Hochschule Bonn-Rhein-Sieg
University of Applied Sciences

A²S
Institute for Artificial Intelligence and Autonomous Systems

b-it Bonn-Aachen International Center for Information Technology

UNIVERSITÄT BONN
RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT BONN

HRL
HUMANOID ROBOTS LAB

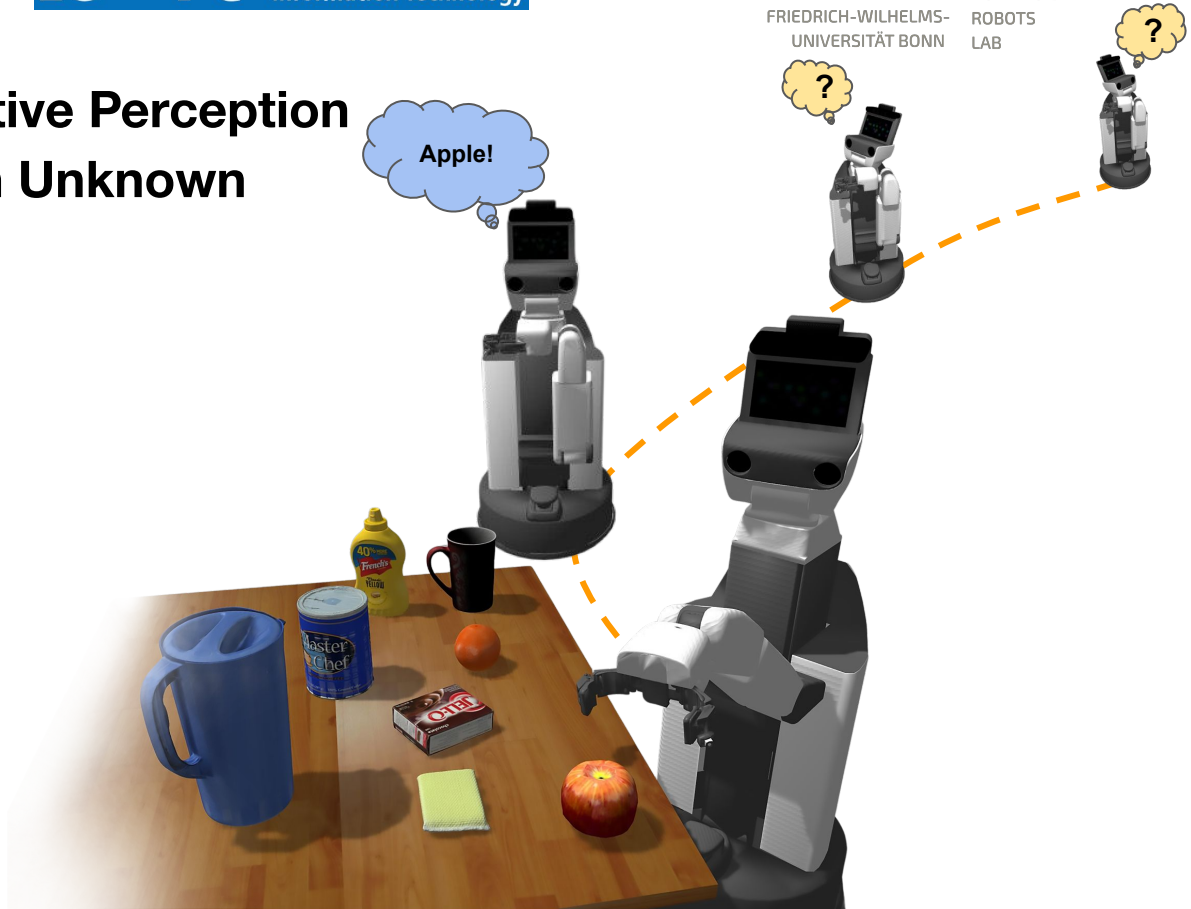
CAPerMoMa: Coupled Active Perception for Mobile Manipulation in Unknown Environments

Hamsa Datta Perur

Master Thesis Defense
28th October, 2024

Supervisors:
Prof. Dr. Sebastian Houben[1]
Prof. Dr. Maren Bennewitz[2]
Rohit Menon, M.Sc.[2]
Dr. Alex Mitrevski[1]

[1] Hochschule Bonn-Rhein-Sieg
[2] University of Bonn



Motivation

Industrial robots [1]

Source : [1] https://blog.robotiq.com/a-brief-history-of-robots-in-manufacturing

Motivation

Industrial robots [1]

Paradigm Shift

Domestic robots [2]

Source : [2] https://www.irobot.com/

Motivation

Industrial robots [1]

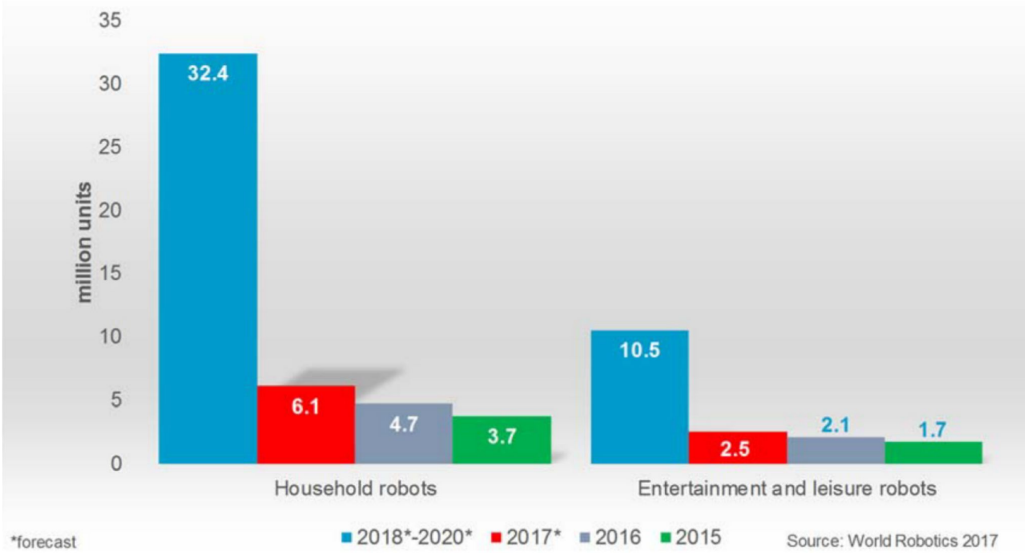
Paradigm Shift

Domestic robots [2]

Paradigm Shift

Domestic robots (mobile manipulator - MoMa)

Motivation



Domestic service robots unit sales 2015-2016, and forecast for 2017-2020 (taken from [3])

Source : [3] J. M. Garcia-Haro, E. D. Oña, J. Hernandez Vicen, S. Martinez, and C. Balaguer, “Service robots in catering applications: A review and future challenges,” Electronics, vol. 10, no. 1, p. 47, 2020.

Motivation

Explore

Navigate

Grasp [4]

Common challenges:

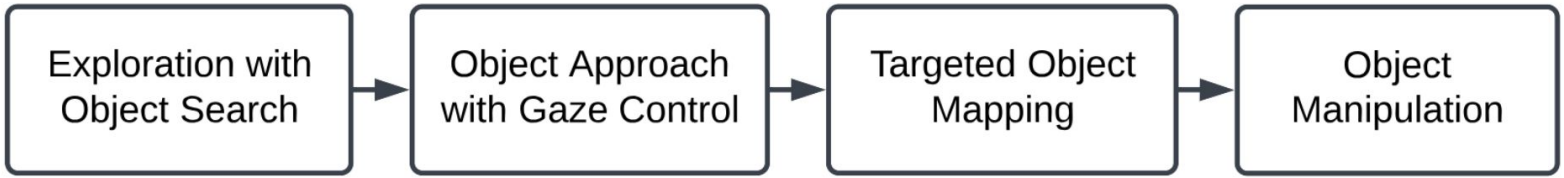
- Recognize objects in Indoor spaces.
- Determine a suitable base placement for manipulation.
- Consistently grasp objects.
- Map target object due to complex geometry.

Source : [4] https://www.youtube.com/watch?v=i6VbOCcqcFk

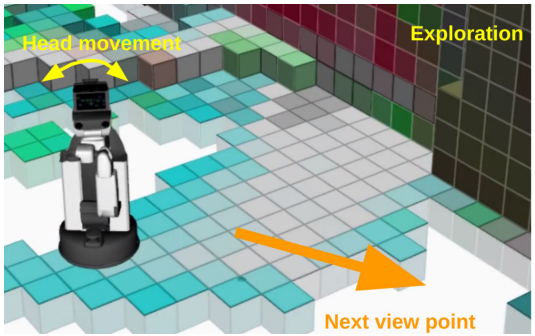
Our Approach

- We propose CAPerMoMa - a system designed to improve **MoMa's (Toyota HSR)** ability to explore indoor environments, detect objects, and execute precise manipulations.
- MoMa is equipped with actively controllable cameras that focus (gaze control) on the target object once it is detected - which we term **"coupled active perception"**.
- The ***'coupled'*** aspect refers to this dual objective of gazing at the object for mapping while planning for a base placement for grasping.

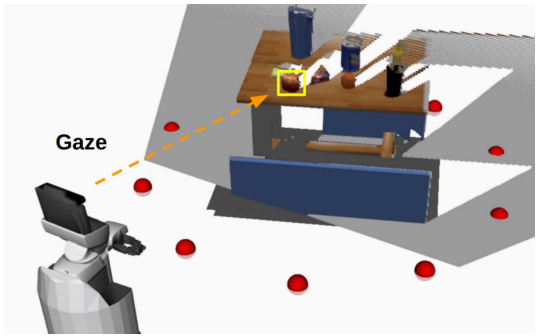
Hypothesis: Coupled approach will help to reduce mapping time (inturn reduce overall time), and the base placement planning will improve the grasp success rate.



