

Affordance RAG: Hierarchical Multimodal Retrieval with Affordance-Aware Embodied Memory for Mobile Manipulation

Ryosuke Korekata^{1,2,3}, Quanting Xie³, Yonatan Bisk³, and Komei Sugiura^{1,2}

Abstract—In this study, we address the problem of open-vocabulary mobile manipulation, where a robot is required to carry a wide range of objects to receptacles based on free-form natural language instructions. This task is challenging, as it involves understanding visual semantics and the affordance of manipulation actions. To tackle these challenges, we propose Affordance RAG, a zero-shot hierarchical multimodal retrieval framework that constructs Affordance-Aware Embodied Memory from pre-explored images. The model retrieves candidate targets based on regional and visual semantics and reranks them with affordance scores, allowing the robot to identify manipulation options that are likely to be executable in real-world environments. Our method outperformed existing approaches in retrieval performance for mobile manipulation instruction in large-scale indoor environments. Furthermore, in real-world experiments where the robot performed mobile manipulation in indoor environments based on free-form instructions, the proposed method achieved a task success rate of 85%, outperforming existing methods in both retrieval performance and overall task success.

I. INTRODUCTION

As robots are increasingly deployed in real-world human environments, such as homes, hospitals, and warehouses, there is a growing demand for systems that can understand and execute flexible, language-driven instructions. The goal of our work is open-vocabulary mobile manipulation (OVMM [1]) guided, where a robot is required to identify and interact with objects and receptacles described in free-form language. Given an instruction and a set of pre-explored environment images, the robot must retrieve the appropriate target object and receptacle, that it can successfully manipulate (e.g., pick and place) in the real-world.

A typical use case of our target problem is a domestic service robot instructed with a natural language instruction such as “Please bring the paper towels to the kitchen counter.” To execute this instruction, the robot must first identify the target object and the receptacle from a set of previously observed images of the environment. Furthermore, when multiple paper towels are present in the environment, the robot is expected to select the one with a higher grasping affordance. Similarly, within the kitchen counter, it should prefer an area that is uncluttered and more suitable for object placement. This task is challenging due to the need for open-vocabulary grounding and affordance-aware reasoning.

A naive approach would involve directly applying a vision-language model (VLM) to evaluate every candidate image

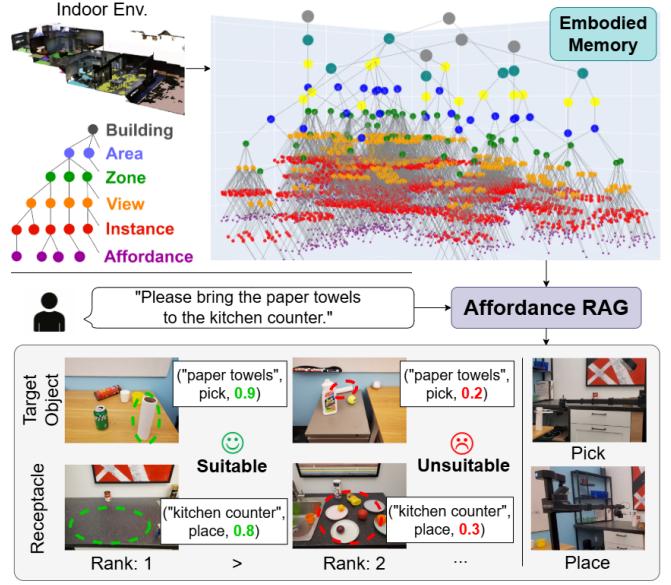


Fig. 1: Overview of Affordance RAG for open-vocabulary mobile manipulation. The robot first constructs an embodied memory based on images collected during pre-exploration of the environment. When a free-form instruction is given, hierarchical multimodal retrieval is performed over the embodied memory to identify the target object and receptacle. To improve task success rates, candidates with higher affordance scores are prioritized during retrieval.

against the instruction. However, this becomes computationally impractical in realistic settings, where hundreds of candidate views or object instances must be considered per scene. Most existing approaches (e.g., [1]–[3]) and multimodal retrieval methods (e.g., [4]–[7]), tend to misidentify visually similar but semantically incorrect candidates (e.g., confusing a lotion bottle with a water bottle). Crucially, these methods lack affordance awareness—retrieving objects that are physically non-manipulable, resulting in downstream execution failures.

To address the limitations, we propose Affordance RAG, a hierarchical multimodal retrieval framework that integrates multi-level semantic representations with robotic affordance-aware reasoning. Fig. 1 shows an overview of the proposed method. The main difference between our method and prior approaches lies in its hierarchical multimodal retrieval framework that fuses regional and visual semantics, and its ability to incorporate affordance-aware reasoning through affordance-centric memory and reranking. While existing methods typically perform flat similarity matching, our method constructs a structured Affordance-Aware Embodied Memory (Affordance Mem) and refines multi-

¹Keio University, ²Keio AI Research Center, ³Carnegie Mellon University. rkorekata@keio.jp

This work was partially supported by Microsoft Corporation as part of the Keio CMU partnership, JSPS Fellows Grant Number JP25KJ2068, JSPS KAKENHI Grant Number 23K03478, and JST Moonshot.

modal retrieval through VLM-based affordance estimation and large language model (LLM)-guided object selection, enabling robust zero-shot mobile manipulation. We introduce Affordance-Aware Reranking to address a fundamental limitation in existing retrieval-based approaches: their inability to distinguish between semantically relevant candidates and those more suitable for execution. Our Affordance-Aware Reranking module filters candidates using VLM-predicted affordances and scores instance-level descriptions with an LLM for relevance, followed by reranking top candidates based on affordance scores to achieve both linguistic consistency and suitability.

The key contributions of this work are three-fold:

- We propose Affordance RAG, a zero-shot hierarchical multimodal retrieval framework that combines regional and visual semantics via Multi-Level Fusion.
- We introduce Affordance Mem by using instance-level Affordance Proposer based on VLMs via visual prompting to estimate robot affordances.
- We introduce Affordance-Aware Reranking that combines affordance prefiltering with LLM-based descriptive instance retrieval, and reranks candidates based on affordance scores to achieve both linguistic relevance and suitability.

