および [4] で提案された手法は，LTRPO タスクにおいて良好な結果が得られているため使用した．評価尺度としては Mean Reciprocal Rank (MRR) および Recall@K (K=1,5) を用いた．両者は，ランキング学習において標準的な評価尺度であるため使用した [7]．本研究では，MRR を主要尺度とした．

表 1 より，YAGAMI データセットにおいて，提案手法は主要尺度である MRR において 25.6%であり，ベースライン手法における最良のスコアを 2.7 ポイント上回った．さらに，提案手法は R@1 および R@5 においてそれぞれ 9.7%および 23.3%であり，ベースライン手法における最良のスコアと比較してそれぞれ 3.1 ポイントおよび 2.1 ポイント上回った．同様に，LTRRIE-2.0 データセットにおいて，提案手法の MRR は 37.9%であり，ベースライン手法における最良のスコアから 3.4 ポイント改善した．R@1 および R@5 においても同様に，提案手法はそれぞれ 11.9%および 37.0%であり，ベースライン手法における最良のスコアをそれぞれ 2.3 ポイントおよび 3.2 ポイント上回った．2 つのデータセットにおけるすべての評価尺度で有意差が認められた（$p < 0.05$）．

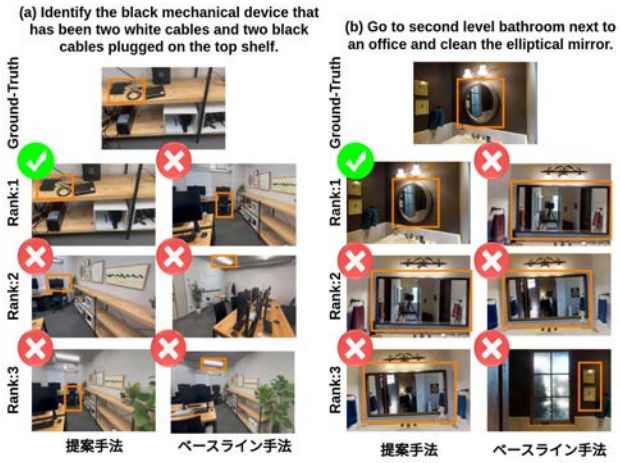

図 3 (a) YAGAMI データセットおよび (b) LTRRIE-2.0 データセットにおける提案手法およびベースライン手法 [4] との定性的比較結果

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定性的比較結果を示す．両手法において，$\boldsymbol{x}_{\text{inst}}$ に対する正解の $\boldsymbol{x}_t^{(n)}$ および上位 3 件の検索結果を示す．(a) では，$\boldsymbol{x}_{\text{inst}}$ として “Identify the black mechanical device that has been two white cables and two black cables plugged on the top shelf.” を入力した場合の結果を示す．$\boldsymbol{x}_{\text{inst}}$ が “that has been two white cables and two black cables plugged on the top shelf” という入れ子構造の複雑な参照表現を含んでいるにもかかわらず，提案手法は正解である $\boldsymbol{x}_t^{(n)}$ を上位にランク付けすることができた．これは，UQE モジュールにおける多様かつ複雑なクエリを考慮したマルチモーダルな言語特徴量が効果的に作用したためだと考えられる．

また，(b) に関して，$\boldsymbol{x}_{\text{inst}}$ として “Go to second level bathroom next to an office and clean the elliptical mirror.” を入力した場合の結果を示す．同様に，ベースライン手法および提案手法において，正解の $\boldsymbol{x}_t^{(n)}$ はそれぞれ 9 位および 1 位にランク付けされた．正解の $\boldsymbol{x}_t^{(n)}$ は楕円形の鏡である一方，ベースライン手法において上位 3 件にランク付けされた $\boldsymbol{x}_t^{(n)}$ はいずれも四角い．また，提案手法では上位 3 件の $\boldsymbol{x}_c^{(n)}$ にはいずれも蛇口および鏡が含まれている．以上から，物体の形状や常識的知識を反映したランク付けが行われており，DSME モジュール内部の dense text を用いてマルチモーダル大規模言語モデルから得られた視覚特徴量が有効であったことが示唆される．

6.3 Ablation study

表 2 に，ablation study における定量的比較結果を示す．ablation study として，以下の条件を定めた．

$\boldsymbol{h}_{\text{lt}}$ ablation. DSME モジュールにおける MLLM に基づく潜在表現 $\boldsymbol{h}_{\text{lt}}$ を除外し，その有効性を調査した．モデル (i) とモデル (iii) を比較した結果，主要尺度である MRR において，YAGAMI データセットおよび LTRRIE-2.0 データセットでそれぞれ 0.5 ポイントおよび 1.4 ポイント減少した．この結果から，dense text に基づいて得られる $\boldsymbol{h}_{\text{lt}}$ が，多様な $\boldsymbol{x}_c^{(n)}$ および $\boldsymbol{x}_t^{(n)}$ を扱ううえで有用であったと示唆される．