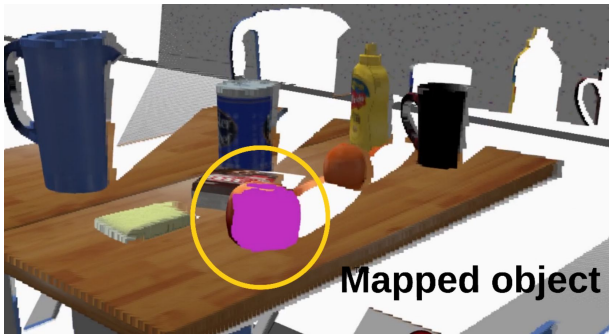
CAPerMoMa: Proposed Approach



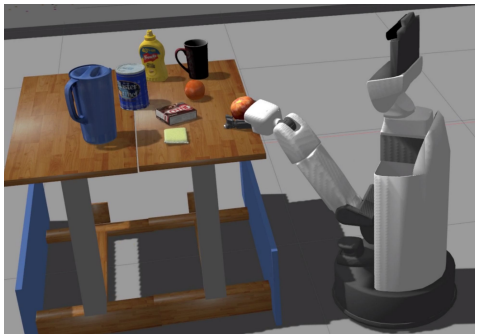
1. Exploration



2. Approach with Gaze Control



3. Object Mapping

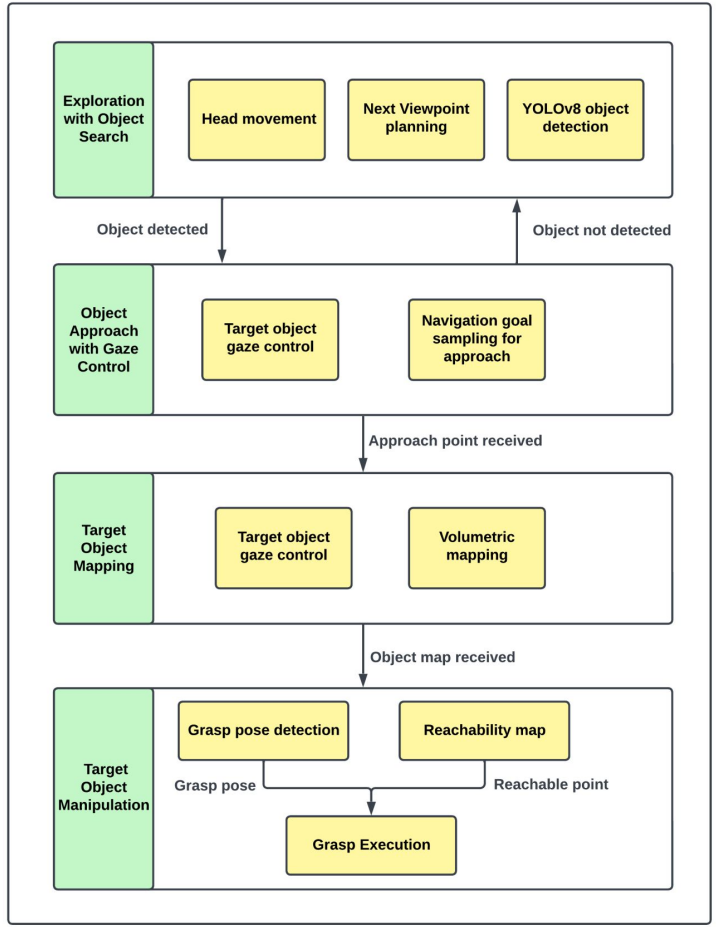


4. Object Manipulation

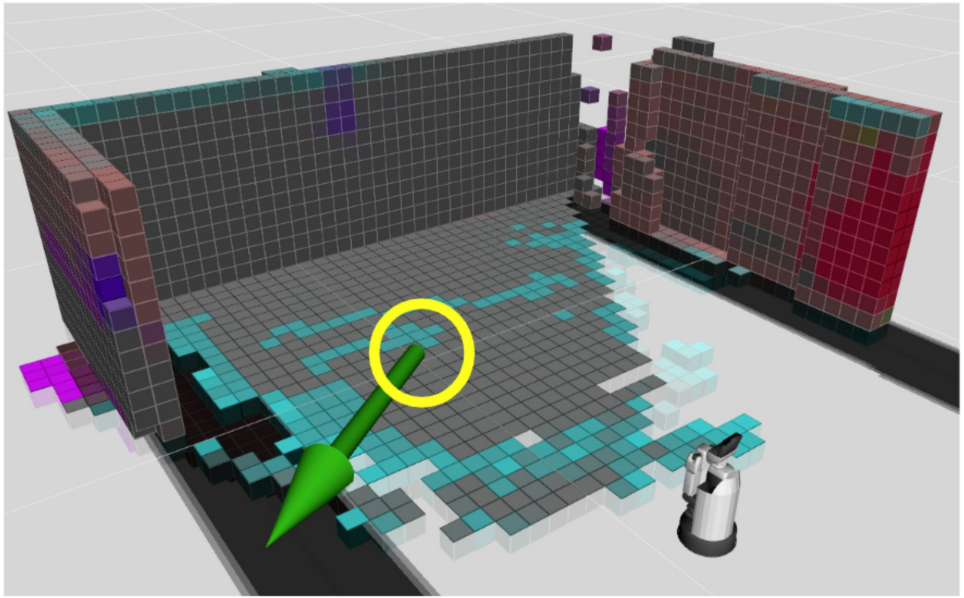
CAPerMoMa: System Overview

Key phases and components:

1. **Exploration with Object Search:**
   - Head movement with entropy-based information gain approach.
   - Next viewpoint planning.
   - Object detection using YOLOv8.
2. **Object Approach with Gaze Control:**
   - Continuous gaze control using YOLOv8 detection output.
   - Navigation goal sampling for base placement approach.
3. **Target Object Mapping:**
   - Volumetric 3D mapping of the target object.
   - Object detection using YOLOv8.
4. **Target Object Manipulation:**
   - Grasp synthesis.
   - Inverse reachability map computation.
   - Base placement computation.
   - Grasp execution.



Exploration: Next Viewpoint Planning



Exploration based on Open3DExplorer [5]

Reference: [5] Z. Wang, H. Chen, and M. Fu, “Whole-body motion planning and tracking of a mobile robot with a gimbal rgb-d camera for outdoor 3d exploration,” Journal of Field Robotics, vol. 41, no. 3, pp. 604–623, 2024.

Exploration: Head Movement

Goal: Move the robot's head towards regions that maximize information gain.

Head movement with entropy-based information gain approach.

$$H = -\sum_{i=0}^{N} p_i \log_2(p_i) \qquad \triangle H = H_{before} - H_{after}$$

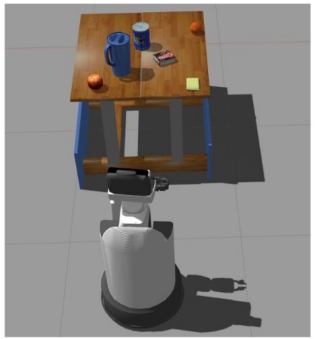
Where:
- N is the number of unique pixel intensity values.
- p_i is the probability calculated using histogram of pixel intensities and normalized.

Pan and Tilt Angles Calculation:

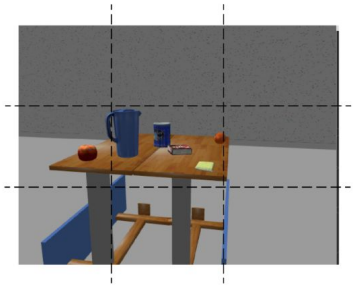
$$\theta_{pan} = \theta_{pan}^{min} + \frac{l}{g-1}(\theta_{pan}^{max} - \theta_{pan}^{min})$$

$$\theta_{tilt} = \theta_{tilt}^{min} + \frac{k}{g-1}(\theta_{tilt}^{max} - \theta_{tilt}^{min})$$

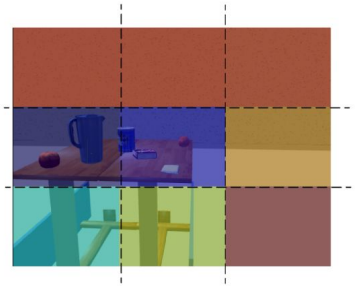
Where:
l, k are the row and column of the grid, respectively. g represents the grid size. The parameter θ represents the pan or tilt angles based on min or max values.



