シーンテキストを考慮した Crosslingual Visual Prompt に基づくマルチモーダル検索

戸倉 健登[1,a)]　後神 美結[1,b)]　雨宮 佳音[1,c)]　八島 大地[1,d)]　勝又 圭[1,e)]　今井 悠人[1,f)]
小松 拓実[1,g)]　是方 諒介[1,h)]　杉浦 孔明[1,i)]

概要

本研究では，自由形式の自然言語指示文に基づき，対象の物体を含む画像を検索するタスクを扱う．本タスクでは物体の視覚的特徴，空間的関係，および画像内の文字情報（シーンテキスト）を統合した理解が要求されるため困難である．そこで，本研究では Crosslingual Visual Prompt およびパッチ境界に頑健な重複パッチ化に基づく画像の詳細理解に加え，指示文の固有表現およびシーンテキストをモデル化する STARE を提案する．また，シーンテキストを含む画像および指示文で構成される GoGetIt ベンチマークおよび TextCaps-test ベンチマークを新規構築した．実験の結果、STARE は標準的なマルチモーダル検索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

労働力不足や高齢化，生活の質向上への関心から，家庭や小売店，都市空間などの屋内外環境においてサービスロボットの重要性が高まっている．これらの環境において，商品ラベルや看板などの画像内の文字情報（シーンテキスト）を用いることで，ロボットに対する明確かつ直感的な指示が可能となる．そのため，シーンテキストと視覚情報の両方を考慮するマルチモーダル検索は，公共空間における物体検索やサービスロボットの物体操作において有用である．

本研究では，ユーザによる物体操作や移動に関する自然言語指示文に基づき，ロボットが事前に撮影した画像群から，対象となる物体を含む画像を検索するマルチモーダル検索タスクを扱う．また，本タスクではシーンテキストを含む画像と含まない画像の両方を検索対象とする．図 1 に，本タスクのユースケースを示す．たとえば，ユーザが “Could you bring me the toothpaste that’s on the shelf where the Thayers Witch Hazel is placed?” と指示した場合，ロボットは事前に撮影した画像群から “Thayers Witch Hazel” 等のシーンテキストを考慮したマルチモーダル検索を行い，指示された操作を実行する．

本タスクでは，物体の視覚的特徴や空間的関係に加え，シーンテキストを考慮する必要がある．特に商品名などの固有表現とそれに対応する視覚的特徴の関係をモデル化は困難である．既存手法 [13, 15] はマルチモーダル検索において良好な結果を得ているが，シーンテキストとその他のマルチモーダル情報との統合が不十分である．

そのため，本研究ではユーザによる自由形式の自然言語指示文に基づき，シーンテキストを考慮して対象の物体を検索するマルチモーダル検索手法 STARE を提案する．STARE では，Crosslingual Visual Prompt (CVP) に基づく MLLM を用いた narrative representation を導入することで，視覚プロンプトとシーンテキストとの混同を軽減し，シーンテキストを捉えた画像特徴量を獲得する．また，OCR で検出されたシーンテキストをプロンプトに含むことで，MLLM による画像説明においてしばしば問題となるハルシネーション [6, 1] を抑制する．さらに，指示文中の固有表現とそれに対応する物体との複雑な関係を効果的にモデル化する Scene Text Reranker (STRR) を導入する．

本研究の主な貢献は以下の通りである：

- CVP に基づくシーンテキストを捉えた narrative representation，パッチ境界に頑健な重複パッチ画像特徴量，および複数粒度の画像特徴量を統合する Scene Text-Aware Visual Encoder (STVE) を導入する．
- 指示文中の固有表現とそれに対応する物体の高度に複雑な関係を効果的にモデル化する STRR を導入する．
- シーンテキストを含む画像および移動や物体操作に関する指示文で構成される GoGetIt ベンチマークおよび TextCaps-test ベンチマークを構築する．

