M1 サーベイレポート

学籍番号: 82518829 氏名: 戸倉 健登

所属: 杉浦孔明研究室

1. 概要

Vision-language Model（VLM）は，ロボティクスのために広く研究されており [1,2]，[3,4] において包括的にレビューされている．多様なマルチモーダル検索手法は [5,6] にまとめられている．さらに，シーンテキストの統合は，視覚表現の曖昧性解消において著しく効果的であることが，[7,8] に記録されている．

2. ロボティクスのための基盤モデル

最近の研究は，大規模言語モデル（LLM）や VLM [1,2,9] などの大規模基盤モデルを活用することで，操作タスクとナビゲーションタスクにわたる顕著な汎化性能を達成できることを示している．大規模な実ロボット軌跡で訓練された Vision-language-action モデル [2,10–13] は，多様な操作タスクにわたるスケーラブルな方策学習を可能にする．例えば，Gemini-Robotics [12] は，複雑な家庭タスクのために高度な視覚理解とロボット能力を統合している．しかし，これらのアプローチはシーンテキストを明示的に組み込んでおらず，テキスト情報が重要な意味的グラウンディングを提供するシーンにおいて，マルチモーダル理解が制限されている．

3. ロボティクスのための RAG

いくつかの既存手法 [14–18] は，強化されたグラウンディングと計画のために，具現化エージェントに RAG を適用している．他のアプローチ [19,20] は，RAG を使用してナビゲーション指示を生成し，それにより vision-language ナビゲーションモデルを改善している．RAG ベースのメモリは，長期的推論のために提案されており [21]，他の手法は模倣方策学習のためにデモンストレーションを検索および融合するために RAG を使用している [22]．これらの進歩にもかかわらず，先行研究は検索信号としてシーンテキストを活用しておらず，これは日常環境において視覚的に類似したオブジェクトを曖昧性解消するために不可欠である．

4. マルチモーダル検索

多数の研究 [23,24] が，自然言語指示に基づくモバイル操作タスクにおける家庭用サービスロボットの性能を評価している．いくつかの既存手法 [25–28] は，環境内のターゲットオブジェクトを識別するためにマルチモーダル検索を実行し，自然言語指示に基づいて関連画像のランク付けリストを出力する．また，マルチモーダル検索を実行してターゲットオブジェクトを見つけ，テキストベースの回答やウェイポイントを出力する手法もある [14]．この分野での急速な進歩にもかかわらず，本質的に曖昧性解消情報を含むシーンテキスト（ラベル，標識など）の活用には重要なギャップが残っている．提案された STARE は，シーンテキストを明示的に組み込む新しいマルチモーダルフレームワークを提案することで，この制限に対処し，実世界環境におけるオブジェクト検索精度の向上を可能にする．

5. シーンテキスト

シーンテキストを含む画像に関する研究は，テキスト認識 [29,30]，テキスト検出 [31–33]，テキスト検索 [33–35]，視覚的質問応答 [36–38]，およびテキスト合成 [39–41] を含む様々なタスクを網羅している．特に，シーンテキスト検索手法は，OCR で抽出されたテキストと位置情報を視覚特徴と統合して，クエリテキストとのアライメントを改善し，より高い検索精度を達成することが多い．Mishra ら [42] は，シーンテキスト検索タスクを導入し，文字を検出および分類してクエリテキストを含む画像の確率を計算する 2 段階パイプラインを提案した．Gomez ら [43] は，テキスト提案を予測し，クエリテキストとの類似性に基づいて画像をランク付けするエンドツーエンドネットワークを提案した．ViSTA [35] は，BERT [44] と線形投影をそれぞれ使用して OCR で抽出されたテキストと位置を埋め込み，マルチモーダル統合のためにトランスフォーマー内の融合トークンを介してテキストと位置を視覚特徴と融合する．これらのアプローチとは対照的に，STARE は，対応するオブジェクトとその属性との関連性を捉えたナラティブ表現にテキストと位置を統合する．さらに，STARE は，シーンテキストと参照されるオブジェクト間の複雑な関係を効率的にモデル化する補助類似度関数を導入する．

