2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

COMPASS: Confined-space Manipulation Planning with Active Sensing Strategy

Qixuan Li[1,*], Chen Le[1,*], Dongyue Huang[3], Jincheng Yu[2,†], Xinlei Chen[1,†]

***Abstract*— Manipulation in confined and cluttered environments remains a significant challenge due to partial observability and complex configuration spaces. Effective manipulation in such environments requires an intelligent exploration strategy to safely understand the scene and search the target. In this paper, we propose COMPASS, a multi-stage exploration and manipulation framework featuring a manipulation-aware sampling-based planner. First, we reduce collision risks with a near-field awareness scan to build a local collision map. Additionally, we employ a multi-objective utility function to find viewpoints that are both informative and conducive to subsequent manipulation. Moreover, we perform a constrained manipulation optimization strategy to generate manipulation poses that respect obstacle constraints. To systematically evaluate method's performance under these difficulties, we propose a benchmark of confined-space exploration and manipulation containing four level challenging scenarios. Compared to exploration methods designed for other robots and only considering information gain, our framework increases manipulation success rate by 24.25% in simulations. Real-world experiments demonstrate our method's capability for active sensing and manipulation in confined environments.**

I. Introduction

Manipulation in confined and cluttered environments remains a significant challenge. In such spaces, manipulator's effectiveness is fundamentally challenged due to perception occlusion and kinematic constraints. On one hand, severe **perception occlusion** to target from surrounding obstacles makes methods assuming full observability [1] [2] not work, necessitating active exploration to incrementally build an understanding of the scene. On the other hand, obstacles compounded by the manipulator's own embodiment, impose tight **kinematic constraints** that drastically limit the robot's reachable workspace, requiring that any planned motion and manipulation must be collision-free with respect to environmental obstacles. This dilemma demands a paradigm shift from simple perception-then-motion pipelines in manipulation tasks to a truly integrated perception-motion-manipulation approach.

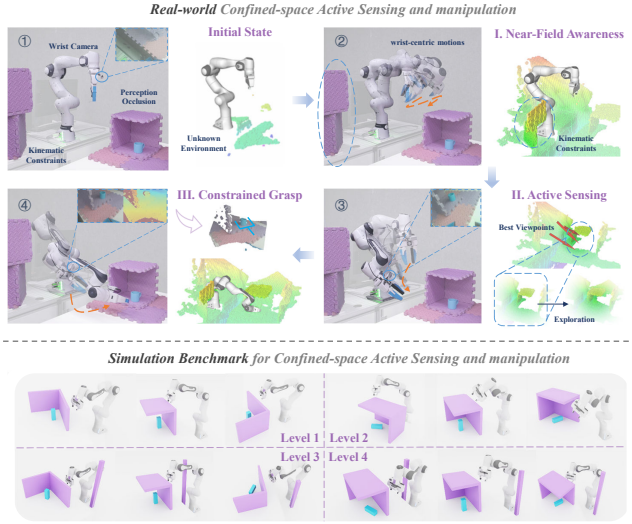

Fig. 1: **Overview of COMPASS**, a framework for active perception and manipulation in confined spaces. (Top) The three-stage pipeline solves a real-world task by sequentially performing: (I) a Near-Field Awareness Scan for safety, (II) MUE-RRT to find the target, and (III) Constrained Grasp optimization. (Bottom) Our principled simulation benchmark systematically evaluates performance across four levels of increasing difficulty of perception occlusion and kinematics constraints.

Current paradigms for robotic manipulation fall short of addressing this challenge. End-to-end learning methods, for instance, often rely on a fixed, third-person camera for a global understanding of the scene [3], [4], a setup that is untenable in confined spaces where such a view is unavailable. They also learn a policy from demonstration data, which typically lacks complex obstacle interactions, making the resulting policies struggle to handle tasks in the presence of obstacles. Additionally, existing planning-based methods typically assume a complete world model and a known target pose [2], [5], ignoring partial observation problem. They focus on finding a path to a pre-defined goal in a totally known environment, rather than on the active information-gathering process required to first explore the space and discover the goal. Both approaches fail to address the critical challenge of how a manipulator incrementally explores and understands a confined space to enable manipulation. To the best of our knowledge, there is no existing work that presents an integrated perception-motion-manipulation approach for confined-space manipulation tasks.

* **Equal Contribution.** † **Corresponding Authors.**
[1] Shenzhen International Graduate School, Tsinghua University, Shenzhen, China. lqx23@mails.tsinghua.edu.cn, le-c25@mails.tsinghua.edu.cn, chen.xinlei@sz.tsinghua.edu.cn.
[2] Department of Electronic Engineering, and the Institute for Embodied Intelligence and Robotics, Tsinghua University, Beijing, China. yu-jc@mail.tsinghua.edu.cn.
[3] The School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. dongyue.huang@ntu.edu.sg.
This research was supported by National Natural Science Foundation of China (62325405), Tsinghua University Initiative Scientific Research Program, Tsinghua-Efort Joint Research Center for EAI Computation and Perception and SunRisingAI Lab, Beijing National Research Center for Information Science, Technology (BNRist), Beijing Innovation Center for Future Chips, and State Key laboratory of Space Network and Communications. This paper was supported by Guangdong Innovative and Entrepreneurial Research Team Program (2021ZT09L197) and Meituan Academy of Robotics Shenzhen.