II. RELATED WORK

A. Language-Guided Mobile Manipulation

Mobile manipulation tasks based on natural language instructions have been widely studied, with several real-world benchmarks (e.g., [1]). While these benchmarks rely on template-based instructions, our work addresses the more challenging task of free-form, OVMM. The OVMM task has been extensively studied in recent work [2], [3], [8]. Approaches that construct 3D scene graphs to represent the environment have been proposed for mobile manipulation tasks (e.g., [9]). Beyond this, scene graphs have also been applied to embodied question answering (EQA [10], [11]), navigation [12]–[14], task planning [15], and semantic mapping [16]. These approaches are constructed based on object detection or semantic segmentation, but rely heavily on object category labels, making them less suitable for the free-form OVMM task considered in this work.

B. Multimodal Retrieval

Text-image retrieval using multimodal foundation models has been extensively studied [4]–[7]. Recent work has actively explored applying these models to robotics, where a robot is given natural language instructions and retrieves target objects from pre-explored environment images to execute manipulation tasks [17]–[20]. In addition, several studies have explored applying the concept of retrieval-augmented generation (RAG) to robotics, including approaches that construct real-world environments as hierarchical embodied memories (e.g., Embodied-RAG [11], [14]). Embodied-RAG is closely related to our work, but differs in that it focuses on what is visible in the scene without explicitly considering robot affordances. In contrast, our approach models high-level robot affordances grounded in atomic

actions. Affordance modeling for object manipulation has been explored at different levels, including predicting object functionality from 3D point clouds [21], estimating contact points from images [22], and representing affordances as high-level atomic actions [23], as in our work.

C. Foundation Models for Robotics

Foundation models as multimodal AI agents hold great promise for a wide range of applications in scene understanding and planning (e.g., [24]). A growing body of research has explored applying LLMs and VLMs to robotic tasks [25]. For example, LLMs and VLMs have been applied to common-sense reasoning (e.g., [26]) and task planning [27]. Several prior works [19], [28] enhance the reasoning capabilities of VLMs by applying visual prompts to input images. In contrast, our method uses visual prompting to generate object-centric descriptions and predict robot-executable affordances from observed images.

III. PROBLEM STATEMENT

In this study, we focus on the Multimodal Retrieval-guided Mobile Manipulation (MRMM) task. In this task, given a natural language instruction, a robot performs mobile manipulation by retrieving the images of the target object and the receptacle from a set of environmental images. The target object and the receptacle are identified based on the user's selections from the top-retrieved images. This task comprises two sub-tasks: *multimodal retrieval* and *action execution*. During the multimodal retrieval phase, it is expected that the images of the target object and the receptacle are ranked highly in their respective retrieved image lists. During the action execution phase, the robot is expected to pick up the target object and transport it to the designated receptacle. The input is defined as $\mathbf{x} = \{\mathbf{x}_{\text{inst}}, X_{\text{img}}\}$, $X_{\text{img}} = \{\mathbf{x}_{\text{img}}^{(j)}\}_{j=1}^{N_{\text{img}}}$, where \mathbf{x}_{inst} and $\mathbf{x}_{\text{img}}^{(j)} \in \mathbb{R}^{3 \times W \times H}$ denote an instruction and an RGB image with width W and height H , respectively. Here, N_{img} denotes the number of candidate images. The output consists of two ranked lists of images corresponding to the target object and the receptacle, respectively.

The terminology used in this paper is defined as follows: The “target object” and “receptacle” are the everyday object and the desired receptacle (piece of furniture), specified in the instruction. We assume that images of the environment have been collected through pre-exploration. This is a realistic setting because mobile service robots are typically deployed to perform tasks repeatedly in the same known environment (e.g., [18], [19]).

IV. PROPOSED METHOD

We propose Affordance RAG, a zero-shot hierarchical multimodal retrieval framework for OVMM. Fig. 2 shows the overview of the proposed method. To understand what kind of robot affordances are executable in an environment, we introduce Affordance-Aware Embodied Memory (Affordance Mem), constructed from observed images through pre-exploration, as shown in Fig 2 (a). Our method utilizes Affordance Mem as a real-world database for RAG to identify

