🦀 MR. CRABS: Mobile RGB-D Camera-LIDAR Robot for Autonomous Bimanual Manipulation & Sensing

Giuse Pham*[1], Matthew Strong*[2], Alex Qiu[1], Joonwon Kang[1], Monroe Kennedy III[1]

***Abstract*— Near-ground bimanual manipulation is seldom studied for manipulation, as a variety of household items require ground interaction with two arms. Additionally, most robots in this domain require moving their *entire body* to view the relevant scene, making tasks, such as environment reconstruction, much slower. To fill these gaps, we present a fully integrated perception system for near-ground bimanual mobile manipulation. This consists of a novel bimanual holonomic robot equipped with dual Intel RealSense depth cameras and a small LIDAR. Unlike other near-ground and bimanual robots, our robot, entitled *MR. CRABS*, is equipped with a pan-and-tilt bottom Realsense, which we show improves mapping and search tasks. Finally, MR. CRABS can perform a proof-of-concept bimanual task for demonstration collection, alleviating the need for teleoperation.**

I. INTRODUCTION

Near-ground manipulation tasks, such as retrieving objects from under furniture, inspecting floor-level infrastructure, or operating in low-clearance environments, are more common than public perception suggests. Unlike tabletop robots, **near-ground workspaces pose significant perception challenges for mobile robots**. Near-ground workspaces suffer from severe viewpoint constraints: cameras mounted at standard heights produce oblique views with heavy occlusion, and the geometry and area of the ground plane within the frames of near-ground robots limit diverse achievable perspectives [1]. To explore a scene, a robot must move inefficiently.

While recent bimanual mobile platforms have often demonstrated ground-level exploration and manipulation capabilities, they suffer from key shortcomings. [2] and [3] only support tabletop tasks. The non-holonomic AhaRobot [4] supports ground tasks, but the mounted camera cannot move in place. YOR [5], the most similar robot to ours, is both holonomic and supports both table-top and bimanual tasks. However, none of its cameras are *active, in-place*, and it does not have LIDAR. Similar to how humans pan and tilt their head to observe a scene without full physical movement, this motivates an *active camera* for near-ground tasks – a feature to quickly improve understanding of a scene.

We present a system that addresses this challenge through three contributions a). **novel bimanual holonomic mobile manipulation platform** equipped with dual Intel RealSense depth cameras, and a top LIDAR; b). a **simple method for active vision information gain** based on an equipped RGB-D pan-and-tilt camera; c). a **quantitative & qualitative** evaluation of the effects of this camera for mapping.

This research was supported by NSF Graduate Research Fellowship No. DGE-2146755.

[·][1]Department of Mechanical Engineering [·][2]Department of Computer Science, Stanford University, Stanford, CA, USA. Emails: {gpham123, mastro1, aqiu34, jwkang, monroek}@stanford.edu.

*Both authors contributed equally to this work.

II. RELATED WORK

Mobile bimanual manipulation platforms. Open-source mobile manipulators have made bimanual data collection broadly accessible. TidyBot++ [6] contributes a holonomic base designed for robot learning, and Mobile ALOHA [2] demonstrates that bimanual mobile tasks can be learned from teleoperated demonstrations on affordable hardware. More recent low-cost designs continue this trend [3]–[5]. Two properties recur across these platforms. First, the workspace of interest is usually at table height, so the arms and sensors are positioned for surfaces roughly a meter off the ground rather than for the floor. Second, the cameras are rigidly mounted, which **ties every change of viewpoint to a change of base pose**. MR. CRABS keeps the open-source, low-cost philosophy of this line of work while targeting the near-ground workspace and decoupling viewpoint from base motion.

Active perception and exploration. The idea that a sensor should be moved deliberately rather than treated as a passive input goes back to the active perception agenda [7], [8] and remains central to active SLAM [9]. Classical exploration drives a robot toward the boundary between known and unknown space [10], [11], while later work selects viewpoints that reduce map or pose uncertainty explicitly [12], [13]. Active perception has since been extended to photorealistic radiance field representations, where the next view is chosen to maximize expected information gain [14]–[18].