To tackle these challenges, we propose COMPASS, a multi-stage framework for exploration and constrained manipulation. It comprises three key stages: a Near-Field Awareness Scan, a Manipulation-Utility Exploration RRT (MUE-RRT) planner, and a constrained manipulation pose generation module. The framework begins with the near-field awareness scan, which executes a series of cautious, wrist-centric motions to build a local collision map. This initial step mitigates the risks of moving in an unknown environment and increases the kinematic feasibility of motion. Once this local map is established, the system employs MUE-RRT to iteratively select the next-best-viewpoint to eliminate perception occlusion. The selection is guided by a multi-objective utility function that balances information gain, manipulability, motion cost, and task heuristics. Concurrently, an asynchronous detection process operates on an observation buffer to identify the target throughout the exploration. Once the target is identified, we perform kinematic-constraint aware manipulation optimization strategy to generates grasp poses that respect all obstacle constraints, leveraging the environmental understanding acquired during the exploration stage.

The proposed framework is validated on a principled and progressively challenging benchmark designed to test the task of manipulation in confined space. Simulations and real-world experiments validate our method's capability to perform confined-space manipulation in typical complex scenarios. Our contributions can be summarized as:

1) We propose COMPASS, a framework for exploration and manipulation in confined spaces. It enables a manipulator with partial observation to safely explore unknown environment and perform constrained manipulation.
2) An exploration planner for manipulators, manipulation-utility exploration RRT (MUE-RRT), is proposed to perform manipulation-centric exploration and scene understanding. It's capable of generating a series of smooth and safe exploration trajectory for manipulation tasks.
3) We introduce a principled and progressively challenging benchmark for confined space manipulation tasks. We demonstrate the effectiveness of the proposed method in simulation and real-world experiments.

II. Related Works

A. Embodied Manipulation

Prevailing approaches to embodied manipulation can be categorized into end-to-end learning-based and planning-based methods. Imitation learning methods, such as VLA [3] and Diffusion Policy [1] [4], directly learn manipulation policies from human demonstrations. Recent foundation model-based frameworks have demonstrated impressive embodied spatial reasoning and slow-thinking capabilities [6]. However, these methods are often constrained by demonstration data, which typically lacks complex obstacle interactions. And their reliance on a static, third-person camera perspective makes them ill-suited for confined spaces. Similarly, planning-based manipulation methods [5] [2] tend to focus on generating the end-effector movement and assume a complete world model and a known target pose. They often overlook the kinematic and environmental constraints imposed on the manipulator's trajectory to reach that pose.

B. Space Exploration

Exploration has been extensively studied in the robotics community. A body of research has focused on Unmanned Aerial Vehicles (UAVs) and ground vehicles [7], [8]. In these studies, the methods often follow a Next-Best-View (NBV) paradigm, aiming to reduce environmental uncertainty by maximizing information gain [9], [10]. Typically, autonomous exploration involves three key stages: generation of candidates viewpoints/trajectories, utility evaluation, motion planning and execution [11]. Directly applying these methods to manipulator-based exploration is challenging, as it is unsafe for a manipulator to perform large-scale movements without sufficient prior understanding of its surrounding environment. Existing works for manipulator exploration [12] [13] often focus on geometric coverage or 3D reconstruction but overlook the necessity of finding viewpoints that are conducive to subsequent grasp execution.

C. Grasping in Confined Spaces

Generating a feasible grasp in a confined space is a multi-faceted challenge. While significant progress has been made in grasp pose generation [14], [15], these methods often decouple grasp quality from the robot's kinematic feasibility, leading to unreachable grasps in cluttered scenes. To address this, the field has moved towards joint grasp and motion planning, with works [16] [17] find a grasp and a path simultaneously in confined space. However, this line of work typically assumes a complete world model, sidestepping the challenge of exploration under uncertainty.

III. Overview Of The Framework

A. Problem Formulation

Define $\mathcal{W} \subset \mathbb{R}^3$ as the work space to be explored and manipulated. Let $\mathcal{W}_{free} \subset \mathcal{W}$ be the no-obstacle subspace. Define viewpoint $\mathbf{v} \in \mathrm{SE}(3)$ to describe the pose of the camera onboard the robot, $\mathbf{v} = [\mathbf{t_v}; \mathbf{R_v}]$ where $\mathbf{t_v} \in \mathcal{W}_{free}$ and $\mathbf{R_v} \in \mathrm{SO}(3)$ respectively denote the position and orientation. Define $\mathcal{Q} \subset \mathbb{R}^d$ as the configuration space (C-space), where d is the DOF of manipulator. Define manipulation pose $\mathbf{e} \in \mathrm{SE}(3)$ to describe the pose of the end effector. Our problem can be formulated as follows.

Exploration problem for manipulator in each cycle: Given current joint configuration $\boldsymbol{q}_{current} \in \mathcal{Q}$ and map $\mathcal{M}_{current}$, find optimal exploration path $\mathcal{T}^* = [\mathbf{v}_1, \mathbf{v}_2, ...]$, which corresponding to a feasible trajectory in C-space $\boldsymbol{\tau} = \{\boldsymbol{q}(t) | t = [0, T]\}$. $\mathcal{T}^*$ is optimal means that it is shortest and cover more surface, $\boldsymbol{\tau}$ is feasible means that it is collision-free and avoids singular configurations.

