

行動実現性を考慮した 階層型マルチモーダル検索に基づく移動マニピュレーション

○是方諒介^{1,2,3}, Quanting Xie³, Yonatan Bisk³, 杉浦孔明^{1,2}
¹ 慶應義塾大学, ² 慶應 AI センター, ³ Carnegie Mellon University

言語指示に基づく移動マニピュレーションにおいて、視覚言語的な正しさのみならず物体操作の実現性が高い対象物体および配置目標を特定することは有益である。本研究では、視覚・領域的な特徴を統合する階層的な Embodied Memory を構築し、VLM を用いて予測された行動実現性を考慮したマルチモーダル検索手法を提案する。実験の結果、提案手法は既存手法を上回る検索性能およびタスク成功率を達成した。

1. はじめに

サービスロボットは家庭、病院、および倉庫などにおいて需要が高まっており、自然言語による指示を理解できれば利便性向上に寄与する。そこで本研究では、Open-Vocabulary Mobile Manipulation (OVMM [1]) の実現を目的とする。具体的には、例えば “Please bring the paper towels to the kitchen counter.” のような指示文が生活支援ロボットに与えられるユースケースを想定する。指示文が与えられたとき、ロボットは実世界において視覚言語的に正しく、行動実現性（例: “pick”, “place”）の高い対象物体および配置目標を特定することが求められる。このように、open-vocabulary なマルチモーダル言語理解と行動実現性の考慮を同時に行う必要がある点で困難である。

本タスクでは環境中に何百と存在する大量の候補画像に対する高速な推論が要求されるため、Vision-Language Model (VLM) を直接適用する手法は非現実的である。OVMM タスクを扱う既存手法 [1, 2] は、物体カテゴリが同一でも参照表現により区別され得る候補の識別が困難である。また、マルチモーダル検索手法 [3-6] に基づく手法においては、ロボットの行動実現性が考慮されておらず物体操作時に失敗してしまう場合がある。

これらの課題に対し、本研究ではロボットの行動実現性を考慮し多階層の視覚言語を統合する階層型マルチモーダル検索手法である Feasible RAG を提案する。図 1 に、提案手法の概要を示す。本手法と既存手法との主要な違いは、物体に対するロボットの行動実現性を考慮した階層的な Feasibility-Aware Embodied Memory (FAMe) を導入する点である。視覚・領域的な特徴を統合する階層型マルチモーダル検索に加え、行動実現性に基づくリランキングを通じて視覚言語的に正しく成功率の高い物体の検索を実現する。本研究の主な貢献は、以下の 3 点である。

- Multi-Level Fusion により視覚・領域的な特徴を統合する、階層型マルチモーダル検索手法である Feasible RAG を提案する。
- 階層的な FAMe を構築するため、Visual Prompt に基づく VLM を用いてインスタンスレベルの行動実現性を推定する Affordance Proposer を導入する。
- アフォーダンスを考慮したフィルタリングおよび Large Language Model (LLM) を用いたインスタンスマッチングに基づき、視覚言語的な正しさと行動実現性向上の両立を可能にする Feasibility-Aware Reranking を導入する。

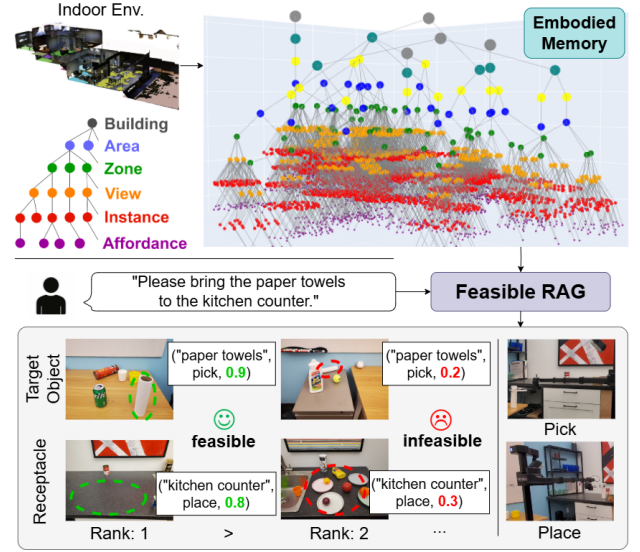


図 1 タスクの概要

2. 問題設定

本研究では、マルチモーダル検索に基づく移動マニピュレーションタスクに取り組む。本タスクでは、自然言語による指示が与えられたうえで、ロボットが環境中の画像群から対象物体および配置目標の画像を検索し、それらに基づき物体操作を実行する。したがって、本タスクはマルチモーダル検索および動作実行の 2 つのサブタスクから構成される。マルチモーダル検索においては、対象物体および配置目標の正解画像が、それぞれの検索結果において高くランク付けられることが期待される。

