Affordance-Guided Coarse-to-Fine Exploration for Base Placement in Open-Vocabulary Mobile Manipulation

Tzu-Jung Lin[1], Jia-Fong Yeh[1], Hung-Ting Su[1], Chung-Yi Lin[1], Yi-Ting Chen[2], Winston H. Hsu[1]

[1]National Taiwan University
[2]National Yang Ming Chiao Tung University

Abstract

In open-vocabulary mobile manipulation (OVMM), task success often hinges on the selection of an appropriate base placement for the robot. Existing approaches typically navigate to proximity-based regions without considering affordances, resulting in frequent manipulation failures. We propose Affordance-Guided Coarse-to-Fine Exploration, a zero-shot framework for base placement that integrates semantic understanding from vision-language models (VLMs) with geometric feasibility through an iterative optimization process. Our method constructs cross-modal representations, namely *Affordance RGB* and *Obstacle Map+*, to align semantics with spatial context. This enables reasoning that extends beyond the egocentric limitations of RGB perception. To ensure interaction is guided by task-relevant affordances, we leverage coarse semantic priors from VLMs to guide the search toward task-relevant regions and refine placements with geometric constraints, thereby reducing the risk of convergence to local optima. Evaluated on five diverse open-vocabulary mobile manipulation tasks, our system achieves an 85% success rate, significantly outperforming classical geometric planners and VLM-based methods. This demonstrates the promise of affordance-aware and multimodal reasoning for generalizable, instruction-conditioned planning in OVMM.

1 Introduction

In open-vocabulary mobile manipulation (OVMM), selecting an appropriate base placement is critical for successful task execution. However, prior navigation systems (Yenamandra et al. 2023; Huang et al. 2023a; Qiu et al. 2024; Tan et al. 2025) often treat the task as complete once the robot reaches a location near the target. In practice, mere proximity does not guarantee effective manipulation. Specifically, determining the base placement presents two key challenges. **First**, the robot must *reason jointly about geometric feasibility and semantic intent*. It needs to identify a collision-free location that maintains appropriate distance and aligns with task-relevant affordances. For example, to open a cabinet, the robot must position itself in front of the cabinet with enough clearance for effective interaction. **Second**, the robot must *reason globally despite limited perceptual input*. A narrow field of view restricts spatial awareness and may cause the robot to overlook suitable placements. For instance, if the area in front of the cabinet is not visible in the RGB image, the robot cannot evaluate it as a potential base location.

Existing approaches struggle to address the aforementioned challenges. In many OVMM (Huang et al. 2023a; Qiu et al. 2024) or navigation (Singh et al. 2023; Huang et al. 2023b) works, once the target object's location is known, classical path planners such as A* (Hart, Nilsson, and Raphael 1968) and RRT* (Karaman and Frazzoli 2011) are used to navigate the robot close to the object. These planners rely solely on geometric heuristics and lack task-level semantic understanding. As a result, they often produce placements that fail to account for task affordances, as illustrated in Figure 1. Conversely, semantic methods that leverage vision-language models (VLMs) can infer high-level task intent (Nasiriany et al. 2024; Sathyamoorthy et al. 2024), but typically overlook geometric feasibility and reachability constraints. Furthermore, these methods often rely on a single RGB image, which limits their ability to reason about occluded or unseen areas.

In this paper, we propose a novel zero-shot framework, *Affordance-Guided Coarse-to-Fine Exploration*, which integrates both semantic reasoning and geometric understanding for effective base placement, especially under limited field of view. Our approach introduces two key innovations: (1) **Cross-modal representations**, specifically *Affordance RGB* and *Obstacle Map+*, which align task intent with spatial layout while mitigating perceptual limitations caused by occluded or unseen regions. (2) **A coarse-to-fine optimization process** that begins with sampling in semantically guided regions and iteratively refines the search toward geometrically feasible placements. This enables the robot to satisfy both semantic and geometric requirements.

Our system requires no task-specific supervision and operates solely on natural language instructions, RGB-D images, and an obstacle map. It generalizes across a wide range of OVMM tasks with varying spatial and semantic demands. Experiments show that our method achieves an overall success rate of 85%, significantly outperforming classical and semantic-only baselines. These results underscore the effectiveness of our key component, i.e. affordance-aware and multimodal reasoning, in enabling instruction-driven base placement for open-vocabulary mobile manipulation.

Our main contributions are:

Figure 1: Examples of failure cases caused by base placements without affordance awareness. Left: The robot cannot open the cabinet because it is not facing the drawer. Middle: The robot cannot grasp the pot handle due to misalignment. Right: The robot fails to place the mug on the shelf as it is not facing the open side. These failures arise from a lack of joint reasoning over task intent and geometric feasibility, leading to semantically misaligned placements that prevent successful manipulation.

- We identify base placement as a critical challenge in OVMM, driven by the need to reason about both semantics and geometry under limited perceptual input.
- We propose a coarse-to-fine strategy that unifies semantic and geometric cues via *Affordance RGB* and *Obstacle Map+*, addressing limited field of view and overcoming the limitations of prior single-focus methods.
- We develop a generalizable zero-shot system that achieves 85% success across tasks, outperforming prior classical and semantic-only approaches.