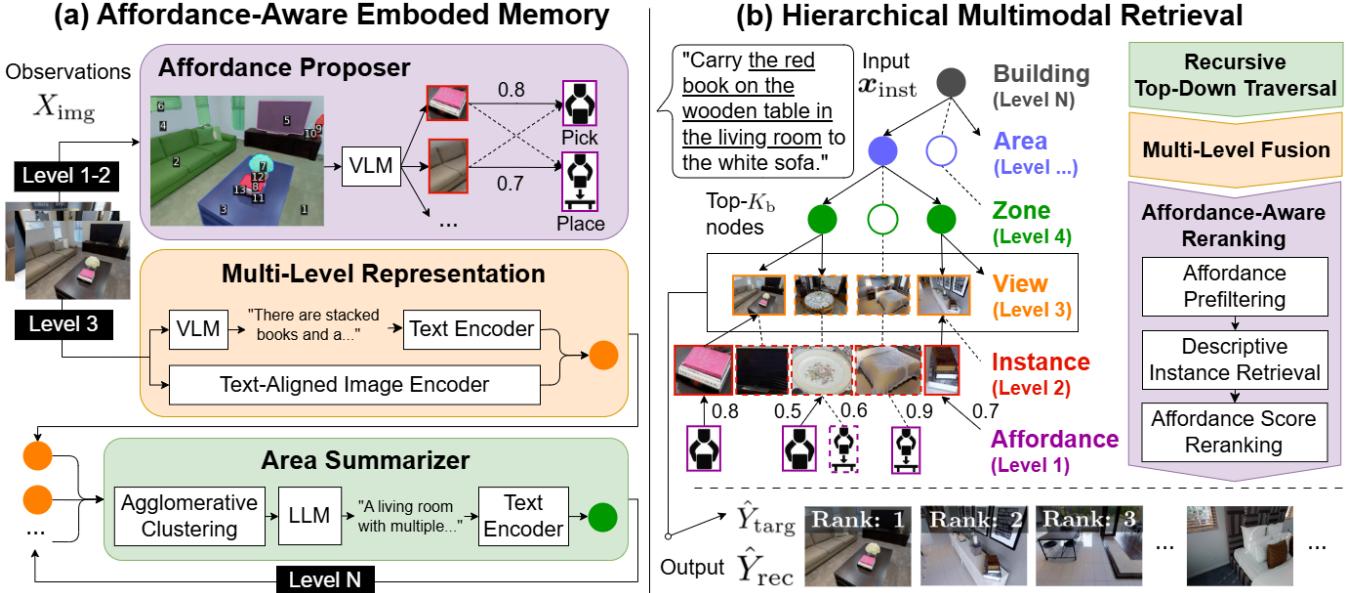


Fig. 2: Overview of the Affordance RAG framework. (a) The robot constructs Affordance-Aware Embodied Memory (Affordance Mem) through pre-exploration. Affordance Mem is constructed from three components: Affordance Proposer, Multi-Level Representation, and Area Summarizer. (b) Upon receiving an instruction, hierarchical multimodal retrieval is performed to identify both the target object and the receptacle. This process consists of three stages: Recursive Top-Down Traversal, Multi-Level Fusion, and Affordance-Aware Reranking.

both the target object and the receptacle, as shown in Fig. 2 (b). While our primary focus is on OVMM, the proposed Affordance Mem is considered to be broadly applicable to other embodied reasoning tasks such as EQA [10], [29].

A. Affordance-Aware Embodied Memory

For robust OVMM, it is important to identify the most suitable option in terms of robot affordance when multiple valid candidates exist. However, existing embodied memories (e.g., [11], [14]) and scene graphs (e.g., [9], [13]) focus solely on what is present in the scene. To address this limitation, we propose a hierarchical embodied memory that spans multiple levels of nodes—from robot affordance, instance, view, zone, and area to building—across levels 1 through N . As illustrated in Fig. 2 (a), Affordance Mem is constructed sequentially in a bottom-up manner by obtaining nodes through the following three components: (1) Affordance Proposal, which predicts instance-level affordances from visual observations; (2) Multi-Level Representation, which obtains view-level regional and visual semantics; and (3) Area Summarization, which aggregates multi-view features to form regional nodes.

1) **Affordance Proposer (Level 1-2):** In this module, we apply object-centric visual prompting to $x_{\text{img}}^{(j)}$ and feed the resulting image into VLMs (e.g., GPT-4o [30]), to extract instance-level representations and robot affordances, which are stored as levels 2 and 1 nodes, respectively. This enables the extraction of descriptive instance-level expressions (e.g., “red metal mug,” “antique wooden side table with vertical slats”) that go beyond category-level nodes (e.g., “cup,” “table”) typically constructed via object detection or semantic segmentation [9], [13]. Specifically, we use SEEM [31] to generate a segmentation mask for $x_{\text{img}}^{(j)}$, and feed both the

original image and a visual prompt image—where segmented regions are overlaid with indexed numbers—into the VLM in parallel. Although previous studies [19], [28] have proposed a similar captioning enhancement approach, our work extends this idea to affordance-aware memory construction. The output is a set of instance-affordance triplets $\mathcal{A}^{(j)} = \{(o_k, a_k, f_k)\}_{k=1}^{N_{\text{af}}}$, where o_k , a_k , f_k , and N_{af} denote the instance representation, the type of robot affordance, the affordance score, and the number of such triplets, respectively. We focus on mobile manipulation tasks and consider “pick” and “place” as the primary robot affordances. However, as the Affordance Proposer is a prompt-based module, it can be easily extended to other atomic actions such as “open” and “close” without any architectural modification.

2) **Multi-Level Representation (Level 3):** In this module, we construct a Multi-Level Representation for $x_{\text{img}}^{(j)}$ by explicitly combining two types of features—high-level regional and low-level visual semantics—which are stored as a level 3 node. Specifically, the former refers to $l_s^{(3,j)} \in \mathbb{R}^{d_t}$, obtained by embedding the image description generated by VLMs using a text encoder (e.g., text-embedding-3-large [32]). The latter, $v^{(3,j)} \in \mathbb{R}^{d_m}$, is obtained using a text-aligned image encoder from a multimodal foundation model (e.g., BEiT-3 [7]). Here, d_t and d_m denote the dimensionality of the respective embeddings. The fusion strategy for combining these features is detailed in Sec. IV-B.2. The generated image descriptions are also used in the Area Summarizer module to perform hierarchical node aggregation.

3) **Area Summarizer (Level N):** In this module, we apply agglomerative clustering to generate level $(i+1)$ nodes by aggregating nodes at level i for $i \geq 3$. Most existing multimodal foundation models [4]–[7] handle each image independently as a feature vector, making it challenging to

capture out-of-view contexts. However, in our target task, instructions often contain referring expressions related to out-of-view objects or room-level semantics (e.g., “... towel near the sink...,” “... in the bedroom...”), which should be considered during retrieval. Therefore, this module performs agglomerative clustering based on both the physical and semantic distances between nodes.

In this module, we repeat the following procedure recursively up to level N : Initially, clustering is performed on the basis of the Euclidean distance between node positions and the cosine similarity between their textual descriptions in the embedding space. Subsequently, for each cluster, we generate a summary of the node descriptions using an LLM, and embed the resulting text into $\mathbf{l}_s^{(i+1,j)} \in \mathbb{R}^{d_t}$ using the text encoder. This hierarchical representation is expected to enable regional and semantically grounded understanding of the environment during retrieval.

B. Hierarchical Multimodal Retrieval

Fig. 2 (b) illustrates the procedure for performing hierarchical multimodal retrieval over Affordance Mem given a natural language instruction. The hierarchical multimodal retrieval process is performed sequentially in a top-down manner, comprising the following three stages: (1) Recursive Top-Down Traversal, which explores the hierarchical memory to progressively narrow down candidate regions; (2) Multi-Level Fusion, which ranks view-level nodes by integrating regional and visual semantics; and (3) Affordance-Aware Reranking, which refines the top-retrieved nodes by evaluating affordance.