入力は、 $\mathbf{x} = \{\mathbf{x}_{\text{inst}}, \mathbf{X}_{\text{img}}\}$, $\mathbf{X}_{\text{img}} = \{\mathbf{x}_{\text{img}}^{(j)}\}_{j=1}^{N_{\text{img}}}$ と定

義される。ここで、 $\mathbf{x}_{\text{inst}}$ は指示文、 $\mathbf{x}_{\text{img}}^{(j)} \in \mathbb{R}^{3 \times W \times H}$ は幅 W 、高さ H の RGB 画像を示す。また、 N_{img} は候補画像の枚数を表す。出力は、対象物体および配置目標についてそれぞれランク付けされた 2 つの画像リストである。

3. 提案手法

図 2 に、提案手法の構造を示す。本研究では、OVMM のための階層型マルチモーダル検索手法である Feasible RAG を提案する。そこで、環境内で実行可能なロボットのアフォーダンスを理解するために、事前探索により得られた観測画像から構築される Feasibility-Aware Embodied Memory (FAMe) を導入する。

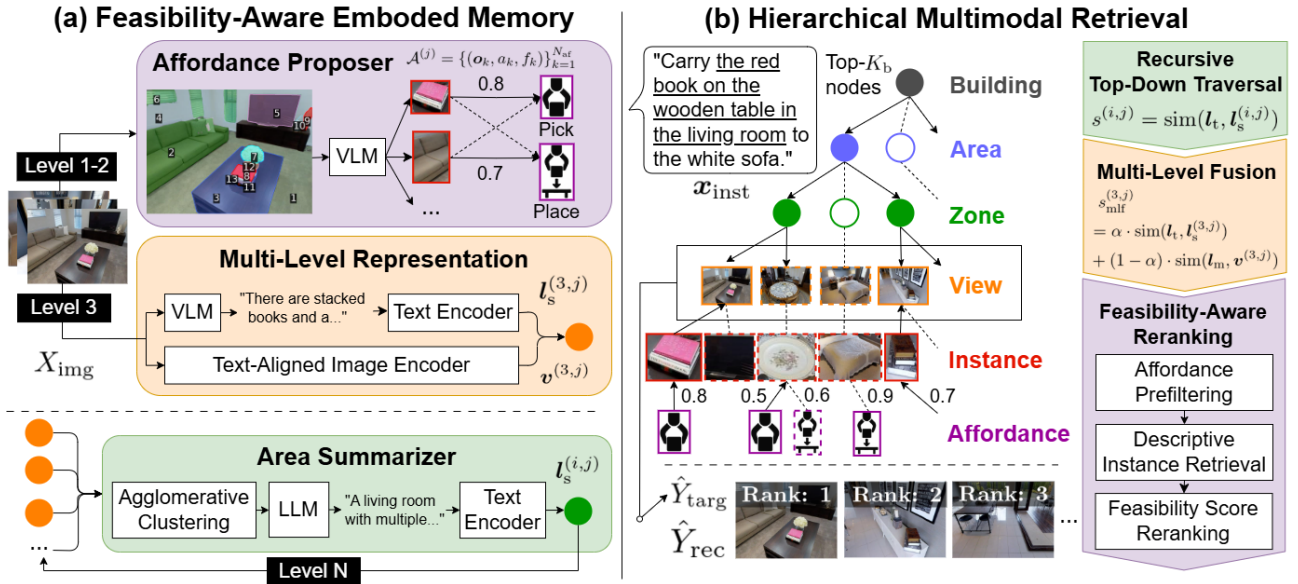


図2 提案手法における (a) Feasibility-Aware Embodied Memory (FAME) の構築および (b) 階層型マルチモーダル検索

3.1 Feasibility-Aware Embodied Memory

3.1.1 Affordance Proposer (Level 1-2)

本研究では、図2(a)に示すように、行動実現性を考慮したロボットのアフォーダンスから、インスタンス、視点、ゾーン、エリア、および建物に至るまでの複数の階層にわたるノードで構成される階層的なFAMEを構築する。本モジュールでは、 $x_{img}^{(j)}$ に対してインスタンスレベルのVisual Promptを適用し、得られた画像をVLM (GPT-4o) に入力することで、各インスタンスに関する記述とロボットのアフォーダンスに関する行動実現性を抽出する。これらはそれぞれ、レベル2およびレベル1のノードとして格納される。本モジュールにより、既存の物体検出やセマンティックセグメンテーションにより構築されるカテゴリレベルのノード（例：“cup”）[7]に留まらず、“red metal mug”のように記述的なインスタンスレベルの表現を獲得することが可能になる。具体的には、SEEM [8]を用いて $x_{img}^{(j)}$ に対するセグメンテーションマスクを生成し、元の画像と、各領域にインデックス番号を重畳したVisual Prompt付きの画像の両方をVLMに並列で入力する。出力は、集合 $\mathcal{A}^{(j)} = \{(o_k, a_k, f_k)\}_{k=1}^{N_{af}}$ である。ここで o_k はインスタンスに関する表現、 a_k はロボットのアフォーダンス、 f_k は行動実現可能性スコア、 N_{af} は組合せ数をそれぞれ表す。

3.1.2 Multi-Level Representation (Level 3)