6. ベンチマーク

マルチモーダル検索とシーンテキストを含むタスクの分野で，いくつかの標準ベンチマークが提案されている．前者のうち，COCO [45] と Flickr30K [46] は，もともと画像キャプショニングベンチマークであり，マルチモーダル検索タスクに広く使用されている．REVERIE [47] と GOAT-Bench [48] は，オブジェクト目標ナビゲーションタスク用に設計されており，LTRRIE [49] は画像のランク付けによるマルチモーダル検索に焦点を当てている．シーンテキストのベンチマークは，様々な vision-language タスクにわたっている．例えば，ST-VQA [37] は，シーンテキストベースの VQA に一般的に使用され，TextCaps [50] と COCO-Text-Captioned [51] は，シーンテキストを考慮した画像キャプショニングに適している．Visual Genome [52] には，シーンテキストと詳細なアノテーションを持つ画像が含まれており，vision-language グラウンディングと推論タスクをサポートしている．最後に，RefText [53] は，参照表現理解のベンチマークであり，画像-オブジェクトペアと参照表現を提供し，その一部はコンテキストの手がかりとしてシーンテキストを活用している．マルチモーダル検索のための多くの既存ベンチマークの中で，ほとんどはシーンテキストを欠いているため，我々のタスクには適していない．シーンテ

キストを含むベンチマークは，通常，マルチモーダル検索用に設計されていないか，ロボットタスクに適していないか，または過度に単純な指示を含んでいる．これらの制限により，シーンテキストと視覚コンテンツにわたるきめ細かい推論を必要とする，ロボットのための実世界の指示ベースの検索タスクを評価するには不十分である．

7. 既存手法との違い

STARE は，2 つの重要な側面で既存のマルチモーダル検索手法と異なる．第一に，ほとんどの先行手法が屋内環境での検索に焦点を当てているのに対し，STARE は広範囲の屋内および屋外環境にわたってタスクを実行し，実世界のシナリオにより適用可能である．第二に，STARE は画像からシーンテキストを抽出および利用するが，既存の手法はシーンテキストを明示的に組み込んでいない．次に，STARE は，シーンテキストを処理する方法において，シーンテキストを扱うまたはシーンテキストを考慮したマルチモーダル検索を実行する既存の手法と異なる．シーンテキストをそのまま使用する先行手法とは対照的に，STARE は，対応するオブジェクトとその属性との関連性を捉えたナラティブ表現にシーンテキストを統合する．また，本手法は，シーンテキストと参照されるオブジェクト間の複雑な関係を効率的にモデル化する補助類似度関数を導入する．

参考文献

[1] Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, et al. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. In *CoRL*, pp. 287–318, 2023.

[2] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, et al. RT-1: Robotics Transformer for Real-World Control at Scale. *RSS*, 2023.

[3] Xuan Xiao, Jiahang Liu, Zhipeng Wang, Yanmin Zhou, Yong Qi, Shuo Jiang, Bin He, and Qian Cheng. Robot Learning in the Era of Foundation Models: A Survey. *Neurocomputing*, Vol. 638, p. 129963, 2025.

[4] Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A Survey on Vision-Language-Action Models for Embodied AI. *arXiv preprint arXiv:2405.14093*, 2024.

[5] Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. Image-text Retrieval: A Survey on Recent Research and Development. In *IJCAI*, pp. 5376–5383, 2022.

[6] Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Shen Tao. Cross-Modal Retrieval: A Systematic Review of Methods and Future Directions. *Proceedings of the IEEE*, pp. 1–39, 2025.

[7] Shangbang Long, Xin He, and Cong Yao. Scene Text Detection and Recognition: The Deep Learning Era. *IJCV*, Vol. 129, pp. 161–184, 2018.

[8] Neeraj Gupta and Anand Jalal. Traditional to transfer learning progression on scene text detection and recognition: a survey. *Artificial Intelligence Review*, Vol. 55, No. 4, pp. 3457–3502, 2022.

[9] Danny Driess, F. Xia, Mehdi Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, et al. PaLM-E: An Embodied Multimodal Language Model. In *ICML*, pp. 8469–8488, 2023.

[10] Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization. In *CoRL*, 2025.

[11] Junjie Wen, Yichen Zhu, Jinming Li, Zhibin Tang, Chaomin Shen, and Feifei Feng. DexVLA: Vision-Language Model with Plug-In Diffusion Expert for General Robot Control. *CoRL*, 2025.

[12] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Baptiste Alayrac, Arenas Gonzalez, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini Robotics: Bringing AI into the Physical World. *arXiv preprint arXiv:2503.20020*, 2025.

[13] Puhao Li, Yingying Wu, Ziheng Xi, Wanlin Li, Yuzhe Huang, Zhiyuan Zhang, Yinghan Chen, Jianan Wang, Song-Chun Zhu, Tengyu Liu, and Siyuan Huang. ControlVLA: Few-shot Object-centric Adaptation for Pre-trained Vision-Language-Action Models. *CoRL*, 2025.