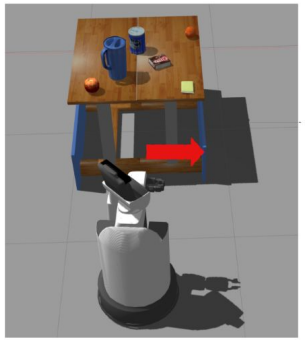
Intial Head Pose



3x3 Sections

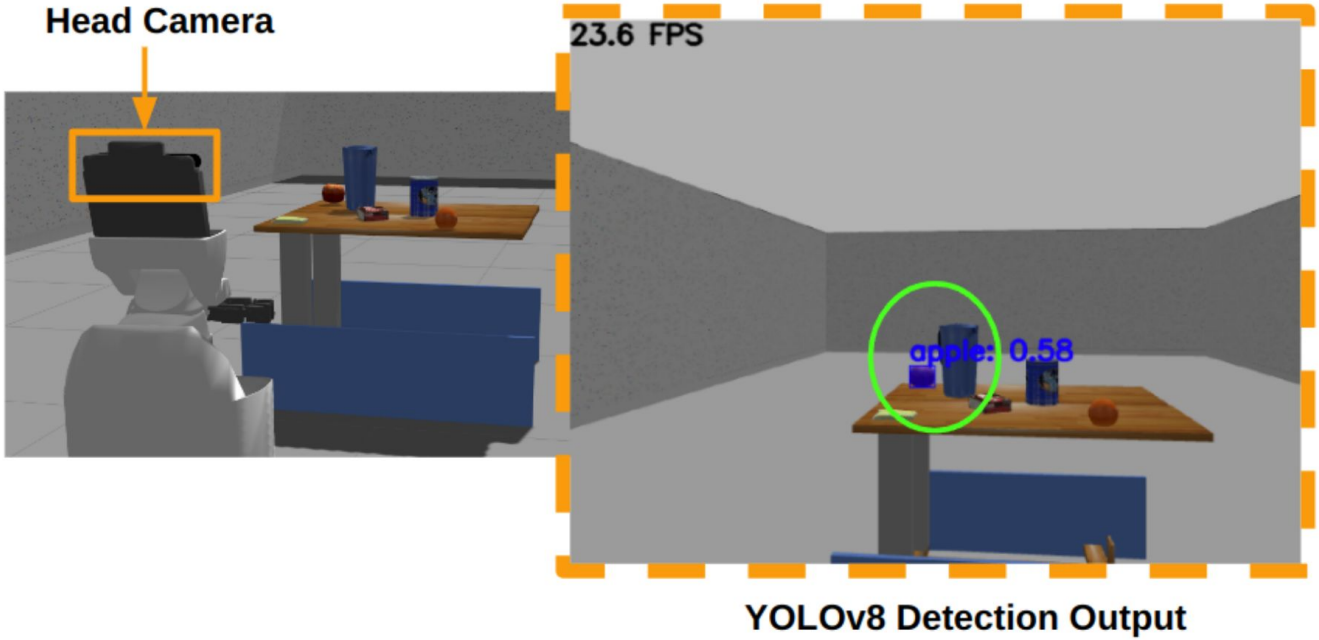


Entropy Color Map



Next Head Pose

Exploration: Object detection



Object detection based on YOLOv8 [6]

Reference: [6] https://github.com/ultralytics/ultralytics

Object Approach: Gaze Control

Goal: Given the 2D pixel coordinates (u, v) and the depth value z from the depth image, we can compute the corresponding 3D point (x, y, z) in the map frame.

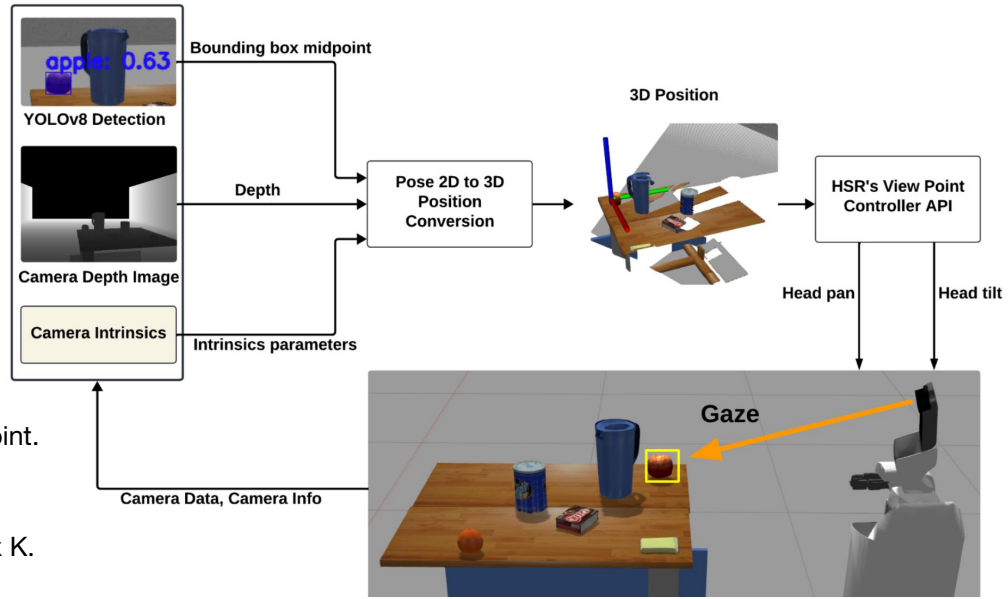
The relationship is derived from the camera projection equation:

$$x = \frac{(u - c_x).z}{f_x}$$

$$y = \frac{(v - c_y).z}{f_y}$$

$$z = z$$

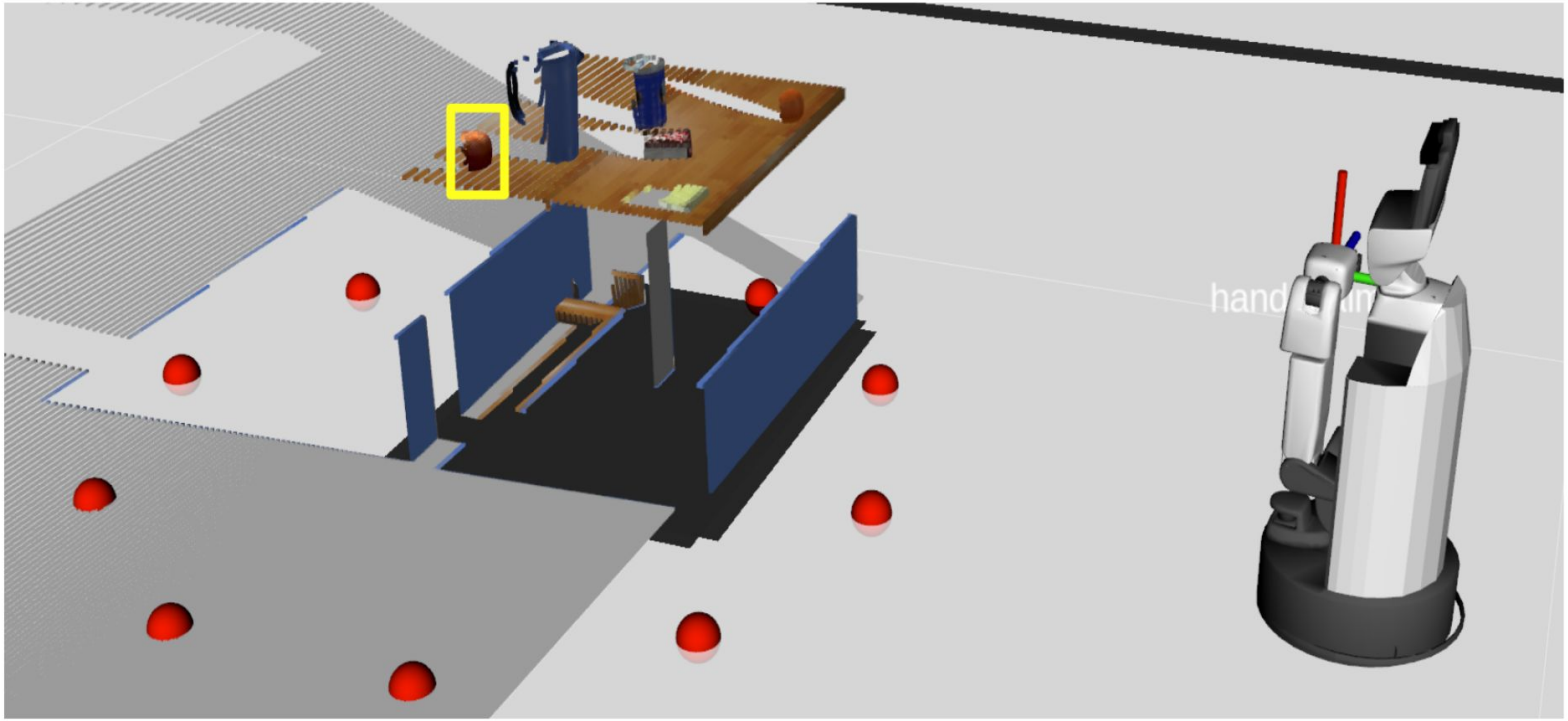
$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$$



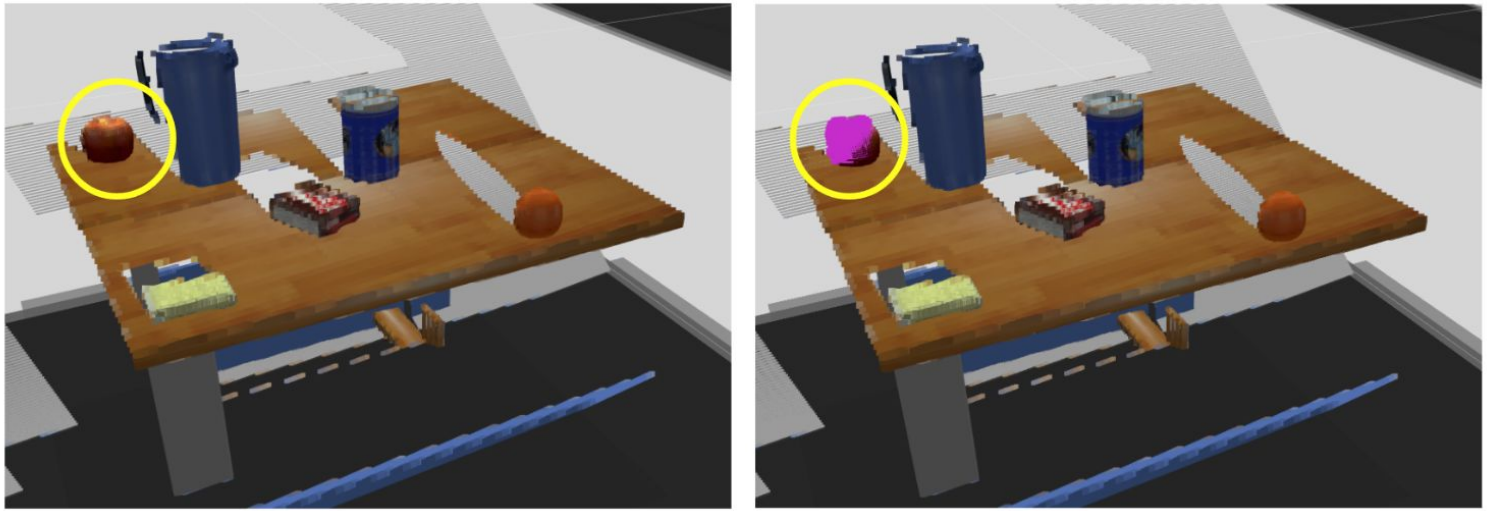
where:

- u and v are the pixel coordinates of the bounding box midpoint.
- z is the depth value from the depth image at the pixel (u, v).
- The matrix K is extracted from the camera info topic in ROS.
- c_x , c_y , f_x , f_y are intrinsic parameters from the camera matrix K.
- f_x and f_y are the focal lengths in the x and y directions.
- c_x and c_y are the coordinates of the optical center (principal point).

