シーンテキストを用いたマルチモーダル検索に基づく日常物体操作

○後神美結，戸倉健登，雨宮佳音，八島大地，勝又圭，今井悠人，小松拓実，是方諒介，杉浦孔明（慶應義塾大学）

サービスロボットによる日常物体操作において，シーンテキストおよび固有表現を考慮した自由形式の指示文を用いることができると便利である．そこで，本研究では，ロボットが事前に収集した画像群から自然言語指示文に基づいて対象物体を含む画像を検索するタスクを扱う．本タスクは，視覚特徴や物体間の空間的関係だけでなく，シーンテキストおよび固有表現の理解を要する．そのため，指示文内の固有表現と対応する物体の複雑な関係をモデル化する手法を提案する．本手法は，新規に構築した屋内外の実世界画像と指示文を含むベンチマークおよび実機実験で標準的なマルチモーダル検索および物体操作の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，少子高齢化，および生活の質向上への関心を背景に，家庭，店舗，および都市空間などの屋内外の環境におけるサービスロボットの重要性が高まっている．これらのロボットに対してユーザが言語指示を与える際，移動先のランドマークとなる看板や，操作物体のパッケージなどに記載された文字情報（シーンテキスト）を用いることは極めて自然である．したがって，実世界で活用可能なロボットは，シーンテキストを含む視覚特徴に基づき対象物体を適切に理解できることが重要である．

本研究では，ユーザによって与えられた自然言語による自由形式の指示文をもとに，対象物体を含む画像を検索するタスクを扱う．本タスクは，屋内外の環境において物体の検索およびロボットによる物体操作を行う場面での活用が想定される．

本タスクは視覚特徴や物体間の空間的関係に加えて，シーンテキストの理解も求められる点で困難である．既存手法 [1,2] はマルチモーダル検索において良好な結果を得ている一方で，シーンテキストや固有表現と視覚特徴の統合が不十分である．この理由として，固有表現と対応する物体との複雑な関係性のモデル化が困難であることが挙げられる．

そこで，本研究ではシーンテキストおよび視覚特徴を用いて検索を行うマルチモーダル検索手法を提案する．本手法では，大規模言語モデル (LLM) が持つ世界知識を用いて指示文内の固有表現から対象物体の視覚特徴を取得する Attribute Description Generation (ADG) を導入する．これにより，指示文中の固有表現と対応する物体間の複雑な関係のモデル化が可能になる．また，シーンテキストの誤認識やハルシネーションの抑制のため，サードパーティOCR で検出されたシーンテキストの左上にシーンテキストの主要な言語とは異なる言語で記されたマークを重畳する Crosslingual Visual Prompt (CVP) を導入する．

本研究の新規性は以下である．

- 固有表現に基づく視覚特徴を含む複数粒度でアラインされた言語特徴を用い，指示文中の複雑な固有表現と物体間の関係をモデル化する Multi-Form Instruction Encoder (MFIE) を提案する．

2. 問題設定

本研究では，ロボットが事前探索によって収集した画像群から，自由形式の自然言語指示に基づいて画像を検索するタスクを扱う．本タスクでは，操作対象または移動の目的になりうる物体を含む，屋内外の多様な環境で撮影された画像を検索対象とする．また，シーンテキストを含む画像と含まない画像の両方を対象とする．本タスクでは，指示文に対応する物体（対象物体）を含む画像（対象画像）を上位にランク付けしたリストを出力することが期待される．

3. 提案手法

本研究では，固有表現やシーンテキストを視覚特徴と統合することで自然言語指示に基づいてマルチモーダル検索を行う手法を提案する．図 1 に提案手法のモデル図を示す．Emb.，MLP，および点線はそれぞれ言語エンコーダ，多層パーセプトロン，およびパッチ化された視覚特徴量を示す．また，Named Entity (NE) Extractor は指示文に含まれる固有表現の抽出，ADG は固有表現の視覚特徴の取得を行っている．提案手法は，MFIE, Scene Text Aware Visual Encoder (STVE)，および Scene Text Reranker (STRR) の 3 つのモジュールから構成される．