The above problem is solved iteratively to select viewpoints and plan trajectories. When target is found, a manipulation pose $\mathbf{g} = (\mathbf{e}, w)$ is generated, where $\mathbf{e}$ is the pose of gripper and $w \in \mathbb{R}$ denotes the gripper width.

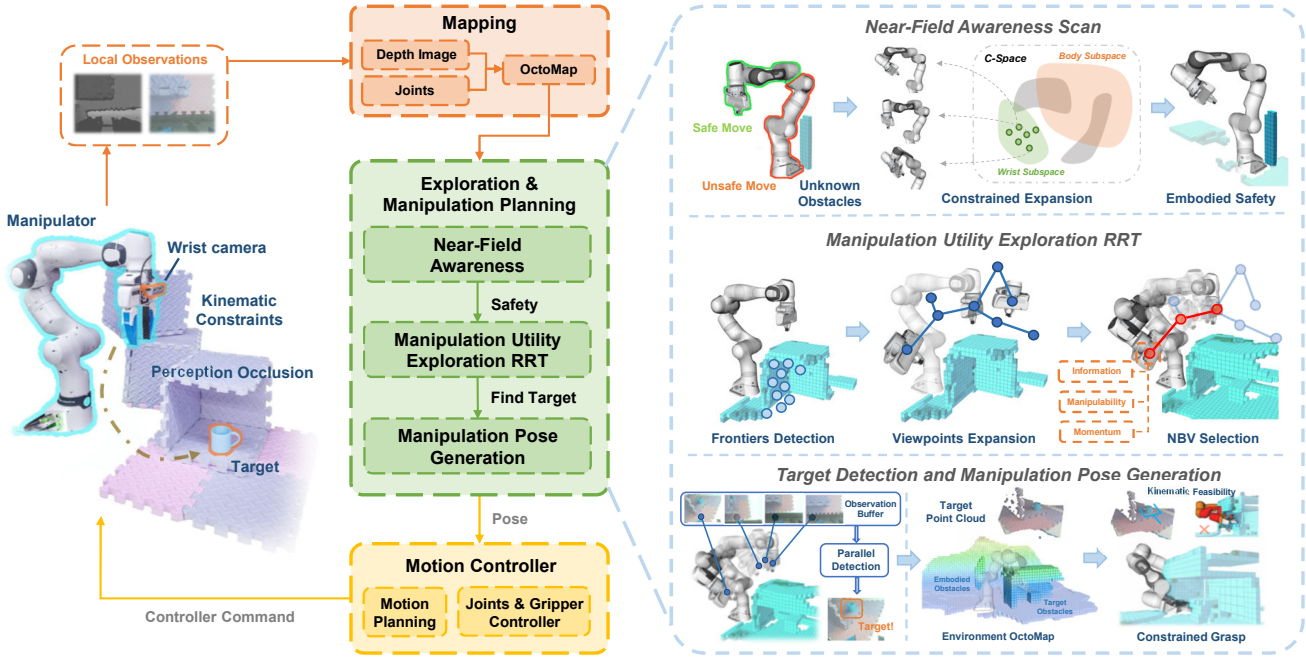

Fig. 2: Overview of system framework of COMPASS. Depth images from the wrist-camera are used to incrementally build an OctoMap. The core multi-stage planner first ensures safety with a Near-Field Awareness Scan, then executes the Manipulation-Utility Exploration RRT to conduct active perception and find the target. Upon target discovery, the Manipulation Pose Generation stage computes a whole-body, collision-free grasp. The Motion Controller plans and executes the trajectory.

B. System Framework

In this paper, we propose a framework for active perception and manipulation in confined spaces, illustrated in Fig. 2. Firstly, a map node incrementally builds an octomap from wrist-camera depth images. Then, an exploration node works in a three-stage manner: (1) Near-Field Awareness Scan serves for embodiment safety; (2) Manipulation-Utility Exploration constructs the environment understanding and detects the target; (3) Manipulation pose generation stage generates the grasp pose in confined space. A motion node communicates with the manipulator and controls it.

IV. Methods

A. Manipulation-Oriented Exploration

Manipulation in confined spaces presents the dual challenges of perception occlusion and kinematic constraints. Obstacles near the manipulator pose significant collision risks, while those surrounding the target cause severe perception occlusions, transforming the task into an active perception and exploration problem. To address this, we propose a multi-stage exploration strategy, the Manipulation-Utility Exploration RRT (MUE-RRT), which operates in two stages: a near-field awareness stage to ensure immediate safety, and a global stage to efficiently detect the target.

1) Near-Field Awareness Scan: To mitigate collision risks in the initially unknown environment, our framework begins with a Near-Field Awareness Scan. The goal is to find a minimal set of joint configurations that maximizes the volumetric coverage of a predefined safety envelope.

We formulate this as a constrained viewpoint set optimization problem. First, we define a target safety volume, Ω_{safe}, as a bounding box around the robot's base. The robot's body configuration, $\boldsymbol{q}_b \in \mathcal{Q}_B$ (i.e., the base and shoulder joints), is held constant at its safe initial posture, $\boldsymbol{q}_{b,init}$. The optimization is then performed over the wrist's configuration space, $\mathcal{Q}_W$ (i.e., the wrist joints). We generate a set of candidate wrist configurations, $\mathcal{Q}_W^{\text{cand}}$, by uniformly sampling this subspace.