本モジュールでは、 $x_{img}^{(j)}$ に対して、高レベルの領域的な特徴と低レベルの視覚的な特徴を明示的に統合するMulti-Level Representationを獲得してレベル3のノードとして格納する。具体的には、前者はVLMにより生成された記述文をtext encoder (text-embedding-3-large) で埋め込むことで特徴量 $l_s^{(3,j)} \in \mathbb{R}^{d_t}$ を得る。後者の特徴量 $v^{(3,j)} \in \mathbb{R}^{d_m}$ は、マルチモーダル基盤モデル (BEiT-3 [6]) のimage encoderを用いて得られる。ここで、 d_t および d_m は、それぞれの埋め込み表現の次元数を表す。これらの特徴を統合した検索に関しては、Sec. 3.2.2 で詳述する。

3.1.3 Area Summarizer (Level N)

本モジュールでは、凝集型クラスタリングを適用してレベル i のノードを集約することにより、 $i \geq 3$ に対してレベル $(i+1)$ のノードを構築する。既存のマルチモーダル基盤モデル [3–6] は、各画像を独立した特徴量として扱っており、視覚外のコンテキストを捉えることが困難である。そこで本モジュールでは、[9]と同様にノード間の物理・意味的距離の両方に基づいて凝集型クラスタリングを行う。この際、各クラスタに対してLLMを用いて要約文を生成し、text encoderで埋め込むことで特徴量 $l_s^{(i+1,j)} \in \mathbb{R}^{d_t}$ を得る。

3.2 階層型マルチモーダル検索

3.2.1 Recursive Top-Down Traversal

図2(b)に、指示文が与えられた際にFAMEに対して階層型マルチモーダル検索を行う手順を示す。第1段階として本モジュールでは、レベル N からレベル3に向けてFAMEを再帰的に探索する。各レベル i において、類似度スコア $s^{(i,j)} = \text{sim}(l_t, l_s^{(i,j)})$ に基づき、上位 K_b 個のノードを選択する。ここで、 $\text{sim}(\cdot, \cdot)$ はコサイン類似度、 $l_t \in \mathbb{R}^{d_t}$ は指示文の言語特徴量、 $l_s^{(i,j)} \in \mathbb{R}^{d_t}$ はレベル i における j 番目のノードの言語特徴量を表す。本モジュールにより、エリアやゾーンといった領域的な特徴に関するフィルタリングが可能となる。本モジュールの出力として、視点レベル（レベル3）のノード集合 $S \subseteq (v^{(3,j)}, l_s^{(3,j)})$ が得られる。

3.2.2 Multi-Level Fusion

本モジュールでは、Multi-Level Representationで得られた視覚・領域的な特徴を用いた複合的なマルチモーダル検索を行う。具体的には、 S の各要素に対して、以下の類似度スコア $s_{mlf}^{(3,j)}$ を計算する。

$$s_{mlf}^{(3,j)} = \alpha \cdot \text{sim}(l_t, l_s^{(3,j)}) + (1 - \alpha) \cdot \text{sim}(l_m, v^{(3,j)})$$