Object Approach: Base Placement Sampling



Object Mapping: Volumetric 3D Mapping



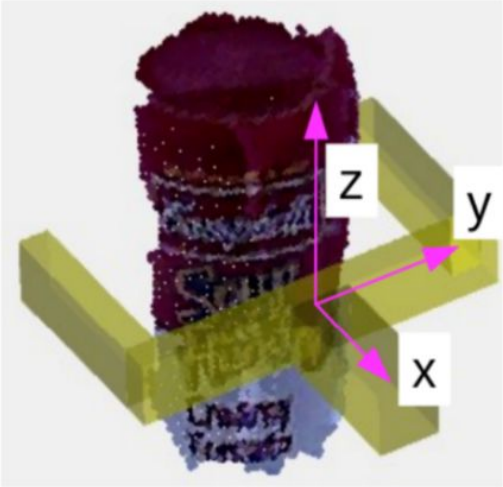
Before **After**

Volumetric Mapping

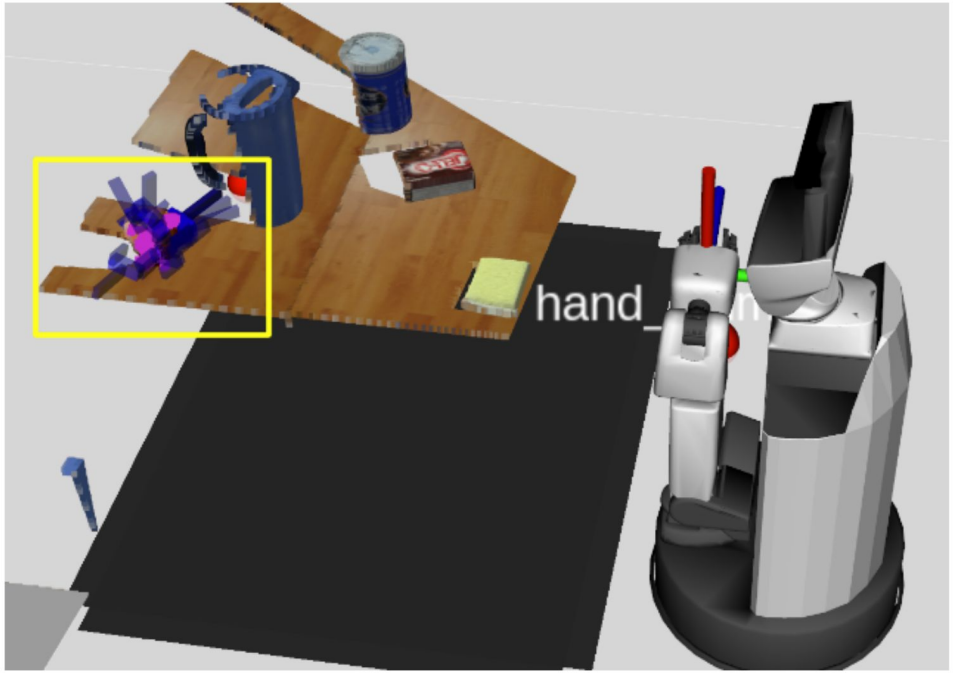
Volumetric 3D Mapping based on VoxBlox++ [7]

Reference: [7] M. Grinvald, F. Furrer, T. Novkovic, J. J. Chung, C. Cadena, R. Siegwart, and J. Nieto, “Volumetric instance-aware semantic mapping and 3d object discovery,” IEEE Robotics and Automation Letters, vol. 4, no. 3, pp. 3037–3044, 2019.

Object Manipulation: Grasp Synthesis



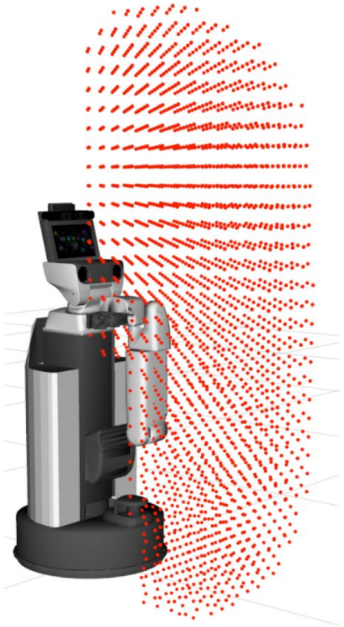
Sample Grasp Pose Detector (GPD) [8] output.(Image from [8])



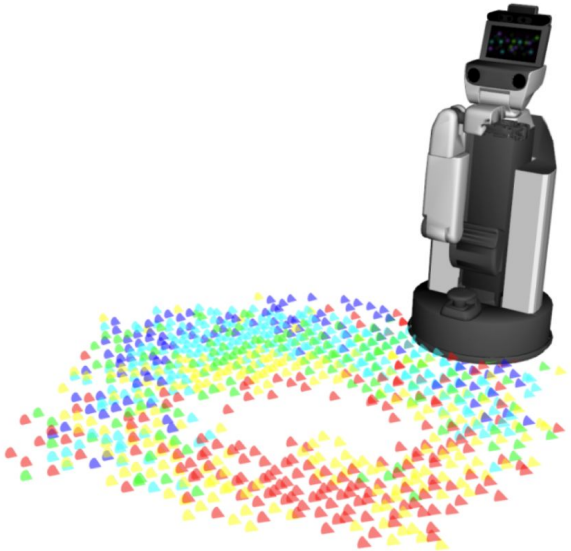
Grasp Synthesis based on GPD [8]

Reference: [8] A. Ten Pas, M. Gualtieri, K. Saenko, and R. Platt, “Grasp pose detection in point clouds,” The International Journal of Robotics Research, vol. 36, no. 13-14, pp. 1455–1473, 2017.

Object Manipulation: (Inverse) Reachability Map



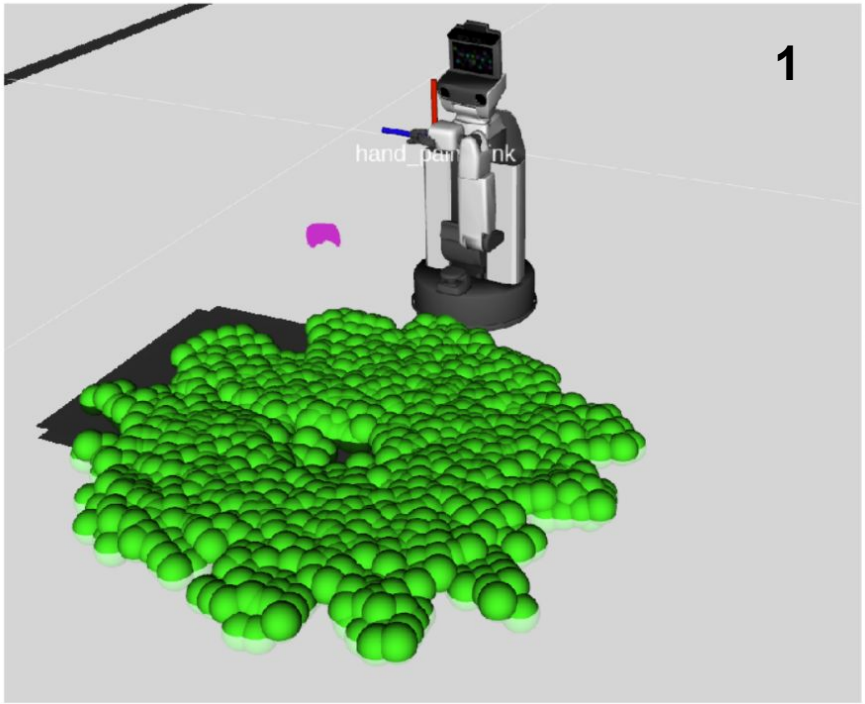
Reachability Map

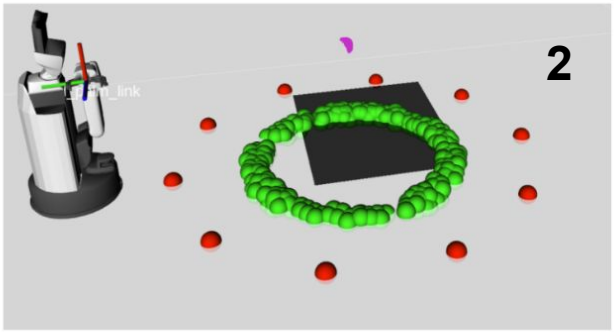


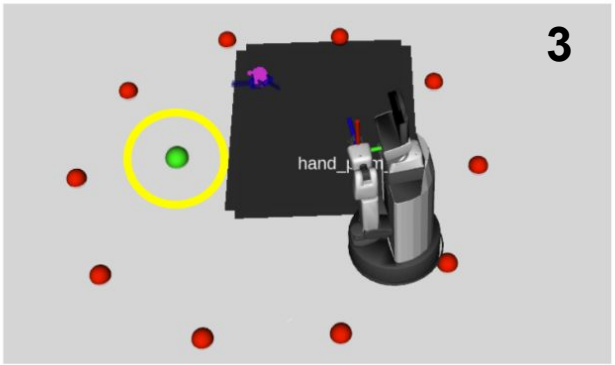
Inverse Reachability Map [9]

Reference: [9] S. Jauhri, J. Peters, and G. Chalvatzaki, “Robot learning of mobile manipulation with reachability behavior priors,” IEEE Robotics and Automation Letters, vol. 7, no. 3, pp. 8399–8406, 2022.