提案手法への入力を，$\boldsymbol{x} = \{\boldsymbol{x}_{\text{txt}}, X_{\text{img}}\}$，$X_{\text{img}} = \{\boldsymbol{x}_{\text{img}}^{(i)}\}_{i=1}^{N_{\text{img}}}$ と定義する．ここで，$\boldsymbol{x}_{\text{txt}} \in \{0,1\}^{V\times L}$ および $\boldsymbol{x}_{\text{img}}^{(i)} \in \mathbb{R}^{3\times W\times H}$ は指示文および i 番目の画像を表す．ただし，N_{img}，H，W，V および L はそれぞれ検索対象となる候補画像の数，画像の高さ，画像の幅，語彙数，および最大トークン長を示している．

3.1 MFIE

MFIE モジュールは，複数粒度でアラインされた言語特徴を用いることで $\boldsymbol{x}_{\text{txt}}$ に含まれる複雑な表現をモデル化する．本モジュールで用いる言語特徴量は，(i) 言語エンコーダで抽出した特徴量，(ii) 画像とアラインされた特徴量，(iii)LLM の世界知識に基づく対象物体の視覚特徴についての説明を統合したものである．画像とアラインされたエンコーダ（例：CLIP [1]）は，一般的なマルチモーダル特徴を効果的に捉えることができる一方で，製品名やブランド名などの固有表現の理解は限定的である．これは，言語エンコーダの学習データセットでは固有表現を十分に網羅していないことに起因する．この課題に対応するため，本研究では LLM の世界知識を活用し，固有表現と対応する物体の説明 $\boldsymbol{x}_{\text{targ}}$ を生成する ADG を導入する．$\boldsymbol{x}_{\text{targ}}$ は対象物体のカテゴリや色などの視覚特徴を含む．

本モジュールでは，指示文に関する多様な特徴量を統合した特徴量 $\boldsymbol{h}_{\text{txt}} \in \mathbb{R}^{d_{\text{txt}}}$ を $\boldsymbol{h}_{\text{txt}} = \text{MLP}\left(f_{\text{txt}}(\boldsymbol{x}_{\text{txt}}), f_{\text{targ}}(\boldsymbol{x}_{\text{targ}})\right)$ と定義する．ここで，