[1] 慶應義塾大学
[a)] tkento1985@keio.jp
[b)] miyu.goko@keio.jp
[c)] kanon-amemiya@keio.jp
[d)] ydaichi1207@keio.jp
[e)] ke59ka77@keio.jp
[f)] ytim8812@keio.jp
[g)] tak3k_1999@keio.jp
[h)] rkorekata@keio.jp
[i)] komei.sugiura@keio.jp

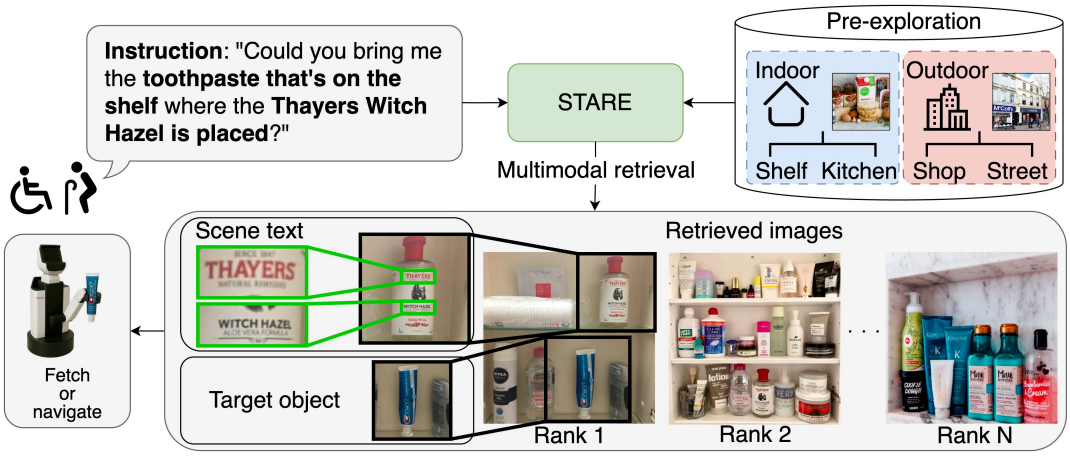

図 1 本タスクのユースケース

2. 問題設定

本タスクでは，ロボットが事前に屋内外の広範な環境からシーンテキストを含む可能性のある画像を撮影する．モデルは，ユーザによる物体操作やナビゲーションに関する自由形式の自然言語指示文に基づき，撮影した画像群から対象の画像を検索する．ここで，シーンテキストを含む画像と含まない画像の両方を検索対象とし，モデルは指示文に適合する画像を出力リストの上位にランク付けすることが求められる．本研究で使用する用語を以下の通り定義する．

- 指示文：物体操作または移動に関する自由形式の自然言語指示文．
- 対象物体および対象画像：指示文に対応する物体と，その物体を含む画像．
- シーンテキスト：画像内に存在する文字情報．

ここで対象画像には，ナビゲーション対象となる物体 (例：建物や看板) または操作可能な日常物体 (例：調理器具やボトル) が含まれていることを前提とする．

3. 提案手法

3.1 STARE

本研究では，自然言語指示文に基づきシーンテキストを考慮して画像を検索するマルチモーダル検索手法 STARE を提案する．図 2 に STARE のモデル図を示す．ここで，MLP，Emb.，および点線はそれぞれ多層パーセプトロン，言語エンコーダ，およびパッチ化された画像特徴量を示す．STARE は, Multi-Form Instruction Encoder (MFIE)，Scene Text Aware Visual Encoder（STVE），Scene Text Reranker（STRR）の 3 つのモジュールから構成されている．提案手法では，Crosslingual Visual Prompt（CVP）とシーンテキストを組み合わせた narrative representation，および指示文中の固有表現とシーンテキストの語彙的な一致に基づくリランキングにより，シーンテキスト用いた視覚情報と自然言語情報の整合性を高める拡張を行う．