1) Recursive Top-Down Traversal: In this module, we recursively traverse the memory from level N down to level 3 by selecting the top- K_b nodes at each level based on the similarity score $s^{(i,j)} = \text{sim}(\mathbf{l}_t, \mathbf{l}_s^{(i,j)})$, where $\text{sim}(\cdot, \cdot)$, $\mathbf{l}_t \in \mathbb{R}^{d_t}$, and $\mathbf{l}_s^{(i,j)} \in \mathbb{R}^{d_t}$ denote cosine similarity, the text embedding of the instruction, and the text embedding of the j -th node at level i , respectively. This hierarchical traversal enables coarse-to-fine filtering based on regional semantics such as areas and zones. Moreover, this module selects appropriate nodes by embedding similarity because previous approaches [11], [14], which utilize LLM for selecting nodes, often struggle with global selection from a large number of candidates, presenting a key limitation. Since node embeddings can be precomputed and stored during pre-exploration, this also enables faster inference than those existing methods. As the output of this module, a set of view-level (level 3) nodes $\mathcal{S} \subseteq \{(\mathbf{v}^{(3,j)}, \mathbf{l}_s^{(3,j)})\}$ is extracted.

2) Multi-Level Fusion : This module performs complementary multimodal retrieval using the high-level regional and low-level visual semantics obtained from the Multi-Level Representation. Specifically, for each element in \mathcal{S} , we compute the similarity score $s_{\text{mlf}}^{(3,j)}$ as follows:

$$s_{\text{mlf}}^{(3,j)} = \alpha \cdot \text{sim}(\mathbf{l}_t, \mathbf{l}_s^{(3,j)}) + (1 - \alpha) \cdot \text{sim}(\mathbf{l}_m, \mathbf{v}^{(3,j)}),$$

where $\alpha \in [0, 1]$ and $\mathbf{l}_m \in \mathbb{R}^{d_m}$ denote a weighting hyperparameter and a feature vector obtained from the text encoder of a multimodal foundation model (e.g., BEiT-3),

respectively. This represents a weighted sum of regional semantics obtained via hierarchical traversal and visual semantics derived from a multimodal foundation model. Prior work has typically utilized either the former (e.g., [11], [14]) or the latter (e.g., [4]–[7]) in isolation. By fusing these regional semantics with visual semantics, Multi-Level Fusion enables complementary multimodal retrieval that considers both global contextual consistency and fine-grained visual similarity.

3) Affordance-Aware Reranking: This module reranks the top- K_r nodes in \mathcal{S} —initially sorted by Multi-Level Fusion—using $\{\mathcal{A}^{(j)}\}$ from levels 1 and 2. The reranking process consists of three steps: Affordance Prefiltering, Descriptive Instance Retrieval, and Affordance Score Reranking (ASR). In the first stage, we prefilter instance nodes based on robot affordances: “pick” for target object image retrieval and “place” for receptacle image retrieval. This allows the method to narrow down candidates based on both object appearance and functional perspective. In the second stage, we provide the descriptive expressions generated by the Affordance Proposer to an LLM, which scores the filtered instance nodes based on their similarity to the instruction. While LLMs are less effective at global retrieval, they excel at matching within a small set of candidates, making this setting well-suited to their strengths. In the third stage, the top- K_f instance nodes are further refined by reranking them based on their affordance scores. Finally, we perform fine-grained reranking over the top- K_r view nodes by prioritizing those that contain the instance nodes selected in this module. This hierarchical multimodal retrieval process is executed separately for the target object and the receptacle, resulting in ranked image lists \hat{Y}_{targ} and \hat{Y}_{rec} , respectively.

V. EXPERIMENTS

A. WholeHouse-MM Benchmark

We introduce the WholeHouse-MM benchmark, constructed from the Matterport3D (MP3D [33]) dataset. In this task, building-scale indoor environments with hundreds of images and human-annotated (not generated) instructions for mobile manipulation are required. Most existing benchmarks addressing open-vocabulary mobile manipulation either use template-based instructions [1] or are not designed for building-scale task execution [19], [20]. Therefore, we collected images from MP3D, a standard dataset widely used for research on navigation and scene understanding in indoor environments, enabling the evaluation of multimodal retrieval at the building scale. While [34], [35] also focus on referring expression comprehension tasks in MP3D environments, they do not handle mobile manipulation instructions. Thus, following [19], [20], the WholeHouse-MM benchmark uses human-annotated instructions containing referring expressions collected via crowdsourcing.

To collect images from each environment, we simulated a pre-exploration phase in MP3D. Since MP3D provided a map of the environment, we captured panoramic views by rotating the camera at each waypoint in 60-degree increments, collecting six images per location. Each environment contained

TABLE I: Quantitative comparison between the proposed method and baseline methods on the WholeHouse-MM benchmark. The best and second-best scores for each metric are indicated in **bold** and underline, respectively. “*” denotes reproduced results.

[%]	Method	Target Object			Receptacle			Overall		
		R@5 ↑	R@10 ↑	R@20 ↑	R@5 ↑	R@10 ↑	R@20 ↑	R@5 ↑	R@10 ↑	R@20 ↑
CLIP [4]	15.7	24.2	33.6	6.2	11.7	21.7	10.9	18.0	27.7	
Long-CLIP [5]	24.6	36.1	48.5	3.5	9.2	19.9	14.0	22.6	34.2	
BLIP-2 [6]	<u>30.2</u>	40.7	48.0	5.2	10.3	19.9	17.7	25.5	34.0	
BEiT-3 [7]	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* [17]	15.1	19.2	30.4	7.3	15.1	23.6	11.2	17.2	27.0	
RelaX-Former [19]	17.9	26.5	37.7	8.8	<u>19.8</u>	<u>28.4</u>	13.3	23.2	33.0	
Embodied-RAG [11]	15.1	18.5	22.8	6.7	11.3	14.4	10.9	14.9	18.6	
Affordance RAG (ours)	32.8	49.9	61.7	14.8	24.3	30.2	23.8	37.1	45.9	

an average of 590 images. The instructions in WholeHouse-MM were collected via crowdsourcing from 116 annotators. The annotators were presented with two images from the environment—one depicting the target object and the other the receptacle—and asked to give instructions for carrying the target object to the receptacle. The target objects and receptacles were obtained by extracting the locations of predefined object categories from REVERIE [34], a standard benchmark for Vision-and-Language Navigation tasks. If a target object or receptacle appeared in multiple viewpoints, those images were also treated as positive.