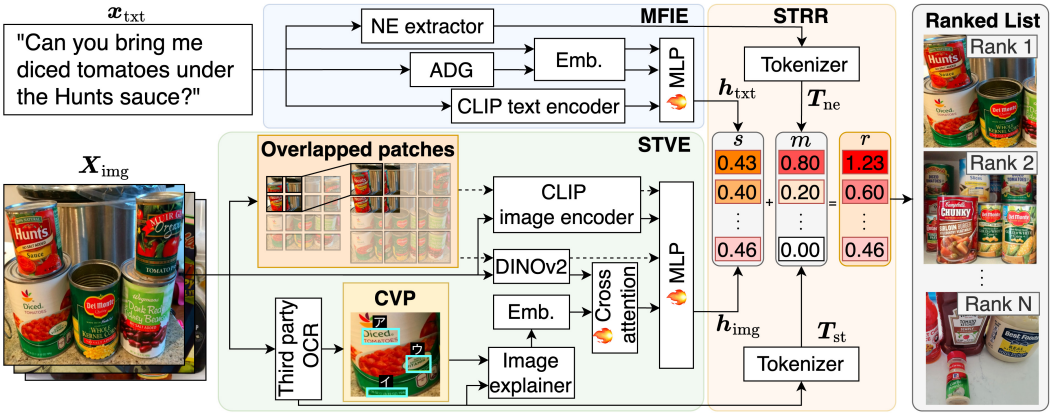

図 1　提案手法のモデル図

d_{txt} は統合した特徴量の次元数，MLP$(\cdot)$ は多層パーセプトロン，$f_{\text{txt}}(\cdot)$ および $f_{\text{targ}}(\cdot)$ はそれぞれ $\boldsymbol{x}_{\text{txt}}$ および $\boldsymbol{x}_{\text{targ}}$ の特徴抽出に使用する言語エンコーダである．本手法では，$f_{\text{txt}}(\cdot)$ として Stella [3] および CLIP 言語エンコーダ，$f_{\text{targ}}(\cdot)$ として Stella を用いる．さらに，MFIE では LLM を活用した NE extractor により $\boldsymbol{x}_{\text{txt}}$ から固有表現 $X_{\text{ne}} = \{\boldsymbol{x}_{\text{ne}}^{(i)}\}_{i=1}^{N_{\text{ne}}}$ を抽出する．本モジュールの出力は $\boldsymbol{h}_{\text{txt}}$ および X_{ne} である．

3.2　STVE

STVE モジュールは，Crosslingual Visual Prompt (CVP) に基づくシーンテキストを考慮した narrative representation [4]，パッチ境界に頑健な重複パッチ化した視覚特徴量，および複数粒度でアラインされた特徴量を統合する．CVP では従来の visual prompt とは異なり，画像内のシーンテキストと区別するため，それらとは異なる言語の文字で記されたマークを用いる．CVP では，OCR により $\boldsymbol{x}_{\text{img}}^{(i)}$ に含まれるシーンテキスト $X_{\text{st}}^{(i)} = \{\boldsymbol{x}_{\text{st}}^{(i,j)}\}_{j=1}^{N_{\text{st}}}$ とそれらの画像内の位置情報を取得し，画像の各シーンテキストのバウンディングボックスの上にマークを重畳することで作成される．CVP および $X_{\text{st}}^{(i)}$ をもとに MLLM でシーンテキストを考慮した画像の説明を生成し，言語エンコーダを用いて言語特徴量を抽出することで narrative representation $\boldsymbol{x}_{\text{nr}}^{(i)}$ を得る．

また，画像の細部を捉えるために，重複パッチ化した視覚特徴量を利用する．標準的なパッチ化で得られる特徴量は，パッチの境界付近に位置する物体やシーンテキストの視覚特徴を正確に捉えることが困難であるため，本手法では重複パッチ化を採用している．重複パッチ化した視覚特徴量 $X_{\text{op}}^{(i)} = \{\boldsymbol{x}_{\text{op}}^{(i,j)}\}_{j=1}^{N_{\text{op}}}$ は，$\boldsymbol{x}_{\text{img}}^{(i)}$ を $N \times M$ セルに分割し，連続する 2×2 セルを 1 パッチとして抽出することで得られる．各パッチを複数の画像エンコーダ（例：CLIP，DINOv2 [5]）で視覚特徴量を抽出し，統合することで得られる．

さらに，複数の層からの視覚特徴量を抽出することで，複数粒度での視覚特徴を捉える．これにより，色や形などの低次の特徴，物体同士の空間的関係などの中次の特徴，および商品名等に基づく視覚特徴の説明などの高次の特徴の複数粒度での情報を含む．

また，cross-attention を用いて複数粒度での視覚特徴量と narrative representation を結合し，最終的な視覚特徴量 $\boldsymbol{x}_{\text{maf}}^{(i)}$ を得る．STVE モジュールの出力 $\boldsymbol{h}_{\text{img}}^{(i)} \in \mathbb{R}^{d_{\text{img}}}$ は $\boldsymbol{h}_{\text{img}}^{(i)} = \text{MLP}(\boldsymbol{x}_{\text{nr}}^{(i)}, X_{\text{op}}^{(i)}, \boldsymbol{x}_{\text{maf}}^{(i)})$ である．ここで，d_{img} は STVE モジュールの出力次元を表す．