2 Related Work

Open-Vocabulary Navigation and Manipulation Recent vision-language models (VLMs) (OpenAI 2023; Team et al. 2023; Liu et al. 2023) have shown strong potential in open-vocabulary navigation and manipulation. Some systems build semantic maps using VLM features to enable language-guided planning. For example, FindAnything (Laina et al. 2025) constructs object-centric submaps, while CLIP-Fields (Singh et al. 2023) and VLMaps (Huang et al. 2023b) embed vision-language features into 3D environments. CLIP-Fields(Singh et al. 2023) also projects weakly-supervised features to generate semantic maps. Methods like USA-Nets (Bolte et al. 2023) and GOAT (Chang et al. 2023) locate targets using language-image matching, then guide navigation using geometric planners like A*. These techniques extend to open-vocabulary mobile manipulation (OVMM). OK-Robot (Huang et al. 2023a) builds a VoxelMap from RGB-D scans and uses CLIP (Radford et al. 2021), SAM (Kirillov et al. 2023), and OWL-ViT (Minderer et al. 2022) for grounding and segmentation. It then queries VLMs to plan collision-free paths. COME-robot (Zhi et al. 2025) incorporates GPT-4V for task planning and failure recovery. Our work differs by focusing specifically on selecting semantically meaningful and physically feasible base placements.

Visual Prompting for Robotics Visual prompting is an emerging tool in robotic VLMs. While Set-of-Mark Prompting (SoM) (Yang et al. 2023) was originally developed for visual grounding tasks, its underlying techniques have been extended to symbolic prompting in systems like PIVOT (Nasiriany et al. 2024), which frames spatial tasks as iterative visual question answering and refines VLM predictions through repeated annotation and selection. CoPa (Huang et al. 2024a), ReKep (Huang et al. 2024b), and MOKA (Zhang et al. 2025) use visual prompts to infer keypoints or constraints for manipulation. KAGI (Lee et al. 2025) utilizes keypoints to define dense rewards in reinforcement learning, while CoNVOI (Sathyamoorthy et al. 2024) applies region-level prompting for navigation. However, these methods rely on single-view RGB inputs, limiting their ability to reason about occluded or unseen regions. In contrast, our approach applies visual prompting directly to obstacle maps, enabling global inference beyond the current view.

Recent Advances in Base Placement Recent work addresses base placement by integrating perception and planning. MoMa-Pos (Shao et al. 2024) optimizes base placement by modeling task-relevant articulated objects using frontal-view images, but requires object-specific modeling and lacks generalization to novel categories. MoMa-Kitchen(Zhang et al. 2025) introduces a large egocentric dataset for affordance prediction, but its performance may be constrained by the limited field of view inherent in egocentric perspectives. Navi2Gaze (Zhu et al. 2024) uses VLMs to choose base locations aligned with object orientation but does not reason about reachability or execution feasibility. Our approach applies affordance prompting directly on obstacle maps, enabling generalization to unseen scenes and occluded configurations.

3 Preliminaries

Pipeline Overview

We consider an open-vocabulary mobile manipulation setting in which a robot is given a high-level natural language instruction ℓ. The instruction is parsed by GPT-4 (OpenAI 2023), a large language model, into a sequence of sub-instructions. Each sub-instruction is represented as a tuple

$(t, \tilde{\ell})$, where t denotes the name of the referenced object and $\tilde{\ell}$ is the corresponding sub-task instruction. For example, the instruction "Put the mug on the shelf" is parsed as: [("mug", "pick up the mug."), ("shelf", "put the mug on the shelf.")].

To focus on base placement, we assume that the 2D position of the target object t, denoted $\mathbf{p}_t \in \mathbb{R}^2$, is directly provided by the simulator. This allows the system to operate given a known object location and a global obstacle map M_{global}, without addressing open-world object grounding.

Given a grounded sub-instruction $(t, \tilde{\ell})$ and access to $\mathbf{p}_t$ and M_{global}, the robot executes the following three-stage pipeline:

1. **Navigation:** Navigate to a coarse waypoint near the object's known position (typically within 1.5 meters) using a path planner and orient the robot to face the object.
2. **Base Placement Selection:** Select an optimized base placement for manipulation by reasoning over local geometric and semantic cues, and move to the selected placement.
3. **Manipulation:** Execute the sub-task $\tilde{\ell}$ using a predefined manipulation primitive such as pick, place, or open.

To support this pipeline, the robot maintains a global 2D occupancy grid map M_{global} and dynamically derives a local egocentric map M_{local} at runtime.

Problem Statement

After coarse navigation, the robot has access to:

- A target object name t and sub-instruction $\tilde{\ell}$ parsed from the high-level command ℓ,
- An obstacle map M_{local} indicating non-navigable regions,
- An RGB I and a depth image D from onboard sensors.