The benchmark consists of 402 instructions and 2,360 images collected from real-world indoor environments. The vocabulary size is 517, with a total of 6,410 words and an average sentence length of 15.9 words. The environments for each split were selected according to [19], [20]. The validation set was used for hyperparameter tuning, while the test set was used for evaluating the performance of the methods.

B. Quantitative Results

Table I shows the quantitative comparison between the proposed method and baseline methods. Since the mobile manipulation instructions include both a target object and a receptacle, we reported multimodal retrieval performance for each component as well as the overall score. Note that in this benchmark, we focus on retrieval performance and omit physical manipulation; therefore, the ASR step was excluded from the proposed method. For results that include this step, please see Sec. VI. We used recall@ K ($K = 5, 10, 20$) as the evaluation metric. The primary evaluation metric was recall@10. We used recall@ K as it is a standard evaluation metric in image retrieval settings [36]. We used eight baseline methods: CLIP (ViT-L/14) [4], Long-CLIP (ViT-L/14) [5], BLIP-2 (ViT-g) [6], BEiT-3 (large) [7], HomeRobot [1], NLMap [17], RelaX-Former [19], and Embodied-RAG [11]. Except for RelaX-Former, all methods were evaluated in a zero-shot setting.

As shown in Table I, our proposed method achieved recall@10 scores of 49.9%, 24.3%, and 37.1% for the target

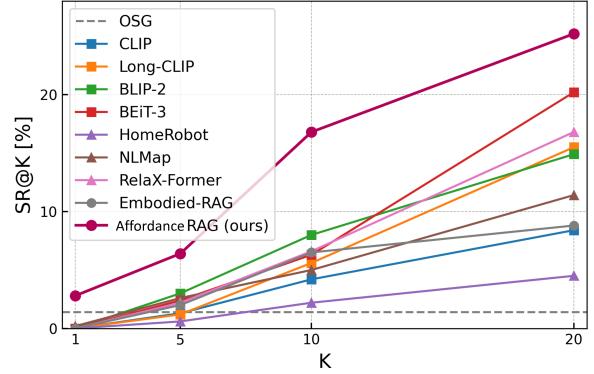


Fig. 3: Comparison of task success rate (SR) on the WholeHouse-MM benchmark. SR@ K denotes the percentage of samples in which both the target object and the receptacle are correctly retrieved within the top- K results.

object, receptacle, and overall metrics, respectively. In contrast, the best-performing baseline method achieved 42.1%, 19.8%, and 28.7%, indicating that our method outperformed it by 7.8, 4.5, and 8.4 points, respectively. Similarly, our method outperformed the best baseline method across all evaluation metrics.

Furthermore, to compare our approach with the Object Goal Navigation method that also utilizes a hierarchical memory (OSG [12]), we present the task success rate (SR) comparison results in Fig. 3. SR@ K denotes the percentage of samples in which both the target object and the receptacle are correctly retrieved within the top- K results. As shown in Fig. 3, OSG obtained an SR of 1.4%, whereas our proposed method achieved 2.8%, 6.4%, 16.8%, and 25.2% for $K = 1, 5, 10, 20$, respectively, the best performance among all methods.

C. Qualitative Results

Fig. 4 shows a successful example from the WholeHouse-MM benchmark. In this example, x_{inst} was “Take a photo frame from the side table in the bedroom and place it on the dining table with a bouquet of flowers.” As shown in Fig. 4 (a), the baseline method ranked an unrelated white table among the top candidates, whereas the proposed method

TABLE II: Results of ablation studies on the WholeHouse-MM benchmark. The best scores for each metric are indicated in **bold**.

[%]	Regional Semantics	Visual Semantics	Affordance-Aware Reranking	Affordance Proposer	Overall		
Method					R@5 ↑	R@10 ↑	R@20 ↑
(a)		✓	✓	✓	22.9	32.1	40.7
(b)	✓		✓	✓	20.3	30.0	39.8
(c)	✓	✓		✓	20.1	29.5	41.5
(d)	✓	✓	✓		20.0	31.2	41.5
(e)	✓	✓	✓	✓	23.8	37.1	45.9

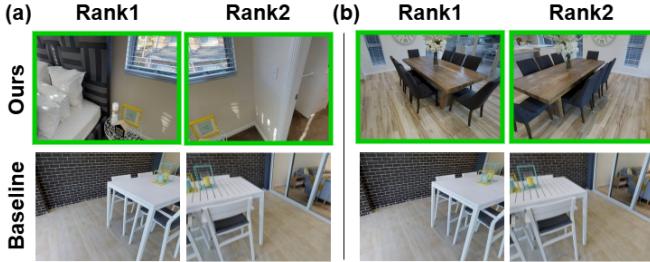


Fig. 4: Qualitative results on the WholeHouse-MM benchmark. The given x_{inst} was “Take a photo frame from the side table in the bedroom and place it on the dining table with a bouquet of flowers.” (a) Target object and (b) receptacle: Top-2 retrieved images are shown for both the our method and the best baseline method (BEiT-3 [7]). The ground-truth image is highlighted with a green border.

successfully ranked the photo frame placed on the side table in the bedroom as the top-1 and top-2 candidates. Similarly, as shown in Fig. 4 (b), the proposed method ranked the dining table with a bouquet of flowers as both the top-1 and top-2 candidates. These results suggest that the proposed hierarchical multimodal retrieval method, which incorporates both regional and visual semantics, is effective.