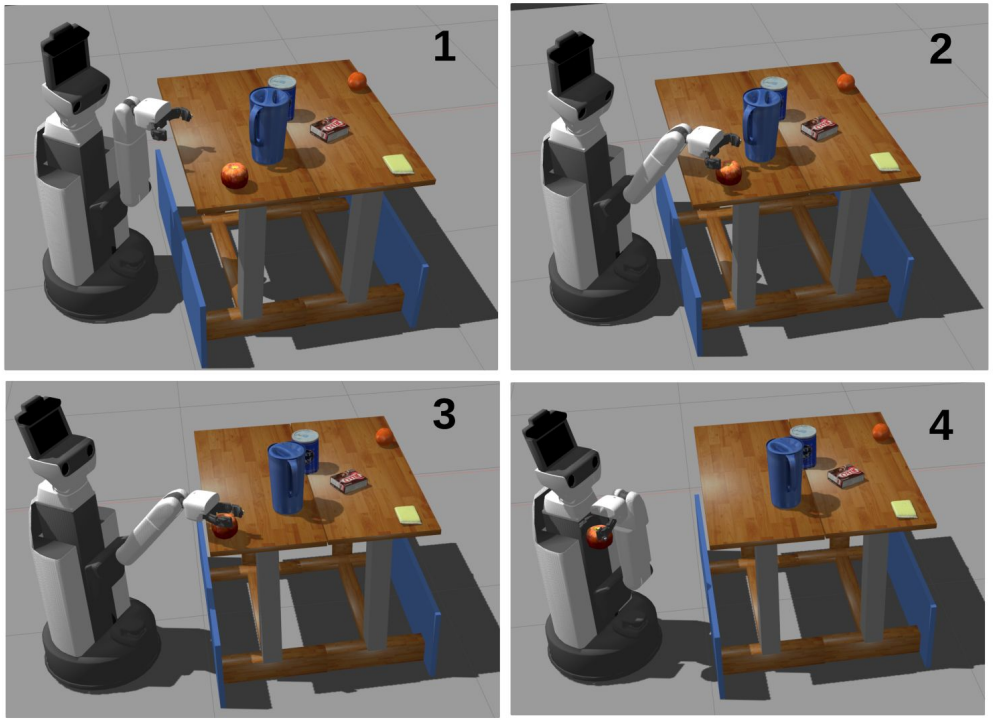
Object Manipulation: Best Base Placement Computation







Object Manipulation: Grasp execution



Evaluation

Experimental Setup:

All experiments were conducted in simulation using the *Gazebo 11* simulator on a high-performance PC. Below are the relevant technical specifications of the system used:

- **Model:** XMG NEO 16 (E23)
- **GPU:** NVIDIA GeForce RTX 4060 Laptop, 8 GB GDDR6
- **Processor:** Intel Core i9-13900HX
- **RAM:** 16GB DDR5-5600 (2 × 8 GB)
- **Storage:** 1TB M.2 Samsung 980, PCIe 3.0 x4 NVMe

Experimental Overview:

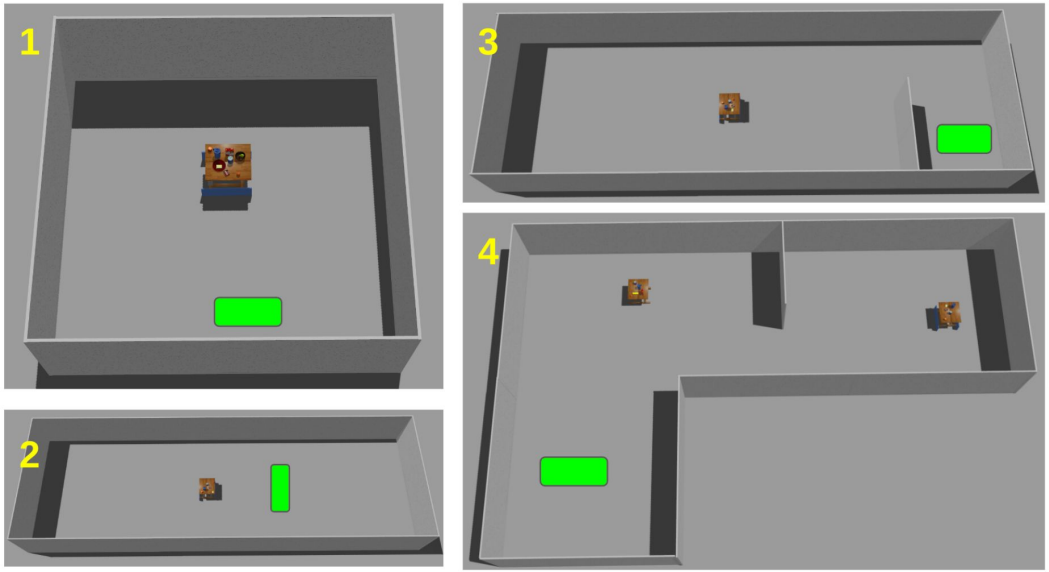
- **Experiment 1:** The primary aim of this experiment is to assess the overall performance of the CAPerMoMa system, with particular focus on evaluating the time taken by each stage of the pipeline.
- **Experiment 2:** The second experiment is an ablation study designed to analyze the importance of key components in the CAPerMoMa system.
- **Experiment 3:** The third experiment evaluates the CAPerMoMa system's performance with different objects

Evaluation Metrics:

- **Total time taken** for each stage of the pipeline and for the entire mobile manipulation task**.**
- **Grasp success or failure,** where a successful grasp is defined as the robot's ability to successfully hold and lift the object from the table.

Experiment 1: Overview

- The first experiment involves four different scenarios with progressively increasing search space.
- In each scenario, the target object (**apple**) is placed in five distinct locations on the table.
- Experiment 1 consists of a total of 20 trials, with each scenario having five trials corresponding to different graspable object positions.



Four different scenarios.

Five different trials.

Experiment 1: Results

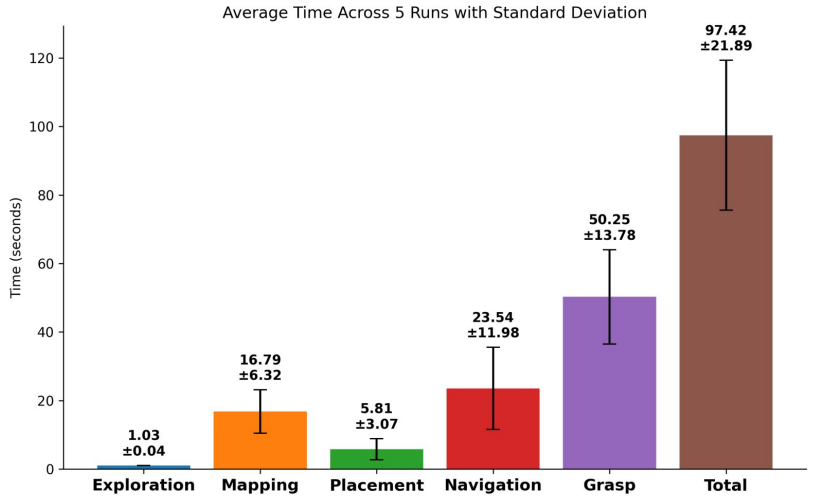
Scenario 1

Key observations:

- Mapping time varied across trials, with Trial 3 having the shortest time of 7.42 seconds.
- Placement time showed notable variation, with Trial 2 achieving the least time at 2.78 seconds.
- Exploration times were consistent across trials, with values around 1.01-1.10 seconds, this is because MoMa's **initial pose was very near** to the table.
- Trial 5 had the lowest total execution time of 78.15 seconds, with the fastest navigation time at 13.53 seconds.

Trial Number	Exploration Time (s)	Mapping Time (s)	Placement Time (s)	Navigation Time (s)	Grasp Time (s)	Total Execution Time (s)
1	1.02	24.29	9.77	17.68	**31.33**	84.10
2	1.10	20.46	2.78	39.90	60.92	125.16
3	1.02	**7.42**	7.01	14.01	53.27	82.73
4	**1.01**	16.42	**2.60**	32.57	64.37	116.97
5	1.02	15.34	6.89	**13.53**	41.36	**78.15**
Average	1.03	16.79	5.81	23.53	50.25	97.42
Standard Deviation	±0.04	±6.32	±3.07	±11.98	±13.78	±21.89

The highlighted are the least time taken for that particular task



Experiment 1: Results

Scenario 1

Grasp outcome: **5/5** successful grasps with no reported failures (**100% success)**

Trial Number	Grasp Outcome	Reason for Failure
1	Success	No Failure
2	Success	No Failure
3	Success	No Failure
4	Success	No Failure
5	Success	No Failure

Experiment 1: Results

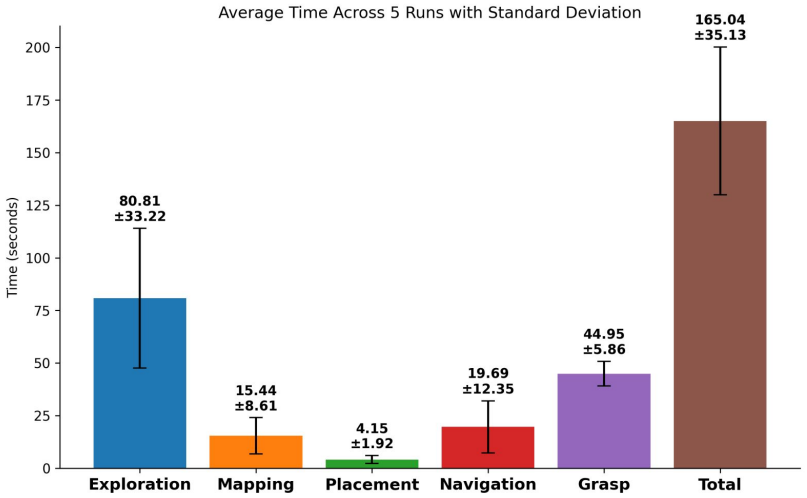
Scenario 2

Key observations:

- The average mapping time was 15.44 seconds. Trial 1 had the fastest mapping time of 2.45 seconds, where MoMa might have found the target object quickly from the far and initiate the mapping early.
- Placement times were more consistent, with an average of 4.15 seconds.
- The average exploration time across the five trials was 80.81 seconds. Trial 4 achieved the lowest exploration time of 39.71 seconds.
- The overall execution time for Scenario 2 averaged 165.04 seconds.