The final waypoints, $\{\boldsymbol{q}^*_{w,1}, \ldots, \boldsymbol{q}^*_{w,k}\} \subset \mathcal{Q}_W$, are computed greedily and then executed sequentially. At each selection step i, we find the wrist configuration that covers the largest remaining unknown volume within Ω_{safe}:

$$\boldsymbol{q}^*_{w,i+1} = \underset{\boldsymbol{q}_w \in \mathcal{Q}_W^{\text{cand}}}{\arg\max}\ \text{Volume}(\Omega_{\text{visible}}(\boldsymbol{q}_{b,init}, \boldsymbol{q}_w) \cap \Omega_{\text{safe}}^{\text{unk},i}) \quad (1)$$

where the visible volume $\Omega_{\text{visible}}(\boldsymbol{q}_{b,init}, \boldsymbol{q}_w)$ is explicitly a function of the fixed body configuration $\boldsymbol{q}_{b,init}$ and the variable wrist configuration $\boldsymbol{q}_w$. $\Omega_{\text{safe}}^{\text{unk},i}$ represents the yet-unseen portion of the safety volume at step i.

2) Global Exploration: The system's perception input is a depth image from the wrist-mounted camera. This image is registered to the world frame using forward kinematics and integrated into the global OctoMap [18], denoted as $\mathcal{M} = \{\mathcal{V}_{free}, \mathcal{V}_{occ}, \mathcal{V}_{unk}\}$. To prevent the robot's own geometry from generating spurious obstacles, its links are dynamically filtered from $\mathcal{V}_{occ}$.

We identify frontiers in the OctoMap, which serve as the primary guidance for exploration. A frontier is defined as the boundary between known free space and unknown space [9]. Formally, a voxel is considered a frontier if it resides in free space and is adjacent to at least one unknown voxel. Detected frontier voxels are then clustered into distinct

regions $\{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_N\}$.

At each decision-making step, we dynamically build a Rapidly-exploring Random Tree (RRT) [19]. Each node in the RRT corresponds to a reachable camera viewpoint, and each connecting edge represents a collision-free trajectory in the workspace. This approach significantly reduces the time required for motion planning during tree expansion. To guide the exploration efficiently, the RRT's growth is biased towards the frontier regions $\{\mathcal{F}_i\}$ and any task-relevant priors. Specifically, our sampling strategy directs new samples towards either the frontiers or the task priors with a certain probability, while performing uniform random sampling otherwise.

Once the RRT is constructed, the Next-Best-View (NBV) is selected by evaluating each node $\boldsymbol{x} = (\mathbf{v}, \boldsymbol{q})$ in the tree using a multi-objective utility function, $\mathcal{U}(\boldsymbol{x})$. This function is designed to holistically assess a candidate viewpoint by normalizing its potential rewards against its associated motion cost, thereby promoting an efficient exploration strategy that maximizes the value gained per unit of effort. The optimal next node, $\boldsymbol{x}^*$, is the one that maximizes this utility:

$$\boldsymbol{x}^* = \underset{\boldsymbol{x} \in \text{RRT}}{\arg\max}\ \mathcal{U}(\boldsymbol{x}) \tag{2}$$

The utility function $\mathcal{U}(\boldsymbol{x})$ is formulated as a sum of cost-normalized reward terms:

$$\mathcal{U}(\boldsymbol{x}) = \boldsymbol{w}^T[\mathcal{G}(\boldsymbol{x})\ \mathcal{D}(\boldsymbol{x})\ \mathcal{M}(\boldsymbol{x})\ \mathcal{H}(\boldsymbol{x})]^T / \mathcal{C}(\boldsymbol{x}) \tag{3}$$

where $\boldsymbol{w} = [w_g\ w_d\ w_m\ w_h]^T$ represents the respective weighting factors for each reward component. These components are defined as follows:

1) Information Gain $\mathcal{G}(\boldsymbol{x})$: The information gain measures the volume of new space observable from viewpoint $\mathbf{v}$. It is defined as the number of previously unknown voxels within the viewpoint's field of view (FOV). Formally, $\mathcal{G}(\boldsymbol{x}) = \int_{v \in \mathcal{V}_{\text{vis}} \cap \mathcal{V}_{\text{unk}}} dv$, where $\mathcal{V}_{\text{vis}}$ is the set of voxels visible from $\mathbf{v}$. This value is computed by raycasting from the virtual camera pose into the current OctoMap [20].

2) Exploration Momentum $\mathcal{D}(\boldsymbol{x})$: To encourage continuous and smooth exploration paths, this term rewards viewpoints that maintain the current direction of exploration. It is calculated as $\mathcal{D}(\boldsymbol{x}) = (\mathbf{t}_{\text{prev}} - \mathbf{t}_{\text{root}}) \cdot (\mathbf{t}_{\boldsymbol{x}} - \mathbf{t}_{\text{root}})$, where $\mathbf{t}_{\boldsymbol{x}}$ is the translational component of the viewpoint, $\mathbf{t}_{\text{root}}$ is that of the RRT's root node, and $\mathbf{t}_{\text{prev}}$ corresponds to the previously selected viewpoint.