Much of this work assumes a sensor that can be placed freely, as on a drone or a manipulator wrist, or accepts the cost of moving the whole robot. Our setting is the constrained middle ground: a ground robot whose viewpoint is otherwise fixed, where a two-degree-of-freedom pan-tilt mount is the cheapest available source of viewpoint diversity.

Mapping and open-vocabulary search. Occupancy mapping remains the practical substrate for exploration, whether as an octree [19] or a GPU-resident signed-distance volume [20], and we use the former to accumulate observed voxels. Searching for a *named* object additionally requires grounding language in the observation stream; object-goal navigation work learns this mapping from experience [21], [22], while promptable segmentation models now supply it zero-shot [23], [24]. We take the latter route, querying SAM3 [24] on each acquired view, which keeps the search free of task-specific training.

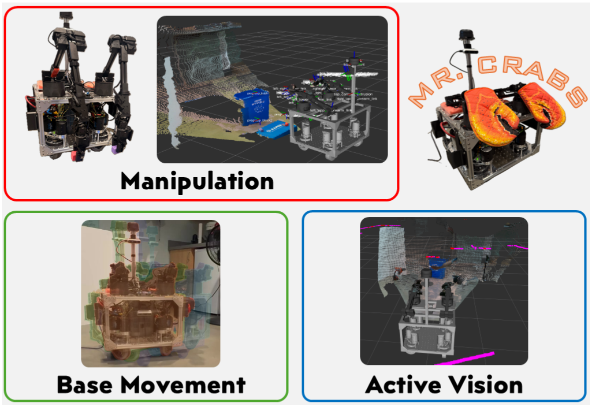

Fig. 1: MR. CRABS: a holonomic mobile TidyBot++ base with dual WidowX-250 arms, a top-mounted LIDAR, and two RealSense depth cameras: one with pan-tilting for active near-ground perception.

III. METHOD

We now present *MR. CRABS*, the bimanual mobile manipulator on a TidyBot++ base. The method is divided into two components: **Hardware Overview** and **Active Perception**.

A. Hardware Overview.

MR. CRABS features a low-cost, modular, smaller mobile base designed from the TidyBot++ hardware package [6]. MR. CRABS is capable of holonomic motion (independently capable of moving in its three DoFs: x, y, θ). Unlike TidyBot++, MR. CRABS uses two Trossen Robotics WidowX-250 arms for bimanual manipulation, given its size and ability to fit on the top of the smaller base. MR. CRABS' perception and mapping capabilities also differ from TidyBot++, relying on two Intel RealSense RGB-D cameras and an RPLiDAR system. The RealSense camera mounted on the base is attached to two Dynamixel motors for panning and tilting motion, and is used to observe the scene closer to the ground, but can be used for objects higher off the ground. We find that this simple design choice leads to improved perception for search and mapping without any base movement. The base has an aluminum extrusion to elevate the RPLiDAR sensor for obstacle detection and mapping. It also holds a static RealSense camera to observe the greater scene.

B. Active Perception.

We formulate object search for a mobile manipulator as greedy selection over a discrete action $\mathcal{A}$ combining base movements $(\Delta\theta, \Delta x, \Delta y)$ with pan-tilt camera poses $(\phi_{\text{pan}}, \phi_{\text{tilt}})$, totaling a set $|\mathcal{A}| = 81$ actions, which consists of every unique combination of 9 total camera poses (3 different pans and tilts) with 9 total combined combinations of base movements. The robot maintains an observed voxel set $\mathcal{V}_t \subset \mathbb{Z}^3$ at resolution $\delta = 5$cm, but this can be tuned depending on the environment. For each candidate action a, we **project the camera frustum** from the hypothetical post-action pose by sampling $N = 400$ rays up to $d_{\text{max}} = 3$m, yielding a predicted visible voxel set $\mathcal{F}(a)$. The information