ここで、 $\alpha \in [0, 1]$ はハイパーパラメータ、 $l_m \in \mathbb{R}^{d_m}$ はマルチモーダル基盤モデル (BEiT-3) のtext encoderから得られる特徴量を表す。既存研究では、第1項（例：[9]）または第2項（例：[3–6]）の一方のみを単独で用いていたが、これらを複合することでより包括的な言語理解が可能になると期待される。

表 1 提案手法およびベースライン手法の定量的比較結果

[%]	Target Object			Receptacle			Overall		
	R@5 ↑	R@10 ↑	R@20 ↑	R@5 ↑	R@10 ↑	R@20 ↑	R@5 ↑	R@10 ↑	R@20 ↑
手法									
CLIP [3]	15.7	24.2	33.6	6.2	11.7	21.7	10.9	18.0	27.7
Long-CLIP [4]	24.6	36.1	48.5	3.5	9.2	19.9	14.0	22.6	34.2
BLIP-2 [5]	<u>30.2</u>	40.7	48.0	5.2	10.3	19.9	17.7	25.5	34.0
BEiT-3 [6]	29.0	<u>42.1</u>	<u>53.8</u>	<u>8.9</u>	15.4	27.6	<u>19.0</u>	<u>28.7</u>	<u>40.7</u>
HomeRobot* [1]	5.9	10.2	12.9	1.7	3.9	8.4	3.8	7.0	10.7
NLMap* [10]	15.1	19.2	30.4	7.3	15.1	23.6	11.2	17.2	27.0
RelaX-Former [11]	17.9	26.5	37.7	8.8	<u>19.8</u>	<u>28.4</u>	13.3	23.2	33.0
Embodied-RAG [9]	15.1	18.5	22.8	6.7	11.3	14.4	10.9	14.9	18.6
Feasible RAG (提案手法)	32.8	49.9	61.7	14.8	24.3	30.2	23.8	37.1	45.9

3.2.3 Feasibility-Aware Reranking

本モジュールでは、Multi-Level Fusionによってランク付けされた S 内の上位 K_r 個のノードに対して、レベル 1 および 2 の $A^{(j)}$ を用いてリランキングを行う。本モジュールは、Affordance Prefiltering, Descriptive Instance Retrieval, および Feasibility Score Reranking (FSR) の 3 段階で構成される。第 1 段階では、アフォーダンスに基づきインスタンスノードをフィルタリングする。具体的には、対象物体および配置目標のどちらを検索するかに応じてそれぞれ “pick” および “place” に相当するノードを対象とする。これにより、視覚的だけでなく機能的な側面から候補物体を絞り込むことが可能となる。第 2 段階では、Affordance Proposerによって生成された記述表現を LLM に入力し、指示文との類似度に基づきインスタンスノードにスコアを付与する。第 3 段階では、上位 K_f のインスタンスノードに対して、行動実現性スコアに基づきリランキングを行う。以上のような階層型マルチモーダル検索は、対象物体および配置目標のそれぞれに対して個別に実行され、最終的にランク付けられた画像リスト $\hat{Y}_{\text{rec}}$ および $\hat{Y}_{\text{rec}}$ がそれぞれ得られる。

4. 実験

4.1 実験設定

本研究では、Matterport3D (MP3D [12]) データセットに基づき構築された WholeHouse-MM ベンチマークを新たに導入する。本タスクでは、数百枚の画像群を含む建物規模の屋内環境および移動マニピュレーション指示文が求められる。OVMMに関する既存のベンチマークの多くは、テンプレートの指示文を用いている [1]、小規模な環境を前提としている [11, 13]。そこで、屋内環境における移動やシーン理解の用途で標準的に用いられる MP3D から大規模に画像を収集した。各環境には、平均して 590 枚の画像が含まれる。MP3D は環境の地図を提供しているため、各 waypoint においてパノラマ画像を 60 度ずつ 6 枚の画像に切り出した。さらに WholeHouse-MM ベンチマークでは、[11, 13] と同様に、クラウドソーシングによって指示文を手でアノテーションした。合計 116 名のアノデータに対し、環境中の 2 枚の画像（対象物体・配置目標）を提示したうえで、free-form の移動マニピュレーション指示文を作成するよう依頼した。

本ベンチマークは、実環境から収集された 2,360 枚の画像および 402 の指示文で構成される。語彙数は 517 語、総単語数は 6,410 語であり、1 文あたりの平均語数は 15.9 語である。検証集合はハイパーパラメータの調整に、テスト集合は手法の評価に用いた。