The goal is to select a base placement $\mathbf{x} \in \mathcal{X}_{\text{free}}$, where $\mathcal{X}_{\text{free}}$ is the set of collision-free placements with sufficient clearance from obstacles:

$$\mathcal{X}_{\text{free}} = \{\mathbf{x} \mid \mathbf{x} \text{ is collision-free and } \text{dist}(\mathbf{x}, \mathcal{O}) \geq 0.4\,\text{m}\}$$

with $\mathcal{O} \subseteq M_{\text{local}}$ denoting obstacle regions. We then solve:

$$\mathbf{x}^* = \arg\max_{\mathbf{x} \in \mathcal{X}_{\text{free}}} \mathbb{P}\left[\text{success} \mid \mathbf{x}, I, D, \tilde{\ell}, t, M_{\text{local}}\right]$$

A trial is successful if the robot can reach a valid end-effector pose $g^* \in \mathcal{G}_t$, defined over the affordance region of object t, and execute the manipulation. This includes:

- Solving IK to find joint configuration θ for g^*,
- Moving the arm to θ without collisions,
- Executing the manipulation primitive (e.g., pick, open),
- Verifying the expected physical outcome.

4 Method

We present **Affordance-Guided Coarse-to-Fine Exploration**, a framework for selecting base placements that are both semantically meaningful and geometrically feasible. The key idea is to leverage large vision models (LVMs) and vision-language models (VLMs) for high-level semantic guidance, while using iterative optimization to achieve spatial precision. Our approach consists of two stages: (1) **Affordance Guidance Projection**, which extracts affordance cues from perception and projects them onto a 2D obstacle map; and (2) **Affordance-Driven Coarse-to-Fine Optimization**, which refines candidate base placements through probabilistic sampling and VLM feedback. An overview of the method is depicted in Figure 2.

Affordance Guidance Projection

To semantically guide robot base planning beyond RGB perception, we introduce a cross-modal projection mechanism that aligns visual-semantic information with spatial geometric maps. While obstacle maps provide valuable collision-aware context, they inherently lack semantic richness. Conversely, VLMs can perform language-grounded affordance reasoning but are limited to RGB inputs and lack explicit spatial awareness. To bridge this modality gap, we extract affordance cues from RGB images using Grounded SAM (Ren et al. 2024) for object segmentation and GPT-4o (OpenAI 2023) for language-conditioned reasoning. These cues are then projected onto the robot's 2D obstacle map, enabling consistent semantic alignment between visual and spatial representations. This cross-modal grounding facilitates more informed base placement decisions. To operationalize this idea, we construct two complementary multimodal representations:

- **Affordance RGB (I_{aff}):** An RGB image overlaid with:
  1. 12 directional arrows with distinct colors, evenly spaced at 30° intervals around the object.
  2. One arrow labeled "A" to indicate the coarse affordance direction suggested by the VLM.
- **Obstacle Map+ (M^{+}_{local}):** A top-down spatial map augmented with:
  1. The segmented target object footprint $\mathcal{R}_t$.
  2. The robot's current base location.
  3. A fan-shaped affordance region $\mathcal{F}_t$ centered on the selected direction (labeled "A"), spanning $\pm 60°$ from that direction.
  4. All 12 directional arrows are rendered with colors consistently matched to those in the RGB image, ensuring cross-modal semantic alignment.

Affordance-Driven Coarse-to-Fine Optimization

To complete a manipulation task such as opening the dishwasher, the robot must select a base placement from which the relevant part of the object is physically reachable to enable successful execution. Relying solely on geometric reasoning may result in infeasible placements, while depending only on vision-language models (VLMs) may produce semantically appropriate but unreachable ones.

To address this, we propose an optimization method that begins from a task-specific affordance keypoint $\mathbf{g}$ (e.g., a handle) and searches the surrounding region for functionally viable base placements. The search is guided by two criteria: (1) the placement must align with task semantics, and (2) it must lie within the robot's reachable workspace.