STARE への入力 $\boldsymbol{x}$ は以下のように定義される：$\boldsymbol{x} = \{\boldsymbol{x}_{\text{txt}}, X_{\text{img}}\}$，$X_{\text{img}} = \{\boldsymbol{x}_{\text{img}}^{(i)}\}_{i=1}^{N_{\text{img}}}$．ここで，$\boldsymbol{x}_{\text{txt}} \in 0,1^{V\times L}$ は指示文，$\boldsymbol{x}_{\text{img}}^{(i)} \in \mathbb{R}^{3\times W\times H}$ は画像を表す．N_{img}, H, W, V, および L はそれぞれ，画像の枚数，高さ，幅，語彙サイズ，および最大トークン長を示す．

3.2 MFIE モジュール

MFIE はテキストエンコーダおよび画像とされたマルチモーダルテキストエンコーダを用いて，指示文から複数粒度の言語特徴を抽出する．モデルの最終的な出力 $\boldsymbol{h}_{\text{txt}} \in \mathbb{R}^{d\text{txt}}$ は以下のように得られる：$\boldsymbol{h}_{\text{txt}} = \text{MLP}(f_{\text{te}}(\boldsymbol{x}_{\text{txt}}), f_{\text{mte}}(\boldsymbol{x}_{\text{txt}}))$．ここで，MLP$(\cdot)$ は多層パーセプトロン，f_{te} は Stella [18]，f_{mte} は CLIP テキストエンコーダ [13] を用いる．

さらに, LLM を用いて $\boldsymbol{x}_{\text{txt}}$ から固有表現 $X_{\text{ne}} = \{\boldsymbol{x}_{\text{ne}}^{(i)}\}_{i=1}^{N_{\text{ne}}}$ を抽出する（図 2 中の ”NE extractor”）．X_{ne} は STRR モジュールにおいて使用される．

3.3 STVE モジュール

STVE は，CVP に基づくシーンテキストを考慮した narrative representation，パッチ境界に頑健な重複パッチ画像特徴量，および複数粒度の画像特徴量を統合する．narrative representation は，画像内の主要な言語と異なる言語（例：カタカナ）を用いた CVP に基づき獲得する．CVP は，OCR で抽出されたシーンテキストとその位置情報に基づき，画像内の各シーンテキスト上部にカタカナを用いたマークを重畳することで作成する．さらに，CVP に加えて各シーンテキストを記述した言語プロンプトを MLLM に入力し，シーンテキストを捉えた詳細な画像説明を得る．ここで，画像説明から言語エンコーダを用いて言語特徴量を抽出をすることで narrative representation $\boldsymbol{x}_{\text{nr}}^{(i)}$ を得る．

さらに，$\boldsymbol{x}_{\text{img}}^{(i)}$ を $N \times M$ のセルに分割し，2×2 セルごとに重複パッチ化をすることで，パッチ境界に存在する