[14] Quanting Xie, So Yeon Min, Pengliang Ji, Yue Yang, Tianyi Zhang, Kedi Xu, Aarav Bajaj, Ruslan Salakhutdinov, Matthew Johnson-Roberson, and Yonatan Bisk. Embodied-RAG: General Non-parametric Embodied Memory for Retrieval and Generation. *arXiv preprint arXiv:2409.18313*, 2025.

[15] Yichen Zhu, Zhicai Ou, Xiaofeng Mou, and Jian Tang. Retrieval-Augmented Embodied Agents. In *CVPR*, pp. 17985–17995, 2024.

[16] Weiye Xu, Min Wang, Wengang Zhou, and Houqiang Li. P-RAG: Progressive Retrieval Augmented Generation For Planning on Embodied Everyday Task. In *ACMMM*, p. 6969–6978, 2024.

[17] Gianluca Monaci, Rafael Rezende, Romain Deffayet, Gabriela Csurka, Guillaume Bono, Herve Dejean, Stephane Clinchant, and Christian Wolf. RANa: Retrieval-Augmented Navigation. *arXiv preprint arXiv:2504.03524*, 2025.

[18] Kuanning Wang, Yuqian Fu, Tianyu Wang, Yanwei Fu, Longfei Liang, Yu-Gang Jiang, and Xiangyang Xue. RAG-6DPose: Retrieval-Augmented 6D Pose Estimation via Leveraging CAD as Knowledge Base. In *IROS*, 2025.

[19] Zihan Wang, Yaohui Zhu, Gim Hee Lee, and Yachun Fan. NavRAG: Generating User Demand Instructions for Embodied Navigation through Retrieval-Augmented LLM. In *ACL*, pp. 8430–8440, 2025.

[20] Sheng Fan, Rui Liu, Wenguan Wang, and Yi Yang. Navigation Instruction Generation with BEV Perception and Large Language Models. In *ECCV*, 2024.

[21] Abrar Anwar, John Welsh, Joydeep Biswas, Soha Pouya, and Yan Chang. ReMEmbR: Building and Reasoning Over Long-Horizon Spatio-Temporal Memory for Robot Navigation. In *ICRA*, pp. 2838–2845, 2025.

[22] Sateesh Kumar, Shivin Dass, Georgios Pavlakos, and Roberto Martin. COLLAGE: Adaptive Fusion-based Retrieval for Augmented Policy Learning. In *CoRL*, 2025.

[23] Sriram Yenamandra, A. Ramachandran, Karmesh Yadav, Austin Wang, Mukul Khanna, Théophile Gervet, Tsung Yang, Vidhi Jain, Alexander Clegg, John Turner, Zsolt Kira, Manolis Savva, Angel Chang, Devendra Chaplot, et al. HomeRobot: Open-Vocabulary Mobile Manipulation. In *CoRL*, pp. 1975–2011, 2023.

[24] Mauricio Matamoros, Viktor Seib, Raphael Memmesheimer, and Dietrich Paulus. RoboCup@Home: Summarizing achievements in over eleven years of competition. In *ICARSC*, pp. 186–191, 2018.

[25] Boyuan Chen, Fei Xia, Brian Ichter, Kanishka Rao, et al. Open-vocabulary Queryable Scene Representations for Real World Planning. In *ICRA*, pp. 11509–11522, 2023.

[26] Gunnar Sigurdsson, Jesse Thomason, Gaurav Sukhatme, and Robinson Piramuthu. RREx-BoT: Remote Referring Expressions with a Bag of Tricks. In *IROS*, pp. 5203–5210, 2023.

[27] R. Korekata, K. Kaneda, S. Nagashima, Y. Imai, and K. Sugiura. DM^{2}RM: Dual-Mode Multimodal Ranking for Target Objects and Receptacles Based on Open-Vocabulary Instructions. *AR*, Vol. 39, No. 5, pp. 243–258, 2025.

[28] Daichi Yashima, Ryosuke Korekata, and Komei 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.

[29] Zhen Zhao, Jingqun Tang, Chunhui Lin, Binghong Wu, Hao Liu, Zhizhong Zhang, Xin Tan, Can Huang, and Yuan Xie. Multi-modal In-Context Learning Makes an Ego-evolving Scene Text Recognizer. In *CVPR*, pp. 15567–15576, 2023.

[30] Jianjun Xu, Yuxin Wang, Hongtao Xie, and Yongdong Zhang. OTE: Exploring Accurate Scene Text Recognition Using One Token. In *CVPR*, pp. 28327–28336, 2024.