Trial Number	Exploration Time (s)	Mapping Time (s)	Placement Time (s)	Navigation Time (s)	Grasp Time (s)	Total Execution Time (s)
1	101.80	**2.45**	2.93	14.48	43.82	165.47
2	73.62	23.33	7.54	**0.44**	**39.14**	144.07
3	124.96	11.66	3.51	26.66	54.18	220.98
4	**39.71**	17.43	3.71	26.12	41.11	**128.09**
5	63.97	22.34	**3.04**	30.77	46.48	166.60
Average	80.81	15.44	4.15	19.69	44.94	165.04
Standard Deviation	±33.22	±8.61	±1.92	±12.35	±5.86	±35.13

The highlighted are the least time taken for that particular task



Experiment 1: Results

Scenario 2

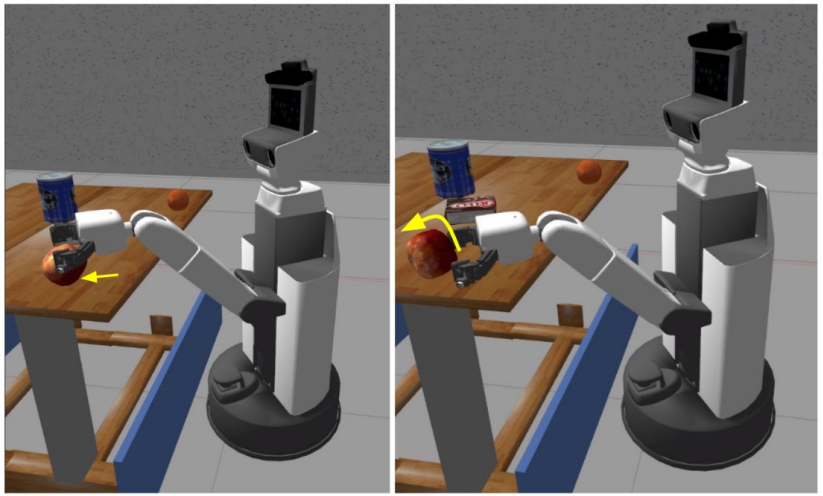
Grasp outcome: **3/5** successful grasps with 2 reported failures (**60% success)**

Trial Number	Grasp Outcome	Reason for Failure
1	Success	No Failure
2	Fail	Object roll over
3	Fail	Early gripper closing
4	Success	No Failure
5	Success	No Failure

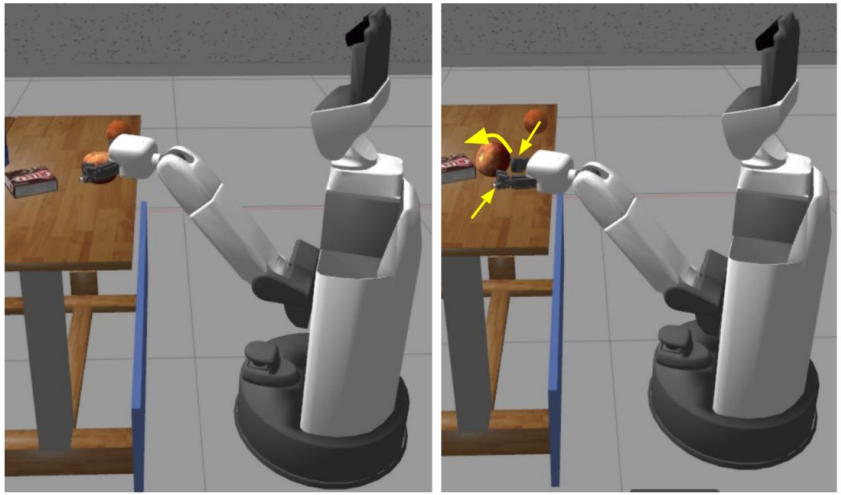
Experiment 1: Results

Scenario 2

Grasp fail scenarios



Object roll over



Early gripper closing

Experiment 1: Results

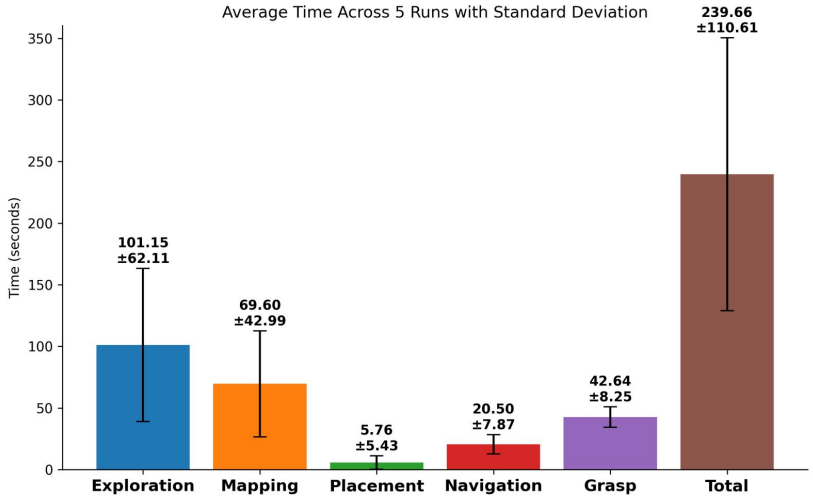
Scenario 3

Key observations:

- The average mapping time was 69.60 seconds. The fastest mapping time was recorded in Trial 1 at 10.90 seconds. The longer mapping time is due to **longer time to query for an instance** during the run.
- Placement times were relatively consistent, with an average of 5.76 seconds.
- The average exploration time across the five trials was 101.15 seconds. The fastest exploration time was 58.79 seconds in Trial 4.
- The overall execution time varied significantly, with an average of 239.66 seconds.

Trial Number	Exploration Time (s)	Mapping Time (s)	Placement Time (s)	Navigation Time (s)	Grasp Time (s)	Total Execution Time (s)
1	89.60	**10.90**	3.14	23.13	**38.22**	**165.00**
2	75.77	54.76	3.31	**10.71**	38.61	183.16
3	210.49	119.05	15.47	31.86	57.26	434.14
4	**58.79**	103.92	3.80	16.50	38.18	221.20
5	71.11	59.35	**3.09**	20.30	40.94	194.80
Average	101.15	69.60	5.76	20.50	42.64	239.66
Standard Deviation	±62.11	±42.99	±5.43	±7.87	±8.25	±110.61

The highlighted are the least time taken for that particular task



Experiment 1: Results

Scenario 3

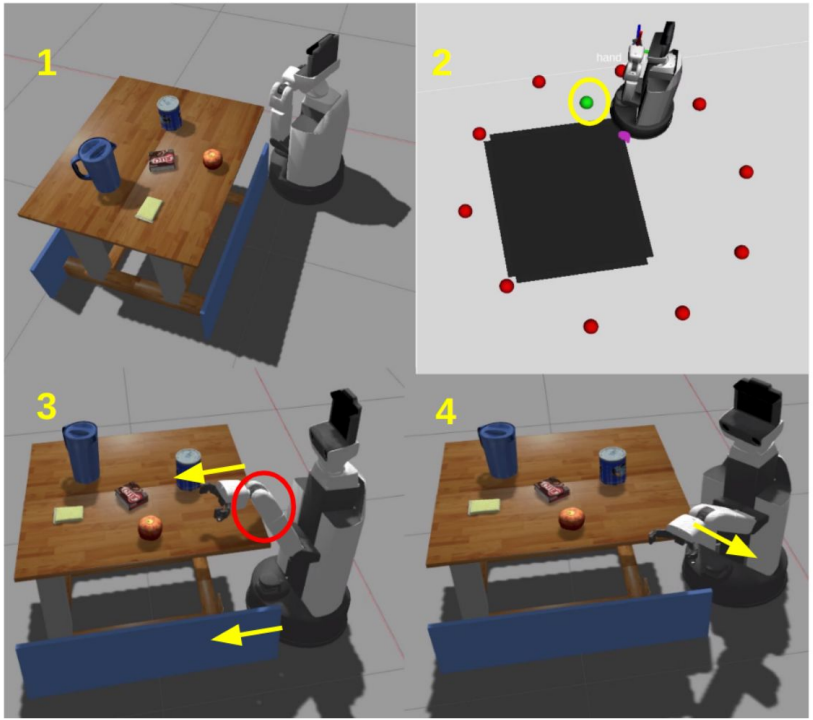
Grasp outcome: **4/5** successful grasps with 1 reported failure (**80% success)**

Trial Number	Grasp Outcome	Reason for Failure
1	Success	No Failure
2	Success	No Failure
3	Fail	Wrong base placement
4	Success	No Failure
5	Success	No Failure

Experiment 1: Results

Scenario 3

Grasp fail scenario



Experiment 1: Results

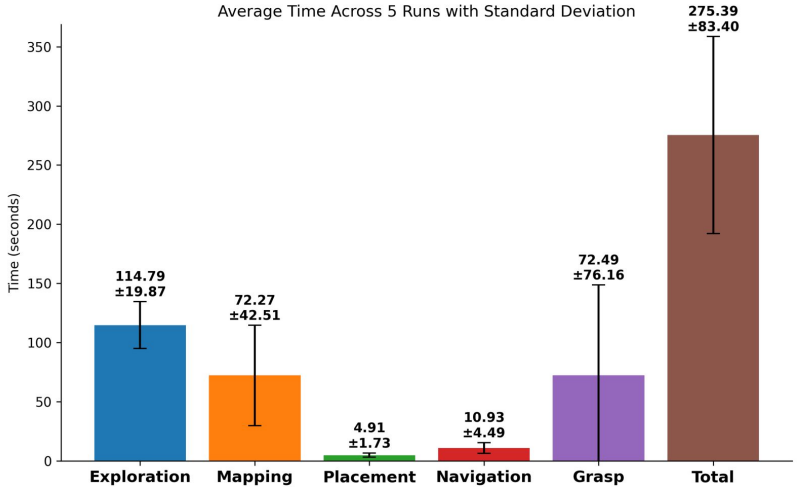
Scenario 4

Key observations:

- The average mapping time was 72.27 seconds. Trial 2 had the fastest mapping time, at 17.83 seconds.
- Placement times were relatively consistent across all trials, with an average of 4.91 seconds.
- The average exploration time across the five trials was 114.79 seconds. The fastest exploration time was recorded in Trial 1 at 91.79 seconds.
- The overall execution time across the trials averaged 275.39 seconds.