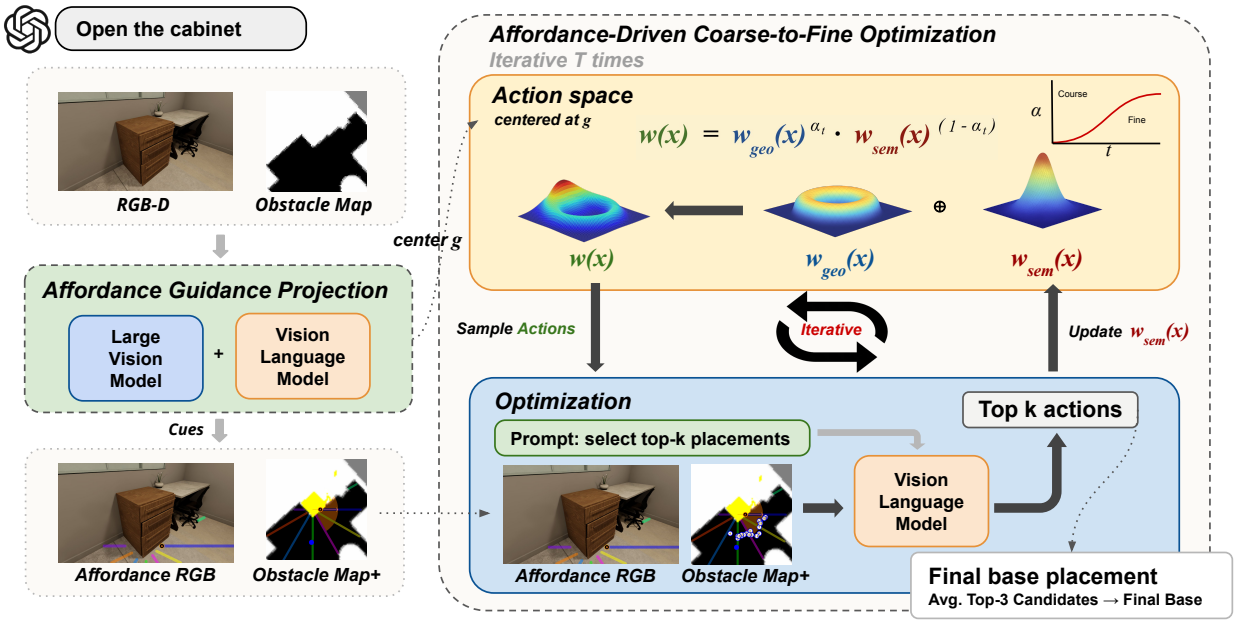


Figure 2: Affordance-Guided Coarse-to-Fine Exploration. The method comprises two key components. (1) To overcome the limitations of single-view perception, it applies *Affordance Guidance Projection*, which uses semantic cues to generate Affordance RGB and Obstacle Map+ from RGB and obstacle maps, enabling global semantic reasoning. (2) To identify base placements that satisfy both semantic relevance and geometric feasibility, it introduces *Affordance-Driven Coarse-to-Fine Optimization*, which leverages the coarse, high-level nature of VLM outputs to explore semantically appropriate regions. As the process iterates, geometric constraints are gradually emphasized, guiding the search toward executable base placements.

Our approach adopts a coarse-to-fine strategy that is realized through an iterative scoring mechanism. Specifically, candidate placements are sampled from a truncated Gaussian distribution centered at $\mathbf{g}$, and each candidate is assigned a score based on a combination of semantic relevance and spatial proximity. As the optimization progresses, the weighting of the scoring function is gradually shifted from semantic alignment to geometric precision. This enables the robot to explore task-relevant placements early on and converge toward physically executable ones in later iterations.

We describe the selection of the affordance keypoint $\mathbf{g}$ using a method similar to (Huang et al. 2024b), followed by details of the sampling and optimization procedure.

Affordance Point Selection To identify the affordance keypoint $\mathbf{g}$, we first extract visual representations from the entire RGB image using DINOv2 (Oquab et al. 2024). Then, Grounded SAM (Ren et al. 2024) is used to segment the target object based on the task instruction. From the extracted DINOv2 features within the segmented region, we apply k-means clustering (with cosine similarity) to produce spatially diverse candidate keypoints. Cluster centroids are projected onto the image, rendered, and annotated. Finally, given the task sub-instruction and annotated image, GPT-4o (OpenAI 2023) selects the keypoint most semantically aligned with the intended interaction. The selected keypoint $\mathbf{g}$ serves as the sampling center for generating candidate base placements.

Iterative Optimization We perform the base selection procedure over T iterative steps. Each iteration consists of three main stages: (1) scoring, (2) sampling, and (3) refinement.

(1) Scoring. At each iteration t, we sample a set of candidate base placements $\{x_i\}_{i=1}^{N}$ from a Gaussian distribution centered at the predicted affordance point $\mathbf{g}$:

$$x_i \sim \mathcal{N}(\mathbf{g}, \sigma^2_{\text{sample}} I), \quad \text{s.t.} \quad \|x_i - \mathbf{g}\| \leq r_{\max}, \quad x_i \in \mathcal{X}_{\text{free}}.$$

Sampling is truncated at a fixed radius $r_{\max}$ and restricted to collision-free regions $\mathcal{X}_{\text{free}}$.

Each candidate is assigned a composite score $w(x)$ that balances geometric and semantic relevance:

$$w(x) = w_{\text{geo}}(x)^{\alpha_t} \cdot w_{\text{sem}}(x)^{1-\alpha_t}, \tag{1}$$

where $\alpha_t \in [0, 1]$ is a time-dependent weighting coefficient that shifts gradually from semantic alignment toward geometric precision.

The *geometric term* encourages sampling at a preferred distance r^* from $\mathbf{g}$:

$$w_{\text{geo}}(x) = \Phi(\|x - \mathbf{g}\|; r^*, \sigma_g),$$

where $\Phi(d; \mu, \sigma)$ measures the cumulative probability mass within a margin δ:

$$\Phi(d; \mu, \sigma) = \text{CDF}_{\mu,\sigma}(d + \delta) - \text{CDF}_{\mu,\sigma}(d - \delta).$$

Here, $\text{CDF}_{\mu,\sigma}(x)$ denotes the cumulative distribution function of a Gaussian distribution with mean μ and standard deviation σ.