3) Manipulability $\mathcal{M}(\boldsymbol{x})$: This term promotes configurations that are far from singularities, ensuring the robot remains poised for future manipulation tasks. It is quantified by the Yoshikawa manipulability index [21], $\mathcal{M}(\boldsymbol{q}) = \sqrt{\det(\boldsymbol{J}(\boldsymbol{q})\boldsymbol{J}(\boldsymbol{q})^T)}$, where $\boldsymbol{J}(\boldsymbol{q})$ is the manipulator's Jacobian matrix.

4) Motion Cost $\mathcal{C}(\boldsymbol{x})$: This represents the total effort required to reach configuration $\boldsymbol{q}$ from the current root state $\boldsymbol{q}_{\text{current}}$. It is calculated as the path length in the robot's joint space, $\mathcal{C}(\boldsymbol{x}) = \int_0^1 ||\dot{\boldsymbol{\tau}}(s)||\, ds$, where $\boldsymbol{\tau}(s)$ is the joint-space trajectory from $\boldsymbol{q}_{\text{current}}$ to $\boldsymbol{q}$. Joint-space distance provides a more accurate measure of robot effort than its Cartesian counterpart.

Algorithm 1: Exploration and Manipulation

Input: Initial robot configuration $\boldsymbol{q}_{init}$
Output: Map $\mathcal{M}_T$ and Manipulation Result $\boldsymbol{\Gamma}$

$\boldsymbol{q}_{current} \leftarrow \boldsymbol{q}_{init}$
$\{\boldsymbol{q}_w^*\}_{i=1:m} \leftarrow$ CalculateWristWaypoints($\boldsymbol{q}_{init}$)
$\mathcal{M}_0 \leftarrow$ ConductNearFieldAwareness($\{\boldsymbol{q}_w^*\}_{i=1:m}$)
$found \leftarrow$ **false**
while $\neg\ found$ **do**
  $I_t \leftarrow$ SenseFromWristCamera($\boldsymbol{q}_{current}$)
  $\mathcal{M}_t \leftarrow$ UpdateMap($\mathcal{M}_{t-1}, I_t, \boldsymbol{q}_{current}$)
  $\mathcal{M}_t \leftarrow$ FilterRobotBody($\mathcal{M}_t, \boldsymbol{q}_{current}$)
  $\mathcal{F}_{list} \leftarrow$ DetectAndClusterFrontiers($\mathcal{M}_t$)
  $(\mathcal{V}, \mathcal{E}) \leftarrow$ ExpansionRRT($\mathcal{M}_t, \mathcal{F}_{list}$)
  $\mathcal{U} \leftarrow$ ComputeUtility($(\mathcal{V}, \mathcal{E}), \mathcal{M}_t$)
  $\boldsymbol{q}_{next} \leftarrow$ SelectNBV($\mathcal{U}$)
  $\boldsymbol{\tau} \leftarrow$ PlanningPath($\boldsymbol{q}_{current}, \boldsymbol{q}_{next}$)
  $\boldsymbol{q}_{current} \leftarrow$ ExecuteTrajectory($\boldsymbol{\tau}$)
  $\mathcal{B}_t \leftarrow \{(I_{t-k}, \mathbf{v}_{t-k}), \ldots, (I_{t-1}, \mathbf{v}_{t-1}), (I_t, \mathbf{v}_t)\}$
  $found, \mathbf{v}^* \leftarrow$ TargetDetection($\mathcal{B}_t$)
$\mathcal{G} \leftarrow$ ManipulationPoseGeneration($I_t, \mathbf{v}^*, \mathcal{M}_t$)
$\boldsymbol{\Gamma} \leftarrow$ ConductManipulation($\mathcal{G}, \mathcal{M}_t$)
return $\mathcal{M}_T, \boldsymbol{\Gamma}$

5) Heuristic Guidance: We integrate task-driven guidance into the core exploration mechanism. This guidance is represented as a set of heuristic 3D points of interest, $\mathcal{P}_h$. This information is incorporated into our planner at two synergistic levels. First, at the sampling stage, we bias the RRT tree expansion process. Second, at the evaluation stage, we enhance the utility function with a heuristic gain term, $\mathcal{H}(\boldsymbol{x})$, defined as: $\mathcal{H}(\boldsymbol{x}) = \max_{\boldsymbol{p}_j \in \mathcal{P}_h} (\boldsymbol{\phi}_{\text{v}} \cdot (\boldsymbol{p}_j - \mathbf{t}_{\text{v}}) / ||\boldsymbol{p}_j - \mathbf{t}_{\text{v}}||)$, where $\boldsymbol{\phi}_{\text{v}}$ is the viewing axis of the viewpoint v, $\boldsymbol{p}_j$ is a heuristic point of interest, and $\mathbf{t}_{\text{v}}$ is the translational position of the viewpoint.