Trial Number	Exploration Time (s)	Mapping Time (s)	Placement Time (s)	Navigation Time (s)	Grasp Time (s)	Total Execution Time (s)
1	**91.79**	98.72	6.17	**8.19**	208.71	413.58
2	121.58	**17.83**	6.79	12.57	39.94	**198.72**
3	99.25	102.54	4.49	14.54	**38.33**	259.15
4	119.30	107.51	**2.33**	14.87	36.67	280.69
5	142.04	34.75	4.77	4.47	38.78	224.82
Average	114.79	72.27	4.91	10.93	72.48	275.39
Standard Deviation	±19.87	±42.51	±1.73	±4.49	±76.16	±83.40

The highlighted are the least time taken for that particular task



Experiment 1: Results

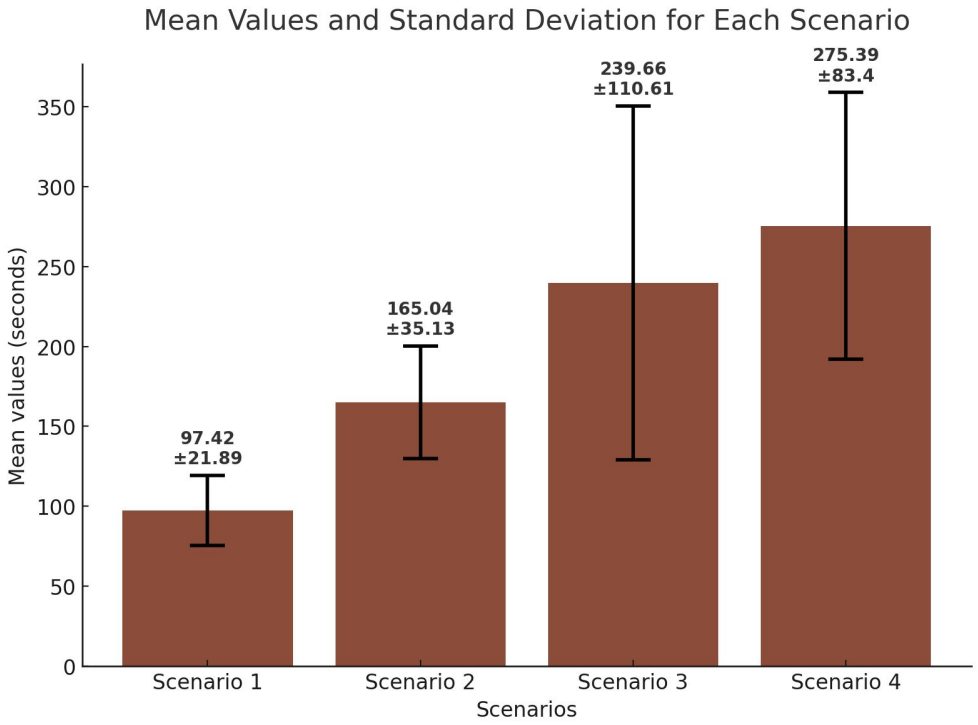
Scenario 4

Grasp outcome: **2/5** successful grasps with 3 reported failure (**40% success)**

Trial Number	Grasp Outcome	Reason for Failure
1	Fail	Wrong pose and grasp planning failed
2	Success	No Failure
3	Fail	Instance ID problem and mapping fail
4	Success	No Failure
5	Fail	Instance ID problem and mapping fail

Experiment 1: Results

Common result comparing all the four scenarios (total execution time)



Experiment 2: Overview

Experiment 2 involves conducting an ablation study to evaluate the significance of key components in the CAPerMoMa system. We use Scenario 2 only as the standard testing environment for this study.

- **Ablation 1 (Disabling Gaze Control):**
  - *Hypothesis*: Disabling gaze control will result in an increase in mapping time.
  - Continuous object detection, facilitated by gaze control, helps depth segmentation and volumetric mapping.
  - Disabling this component, the object detector may struggle to consistently identify the target object, leading to slower mapping completion
- **Ablation 2 (Disabling Reachability Analysis):**
  - *Hypothesis*: Without reachability analysis, the system will fail to manipulate objects positioned far from the robot's current location.

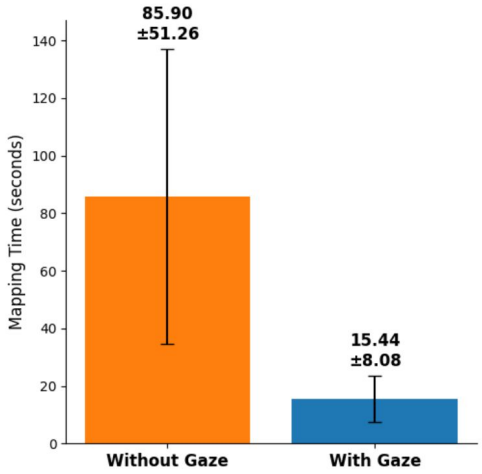
Experiment 2: Results

Ablation 1 (Disabling Gaze Control)

- Comparing the mapping time: with and without gaze control.
- Compared against **Scenario 2** from **experiment 1**.
- A significant reduction in mapping time—**~ 80%**—when gaze control is utilized.
- This continuous focus facilitates uninterrupted object tracking and mapping.
- Without gaze control, MoMa system may intermittently lose sight of the object, causing delays in the mapping process.

Trial Number	Exploration Time (s)	Mapping Time (s)	Placement Time (s)	Navigation Time (s)	Grasp Time (s)	Total Execution Time (s)
1	66.89	81.02	5.40	22.81	41.06	217.18
2	61.06	152.65	5.47	13.25	**38.80**	271.24
3	46.17	**20.19**	**2.80**	11.94	46.88	**127.98**
4	**36.02**	128.52	3.26	**9.93**	44.97	222.70
5	72.98	47.14	7.05	27.53	56.79	211.51
Average	56.62	85.90	4.80	17.09	45.70	210.12
Standard Deviation	±15.19	±51.26	±1.71	±6.70	±6.65	±52.02

The highlighted are the least time taken for that particular task

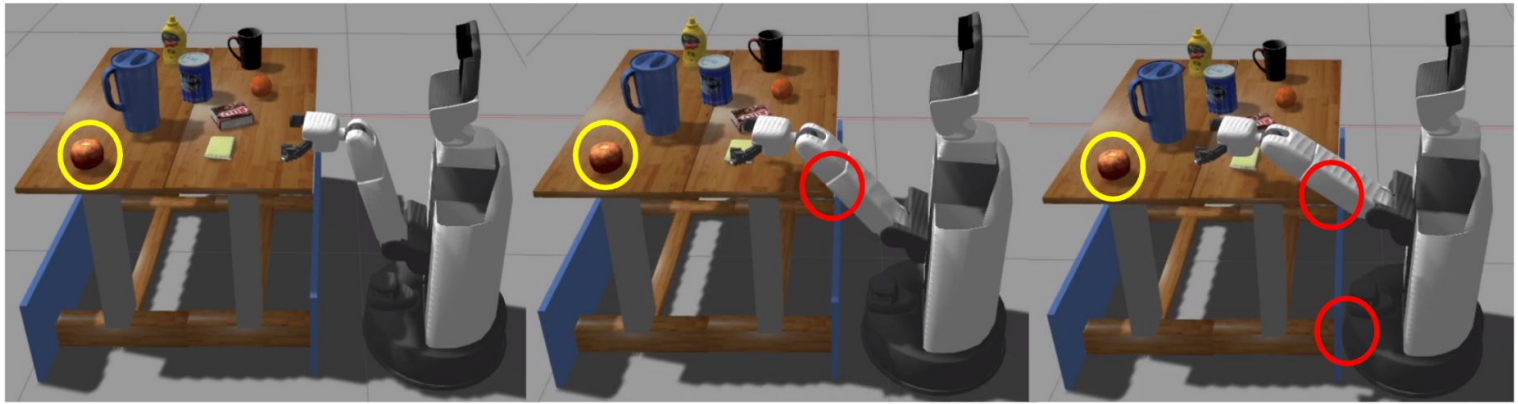


Experiment 2: Results

Ablation 2 (Disabling Reachability Analysis)

- Results indicate that in **4/5**, the robot failed to grasp the target object (**20% grasp success**).
- Failure due to the **object being unreachable**, likely because the robot's base was not positioned optimally to extend its manipulator fully toward the target.