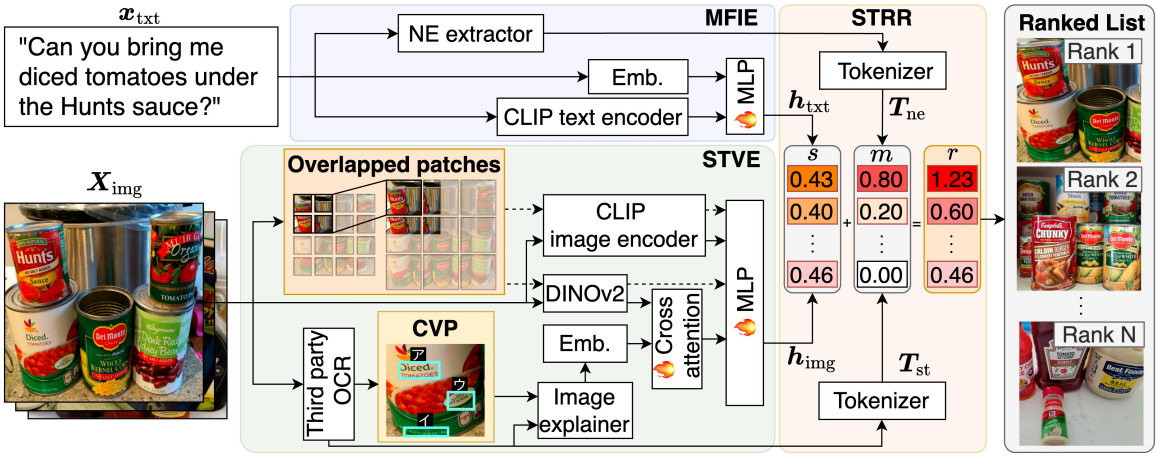

図 2 提案手法のモデル図

シーンテキストや物体に頑健な高解像度化を行う．このようにして得たパッチ群から画像エンコーダ（例：CLIP や DINOv2 [11]）を用いて特徴量抽出を行い，transformer を用いて統合することで $X_{\text{op}}^{(i)} = \{\boldsymbol{x}_{\text{op}}^{(i,j)}\}_{j=1}^{N_{\text{op}}}$ を得る．

また，DINOv2 の複数層から出力される複数粒度の画像特徴量を用いることで，色や形状などの低次の視覚情報から，位置関係や商品名などの高次の視覚情報までを捉える．これらの画像特徴量を $\boldsymbol{x}_{\text{nr}}^{(i)}$ と cross attention によって統合することで，多様な粒度の視覚情報を含む画像特徴量 $\boldsymbol{x}_{\text{maf}}^{(i)}$ を獲得する．STVE モジュールの最終的な出力 $\boldsymbol{h}_{\text{img}}^{(i)} \in \mathbb{R}^{d_{\text{img}}}$ は $\boldsymbol{h}_{\text{img}}^{(i)} = \text{MLP}(\boldsymbol{x}_{\text{nr}}^{(i)}, X_{\text{op}}^{(i)}, \boldsymbol{x}_{\text{maf}}^{(i)})$ として計算される．

3.4 STRR モジュール

STRR モジュールでは，指示文中の固有表現とそれに対応する物体の複雑な関係を高速にモデル化する．本タスクでは，商品名や店名などのシーンテキストが指示文中に固有表現として含まれる場合があり，固有表現と物体間の関係をモデル化することが求められる．既存手法 [13] では，$\boldsymbol{h}_{\text{txt}}$ と $\boldsymbol{h}_{\text{img}}^{(i)}$ 間のコサイン類似度を計算しており，これらの関係性を捉えるには不十分である．そこで，本研究では以下に定義する新たな類似度スコア $s(\cdot,\cdot)$ を導入する：

$$s(\boldsymbol{x}_{\text{txt}}, \boldsymbol{x}_{\text{img}}^{(i)}) = \frac{\boldsymbol{h}_{\text{txt}} \cdot \boldsymbol{h}_{\text{img}}^{(i)}}{\|\boldsymbol{h}_{\text{txt}}\| \|\boldsymbol{h}_{\text{img}}^{(i)}\|} + \frac{w \cdot |T_{\text{ne}} \cap T_{\text{st}}^{(i)}|}{\max_j |T_{\text{ne}} \cap T_{\text{st}}^{(i)}|} \quad (1)$$

ここで，T_{ne} および $T_{\text{st}}^{(i)}$ は，固有表現およびシーンテキストをトークン化した集合であり，ノイズへの頑健性を高めるために小文字化，記号削除などの正規化を施してから一致判定を行う．$s(\cdot,\cdot)$ における第 2 項は，T_{ne} および $T_{\text{st}}^{(i)}$ の語彙的一致のみを評価するため，$s(\cdot,\cdot)$ および既存の類似度スコア計算は同等の速度である．また，$\boldsymbol{x}_{\text{txt}}$ から獲得した X_{ne} を用いることで，指示文理解に不要なストップワードなどの単語を取り除き，意味的な一致のみに基づくリランキングを行う．本モデルの最終的な出力 $\hat{Y}$ は，$s(\cdot,\cdot)$ に基づきランキングした画像リストである．