$\boldsymbol{h}_{\text{st}}$ ablation. UQE モジュールにおける PromCSE [5] に基づく特徴量 $\boldsymbol{h}_{\text{st}}$ を取り除き，その有用性を検証した．同様に，モデル (ii) と (iii) を比較すると，モデル (ii) における MRR は YAGAMI データセットおよび LTRRIE-2.0 データセットでそれぞれ 1.5 ポイントおよび 2.8 ポイント低下した．以上から，$\boldsymbol{h}_{\text{st}}$ の導入によって，$\boldsymbol{x}_{\text{inst}}$ の複雑さおよび曖昧さに対して頑健な特徴抽出が可能になったと考えられる．

7. おわりに

本研究では，物体操作に関する指示文や物体に関する疑問文など，open-vocabulary かつ多様な自然言語文に基づき，移動ロボットが屋内環境で撮影した物体の画像を検索する LTRPO タスク [1] を扱った．大規模な屋内環境から収集されたデータセット [4] における実験の結果，標準的な評価尺度において提案手法が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 CREST，NEDO の助成を受けて実施されたものである．

参 考 文 献

[1] 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.
[2] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, et al., “Learning Transferable Visual Models from Natural Language Supervision,” ICML, pp.8748–8763, 2021.
[3] N. Vo, L. Jiang, C. Sun, K. Murphy, L. Li, L. Fei-Fei, et al., “Composing Text and Image for Image Retrieval - an Empirical Odyssey,” CVPR, pp.6439–6448, 2019.
[4] 今井悠人，兼田寛大，是方諒介，杉浦孔明，“マルチモーダル基盤モデルと緩和対照損失を用いた大規模屋内検索エンジン，” 第 38 回人工知能学会全国大会資料，2024．3O5-OS-16c-04.
[5] Y. Jiang, et al., “Improved Universal Sentence Embeddings with Prompt-based Contrastive Learning and Energy-based Learning,” EMNLP, pp.3021–3035, 2022.
[6] S. Uppal, S. Bhagat, et al., “Multimodal Research in Vision and Language: A Review of Current and Emerging Trends,” Information Fusion, vol.77, pp.149–171, 2022.
[7] M. Cao, S. Li, J. Li, L. Nie, and M. Zhang, “Image-text Retrieval: A Survey on Recent Research and Development,” IJCAI, pp.5410–5417, 2022.
[8] J. Hatori, Y. Kikuchi, S. Kobayashi, et al., “Interactively Picking Real-World Objects with Unconstrained Spoken Language Instructions,” ICRA, pp.3774–3781, 2018.
[9] R. Korekata, et al., “Switching Head-Tail Funnel UNITER for Dual Referring Expression Comprehension with Fetch-and-Carry Tasks,” IROS, pp.3865–3872, 2023.
[10] G. Sigurdsson, J. Thomason, G. Sukhatme, and R. Piramuthu, “RREx-BoT: Remote Referring Expressions with a Bag of Tricks,” IROS, pp.5203–5210, 2023.
[11] Y. Bu, L. Li, J. Xie, Q. Liu, Y. Cai, Q. Huang, and Q. Li, “Scene-Text Oriented Referring Expression Comprehension,” IEEE TMM, vol.25, pp.7208–7221, 2023.
[12] Y. Sun, Y. Qiu, Y. Aoki, and H. Kataoka, “Guided by the Way: The Role of On-the-route Objects and Scene Text in Enhancing Outdoor Navigation,” ICRA, 2024.
[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” arXiv preprint arXiv:1907.11692, 2019.
[14] B. Zhang, P. Zhang, X. Dong, Y. Zang, and J. Wang, “Long-CLIP: Unlocking the Long-Text Capability of CLIP,” arXiv preprint arXiv:2403.15378, 2024.
[15] S. Schuster, et al., “Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks,” LREC, pp.2371–2378, 2016.
[16] A. Rohrbach, L. Hendricks, K. Burns, T. Darrell, and K. Saenko, “Object Hallucination in Image Captioning,” EMNLP, pp.4035–4045, 2018.
[17] A. Shtedritski, C. Rupprecht, and A. Vedaldi, “What does CLIP know about a red circle? Visual prompt engineering for VLMs,” ICCV, pp.11987–11997, 2023.
[18] H. Liu, C. Li, Q. Wu, and Y. Lee, “Visual Instruction Tuning,” NeurIPS, vol.36, p.34892–34916, 2024.
[19] A. Oord, Y. Li, and O. Vinyals, “Representation Learning with Contrastive Predictive Coding,” arXiv preprint arXiv:1807.03748, 2018.
[20] Z. Lin, et al., “Relaxing Contrastiveness in Multimodal Representation Learning,” WACV, pp.2227–2236, 2023.