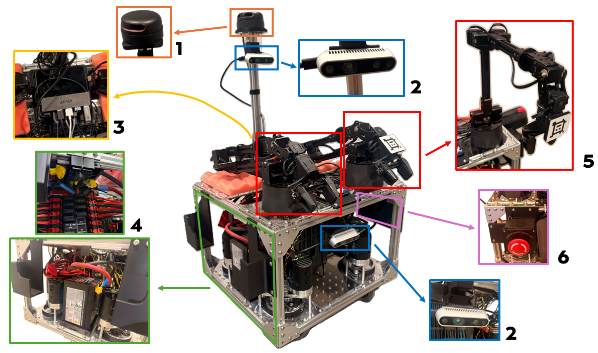

Fig. 2: **Hardware overview of MR. CRABS**: (1) RPLiDAR for mapping and obstacle avoidance, (2) two realsense cameras for navigation and object searching, (3) Intel NUC with USB hubs for sensor input, (4) base movement sub-system, including battery, swerve modules and a breaker to e-stop the base, (5) dual WidowX-250 arms for bimanual manipulation, and (6) an e-stop for the arms.

Algorithm 1 Greedy active view selection

$\mathcal{V}_t \leftarrow \emptyset$; targets $\mathcal{T}$ ▷ empty map, queries
while $\mathcal{T} \neq \emptyset$ **do** ▷ until all found
  $\mathcal{V}_t \leftarrow \mathcal{V}_t \cup \textsc{Scan}()$ ▷ fuse RGB-D view
  $\mathcal{T} \leftarrow \mathcal{T} \setminus \textsc{Sam3}(\mathcal{T})$ ▷ drop found targets
  for all $a \in \mathcal{A}$ **do** ▷ 81 candidates
    $\mathcal{F}(a) \leftarrow \textsc{Cast}(a)$ ▷ N rays to $d_{\max}$
    $g(a) \leftarrow |\mathcal{F}(a) \setminus \mathcal{V}_t| \cdot \rho$ ▷ unseen voxels
    $g(a) \leftarrow \textsc{Adjust}(g(a))$ ▷ saturation / novelty
  end for
  execute $a^* = \arg\max_a g(a)/c(a)$ ▷ gain/sec
end while

gain is: $g(a) = |\mathcal{F}(a) \setminus \mathcal{V}_t| \cdot (\rho)$, where the ρ is a quadratic multiplier that upweights actions into unexplored regions, specifically by using the amount of non-viewed voxels. Pan-tilt actions at saturated poses are penalized, while base movements into fresh poses receive a novelty bonus. The robot selects $a^* = \arg\max_a \; g(a)/c(a)$, where $c(a)$ is the estimated execution time, then queries SAM3 [24] for each target object mask. When such a mask is found (covering more than 100 pixels in the image), the object is considered seen.

Dividing by $c(a)$ is what makes the pan-tilt mount worth having on a mobile manipulator. Naturally, gain alone would often favor driving the base, since translating the robot exposes more unseen space than rotating a camera about a fixed point. Normalizing by the time each action takes turns the objective into a **rate of information per second**, and a pan-tilt sweep that costs a fraction of a second competes well against a base motion that costs several seconds and more energy. Such a policy therefore favors camera motion first and moves the base only once the camera's information gain is exhausted. The procedure is summarized in Alg. 1.

The action set is deliberately small and the horizon is one step, which keeps selection inexpensive enough to run between motions on the onboard computer. The cost of that choice is that the robot cannot plan a sweep whose value only shows up after several actions. We discuss this in Sec. V, and extending the gain to whole trajectories is future work.

IV. Experiments And Results

A. Robot Task

As a proof of concept, MR. CRABS autonomously picks up two blocks in each gripper and places them in a trash bin. Using the combined point cloud, we employ SAM3 to compute object masks. Principal Component Analysis (PCA) on the segmented point cloud extracts the object's major axis, and the gripper is oriented perpendicular to it to grasp across the narrow dimension. Future work will employ grasping algorithms such as GraspNet [25]. We utilize J-PARSE [26] for safe entry in and out of kinematic singularity while retrieving objects at the edge of the workspace. Across a set of 10 trials, we see a strong 80% success rate. Such tasks can be used as augmented demonstrations for a robot learning policy, where only successful tasks are used.

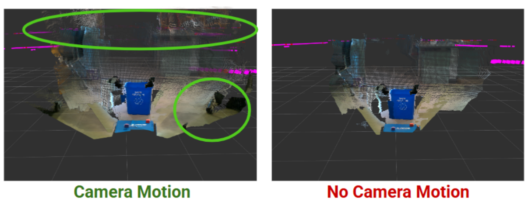

Fig. 3: Point cloud reconstruction with (left) and without (right) pan-tilt camera motion. Green circles highlight additional coverage from active sweeping.