Method	Throw the Can into Trash	Move Pot Near Red Mug	Put Mug on Shelf	Open Cabinet	Open Dishwasher	Total Success
Object Center + A*	**20/20**	9/20	8/20	5/20	5/20	47%
Object Center + RRT*	19/20	8/20	3/20	10/20	10/20	50%
Affordance Point + A*	16/20	10/20	13/20	10/20	9/20	58%
Affordance Point + RRT*	18/20	10/20	10/20	11/20	12/20	61%
Pivot (I)	0/20	2/20	1/20	**17/20**	6/20	26%
Pivot (M^+_{local}, I_{aff})	2/20	3/20	2/20	10/20	6/20	23%
Our method	17/20	**18/20**	**17/20**	16/20	**17/20**	**85%**

Table 1: Success Rates Across Five Mobile Manipulation Tasks

The *semantic term* encourages alignment with an evolving semantic center μ_t:

$$w_{\text{sem}}(x) = \begin{cases} \Phi(\|x - \mu_t\|; 0, \sigma_s), & \text{if } \mu_t \text{ is defined,} \\ 1, & \text{otherwise.} \end{cases}$$

To implement a smooth transition from semantic exploration to geometric refinement, we use a sigmoid schedule:

$$\alpha_t = \alpha_{\max} \left(1 + e^{-\gamma(t - T/2)}\right)^{-1}, \tag{2}$$

where $\alpha_{\max}$ is the maximum geometric weight, γ controls the steepness, and $T/2$ is the inflection point. This design prioritizes VLM-based semantic reasoning in early stages—when coarse, high-level decisions are most useful—and gradually shifts toward precise spatial optimization in later iterations.

(2) Sampling. The normalized weights define a discrete probability distribution:

$$p(x_i) = \frac{w(x_i)}{\sum_{j=1}^{N} w(x_j)}. \tag{3}$$

We sample N_{sample} candidates from this distribution. Each sampled location is projected onto the local obstacle map M^+_{local} and assigned a unique index for identification. The indexed map, along with the Affordance RGB image I_{aff} and the sub-instruction $\tilde{\ell}$, is submitted to the VLM as a joint multimodal prompt for semantic ranking.

(3) Refinement. For iterations $t < T$, the VLM ranks the sampled candidates and returns the top-k semantically relevant points, denoted as $\{x^{(1)}, \ldots, x^{(k)}\}$. The semantic center is then updated as $\mu_t = \frac{1}{k}\sum_{i=1}^{k} x^{(i)}$. To encourage convergence, we progressively reduce σ_s across iterations.

At the final iteration ($t = T$), we still perform VLM-based ranking, but omit the semantic center update. Instead, we directly take the top-5 VLM-ranked candidates, remove the two furthest from their mean, and compute the final base placement by averaging the remaining three.

5 Experiments

We evaluate our approach in a simulated mobile manipulation environment, focusing on base placement for diverse open-vocabulary tasks. This section details the experimental setup, task definitions, and comparative performance analysis, including ablation studies to assess the contribution of each component.

Experimental Setup

All experiments are conducted in NVIDIA Isaac Sim using the TIAGo++ mobile manipulation platform, with the left arm employed throughout this study. The robot is equipped with a differential-drive base and a 7-DOF manipulator, which is controlled via inverse kinematics (IK) through the left arm–torso kinematic chain. Perception is enabled by a head-mounted RGB-D camera with a resolution of 1280×720. Known intrinsic and extrinsic parameters are used to project observations into both robot-centric and world coordinate frames.

Task Description

We evaluate five open-vocabulary mobile manipulation (OVMM) tasks representative of common household scenarios: *(1) Throw the can into the trash bin (2) Move the pot near the red mug (3) Put the mug on the shelf (4) Open the cabinet (5) Open the dishwasher*

Tasks (1)–(3) are inspired by standard pick-and-place setups frequently explored in prior OVMM work (Yenamandra et al. 2023; Huang et al. 2023a; Qiu et al. 2024; Zhi et al. 2025). However, we argue that OVMM should extend beyond pick-and-place to capture a broader range of interactions relevant to real-world deployment. Consequently, tasks (4) and (5) incorporate articulation-based actions to open objects, thereby expanding the diversity of interactions. These tasks cover a diverse range of object types with varying spatial and directional constraints. We include: (i) simple objects with relaxed requirements (e.g., cans and bins) that can be approached from most directions; (ii) objects like mugs, pots, and shelves that require alignment with specific affordance regions such as handles or flat surfaces; and (iii) articulated objects such as cabinets and dishwashers that require correct approach angles for successful manipulation.

Each task is executed 20 times with randomized object placements, orientations, or initial robot base positions to evaluate robustness. A fixed random seed is used to ensure reproducibility.

Baseline Methods

We compare our approach against four baselines, including classical geometric planners, keypoint-guided methods, and VLM-based prompting strategies. Except for *Object Center + A*/RRT**, all baselines share our coarse navigation setup: using A* to approach within 1.5 m of the object, facing it.

Setting	Throw the Can into Trash	Move Pot Near Red Mug	Put Mug on Shelf	Open Cabinet	Open Dishwasher	Total Success
$\alpha = 0$	10/20	11/20	5/20	6/20	11/20	43%
$\alpha = 0.5$	18/20	14/20	14/20	15/20	15/20	76%
$\alpha = 1$	**20/20**	14/20	**18/20**	12/20	16/20	79%
Increasing α_t (Ours)	17/20	**18/20**	17/20	**16/20**	**17/20**	**85%**

Table 2: Success Rates Under Different α Settings

- **Object Center + A*/RRT***: Classical path planners that select a base placement at a fixed distance from the center of the target object, based solely on collision avoidance and proximity.
- **Affordance Point + A*/RRT***: A VLM selects an affordance point $\mathbf{g}$ using the same procedure as our method, and a classical planner (A*/RRT*) computes a base placement at a fixed distance from $\mathbf{g}$ based on collision-free feasibility.
- **Pivot (I)**: Based on PIVOT (Nasiriany et al. 2024), this method uses a VLM to iteratively update the placement action space, using only the RGB image and prompt.
- **Pivot (M^+_{local}, I_{aff})**: A variant of the above method that uses the same multimodal inputs as our approach. The VLM selects base placements using both M^+_{local} and I_{aff}.