Trial Number	Grasp Outcome	Comments
1	Fail	Target object unreachable
2	Fail	Base hitting the table
3	Success	No Failure
4	Fail	Base hitting the table
5	Fail	Target object unreachable

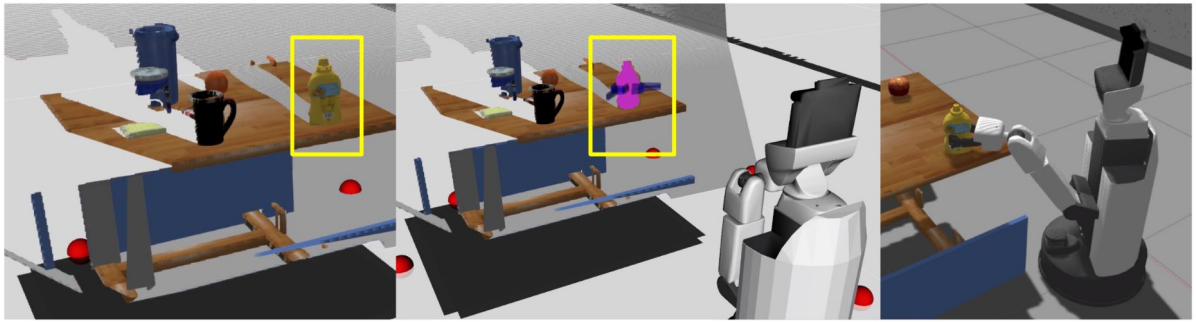


Experiment 3: Overview

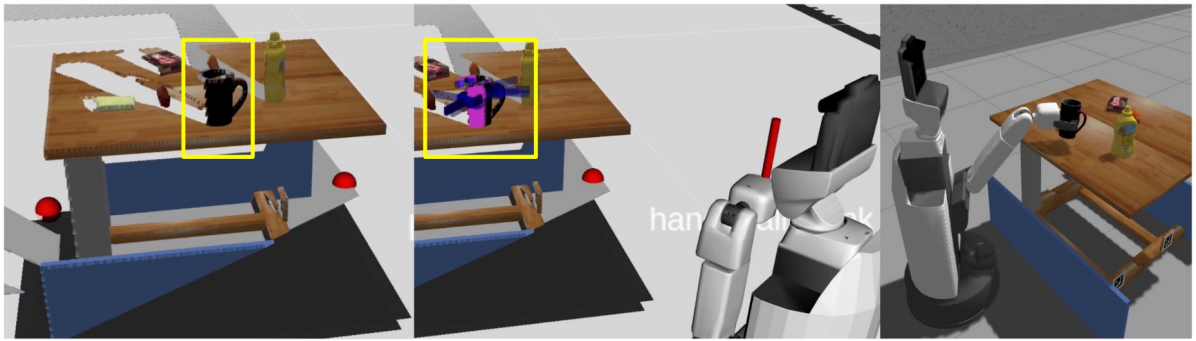
- Experiment 3, we evaluate the CAPerMoMa system's performance using different objects to assess its **adaptability** to new objects.
- Below figure depicts the **two new objects** utilized in this experiment. These objects serve as the new targets to be tested using CAPerMoMa system.

Mustard Bottle **Coffee Cup**

Experiment 3: Results



Grasping sequence for the mustard bottle.



Grasping sequence for the coffee cup.

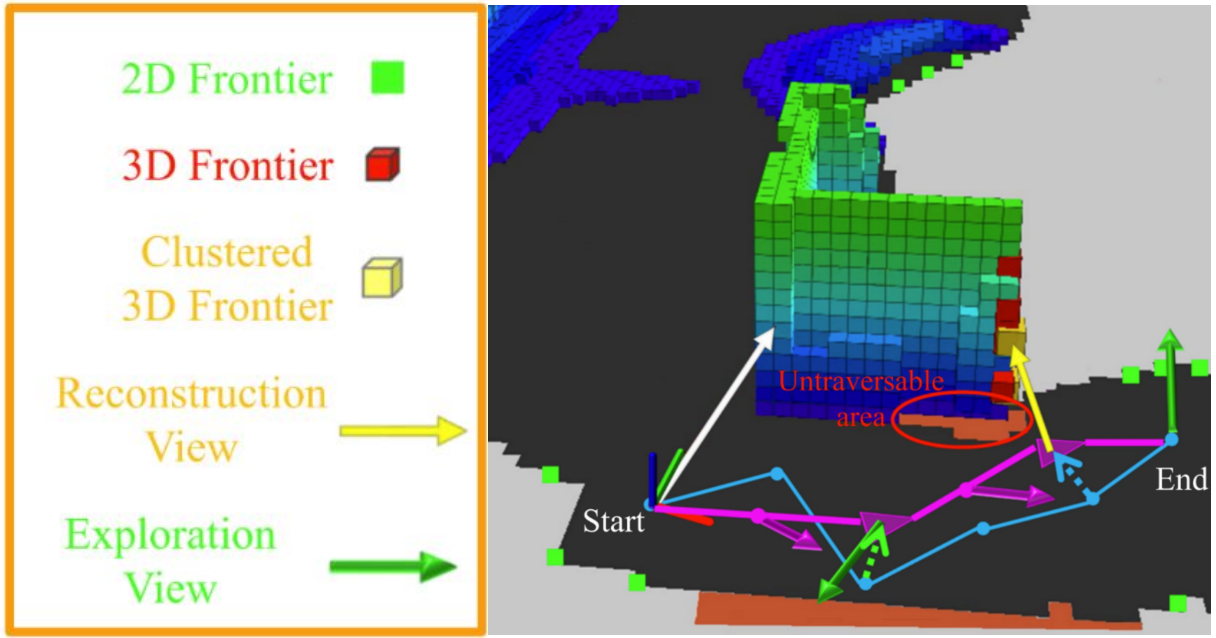
Conclusions

- **Summary**
  - In experiment 1, the system demonstrated the best average **mapping time under 20 seconds** (Scenario 1), achieving a **grasp success rate of 70%** across 20 trials.
  - Experiment 2 highlighted that gaze control **significantly reduced mapping time**. Integrating reachability maps **improved grasp success** by optimizing base placement.
  - Experiment 3 demonstrated that CAPerMoMa **adapts well to new objects** as well.
- **Contributions**
  - CAPerMoMa System: we introduced CAPerMoMa, a system capable of searching, mapping, navigating, and grasping objects in a tabletop mobile manipulation scenario.
    - **Entropy-based head movement, continuous gaze control, reachability with navigability.**
- **Future work**
  - Adopt a receding horizon formulation instead of state machines.
  - Improvements in the exploration strategy. For example, using an object-based information gain approach, where the information gain is computed with the target object in mind rather than the entire scene.
  - Incorporate safety and fault-tolerant mechanisms into the system.

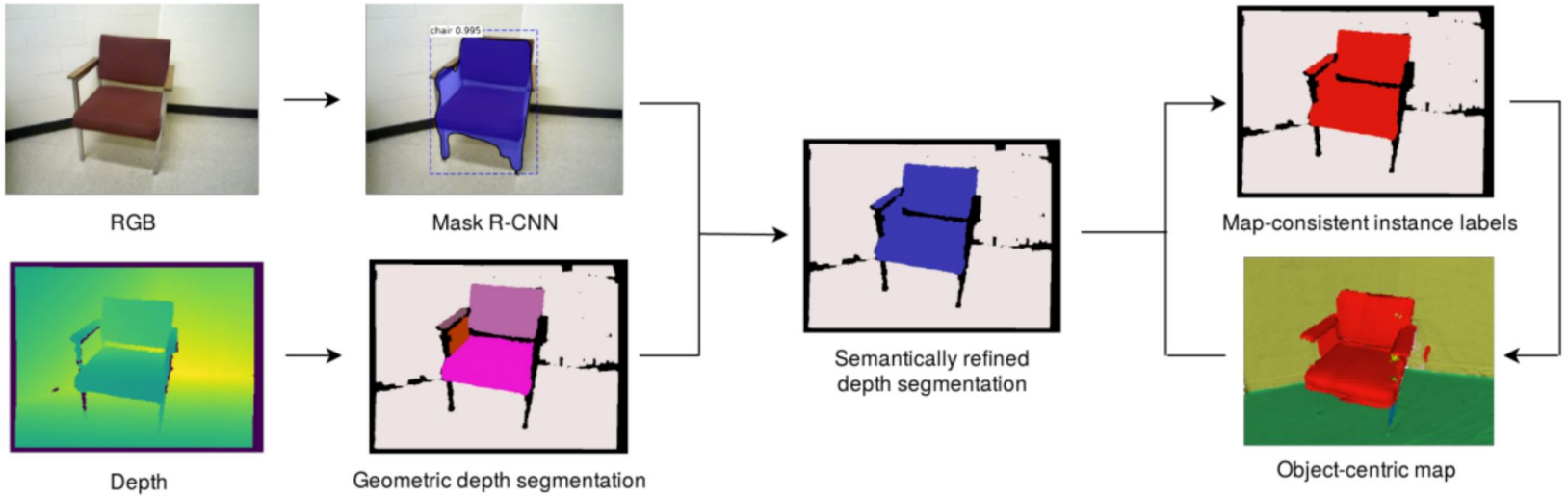
Thank you!

Additional Slides

3D Exploration: Open3DExplorer



3D Mapping: Voxblox++



Grasp Synthesis: Grasp Pose Detector (GPD)

Algorithm 1 Grasp Pose Detection

Input: a viewpoint cloud, $\mathbb{C}$; a region of interest, $\mathcal{R}$; a hand, Θ; a positive integer, N

Output: a set of 6-DOF grasp candidates, $H \subset \mathcal{R}$

1: $\mathbb{C}' = \text{PreprocessCloud}(\mathbb{C})$
2: $\mathcal{R} = \text{GetROI}(\mathbb{C}')$
3: $S = \text{Sample}(\mathbb{C}', \mathcal{R}, \Theta, N)$
4: $I = \text{Encode}(S, \mathbb{C}', \mathcal{R}, \Theta)$
5: $H = \text{Score}(I)$
6: $g = \text{SelectGrasp}(S, H)$