D. Ablation Studies

To validate our framework, we conducted ablation studies on the construction of Affordance Mem and the hierarchical multimodal retrieval strategy. Table II shows the ablation results on the WholeHouse-MM benchmark.

Multi-Level Fusion ablation: To investigate the effectiveness of the node representation, we conducted ablation studies by removing the features related to regional and visual semantics. Specifically, (a) for the former, we excluded $l_s^{(3,j)}$, which was obtained via hierarchical retrieval; and (b) for the latter, we excluded $v^{(3,j)}$, derived from the text-aligned image encoder. As shown in Table II, Methods (a) and (b) achieved recall@10 scores of 32.1% and 30.0%, respectively, which were 5.0 and 7.1 points lower than Method (e). These results support our design of Multi-Level Fusion, where high-level semantic reasoning (regional semantics) and low-level perceptual grounding (visual semantics) jointly contribute to robust instruction-grounded retrieval.

Affordance-Aware Reranking ablation: To investigate the effectiveness of reranking using instance- and affordance-level nodes in Affordance Mem, we evaluated a variant without the reranking step. As shown in Table II, Method (c) achieved a recall@10 of 29.5%, which was 7.6 points lower than Method (e). This result suggests that reranking based

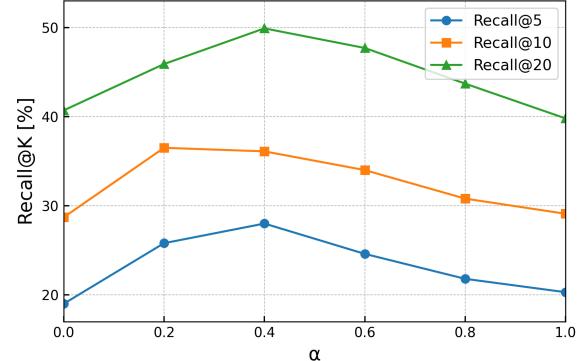


Fig. 5: Sensitivity analysis of the weight α in Multi-Level Fusion. The hyperparameter α balances the contribution between regional semantics and visual semantics.

on instance- and affordance-level nodes in Affordance Mem enables more instruction-consistent multimodal retrieval.

Affordance Proposer ablation: To validate the effectiveness of descriptive affordance proposals generated via visual prompting, we implemented a method that performed LLM-based reranking using only image captions generated by VLMs. Table II shows that Method (d) achieved a recall@10 of 31.2%, which was 5.9 points lower than Method (e). This result suggests that beyond relying solely on VLM-generated captions, incorporating structured information about nodes is beneficial for reranking.

Impact of Multi-Level Fusion weight α : Fig. 5 shows the sensitivity analysis of the weight α in Multi-Level Fusion. We investigated the effect of α on recall@ K by varying it from 0.0 to 1.0 in increments of 0.2. The results show that performance degrades when the retrieval was biased toward either regional semantics or visual semantics. This suggests that the two representations play complementary roles in multimodal retrieval, and appropriately balancing them leads to a more comprehensive understanding.

VI. REAL-WORLD EXPERIMENTS

We validated our method through real-world experiments using a mobile manipulator. In particular, we aimed to evaluate the effectiveness of ASR on the task success rate by executing pick-and-place actions in the real world, where the suitability of object grasping and placement varies across objects.

A. Settings & Implementation

We used a Hello Robot Stretch 2 [37] equipped with a DexWrist for mobile manipulation. This robot is commonly

TABLE III: Quantitative comparison in the real-world experiments. The best scores for each metric are indicated in **bold**. The scores are the results from 40 trials.

Method [%]	R@5 ↑	SR ↑
BEiT-3 [7]	79	45
Affordance RAG (w/o ASR)	94	70
Affordance RAG (full)	94	85

used as a standard platform for OVMM tasks [1]–[3]. The environment used in our experiments was a $5.0 \times 7.0 \text{ m}^2$ indoor room consisting of office and kitchen areas, containing 10 different pieces of furniture. We used the 20 everyday objects as target objects. Among them, 14 objects were randomly selected from the YCB objects [38], which are widely used in manipulation research, and the remaining 6 were composed of commonly used household objects. In our experiments, we assumed that these objects were initially placed on randomly selected pieces of furniture.

Firstly, to construct Affordance Mem, the robot first performed a pre-exploration phase. During this phase, the robot captured RGB images using an Intel RealSense D435i camera from a pre-defined viewpoint that allowed observation of the entire environment. Path planning and navigation followed standard map-based approaches. The robot constructed a 2D map using Hector SLAM [39]. Next, the user provided a free-form instruction. The user were asked to provide an instruction that required transporting a randomly selected object in the environment to a randomly selected piece of furniture. We conducted 40 trials in total, using a different instruction for each trial.

After receiving the instruction, the robot performed the following steps. First, the robot retrieved images of the target object and receptacle from Affordance Mem and presented the top-5 retrieved images for each to the user. In real-world experiments, when performing affordance prediction based on a VLM, we provided a prompt that included information about the robot’s embodiment (e.g., gripper shape and width, arm length). The inference took 0.12 seconds per instruction and used 16 GB of VRAM on an NVIDIA Geforce RTX 3090. Next, the robot navigated to the location where the user-selected target object image had been captured, and performed the grasping action. The grasping point was determined as the median of the point cloud corresponding to the segmented mask of the target object, obtained from the depth image and the segmentation produced by SAM [40]. Similarly, the robot transported the target object to the location where the user-selected receptacle image had been captured and placed the target object. Since motion generation for grasping, placement, and navigation is out of the scope of this study, we adopted heuristic-based methods.