3.3　STRR

STRR モジュールは，指示文の固有表現と対応する物体間の複雑な関係を高速にモデル化する．本タスクでは，指示文やシーンテキストに製品名や店舗名等の多数の視覚特徴と関連する物体を指す固有表現が含まれている場合があり，それらの対応関係を正確にモデル化する必要がある．既存手法では，$\boldsymbol{h}_{\text{txt}}$ と $\boldsymbol{h}_{\text{img}}^{(i)}$ の間のコサイン類似度 [1] を用いるのが一般的であるが，前述の関係性を捉えるには不十分である．さらに，検索タスクでは高速に類似度計算を行うことも求められる．

そこで，本研究ではコサイン類似度に加えて，固有表現と対応する物体の複雑な関係を考慮した項を組み合わせた新たな類似度スコア $s(\cdot,\cdot)$ を以下のように定義する．

$$s(\boldsymbol{x}_{\text{txt}}, \boldsymbol{x}_{\text{img}}^{(i)}) = \frac{\boldsymbol{h}_{\text{txt}} \cdot \boldsymbol{h}_{\text{img}}^{(i)}}{\|\boldsymbol{h}_{\text{txt}}\| \|\boldsymbol{h}_{\text{img}}^{(i)}\|} + w \cdot \frac{|T_{\text{ne}} \cap T_{\text{st}}^{(i)}|}{\max_j |T_{\text{ne}} \cap T_{\text{st}}^{(i)}|}$$

ここで，T_{ne} および $T_{\text{st}}^{(i)}$ はそれぞれ X_{ne} および $X_{\text{st}}^{(i)}$ をトークン化したものであり，w は学習可能な重みパラメータである．$s(\cdot,\cdot)$ における第 2 項により，コサイン類似度と同様の処理速度を維持しつつ，X_{ne} と正規化処理を適用した $X_{\text{st}}^{(i)}$ の語彙的な一致を評価することが可能である．また，T_{ne} を用いることで，ストップワードなどの指示文理解に不要な単語を排除し，意味的に関連する単語の一致のみに基づいてリランキングを行うことが可能となる．本モデルの出力 $\hat{Y}$ は，X_{img} を $s(\cdot,\cdot)$ に基づいてランク付けしたリストである．

4.　実験

4.1　実験設定

提案手法の評価を行うために GoGetIt および TextCaps-test ベンチマークを構築した．既存のマルチモーダル検索タスクのデータセット [10,11] はロボットによる物体操作や移動タスクに適しておらず，RefText データセット [12] はシーンテキストを含むものの，一部の指示文は 1 語のみで構成されるなど，極めて単純であるという問題がある．よって，RefText データセットのサンプルにより長く，ロボットを用いたタスクを想定した指示文を付与することで GoGetIt ベンチマークを作成した．一方で，RefText データセットはラ