4.2 実験結果

表 1 に、提案手法およびベースライン手法の定量的比較結果を示す。評価指標として、 $\text{recall}@K$ ($K = 5, 10, 20$) を用いた。主要評価指標は $\text{recall}@10$ である。対象物体、配置目標、および全体での評価値を報告する。なお、本ベンチマークでは検索性能の評価に焦点を当てており、ロボットの物理的な操作は行わないため、提案手法における FSR は除外した。FSR に関する実験結果は、5. 章を参照されたい。ベースライン手法として、代表的なマルチモーダル基盤モデル [3–6]、OVMM タスクにおいて良好な結果が得られている手法 [1, 10, 11]、および階層的な Embodied Memory を用いる手法 [9] の合計 8 つを用いた。

表 1 より、提案手法は対象物体、配置目標、および全体での $\text{recall}@10$ において、それぞれ 49.9%、24.3%、および 37.1% を達成した。一方で、ベースライン手法における最良値はそれぞれ 42.1%、19.8%、および 28.7% であり、提案手法はこれらをそれぞれ 7.8、4.5、および 8.4 ポイント上回った。同様に、すべての評価指標において提案手法はベースラインを上回った。

5. 実機実験

5.1 実験設定

本実験には、DexWrist を搭載した Hello Robot Stretch 2 [14] を使用した。本ロボットは、OVMM タスクにおける標準的なプラットフォームとして広く用いられている [1, 2]。実験はオフィスおよびキッチンから構成される、広さ $5.0 \times 7.0 \text{ m}^2$ の屋内環境を用い、10 種類の異なる家具が設置した。対象物体として、20 種類の日常物体を用いた。

まず、ロボットは FAME を構築するために環境中の viewpoint を探索して巡回して、Intel RealSense D435i カメラを用いて RGB 画像を収集した。経路計画および移動には、標準的な地図ベースの手法を用いた。次に、ユーザにより無作為に選ばれた物体を家具まで運搬する free-form の指示文が与えられた。合計で 40 回の試行を行い、各試行ごとに異なる指示文が用いられ

表 2 実機実験における定量的結果

手法 [%]	R@5 ↑	SR ↑
BEiT-3 [6]	79	45
Feasible RAG (w/o FSR)	94	70
Feasible RAG (full)	94	85

た. 指示文が与えられると, 対象物体および配置目標の画像が検索され, それぞれ上位 5 件の画像がユーザーに提示された. 最後に, ロボットはユーザーが選択した画像に基づき物体把持, 運搬, および物体配置動作を行った. なお, これらの動作に関する軌道生成はヒューリスティックに行われた.

5.2 実験結果

表 2 に, 実機実験における定量的結果を示す. 本実験では, 表 1 において最良のベースライン手法であった BEiT-3 [6] との比較を行った. 評価指標としては, recall@5 およびタスク成功率 (SR) を用いた. 表 2 に示すように, 提案手法は recall@5 において 94% を達成し, ベースライン手法を 15 ポイント上回った. 同様に, 提案手法は SR において 85% を達成し, ベースライン手法を 40 ポイント上回った. さらに, 提案手法における FSR の有効性を検証するため, 本モジュールを除いた ablation study を行った. 表 2 より, 提案手法は FSR を含む場合に 85% の SR を達成し, FSR を除いた場合の 70% と比較して 15 ポイント改善した. 以上より, FSR を用いて把持・配置の行動実現性の高い物体を優先的に選択することで, SR の向上に寄与することが示唆される.

図 3 に, 実機実験における成功例を示す. 正解画像および意味的には正しいが行動実現性が低い画像は, それぞれ緑色およびオレンジ色の枠線で示される. 本例における x_{inst} は, “Please deliver a cup to the desk that has some coffee powder on it.” であった. 結果より, 提案手法はキッチンカウンターに直立する緑色のカップを 1 位にランク付けた一方, FSR を除いた場合は冷蔵庫の上の奥に置かれたカップや倒れたカップなど, 行動実現性の低い候補を上位にランク付けた. 同様に, 提案手法は整理された机を 1 位にランク付けたのに対し, FSR を除いた場合は散らかった机をより上位にランク付けた. 以上より, FSR は曖昧な指示文が与えられた際により行動実現性が高い物体を上位に提示することで SR の向上に寄与すると考えられる.