4. 実験

4.1 実験設定

GoGetIt および TextCaps-test ベンチマーク

本研究では，新たに 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データセット [14, 10] はロボットによる物体操作や移動に関する指示に適さない．また，RefText データセット [2] の指示文はシーンテキストなどを用いて対象物を特定するが，指示文が非常に単純であり，一語のみの指示文も多く含まれている．そのため，より長文で実用的な指示文を付与した GoGetIt ベンチマークを構築した．

RefText データセットには，RefText データセットを構築する COCO[9] などのデータセットで訓練されたマルチモーダル基盤モデルと公平な比較が困難であるという重大な問題がある．RefText データセットは元データセットのスプリットを考慮せずデータをランダムにシャッフルして分割しており，データリークの懸念があるためである．そのため，本研究では MLLM と公平に比較可能な TextCaps-test ベンチマークを作成した．また，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 [7] を用いて，シーンテキストを含まない設定におけるモデルの性能も評価した．LTRRIE データセットを構成する指示文および画像は，それぞれ REVERIE [12] および Matterport3D [4] から収集されたものである．

ベースライン手法

本研究では，CLIP（ViT-L/14）[13]，BLIP-2（ViT-g）[8]，BEiT-3（large）[15]，Long-CLIP（ViT-L/14）[17]，NLMap [5]，および RelaX-Former [16] をベースライン手法として用いた．また，CLIP および Long-CLIP のテキスト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て用いた．

表 1　提案手法とベースライン手法の定量的比較結果

[%]	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 [13] (frozen)	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 (fine-tuned)	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 [17] (frozen)	-	-	-	-	-	-	86.0	90.3	94.8	61.6	79.3	91.0
Long-CLIP (fine-tuned)	-	-	-	-	-	-	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 [8]	-	-	-	-	-	-	86.0	90.3	94.8	61.6	79.3	91.0
BEiT-3 [15]	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* [5]	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 [16]	-	-	-	-	-	-	62.3 (±1.2)	73.7 (±0.7)	84.9 (±0.9)	66.6 (±0.9)	81.7 (±0.7)	92.3 (±0.5)
STARE (提案手法)	**90.2** (±1.2)	**94.1** (±0.6)	**96.9** (±0.2)	**87.6** (±0.2)	**91.7** (±0.2)	**95.2** (±0.2)	**91.8** (±1.0)	**95.2** (±0.6)	**98.2** (±0.6)	69.3 (±1.1)	**84.6** (±0.8)	**94.4** (±0.6)

4.2　定量的比較結果

表 1 に，STARE およびベースライン手法の定量的比較結果を示す．本研究では，マルチモーダル検索において標準的な評価指標である recall@K（K=5,10,20）を用いて評価を行い，recall@10 を主要な評価指標とした [3]．CLIP（frozen），Long-CLIP（frozen），BLIP-2，BEiT-3，および NLMap は複数の試行において一貫した結果が得られたため，1 回の試行結果を示す．その他の手法は，5 回の試行における平均と標準偏差を示す．表中の太字は各評価指標において最も高い数値を示す．ここで，GoGetIt ベンチマークではデータリークの問題から Long-CLIP，BLIP-2，および RelaX-Former の結果は示さない．

表 1 より，STARE は GoGetIt ベンチマークの RefText サブセット，Instruction サブセット，TextCaps-test ベンチマーク，および LTRRIE データセットにおける recall@10 は，それぞれ 94.1 %，91.7 %，95.2 %，および 84.6 %であった．これは，ベースライン手法の最良値と比較してそれぞれ 20.0 ポイント，7.8 ポイント，3.4 ポイント，および 0.3 ポイント上回る結果であった．