B. Active Search for Multiple Objects

We highlight trials of the ability of the pan-tilt camera motion to assist the base in finding semantically meaningful objects faster. We construct a simple experiment where the robot has to select the highest voxel information gain to find a Stanford bunny and a water bottle on opposite sides of the robot. We run SAM3 on each view (both top and bottom RealSense cameras) until both objects are detected. On 5 trials' average, MR. CRABS finds both objects in 11.7 seconds with camera and base motion, and compared to 24.3 seconds with base motion.

TABLE I: Voxel Coverage Comparison (Sweep vs Fixed Camera)

	Fixed Camera	Sweep Camera	Ratio
Voxels	39,196	196,515	5.0x
Volume	1.06 m^3	5.31 m^3	5.0x
OctoMap nodes	51,453	554,044	10.8x

1) Straight Line Mapping: To highlight the robot's usefulness in moving and using the Pan-Tilt motion, we perform an isolated experiment to have the robot move approximately over 1 meter in the forward direction with OctoMap [19] and a voxel size of 0.03m. In one ablation, we run with no camera motion; in the other, with consistent camera motion. In Table I, we find that voxel coverage consists of about 5 times more voxels and volume. A top-down view of the voxels is shown in Fig. 4 in ground slices, demonstrating a much broader view of the environment, while moving in only one direction.

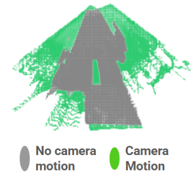

Fig. 4: Voxel coverage with a fixed (gray) vs. panning (green) camera, showing the additional environment coverage gained through camera motion.

V. Limitations

MR. CRABS is not built for higher-ground manipulation. This limits the long-horizon tasks it can complete, such as picking up an item of clothing from the floor and hanging it somewhere higher.

Our evaluation is a set of proof-of-concept demonstrations rather than a benchmark. Each condition is run 5 or 10 times in a single environment, which is enough to show that the information gain metric with the pivoting camera is a promising first step. Because our method uses a 3D-aware representation and does not rely heavily on color, we expect it to generalize to other environments.

The information gain is computed over a discrete set of actions rather than a continuous space. In a larger environment, a robot must plan the **most informative path** to navigate around an environment, while simultaneously avoiding obstacles. A fully autonomous robot in such environments could serve for mapping, data collection, or even real2sim scene construction. The gain is also purely geometric: $g(a)$ counts newly visible voxels regardless of whether they are likely to contain the target, so conditioning it on where the target plausibly lies could find objects sooner. Finally, the pan-tilt mount widens the camera's field of view but does not raise it, so tall or shelved objects remain outside its useful range.

VI. Conclusion

In this work, we presented **MR. CRABS**, a bimanual, holonomic mobile manipulator with a sensing suite for more efficient active perception. The design of the hardware is selected in such a way as to allow for rapid exploration of a scene for manipulation, mapping, and search. Future work focuses in two directions: a full-body teleoperation interface and autonomous mapping of an environment, especially to alleviate the burden of a human operator; and **active** perception during manipulation for difficult close-to-floor manipulation tasks, such as occlusion handling and cluttered scene manipulation.

References

[1] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on robotics*, vol. 32, no. 6, pp. 1309–1332, 2017.

[2] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," *arXiv preprint arXiv:2401.02117*, 2024.

[3] A. Shaw, C. Liu, J. Costa, R. Gray, A. Skowronek, K. Diaz, N. Bui, and N. Correll, "Cutting the cord: System architecture for low-cost, gpu-accelerated bimanual mobile manipulation," *arXiv preprint arXiv:2603.09051*, 2026.

[4] H. Cui, Y. Yuan, Y. Zheng, and J. Hao, "Aharobot: A low-cost open-source bimanual mobile manipulator for embodied ai," *arXiv preprint arXiv:2503.10070*, 2025.

[5] M. H. Anjaria, M. E. Erciyes, V. Ghatnekar, N. Navarkar, H. Etukuru, X. Jiang, K. Patel, D. Kabra, N. Wojno, R. A. Prayage *et al.*, "Yor: Your own mobile manipulator for generalizable robotics," *arXiv preprint arXiv:2602.11150*, 2026.