B. Target Detection and Manipulation Pose Generation

To facilitate target detection during manipulator motion, our framework employs an asynchronous detection process. This module addresses the latency mismatch between the robot's continuous motion and the computationally intensive inference of detection models, such as the YOLO [22] model used in this paper. The detection process maintains a spatio-temporal observation buffer, $\mathcal{B}_t$, which stores a history of synchronized image-pose pairs over a sliding time window:

$$\mathcal{B}_t = \{(I_{t-k}, \mathbf{v}_{t-k}), \ldots, (I_{t-1}, \mathbf{v}_{t-1}), (I_t, \mathbf{v}_t)\} \tag{4}$$

where I_i is the RGB image from the wrist camera and $\mathbf{v}_i$ is the corresponding camera pose at time i. The parameter k denotes the buffer size. By processing this buffer in batches, the detection process can perform robust, temporally-consistent inference, ensuring that transient views of the target are not missed due to processing delays. To ensure high-fidelity grasp perception, we select the optimal observation $(I^*, \mathbf{v}^*)$

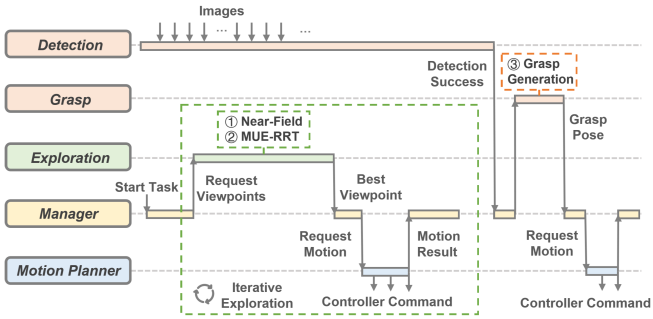

Fig. 3: The runtime logic of a successful exploration and manipulation in confined space.

from the subset of the buffer containing successful detections, $\mathcal{B}_{\text{det}} \subseteq \mathcal{B}_t$. This selection is based on a perceptual quality score, S_{obs}, defined as:

$$S_{\text{obs}}(I_i) = -\|\boldsymbol{p}_{\text{bbox}}(I_i) - \boldsymbol{p}_{\text{center}}\|_2 \tag{5}$$

where $\boldsymbol{p}_{\text{bbox}}(I_i)$ is the center of the target's 2D bounding box in image I_i, and $\boldsymbol{p}_{\text{center}}$ is the image center. This score prioritizes the viewpoint where the target is most centrally located, thereby maximizing view quality for the subsequent grasp pose generation.

The grasp pose is generated at the optimal viewpoint $\mathbf{v}^*$. In this paper, we employ the GraspNet [14] detection network, which processes the depth data from image I^* to generate a set of candidate grasps, $\mathcal{G}_{\text{cand}} = \{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_m\}$. Each candidate grasp $\mathbf{g}_i$ is defined as a tuple $\mathbf{g}_i = (\mathbf{e}_i, w_i, s_i)$, where $\mathbf{e}_i \in \text{SE}(3)$ is the end-effector pose, w_i is the gripper width, and $s_i \in (0, 1]$ is the grasp quality score. This score s_i reflects only the geometric quality of the grasp (e.g., force closure as in [14]) and remains naive to a critical aspect: the kinematic feasibility of the manipulator.

To address this limitation, we perform a pose refinement for each grasp pose candidate. This process can be viewed as a constrained optimization problem:

$$\begin{aligned} \max_{\mathbf{g}_i \in \mathcal{G}_{\text{cand}}} \quad & s_i \\ \text{s.t.} \quad & \text{IK}(\mathbf{e}_i) \in \mathcal{Q}_{free} \\ & \angle(\boldsymbol{z}_{\mathbf{g}_i}, \boldsymbol{n}) \leq \delta \\ & \text{FK}(\boldsymbol{\tau}(\boldsymbol{q}_{cur}, \boldsymbol{q}_{grasp})) \subset \mathcal{M}_{free} \end{aligned} \tag{6}$$

where s_i defines the analytic force-closure quality evaluated under varying friction coefficients [14]. The term $\text{IK}(\mathbf{e}_i)$ denotes the inverse kinematics solution for the grasp pose. The second constraint enforces a context-aware approach direction for the gripper, where $\boldsymbol{z}_{\boldsymbol{g}_i}$ represents the approach vector corresponding to the i-th grasp candidate $\boldsymbol{g}_i$ (typically the z-axis of end effector frame), and the desired approach vector $\boldsymbol{n}$ is determined by analyzing the object's orientation. Finally, the entire workspace trajectory, obtained by applying forward kinematics $\text{FK}(\cdot)$ to the planned joint-space path $\boldsymbol{\tau}(\boldsymbol{q}_{cur}, \boldsymbol{q}_{grasp})$, is required to be within the free space of the current map, $\mathcal{M}_{free}$.

C. Runtime Logic

The overall system is implemented as a set of interacting, asynchronous nodes for exploration, motion planning, target detection, and grasp generation, all coordinated by a manager node using a Finite State Machine (FSM). The runtime logic is illustrated in Fig. 3. Upon receiving a *start task* signal, the manager initiates an iterative exploration loop by requesting the next-best-viewpoint from the exploration node. It then commands the motion planner to generate and execute a collision-free trajectory to that viewpoint. This perception-action loop continues, with the detection node asynchronously processing incoming images, until the target is found. Upon successful detection, the manager transitions to the final phase: it requests a refined grasp pose from the grasp generation node and commands the motion planner to execute the grasp, completing the task.