Comparison Results

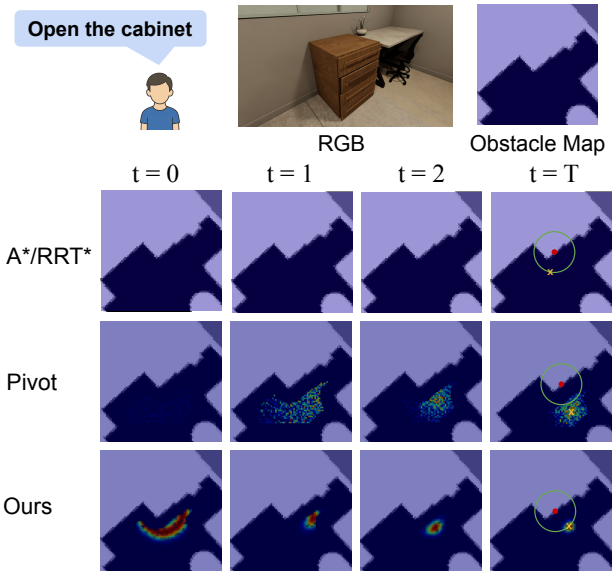
Table 1 summarizes success rates across five semantic manipulation tasks. Our method achieves the highest overall success rate of 85%, significantly outperforming all baselines. It excels on direction-sensitive tasks such as *Open the cabinet*, while remaining robust on less constrained ones like *Throw the can into the trash bin*. Figure 3 presents qualitative comparisons of base placements generated by different baselines and our method for the *Open the cabinet* task, illustrating the importance of both semantic understanding and geometric feasibility.

Object Center + A*/RRT* Classical geometric planners such as A* and RRT* rely solely on spatial proximity. They perform well in tasks where simply reaching a reasonable distance from the target is sufficient, such as *throwing the can into the trash bin* but fail when specific approach angles or semantic understanding are required. Due to their lack of semantic reasoning, success rates drop to 47% and 50%.

Affordance Point + A*/RRT* The limited performance of prior methods may be due to their tendency to approach the object as a whole, rather than targeting the manipulable part. To address this, we incorporate VLM-guided affordance point $\mathbf{g}$ to guide base placement. This yields modest improvements, particularly on spatially structured tasks, with overall success rates reaching up to 61%. However, these methods still lack the ability to reason about directional constraints and reachability. Even when $\mathbf{g}$ lies near a handle, approaching from the side often results in failure. Mislocalization to occluded or non-manipulable regions further reduces effectiveness.

Pivot (I) and Pivot (M^+_{local}, I_{aff}) To further enhance semantic reasoning, we consider VLM-driven prompting strategies. Pivot (I), which relies solely on RGB input, lacks geometric knowledge and often proposes placements that are either at incorrect distances or physically inaccessible. As a result, its performance is limited to 26%. The multimodal variant, Pivot (M^+_{local}, I_{aff}), incorporates affordance RGB and obstacle maps, which partially address perceptual gaps. Yet it still struggles to propose consistently feasible placements, with only a 23% success rate.

Our method Our method addresses the limitations of egocentric perception by projecting affordance cues onto obstacle maps through a cross-modal representation. We then perform a coarse-to-fine optimization that balances semantic intent and geometric feasibility. This approach leads to robust base placements and achieves an overall success rate of 85%, significantly outperforming all baselines.



Green circle: radius r* centered at the affordance location. **Yellow cross:** predicted base blacement. **t**: iteration step. **T**: final step.

Figure 3: Base placement distribution evolution for the task "Open the cabinet." The A*/RRT* baseline (top row) selects a base placement at an oblique angle in front of the cabinet, which is not ideal. The Pivot baseline (second row) selects a region in front of the cabinet but fails due to excessive distance from the target. Our method (bottom row) converges to a base placement that is both feasible and semantically appropriate.

Alpha Comparison

We compare four configurations of the coefficient α_t in Eq.(1), including fixed values of 0, 0.5, and 1, as well as a sigmoid schedule defined in Eq.(2), which gradually increases α_t during optimization. This coefficient controls the weighting between semantic and geometric factors, allowing us to observe how different emphases affect task performance. The results confirm that the coarse-to-fine strategy yields the best overall outcome.

As shown in Table 2, setting $\alpha = 0$ (semantics only) results in sampling concentrated around task-relevant regions but often yields physically infeasible candidates, such as those beyond the robot's reachable workspace. When $\alpha = 0.5$ (balanced semantic and geometric weights), the introduction of geometric constraints effectively filters out physically invalid candidates, leading to a substantial improvement in success rates. However, maintaining a fixed trade-off between semantic and geometric weights can still result in suboptimal placements, particularly in scenarios where the two objectives conflict. With $\alpha = 1$ (geometry only), sampling concentrates on a ring around the affordance point $\mathbf{g}$, favoring geometrically feasible placements. However, these sampled candidates may lack semantically appropriate options or include only a few such candidates. As a result, the VLM is more likely to select semantically incorrect base placements, causing the optimization to converge prematurely to geometrically valid but semantically misaligned regions and ultimately resulting in task failure.