[6] J. Wu, W. Chong, R. Holmberg, A. Prasad, Y. Gao, O. Khatib, S. Song, S. Rusinkiewicz, and J. Bohg, "Tidybot++: An open-source holonomic mobile manipulator for robot learning," in *8th Conference on Robot Learning (CoRL 2024)*, Munich, Germany, 2024. [Online]. Available: http://tidybot2.github.io

[7] R. Bajcsy, "Active perception," *Proceedings of the IEEE*, vol. 76, no. 8, pp. 966–1005, 1988.

[8] R. Bajcsy, Y. Aloimonos, and J. K. Tsotsos, "Revisiting active perception," *Autonomous Robots*, vol. 42, pp. 177–196, 2018.

[9] J. A. Placed, J. Strader, H. Carrillo, N. Atanasov, V. Indelman, L. Carlone, and J. A. Castellanos, "A survey on active simultaneous localization and mapping: State of the art and new frontiers," *IEEE Transactions on Robotics*, vol. 39, no. 3, pp. 1686–1705, 2023.

[10] B. Yamauchi, "A frontier-based approach for autonomous exploration," in *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation*. IEEE, 1997, pp. 146–151.

[11] C. Dornhege and A. Kleiner, "A frontier-void-based approach for autonomous exploration in 3d," *Advanced Robotics*, vol. 27, no. 6, pp. 459–468, 2013.

[12] C. Papachristos, S. Khattak, and K. Alexis, "Uncertainty-aware receding horizon exploration and mapping using aerial robots," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 4568–4575.

[13] G. Georgakis, B. Bucher, A. Arapin, K. Schmeckpeper, N. Matni, and K. Daniilidis, "Uncertainty-driven planner for exploration and navigation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 11 295–11 302.

[14] W. Jiang, B. Lei, and K. Daniilidis, "Fisherrf: Active view selection and uncertainty quantification for radiance fields using fisher information," in *ECCV*, 2024.

[15] X. Pan, Z. Lai, S. Song, and G. Huang, "Activenerf: Learning where to see with uncertainty estimation," in *ECCV*. Springer, 2022, pp. 230–246.

[16] Z. Yan, H. Yang, and H. Zha, "Active neural mapping," in *ICCV*, 2023.

[17] M. Strong, B. Lei, A. Swann, W. Jiang, K. Daniilidis, and M. Kennedy III, "Next best sense: Guiding vision and touch with FisherRF for 3D Gaussian splatting," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 3204–3210.

[18] T. Chen, A. Dai, M. Adang, G. Gao, and M. Schwager, "Coverage optimization for camera view selection," *arXiv preprint arXiv:2604.05259*, 2026.

[19] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "OctoMap: An efficient probabilistic 3D mapping framework based on octrees," *Autonomous Robots*, 2013, software available at https://octomap.github.io. [Online]. Available: https://octomap.github.io

[20] A. Millane, H. Oleynikova, L. Lanegger, A. Dosovitskiy, C. Cadena, R. Siegwart, and J. Nieto, "nvblox: GPU-accelerated incremental signed distance field mapping," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024.

[21] D. S. Chaplot, D. Gandhi, S. Gupta, A. Gupta, and R. Salakhutdinov, "Learning to explore using active neural slam," *arXiv preprint arXiv:2004.05155*, 2020.

[22] S. K. Ramakrishnan, Z. Al-Halah, and K. Grauman, "Occupancy anticipation for efficient exploration and navigation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 2020, pp. 400–418.

[23] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "Sam 2: Segment anything in images and videos," 2024. [Online]. Available: https://arxiv.org/abs/2408.00714

[24] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang *et al.*, "Sam 3: Segment anything with concepts," *arXiv preprint arXiv:2511.16719*, 2025.

[25] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 444–11 453.

[26] S. Guptasarma, M. Strong, H. Zhen, and M. Kennedy III, "J-parse: Jacobian-based projection algorithm for resolving singularities effectively in inverse kinematic control of serial manipulators," *arXiv preprint arXiv:2505.00306*, 2025.