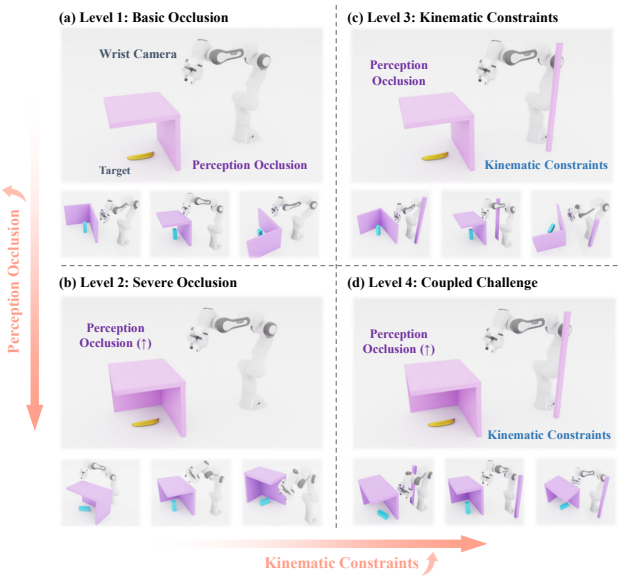

Fig. 4: The four challenging scenarios of our proposed benchmark, designed to systematically increase difficulty by combining Perception Occlusion (Levels 1 & 2) and Kinematic Constraints (Levels 3 & 4). Level 4 represents the fully coupled challenge.

V. Benchmark for Confined-Space Manipulation

We design a comprehensive benchmark consisting of procedurally generated, confined-space scenes and performance metrics. This benchmark is structured to evaluate an algorithm's ability to handle the two fundamental challenges: perception occlusion and kinematic constraints.

A. Benchmark Scenarios

We design a benchmark of four scene categories with progressively increasing difficulty. For each category, we procedurally generate 20 unique environments in the Isaac Sim simulator [23], totaling 80 challenging test cases (Fig. 4). These categories are constructed to systematically isolate and combine the core challenges, as summarized in Table I.

Level 1 (Basic Occlusion): These scenes feature moderate clutter where the target, though occluded by simple obstacles, can be found via monotonic exploration paths.

Level 2 (Severe Occlusion): Target accessibility is reduced compared to *Level 1*. The object is hidden, requiring non-monotonic paths (e.g., looking behind an obstacle and then

backing out). This category is designed to test an algorithm's ability to find the target through global exploration.

Level 3 (High Kinematic Constraint): These scenes feature the same level of occlusion as *Level 1*, but with additional obstacles introduced near the manipulator's base. This setup specifically challenges the planner's awareness of its full-body kinematics.

Level 4 (Coupled Challenge): This level combines the severe occlusion of *Level 2* with the high kinematic constraints of *Level 3*. Success in this level requires an algorithm to plan a global, potentially non-monotonic exploration trajectory to handle severe occlusion, and subsequently conduct manipulation under high kinematic constraints.

TABLE I: Simulation environment characteristics

Level	Perception Occlusion	Kinematic Constraints
Level 1	Moderate (+)	Low (+)
Level 2	High (++)	Low (+)
Level 3	Moderate (+)	High (++)
Level 4	High (++)	High (++)

B. Performance Metrics

We evaluate performance across three dimensions in solving the coupled challenge in exploration and manipulation.

- Exploration Efficiency:
  - Explored Volume over Time: The proportion of the target's bounding box explored versus time. Steeper slopes indicate faster exploration, and higher final values reflect greater coverage.
  - Time to Find Target (TFT): The time elapsed until the target is successfully located. This metric reflects exploration efficiency; a lower value is better.
  - Path Length to Find Target (PFT): The distance traveled by the end-effector before the target is found; a lower value is better.
- Motion Performance:
  - Motion Planning Success Rate (MPSR): The fraction of successful motion planning attempts throughout a task. This metric indicates the kinematic feasibility of the viewpoints; a higher value is better.
  - Average Manipulability (AM): The average manipulability index [21] during the exploration process. This metric quantifies the manipulator's distance from singular configurations; a higher value is better.
- Manipulation Quality:
  - Grasp Pose Quality (GPQ): A score that reflects the quality of a generated grasp pose, calculated as follows: $\mathcal{S} = s_{\mathbf{g}} \cdot \mathbb{I}(\mathrm{IK}(\mathbf{e_g}) \in \mathcal{Q}_{free}) \cdot \cos(\angle(\boldsymbol{z_g}, \boldsymbol{n}))$, where $\mathbb{I}(\cdot)$ is an indicator function that equals 1 if its argument is true, and 0 otherwise. The remaining symbols are defined as in Eq. 6. This metric holistically quantifies the grasp's geometric quality and kinematic reachability; a higher value is better.
  - Overall Success Rate (SR): The success rate of the entire pipeline, from initial exploration to a successful final grasp. A higher value is better.

VI. Experiments

A. Simulation Experiments

We evaluate our method through simulation experiments in the proposed benchmark environment. These experiments are performed in an Isaac Sim [23] simulation environment, featuring a Franka Panda manipulator equipped with a wrist-mounted depth camera. Motion planning is executed using MoveIt [24], and raw grasp poses are generated by GraspNet [14]. All simulations are performed in the Robot Operating System (ROS). We compare our proposed method against the following baselines:

- Fixed-Views (FV): A baseline that employs a pre-programmed, open-loop sequence of viewpoints for perception. The resulting trajectory is non-adaptive and does not react to sensory input.
- Information-Gain Exploration (IG): Based on the method by Isler et al. [25], this approach greedily selects the single viewpoint that maximizes information gain.
- Geometric Sampling-based Exploration (GSE): We adapt a sampling-based exploration method from the mobile robotics domain [9], [12]. This approach utilizes an RRT for spatial expansion but selects the next-best-view based solely on geometric information gain.