表 1　提案手法およびベースライン手法の定量的結果（「*」,「†」,「‡」はそれぞれ再現実装，frozen，fine-tuned を示す）

[%]	GoGetIt (RefText)			GoGetIt (Instruction)			TextCaps-test			LTRRIE		
手法	R@5 ↑	R@10 ↑	R@20 ↑	R@5 ↑	R@10 ↑	R@20 ↑	R@5 ↑	R@10 ↑	R@20 ↑	R@5 ↑	R@10 ↑	R@20 ↑
CLIP [1] †	63.2	72.8	82.3	74.3	83.3	92.1	79.8	86.3	91.5	56.1	71.0	84.6
CLIP ‡	63.5 (±0.3)	74.1 (±0.3)	83.3 (±0.2)	73.9 (±0.7)	83.9 (±0.3)	90.9 (±0.3)	82.0 (±0.6)	90.1 (±0.4)	93.5 (±0.3)	56.8 (±0.7)	72.3 (±0.4)	84.6 (±0.3)
Long-CLIP [6] †	-	-	-	-	-	-	86.0	90.3	94.8	61.6	79.3	91.0
Long-CLIP ‡	-	-	-	-	-	-	87.2 (±0.1)	91.8 (±0.3)	95.9 (±0.1)	69.9 (±0.3)	84.3 (±0.1)	94.1 (±0.1)
BLIP-2 [7]	-	-	-	-	-	-	86.0	90.3	94.8	61.6	79.3	91.0
BEiT-3 [2]	54.4	65.3	76.6	63.7	79.5	89.9	76.5	84.8	91.3	59.9	76.6	88.3
NLMap* [8]	50.9	60.1	70.5	61.0	73.5	86.3	70.0	78.8	85.8	50.9	66.4	78.8
RelaX-Former [9]	-	-	-	-	-	-	62.3 (±1.2)	73.7 (±0.7)	84.9 (±0.9)	66.6 (±0.9)	81.7 (±0.7)	92.3 (±0.5)
提案手法	**91.1** (±0.5)	**95.0** (±0.3)	**97.3** (±0.1)	**88.5** (±0.5)	**92.6** (±0.6)	**95.4** (±0.4)	**93.3** (±0.8)	**96.1** (±0.3)	**99.0** (±0.2)	**72.1** (±0.8)	**87.1** (±0.6)	**94.8** (±0.4)

ンダムに分割される [12] ため，COCO [13] などのデータセットに RefText のテスト集合が含まれる可能性がある．よって，これらのデータセットで事前学習されたマルチモーダル基盤モデルを GoGetIt ベンチマークで公正に比較することは困難である．そこで，TextCaps ベンチマーク [10] をもとに，ロボットを用いたタスクでの応用を考慮した TextCaps-test ベンチマークを構築した．

さらに，シーンテキストを含まない画像に対する手法の評価のため，屋内環境を対象としたマルチモーダル検索データセットである LTRRIE データセット [14] を用いた．本データセットの指示文および画像はそれぞれ REVERIE データセット [15] および Matterport3D シミュレータ [16] から収集されている．

ベースライン手法として，CLIP (ViT-L/14) [1]，BLIP-2 (ViT-g) [7]，BEiT-3 (large) [2]，Long-CLIP (ViT-L/14) [6]，NLMap [8]，および RelaX-Former [9] を用いた．また，CLIP (ViT-L/14) および Long-CLIP (ViT-L/14) の言語エンコーダおよび画像エンコーダを fine-tuning したモデルもベースライン手法とした．

5. 実験結果

5.1 定量的結果

表 1 に，提案手法およびベースライン手法の定量的結果を示す．評価指標としてマルチモーダル検索 [17] における標準的な評価指標である recall@$K(K = 5, 10, 20)$ を使用し，特に recall@10 を主要評価指標とした．CLIP (frozen)，Long-CLIP (frozen)，BLIP-2，BEiT-3，および NLMap はゼロショット転移設定で実験を行った．これらのモデルは異なる試行同士の結果に差が生じなかったため，1 回の試行によるスコアを報告している．これら以外の手法については，5 回の試行から得られた平均値と標準偏差を報告している．各尺度における最高スコアは太字で示している．なお，Long-CLIP，BLIP-2，および RelaX-Former は，GoGetIt ベンチマークのテストセットとこれらの学習データとの間にデータリークの懸念があるため，GoGetIt ベンチマークにおけるスコアは報告していない．

表 1 に示す通り，提案手法における recall@10 は GoGetIt ベンチマークの RefText サブセット，同ベンチマークの Instruction サブセット，TextCaps-test ベンチマーク，および LTRRIE データセットにおいて，それぞれ 95.0%，92.6%，96.1%，および 87.1%であり，それぞれのデータセットにおける最良のベースライン手法のスコアを，それぞれ 20.9，8.3，4.3，および 2.8 ポイント上回った．また，他のすべての評価指標においても，提案手法は全てのベースライン手法を上回った．