Prior experiments showed that fixed weighting strategies are often insufficient. Semantic-only configurations tend to propose placements that are not physically operable on the target object, while geometry-only configurations risk overlooking the task intent. Even a balanced weighting can lead to suboptimal results when semantic and geometric objectives conflict. To overcome these issues, our approach prioritizes semantic alignment in the early stages, guiding the search toward task-relevant regions, and gradually shifts toward geometric precision to ensure physical feasibility.

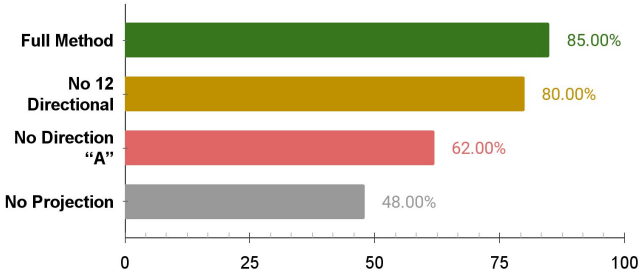


Figure 4: Projection Module Ablation. The full method achieves 85% success. Removing the 12 arrows causes a small drop (80%), while removing the main arrow "A" leads to a larger drop (62%). Without projection, performance drops most significantly to 48%.

Ablation on Affordance Guidance Projection

We conducted an ablation study to evaluate the role of each component in our affordance-guidance projection module by comparing four system variants: (1) **Full Method**, which includes all components and projections; (2) **No 12 Arrows**, which retains "A" and $\mathcal{F}_t$, but removes the 12 auxiliary directional arrows; (3) **No Direction "A"**, which removes the coarse affordance directional arrow "A", omitting the corresponding fan-shaped region $\mathcal{F}_t$; and (4) **No Projection**, which disables the entire projection mechanism and uses only RGB and raw obstacle maps.

As shown in Figure 4, the Full Method achieves the highest success rate at 85%. Removing the 12 additional arrows while keeping "A" results in only a modest drop (to 80%), indicating that while multi-directional cues provide some benefit, the single "A" arrow carries the main guidance signal. Eliminating the coarse arrow "A" reduces performance further (to 62%), showing its critical role in base placement guidance. Finally, removing the projection leads to the largest performance drop (from 85% to 48%), confirming the importance of spatially grounding semantic cues.

Although current VLMs demonstrate strong semantic reasoning capabilities over RGB images, these findings suggest that they have limited ability to automatically transform such semantic understanding into spatially grounded reasoning. This limitation underscores the need for our explicit projection mechanism, which enables VLMs to better associate semantic intent with spatial context and perform more effective reasoning.

6 Conclusion

We identify a key challenge in open-vocabulary mobile manipulation (OVMM), where task failures are often caused by poor base placement. To address this, we propose a cross-modal representation that integrates RGB images with obstacle maps, overcoming the limitations of single-view perception. Our coarse-to-fine planning strategy balances semantics and geometry, enabling the robot to determine base placements that are both task-relevant and physically valid. The proposed method achieves an 85% success rate across five open-vocabulary tasks.

Limitations and Future Work. While generally effective, the predicted placements may exhibit limited geometric precision compared to geometry-based methods, particularly in tasks that require accurate distance estimation. To help the vision-language model better understand semantics and spatial context, we use affordance-guided projection to align the RGB input with the obstacle map. However, the system can still produce incorrect reasoning in some cases. Nevertheless, we believe this limitation will diminish as VLMs continue to improve. Furthermore, although our method can identify base placements that are both semantically appropriate and geometrically feasible, the manipulator arm may still collide with obstacles during execution in more confined or cluttered environments. In future work, we aim to incorporate arm trajectory feasibility into the optimization process to ensure fully executable end-to-end motion plans.

Acknowledgements

This work was supported in part by National Science and Technology Council, Taiwan, under Grant NSTC 113-2634-F-002-007. We are grateful to the National Center for High-performance Computing.

References

Bolte, B.; Wang, A.; Yang, J.; Mukadam, M.; Kalakrishnan, M.; and Paxton, C. 2023. Usa-net: Unified semantic and affordance representations for robot memory. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1–8. IEEE.

Chang, M.; Gervet, T.; Khanna, M.; Yenamandra, S.; Shah, D.; Min, S. Y.; Shah, K.; Paxton, C.; Gupta, S.; Batra, D.; Mottaghi, R.; Malik, J.; and Chaplot, D. S. 2023. GOAT: GO to Any Thing. arXiv:2311.06430.

Hart, P. E.; Nilsson, N. J.; and Raphael, B. 1968. A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics*, 4(2): 100–107.

Huang, H.; Lin, F.; Hu, Y.; Wang, S.; and Gao, Y. 2024a. CoPa: General Robotic Manipulation through Spatial Constraints of Parts with Foundation Models. arXiv:2403.08248.