6. おわりに

本研究では, 環境の画像群から対象物体および配置目標を検索して移動マニピュレーションを行うタスクを扱った. 我々は, 行動実現性を考慮した FAME に基づき, 視覚・領域的な特徴を統合する階層型マルチモーダル検索手法である Feasible RAG を提案した. 新たに構築した WholeHouse-MM ベンチマークおよび実機実験において, 提案手法は標準的な評価指標で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 ムーンショット, JSPS 特別研究員奨励費 JP25KJ2068, および慶應義塾大学と Carnegie Mellon University とのパートナーシップの一環として Microsoft Corporation の助成を受けて実施されたものである.

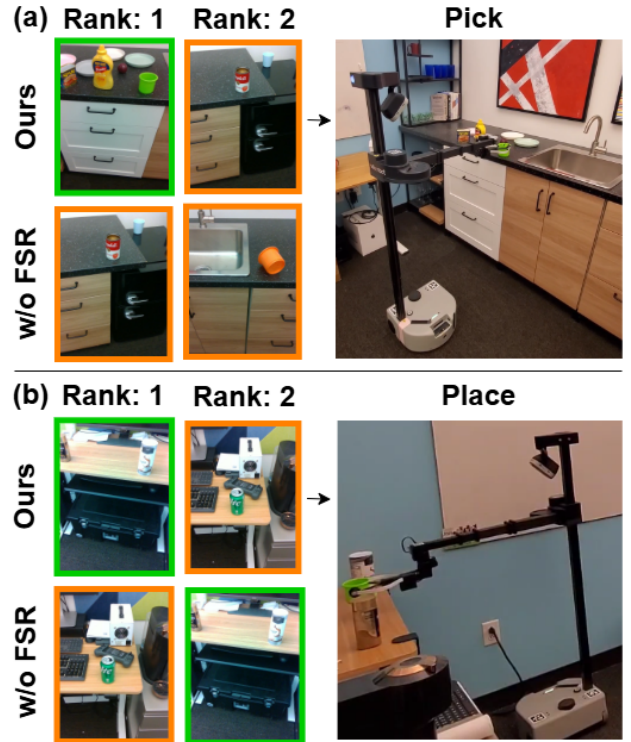


図 3 実機実験における定性的結果

参考文献

- [1] S. Yenamandra, et al., “HomeRobot: Open-Vocabulary Mobile Manipulation,” CoRL, pp.1975–2011, 2023.
- [2] P. Liu, Y. Orru, C. Paxton, et al., “OK-Robot: What Really Matters in Integrating Open-Knowledge Models for Robotics,” arXiv preprint arXiv:2401.12202, 2024.
- [3] 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.
- [4] B. Zhang, P. Zhang, et al., “Long-CLIP: Unlocking the Long-Text Capability of CLIP,” ECCV, pp.310–325, 2024.
- [5] 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.
- [6] 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.
- [7] D. Honerkamp, M. Büchner, F. Despinoy, T. Welschhold, and A. Valada, “Language-Grounded Dynamic Scene Graphs for Interactive Object Search with Mobile Manipulation,” IEEE RA-L, vol.9, no.10, pp.2377–3766, 2024.
- [8] X. Zou, J. Yang, H. Zhang, et al., “Segment Everything Everywhere All at Once,” NeurIPS, pp.19769–19782, 2023.
- [9] Q. Xie, S. Min, P. Ji, Y. Yang, et al., “Embodied-RAG: General Non-parametric Embodied Memory for Retrieval and Generation,” arXiv preprint arXiv:2409.18313, 2024.
- [10] B. Chen, F. Xia, B. Ichter, K. Rao, K. Gopalakrishnan, et al., “Open-vocabulary Queryable Scene Representations for Real World Planning,” ICRA, pp.11509–11522, 2023.
- [11] 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.
- [12] A. Chang, et al., “Matterport3D: Learning from RGB-D Data in Indoor Environments,” 3DV, pp.667–676, 2017.
- [13] R. Korekata, K. Kaneda, S. Nagashima, et al., “DM²RM: Dual-Mode Multimodal Ranking for Target Objects and Receptacles Based on Open-Vocabulary Instructions,” Advanced Robotics, vol.39, no.5, pp.243–258, 2025.
- [14] C. Kemp, A. Edsinger, et al., “The Design of Stretch: A Compact, Lightweight Mobile Manipulator for Indoor Human Environments,” ICRA, pp.3150–3157, 2022.