5.2 定性的結果

図 2 は，提案手法およびベースライン手法 [1, 6] の定性的結果を示す．図 2(a) および (b) は，それぞれ GoGetIt ベンチマークの Instruction サブセットおよび TextCaps-test ベンチマークにおけるサンプルである．各サンプルに対して，対象画像および各手法によって検索された上位 3 件の画像を示している．また，上位 3 件内の対象画像は緑色の枠に囲まれている．

図 2(a) における $\boldsymbol{x}_{\text{txt}}$ は “Pass me the red container of Sun-Maid raisins on the kitchen counter.” であり，対象物体は “Sun-Maid raisins” と記載された赤色の容器である．ベースライン手法は対象物体を含まない画像を上位 3 件にランク付けしたが，提案手法は対象画像を 1 位に正しくランク付けした．さらに提案手法は，対象物体を含む画像を 2 位にランク付けしている．

図 2(b) の $\boldsymbol{x}_{\text{txt}}$ は “Buy the red Orbit candy at the kiosk.” である．この例では，ベースライン手法による上位 3 件の画像にキオスクや赤色は含まれていたが，赤い Orbit キャンディは含まれていなかった．一方，提案手法は赤い Orbit キャンディを含む対象画像を 1 位にランク付けした．これらの例は，提案手法がシーンテキスト，物体間の空間的関係，および参照表現を含む指示文に基づいて検索できることを示している．

6. 実機実験

サービスロボットを用いて，提案手法の性能をゼロショット転移設定で検証した．12 種類の異なる環境構成において，各構成ごとに 5 から 10 エピソード，合計で 100 エピソードの実験を実施した．事前探索では，構築済みのマップ上に事前に設定した 15 箇所を巡回し，RGB 画像を撮影した．その後，シーンテキストを考慮した物体操作や移動タスクを対象とする自由形式の英語指示文が与えられた．

本タスクをマルチモーダル検索と物体把持の 2 つのサブタスクに分割して評価を行った．マルチモーダル検索では，ロボットが事前に収集した画像を用いて提案手法およびベースライン手法の性能を recall@5 で評

図 2　提案手法およびベースライン手法の定性的結果

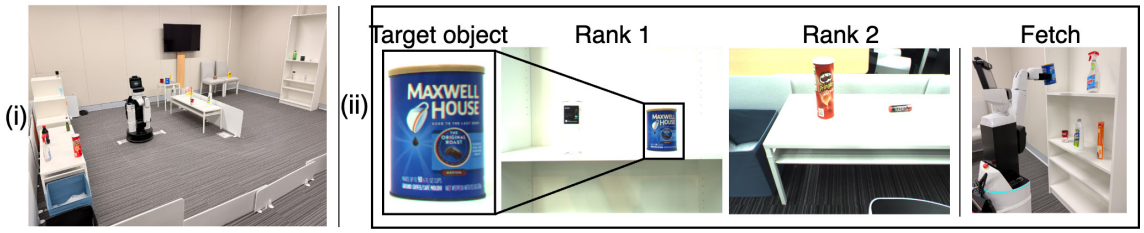

図 3　(i) 実機実験の環境および (ii) 定性的結果

価した．物体把持は，マルチモーダル検索とは独立に実行し，検索結果において対象画像が 5 位以内にランク付けされ，かつロボットが対応する物体を把持できた場合にのみ成功と判定した．評価指標は，マルチモーダル検索およびロボット操作タスクにおける標準的な評価指標である recall@5 および成功率 (SR) を用いた．

表 2 に，提案手法およびベースライン手法 [1, 2, 6–9] の実機実験における定量的結果を示す．表 2 より，提案手法の recall@5 および SR はそれぞれ 88%および 80%であり，ベースライン手法をそれぞれ 12 ポイントおよび 10 ポイント上回った．

図 3 (ii) に，実機実験における成功例を示す．ロボットに与えられた $\boldsymbol{x}_{\text{txt}}$ は “Pass me the Maxwell can.” であり，提案手法は対象画像を 1 位に正しくランク付けした．また，2 位の画像には “Maxwell” というシーンテキストは含まれていないが，缶が写っており，指示内容には関連している．また，対象物体の把持にも成功した．