B. Quantitative Results

Table III shows the quantitative results of the real-world experiments. We used recall@5 and SR as evaluation metrics in our experiments. A trial was considered successful only if the method retrieved a correct image for both the target

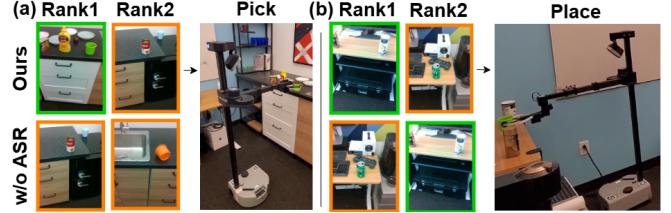


Fig. 6: Qualitative results of the real-world experiments. The given x_{inst} was “Please deliver a cup to the desk that has some coffee powder on it.” (a) Target object and (b) receptacle: Top-2 retrieved images and real-world execution results are shown for both the proposed method and the variant without Affordance Score Reranking (ASR). The ground-truth image and semantically correct but less suitable image are highlighted with green and orange borders, respectively.

object and the receptacle within the top-5 results, and the robot executed the pick and place actions without failure. Recall@5 indicates the overall score for both the target object and the receptacle. We used BEiT-3 [7] as a baseline method because it was the best-performing baseline in the simulated experiments, as shown in Table I.

As shown in Table III, the proposed method achieved a recall@5 of 94%, outperforming the baseline method by 15 points. Similarly, the proposed method achieved an SR of 85%, which is 40 points higher than the baseline method. These results demonstrate that the proposed method can be successfully integrated into a real-world robot to enable open-vocabulary mobile manipulation. The recall@5 scores in Table I are generally lower than those in Table III due to the larger building-scale search space of MP3D.

To validate the effectiveness of ASR in the proposed method, we conducted an ablation experiment by removing this step. As shown in Table III, the proposed method achieved an SR of 85%, compared to 70% without ASR, demonstrating a 15-point improvement. These results suggest that when instructions are ambiguous and allow for multiple valid interpretations, ASR helps prioritize candidates with higher pick-and-place suitability, thereby improving SR. The identical recall@5 scores for both methods are due to reranking being applied only within the top-5 retrieved candidates.

C. Qualitative Results

Fig. 6 shows a successful example from the real-world experiments. In this example, x_{inst} was “Please deliver a cup to the desk that has some coffee powder on it.” As shown in the results, our proposed method ranked a green cup placed upright on the kitchen counter as the top candidate due to its high suitability, while the variant without ASR ranked less suitable options higher, such as a blue cup placed deep on top of the refrigerator and an orange cup lying on its side. Similarly, the receptacle was a desk with coffee powder on it; however, two desks were present in the environment, separated by a coffee machine. According to the results, the proposed method ranked the tidier desk with fewer objects as the top candidate, while the variant without ASR ranked the more cluttered desk higher. Based on the retrieval results, the robot was able to grasp the green cup and successfully transport it to the desk with coffee powder. This suggests

that ASR improved SR by promoting the selection of more suitable objects in response to ambiguous instructions.

VII. CONCLUSIONS

In this study, we address the task of MRMM, where a robot performs mobile manipulation based on language instructions by retrieving target object and receptacle images from environmental images. We proposed Affordance RAG, a zero-shot hierarchical multimodal retrieval framework that combines regional and visual semantics via Multi-Level Fusion based on Affordance Mem. Affordance RAG outperformed the baseline methods in terms of standard metrics on the newly built WholeHouse-MM benchmark. Furthermore, in real-world experiments, the proposed method achieved a task success rate of 85%, outperforming existing methods in both retrieval performance and overall task success. One limitation of the proposed method is that all relevant visual information is assumed to be covered by the observed images, which may not hold in highly occluded scenes. In addition, our current framework focuses on visual and spatial affordance reasoning and does not explicitly consider physical or kinematic constraints of robot actions. As future work, we plan to address both limitations by combining our approach with active exploration and physical reasoning.