4.3　定性的結果

図 3 に STARE およびベースライン手法の定性的結果を示す．図 3(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．各サンプルは，対象画像および各モデルにより検索された上位 3 件の画像を示し，緑色の枠は対象画像を示す．

図 3(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” と書かれた赤い容器である．ベースライン手法は $\boldsymbol{x}_{\text{txt}}$ に対象物体を含まない画像を上位にランク付けしたのに対し，STARE は対象物体を含む画像を 1 位および 2 位にランキングした．

図 3　提案手法とベースライン手法の定性的結果

図 3(b) における $\boldsymbol{x}_{\text{txt}}$ は “Buy the green Orbit candy at the kiosk.” であり，ベースライン手法は “kiosk” や “red” に関する画像を上位にランキングしたが，対象物体を含む画像を 3 位以内にランキングしなかった．一方で，STARE は対象画像を 1 位にランキングした．したがって，STARE ではシーンテキスト，空間的関係，および参照表現を考慮したマルチモーダル検索が可能であることが示唆される．

5.　おわりに

本研究では，シーンテキストを考慮した narrative representation を導入し，固有表現とそれに対応する物体との複雑な関係を効果的にモデル化する STARE を提案した．また，シーンテキストを含む画像および移動または物体操作に関する指示文で構成される GoGetIt および TextCaps-test ベンチマークを構築した．実験の結果，STARE はマルチモーダル検索における標準的な評価尺度において，複数のベンチマークで既存手法を上回った．

謝辞

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

参考文献

[1] Biten, A., Bigorda, L. and Karatzas, D.: Let there be a clock on the beach: Reducing Object Hallucination in Image Captioning, *WACV*, pp. 2473–2482 (2021).

[2] Bu, Y. et al.: Scene-Text Oriented Referring Expression Comprehension, *IEEE TMM*, Vol. 25, pp. 7208–7221 (2023).

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

[4] Chang, A., Dai, A., Funkhouser, T., Halber, M., Nießner, M., Savva, M., Song, S., Zeng, A. and Zhang, Y.: Matterport3D: Learning from RGB-D Data in Indoor Environments, *3DV*, pp. 667–676 (2017).

[5] Chen, B., Xia, F., Ichter, B., Rao, K. et al.: Open-vocabulary Queryable Scene Representations for Real World Planning, *ICRA*, pp. 11509–11522 (2023).

[6] Jiang, C., Xu, H., Dong, M., Chen, J., Ye, W., Yan, M., Ye, Q., Zhang, J., Huang, F. and Zhang, S.: Hallucination Augmented Contrastive Learning for Multimodal Large Language Model, *CVPR*, pp. 27036–27046 (2024).

[7] 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).

[8] Li, J., Li, D., Savarese, S. and Hoi, S.: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, *ICML*, pp. 19730–19742 (2023).

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

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

[11] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A. et al.: DINOv2: Learning Robust Visual Features without Supervision, *arXiv preprint arXiv:2304.07193* (2023).

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

[13] Radford, A., Kim, J., Hallacy, C., Ramesh, A. et al.: Learning Transferable Visual Models From Natural Language Supervision, *ICML*, pp. 8748–8763 (2021).

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

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

[16] Yashima, D., Korekata, R. and Sugiura, K.: Open-Vocabulary Mobile Manipulation Based on Double Relaxed Contrastive Learning With Dense Labeling, *IEEE RA-L*, Vol. 10, No. 2, pp. 1728–1735 (2025).

[17] Zhang, B., Zhang, P., wen Dong, X., Zang, Y. and Wang, J.: Long-CLIP: Unlocking the Long-Text Capability of CLIP, *ECCV*, pp. 310–325 (2024).

[18] Zhang, D., Li, J., Zeng, Z. and Wang, W.: Jasper and Stella: distillation of SOTA embedding models , *arXiv preprint arXiv:2412.19048* (2024).