表 2　実機実験の定量的結果（「*」は再現実装を示す）

手法	R@5 ↑ [%]	SR ↑ [%]
CLIP [1] (frozen)	71	64
CLIP (fine-tuned)	74	65
Long-CLIP [6] (frozen)	70	66
Long-CLIP (fine-tuned)	76	70
BLIP-2 [7]	72	66
BEiT-3 [2]	66	62
NLMap* [8]	71	65
RelaX-Former [9]	60	54
提案手法	**88**	**80**

7.　おわりに

本研究では，固有表現と対応する物体との複雑な関係をモデル化する手法を提案した．また，ロボットの物体操作や移動に関する指示文およびシーンテキストを含む画像から構成される GoGetIt および TextCaps-test ベンチマークを新たに構築した．実験の結果，提案手法は標準的なマルチモーダル検索および物体操作の評価指標においてすべてのベースライン手法を上回り，実環境においても高いゼロショット転移能力を示した．

謝辞

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

参　考　文　献

[1] A. Radford, J. Kim, C. Hallacy, A. Ramesh, et al., “Learning Transferable Visual Models From Natural Language Supervision,” ICML, pp.8748–8763, 2021.

[2] W. Wang, H. Bao, L. Dong, J. Bjorck, et al., “Image as a Foreign Language: BEIT Pretraining for Vision and Vision-Language Tasks,” CVPR, pp.19175–19186, 2023.

[3] D. Zhang, J. Li, Z. Zeng, and W. Wang, “Jasper and Stella: distillation of SOTA embedding models ,” arXiv preprint arXiv:2412.19048, 2024.

[4] M. Goko, M. Kambara, et al., “Task Success Prediction for Open-Vocabulary Manipulation Based on Multi-Level Aligned Representations,” CoRL, pp.3242–3263, 2024.

[5] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, et al., “DINOv2: Learning Robust Visual Features without Supervision,” arXiv preprint arXiv:2304.07193, 2023.

[6] B. Zhang, P. Zhang, X. wenDong, Y. Zang, and J. Wang, “Long-CLIP: Unlocking the Long-Text Capability of CLIP,” ECCV, pp.310–325, 2024.

[7] 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.

[8] B. Chen, F. Xia, B. Ichter, K. Rao, et al., “Open-vocabulary Queryable Scene Representations for Real World Planning,” ICRA, pp.11509–11522, 2023.

[9] D. Yashima, R. Korekata, and K. Sugiura, “Open-Vocabulary Mobile Manipulation Based on Double Relaxed Contrastive Learning With Dense Labeling,” IEEE RA-L, vol.10, no.2, pp.1728–1735, 2025.

[10] O. Sidorov, R. Hu, M. Rohrbach, and A. Singh, “TextCaps: A Dataset for Image Captioning with Reading Comprehension,” ECCV, pp.742–758, 2020.

[11] A. Mafla, S. Rezende, L. Gómez, D. Larlus, and D. Karatzas, “StacMR: Scene-Text Aware Cross-Modal Retrieval,” WACV, pp.2219–2229, 2021.

[12] Y. Bu, et al., “Scene-Text Oriented Referring Expression Comprehension,” IEEE TMM, vol.25, pp.7208–7221, 2023.

[13] Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and L. Zitnick, “Microsoft COCO: Common Objects in Context,” ECCV, pp.740–755, 2014.

[14] 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.

[15] Y. Qi, Q. Wu, P. Anderson, X. Wang, et al., “REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments,” CVPR, pp.9979–9988, 2020.

[16] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Nießner, M. Savva, et al., “Matterport3D: Learning from RGB-D Data in Indoor Environments,” 3DV, pp.667–676, 2017.

[17] M. Cao, S. Li, J. Li, L. Nie, and M. Zhang, “Image-text Retrieval: A Survey on Recent Research and Development,” IJCAI, pp.5376–5383, 2022.