REFERENCES

- [1] S. Yenamandra, A. Ramachandran, K. Yadav, A. Wang, M. Khanna, T. Gervet, T. Yang, V. Jain, A. Clegg, *et al.*, “HomeRobot: Open-Vocabulary Mobile Manipulation,” in *CoRL*, 2023, pp. 1975–1911.
- [2] P. Liu, Y. Orru, C. Paxton, N. M. M. Shafullah, and L. Pinto, “OK-Robot: What Really Matters in Integrating Open-Knowledge Models for Robotics,” *arXiv preprint arXiv:2401.12202*, 2024.
- [3] P. Liu, Z. Guo, M. Warke, S. Chintala, C. Paxton, *et al.*, “DynaMem: Online Dynamic Spatio-Semantic Memory for Open World Mobile Manipulation,” *arXiv preprint arXiv:2411.04999*, 2024.
- [4] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, A. Askell, P. Mishkin, *et al.*, “Learning Transferable Visual Models From Natural Language Supervision,” in *ICML*, 2021, pp. 8748–8763.
- [5] B. Zhang, P. Zhang, X. Dong, Y. Zang, and J. Wang, “Long-CLIP: Unlocking the Long-Text Capability of CLIP,” in *ECCV*, 2024, pp. 310–325.
- [6] J. Li, D. Li, S. Savarese, and S. Hoi, “BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models,” in *ICML*, 2023, pp. 19730–19742.
- [7] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, *et al.*, “Image as a Foreign Language: BEIT Pretraining for Vision and Vision-Language Tasks,” in *CVPR*, 2023, pp. 19175–19186.
- [8] R. Korekata, M. Kambara, Y. Yoshida, S. Ishikawa, *et al.*, “Switching Head-Tail Funnel UNITER for Dual Referring Expression Comprehension with Fetch-and-Carry Tasks,” in *IROS*, 2023, pp. 3865–3872.
- [9] D. Honerkamp, M. Büchner, F. Despinoy, *et al.*, “Language-Grounded Dynamic Scene Graphs for Interactive Object Search with Mobile Manipulation,” *IEEE RA-L*, vol. 9, no. 10, pp. 2377–3766, 2024.
- [10] Y. Yang, H. Yang, J. Zhou, P. Chen, H. Zhang, Y. Du, and C. Gan, “3D-Mem: 3D Scene Memory for Embodied Exploration and Reasoning,” in *CVPR*, 2025, pp. 17294–17303.
- [11] Q. Xie, S. Min, P. Ji, Y. Yang, T. Zhang, K. Xu, A. Bajaj, *et al.*, “Embodied-RAG: General Non-parametric Embodied Memory for Retrieval and Generation,” *arXiv preprint arXiv:2409.18313*, 2024.
- [12] J. Loo, Z. Wu, and D. Hsu, “Open Scene Graphs for Open World Object-Goal Navigation,” *IJRR*, 2025.
- [13] A. Werby, C. Huang, M. Büchner, A. Valada, and W. Burgard, “Hierarchical Open-Vocabulary 3D Scene Graphs for Language-Grounded Robot Navigation,” in *RSS*, 2024.
- [14] Z. Wang, Y. Zhu, G. Lee, and Y. Fan, “NavRAG: Generating User Demand Instructions for Embodied Navigation through Retrieval-Augmented LLM,” *arXiv preprint arXiv:2502.11142*, 2025.
- [15] Q. Gu, A. Kuwajerwala, S. Morin, K. Jatavallabhula, B. Sen, A. Agarwal, C. Rivera, *et al.*, “ConceptGraphs: Open-Vocabulary 3D Scene Graphs for Perception and Planning,” in *ICRA*, 2024, pp. 5021–5028.
- [16] N. Hughes, Y. Chang, and L. Carlone, “Hydra: A Real-time Spatial Perception System for 3D Scene Graph Construction and Optimization,” in *RSS*, 2022.
- [17] B. Chen, F. Xia, B. Ichter, K. Rao, K. Gopalakrishnan, M. Ryoo, *et al.*, “Open-vocabulary Queryable Scene Representations for Real World Planning,” in *ICRA*, 2023, pp. 11509–11522.
- [18] G. Sigurdsson, J. Thomason, G. Sukhatme, and R. Piramuthu, “RREx-BoT: Remote Referring Expressions with a Bag of Tricks,” in *IROS*, 2023, pp. 5203–5210.
- [19] 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.
- [20] R. Korekata, *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.
- [21] A. Delitzas, A. Takmaz, F. Tombari, R. Sumner, *et al.*, “SceneFun3D: Fine-Grained Functionality and Affordance Understanding in 3D Scenes,” in *CVPR*, 2024, pp. 14531–14542.
- [22] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak, “Affordances from Human Videos as a Versatile Representation for Robotics,” in *CVPR*, 2023, pp. 13778–13790.
- [23] H. Jiang, B. Huang, R. Wu, Z. Li, S. Garg, H. Nayyeri, S. Wang, and Y. Li, “RoboEXP: Action-Conditioned Scene Graph via Interactive Exploration for Robotic Manipulation,” in *CoRL*, 2024.
- [24] C. Song, V. Blukis, J. Tremblay, S. Tyree, Y. Su, and S. Birchfield, “RoboSpatial: Teaching Spatial Understanding to 2D and 3D Vision-Language Models for Robotics,” in *CVPR*, 2025, pp. 15768–15780.
- [25] Y. Hu, Q. Xie, V. Jain, J. Francis, J. Patrikar, N. Keetha, S. Kim, Y. Xie, *et al.*, “Toward General-Purpose Robots via Foundation Models: A Survey and Meta-Analysis,” *arXiv preprint arXiv:2312.08782*, 2023.
- [26] D. Driess, F. Xia, M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, *et al.*, “PaLM-E: An Embodied Multimodal Language Model,” in *ICML*, 2023, pp. 8469–8488.
- [27] A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, *et al.*, “Do As I Can, Not As I Say: Grounding Language in Robotic Affordances,” in *CoRL*, 2023, pp. 287–318.
- [28] J. Yang, H. Zhang, F. Li, X. Zou, C. Li, and J. Gao, “Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V,” *arXiv preprint arXiv:2310.11441*, 2023.
- [29] A. Majumdar, A. Ajay, X. Zhang, P. Putta, S. Yenamandra, M. Henaff, S. Silwal, P. Mcvay, *et al.*, “OpenEQA: Embodied Question Answering in the Era of Foundation Models,” in *CVPR*, 2024, pp. 16488–16498.
- [30] OpenAI, “GPT-4o: Optimized Generative Pre-trained Transformer 4,” <https://openai.com>, 2024, accessed: Jun. 2025.
- [31] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Wang, L. Wang, J. Gao, and Y. Lee, “Segment Everything Everywhere All at Once,” in *NeurIPS*, 2023, pp. 19769–19782.
- [32] OpenAI, “text-embedding-3-large,” <https://platform.openai.com/docs/models/embeddings>, 2024, accessed: Jun. 2025.
- [33] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Nießner, M. Savva, S. Song, A. Zeng, and Y. Zhang, “Matterport3D: Learning from RGB-D Data in Indoor Environments,” in *3DV*, 2017, pp. 667–676.
- [34] Y. Qi, Q. Wu, P. Anderson, X. Wang, W. Wang, C. Shen, and A. Hengel, “REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments,” in *CVPR*, 2020, pp. 9979–9988.
- [35] F. Zhu, X. Liang, Y. Zhu, Q. Yu, X. Chang, and X. Liang, “SOON: Scenario Oriented Object Navigation with Graph-based Exploration,” in *CVPR*, 2021, pp. 12689–12699.
- [36] T. Liu, “Learning to Rank for Information Retrieval,” *FNTIR*, vol. 3, no. 3, pp. 225–331, 2009.
- [37] C. Kemp, A. Edsinger, H. Clever, and B. Matulevich, “The Design of Stretch: A Compact, Lightweight Mobile Manipulator for Indoor Human Environments,” in *ICRA*, 2022, pp. 3150–3157.
- [38] B. Calli, A. Walsman, A. Singh, S. Srinivasa, *et al.*, “Benchmarking in Manipulation Research: Using the Yale-CMU-Berkeley Object and Model Set,” *IEEE RAM*, vol. 22, no. 3, pp. 36–52, 2015.
- [39] S. Kohlbrecher, O. Von Stryk, J. Meyer, and U. Klingauf, “A Flexible and Scalable SLAM System with Full 3D Motion Estimation,” in *SSRR*, 2011, pp. 155–160.
- [40] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. Berg, W. Lo, P. Dollar, and R. Girshick, “Segment Anything,” in *ICCV*, 2023, pp. 4015–4026.