Huang, J.; Liang, J.; Shi, B.; Yang, Y.; Liu, Y.; Driess, D.; Toussaint, M.; Fu, K.; Lin, K.; Liu, Z.; et al. 2023a. OK-Robot: Zero-shot object navigation using multimodal world models. In *Conference on Robot Learning (CoRL)*.

Huang, J.; Yang, Y.; Driess, D.; et al. 2023b. VLMaps: Grounding large language models with spatial maps for navigation. In *Conference on Robot Learning (CoRL)*.

Huang, W.; Wang, C.; Li, Y.; Zhang, R.; and Fei-Fei, L. 2024b. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. *arXiv preprint arXiv:2409.01652*.

Karaman, S.; and Frazzoli, E. 2011. Sampling-based algorithms for optimal motion planning. *The international journal of robotics research*, 30(7): 846–894.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment Anything. arXiv:2304.02643.

Laina, S. B.; Boche, S.; Papatheodorou, S.; Schaefer, S.; Jung, J.; and Leutenegger, S. 2025. FindAnything: Open-Vocabulary and Object-Centric Mapping for Robot Exploration in Any Environment. arXiv:2504.08603.

Lee, O. Y.; Xie, A.; Fang, K.; Pertsch, K.; and Finn, C. 2025. Affordance-Guided Reinforcement Learning via Visual Prompting. arXiv:2407.10341.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. arXiv:2304.08485.

Minderer, M.; Gritsenko, A.; Stone, A.; Neumann, M.; Weissenborn, D.; Dosovitskiy, A.; Mahendran, A.; Arnab, A.; Dehghani, M.; Shen, Z.; Wang, X.; Zhai, X.; Kipf, T.; and Houlsby, N. 2022. Simple Open-Vocabulary Object Detection with Vision Transformers. arXiv:2205.06230.

Nasiriany, S.; Xia, F.; Yu, W.; Xiao, T.; Liang, J.; Dasgupta, I.; Xie, A.; Driess, D.; Wahid, A.; Xu, Z.; et al. 2024. Pivot: Iterative visual prompting elicits actionable knowledge for vlms. *arXiv preprint arXiv:2402.07872*.

OpenAI. 2023. GPT-4 Technical Report. https://arxiv.org/abs/2303.08774. Accessed June 6, 2025.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y.; Li, S.-W.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2024. DINOv2: Learning Robust Visual Features without Supervision. arXiv:2304.07193.

Qiu, D.; Ma, W.; Pan, Z.; Xiong, H.; and Liang, J. 2024. Open-vocabulary mobile manipulation in unseen dynamic environments with 3d semantic maps. *arXiv preprint arXiv:2406.18115*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; Zeng, Z.; Zhang, H.; Li, F.; Yang, J.; Li, H.; Jiang, Q.; and Zhang, L. 2024. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. arXiv:2401.14159.

Sathyamoorthy, A. J.; Weerakoon, K.; Elnoor, M.; Zore, A.; Ichter, B.; Xia, F.; Tan, J.; Yu, W.; and Manocha, D. 2024. CoNVOI: Context-aware Navigation using Vision Language Models in Outdoor and Indoor Environments. arXiv:2403.15637.

Shao, B.; Cao, N.; Ding, Y.; Wang, X.; Gu, F.; and Chen, C. 2024. MoMa-Pos: An Efficient Object-Kinematic-Aware Base Placement Optimization Framework for Mobile Manipulation. arXiv:2403.19940.

Singh, A. M.; Yang, J.; Chen, A.; Wu, J.; and Finn, C. 2023. CLIP-Fields: Weakly supervised semantic fields for robotic manipulation. In *Conference on Robot Learning (CoRL)*.

Tan, S.; Zhou, D.; Shao, X.; Wang, J.; and Sun, G. 2025. Language-Conditioned Open-Vocabulary Mobile Manipulation with Pretrained Models. arXiv:2507.17379.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Yang, J.; Zhang, H.; Li, F.; Zou, X.; Li, C.; and Gao, J. 2023. Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V. arXiv:2310.11441.

Yenamandra, S.; Ramachandran, A.; Yadav, K.; Wang, A.; Khanna, M.; Gervet, T.; Yang, T.-Y.; Jain, V.; Clegg, A. W.; Turner, J.; et al. 2023. Homerobot: Open-vocabulary mobile manipulation. *arXiv preprint arXiv:2306.11565*.

Zhang, P.; Gao, X.; Wu, Y.; Liu, K.; Wang, D.; Wang, Z.; Zhao, B.; Ding, Y.; and Li, X. 2025. MoMa-Kitchen:

A 100K+ Benchmark for Affordance-Grounded Last-Mile Navigation in Mobile Manipulation. arXiv:2503.11081.

Zhi, P.; Zhang, Z.; Zhao, Y.; Han, M.; Zhang, Z.; Li, Z.; Jiao, Z.; Jia, B.; and Huang, S. 2025. Closed-Loop Open-Vocabulary Mobile Manipulation with GPT-4V. arXiv:2404.10220.

Zhu, J.; Du, Z.; Xu, H.; Lan, F.; Zheng, Z.; Ma, B.; Wang, S.; and Zhang, T. 2024. Navi2Gaze: Leveraging Foundation Models for Navigation and Target Gazing. arXiv:2407.09053.