[31] Min Liang, Wei Ma, Xiaobin Zhu, Jingyan Qin, and Xu-Cheng Yin. Layoutformer: Hierarchical text detection towards scene text understanding. In *CVPR*, pp. 15665–15674, 2024.

[32] Sergi Garcia-Bordils, Dimosthenis Karatzas, and Marçal Rusiñol. STEP – Towards Structured Scene-Text Spotting. In *WACV*, pp. 883–892, 2024.

[33] Jinzhi Zheng, Heng Fan, and Libo Zhang. Kernel Adaptive Convolution for Scene Text Detection via Distance Map Prediction. In *CVPR*, pp. 5957–5966, 2024.

[34] Gangyan Zeng, Yuan Zhang, Jin Wei, Dongbao Yang, Peng Zhang, Yiwen Gao, Xugong Qin, and Yu Zhou. Focus, Distinguish, and Prompt: Unleashing CLIP for Efficient and Flexible Scene Text Retrieval. In *MM*, pp. 2525–2534, 2024.

[35] Mengjun Cheng, Yipeng Sun, Longchao Wang, Xiongwei Zhu, Kun Yao, Jie Chen, Guoli Song, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. ViSTA: Vision and Scene Text Aggregation for Cross-Modal Retrieval. In *CVPR*, pp. 5174–5183, 2022.

[36] Ali Biten, Ron Litman, Yusheng Xie, Srikar Appalaraju, and R. Manmatha. LaTr: Layout-Aware Transformer for Scene-Text VQA. In *CVPR*, pp. 16527–16537, 2021.

[37] Ali Furkan Biten, Rubèn Tito, Andres Mafla, Lluis Gomez, Marçal Rusiñol, C.V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Scene Text Visual Question Answering. In *ICCV*, pp. 4290–4300, 2019.

[38] Difei Gao, Ke Li, Ruiping Wang, Shiguang Shan, and Xilin Chen. Multi-Modal Graph Neural Network for Joint Reasoning on Vision and Scene Text. In *CVPR*, pp. 12746–12756, 2020.

[39] Jialei Cui, Jianwei Du, Wenzhuo Liu, and Zhouhui Lian. TextNeRF: A Novel Scene-Text Image Synthesis Method based on Neural Radiance Fields. In *CVPR*, pp. 22272–22281, 2024.

[40] Chen Duan, Pei Fu, Shan Guo, Qianyi Jiang, and Xiaoming Wei. ODM: A Text-Image Further Alignment Pre-training Approach for Scene Text Detection and Spotting. In *CVPR*, pp. 15587–15597, 2024.

[41] Joshua Santoso and Christian Williem Simon. On Manipulating Scene Text in the Wild with Diffusion Models. In *WACV*, pp. 5202–5211, 2024.

[42] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. OCR-VQA: Visual Question Answering by Reading Text in Images. In *ICDAR*, pp. 947–952, 2019.

[43] Lluis Gomez, Andres Mafla, Marçal Rusiñol, and Dimosthenis Karatzas. Single Shot Scene Text Retrieval. In *ECCV*, pp. 700–715, 2018.

[44] Jacob Devlin, Ming Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, pp. 4171–4186, 2019.

[45] Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, pp. 740–755, 2014.

[46] Peter Young, et al. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. *TACL*, Vol. 2, pp. 67–78, 2014.

[47] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, et al. REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments. In *CVPR*, pp. 9979–9988, 2020.

[48] Mukul Khanna, Ram Ramrakhya, Gunjan Chhablani, Sriram Yenamandra, Theophile Gervet, Matthew Chang, Zsolt Kira, Singh Chaplot, Dhruv Batra, and Roozbeh Mottaghi. GOAT-Bench: A Benchmark for Multi-Modal Lifelong Navigation. In *CVPR*, pp. 16373–16383, 2024.

[49] Kanta 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.

[50] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. TextCaps: A Dataset for Image Captioning with Reading Comprehension. In *ECCV*, pp. 742–758, 2020.

[51] Andres Mafla, S. Rezende, Lluis Gómez, Diane Larlus, and Dimosthenis Karatzas. StacMR: Scene-Text Aware Cross-Modal Retrieval. In *WACV*, pp. 2219–2229, 2021.

[52] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li Li, David Shamma, Michael Bernstein, and Li Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *IJCV*, Vol. 123, No. 1, pp. 32–73, 2017.

[53] Yuqi Bu, et al. Scene-Text Oriented Referring Expression Comprehension. *IEEE TMM*, Vol. 25, pp. 7208–7221, 2023.