The performance of all methods is evaluated across the 80 generated scenes, with 10 trials conducted for each scene. A trial concludes when the exploration is reported as complete, the manipulation process finishes, or a predefined time limit is reached. All methods are tested on a computer equipped with an Intel i7-14650HX CPU and a RTX 4060 GPU.

The first two rows of Table II present the quantitative results for exploration efficiency across all scenarios in our benchmark. Our method consistently achieves higher efficiency in detecting the target, both in terms of time and path length. It is worth noting that since unsuccessful trials are assigned the maximum time limit (i.e., 200s), a lower average time is also indicative of a higher success rate.

Fig. 5 illustrates the average Explored Volume over Time for each method across the four categories depicted in Fig. 4. As shown, our approach consistently achieves the highest exploration efficiency. In contrast, FV only attains a fixed and limited coverage, while IG is prone to becoming trapped in local optima and subsequently fails to progress. By integrating RRT with goal-directed guidance, our method enables the most efficient exploration.

The third and fourth rows of Table II present the quantitative results for motion performance. Compared to the baselines, our method exhibits a higher motion planning success rate during exploration and consistently maintains a higher Average Manipulability. Furthermore, as the environmental kinematic constraints become more complex, the baselines increasingly struggle to find valid motions. This is a direct consequence of their viewpoint selection strategies: the FV is non-adaptive to obstacle configurations, while the Information-Gain based methods greedily pursues perceptual rewards without considering kinematic feasibility.

TABLE II: Quantitative Performance Comparison Of All Methods In Difficulty Levels.

Metric	Level1				Level2				Level3				Level4			
	FV	IG	GSE	Proposed	FV	IG	GSE	Proposed	FV	IG	GSE	Proposed	FV	IG	GSE	Proposed
TFT (s) ↓	**83.5**	93.9	113.7	85.7	166.9	164.3	161.5	**133.9**	106.4	153.6	111.0	**95.0**	173.1	173.2	158.2	**131.4**
PFT (m) ↓	4.1	4.3	4.7	**2.9**	8.4	8.2	7.1	**5.4**	5.3	7.7	4.4	**3.6**	8.7	8.5	7.0	**5.5**
MPSR (%) ↑	66.5	39.5	59.3	**71.3**	62.5	36.4	62.3	**64.1**	**61.6**	34.7	60.6	60.3	59.7	41.8	59.6	**64.9**
AM ↑	0.042	0.067	0.064	**0.070**	0.033	0.062	0.065	**0.069**	0.035	0.061	0.064	**0.067**	0.029	0.058	0.059	**0.066**
GPQ ↑	-	0.578	0.670	**0.676**	-	0.544	**0.603**	0.501	-	0.594	0.617	**0.634**	-	0.609	0.584	**0.613**
SR(%) ↑	-	40.0	47.5	**80.0**	-	27.5	42.5	**67.0**	-	30.0	47.5	**70.0**	-	25.0	45.0	**62.5**

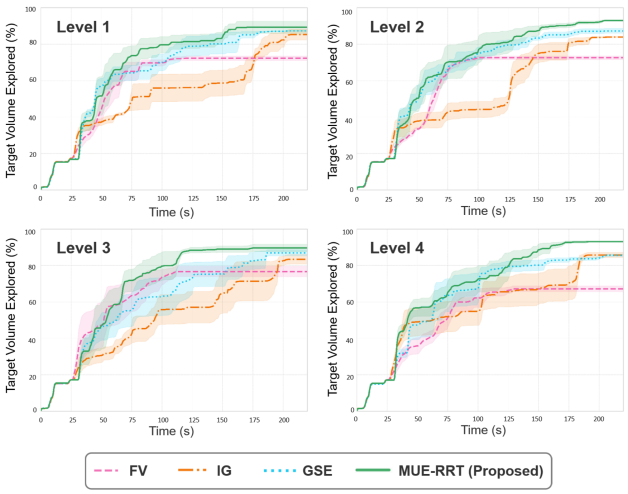

Fig. 5: Target volume exploration performance over time across the four benchmark difficulty levels. Solid lines denote the mean value, with shaded areas representing the standard deviation.

The last two rows of Table II present grasp performance. The improvements in exploration and motion directly translate to manipulation quality, with our method achieving an average overall success rate improvement of 39.25% over IG and 24.25% over GSE. Since we employed the same constrained pose generation strategy for all methods, the gap in grasp pose quality between the different approaches is relatively small. The primary difference in overall performance stems from the exploration phase.

B. Real-World Experiments

To validate the feasibility and efficiency of our proposed framework, we conduct experiments on a real robotic platform in a confined indoor environment. The platform consists of a 7-DOF Franka Emika FR3 arm equipped with a wrist-mounted RealSense D435i depth camera. We physically recreate a subset of our benchmark scenarios from *Level 4*, which feature both severe perception occlusion and tight kinematic constraints. The robot's task is to locate and retrieve a target object from within a cluttered container, mirroring the objectives of our simulation benchmark. In such environments, planning reactively based on local observations is insufficient and will lead to collisions.

The experimental process is illustrated in Fig. 6. Initially, the robot has no prior knowledge of the environment (Fig. 6 I). The robot then conducts the active exploration phase, autonomously generating a sequence of viewpoints to actively build a map of the confined space and search for the target (Fig. 6 II). Once the target is detected, the system transitions

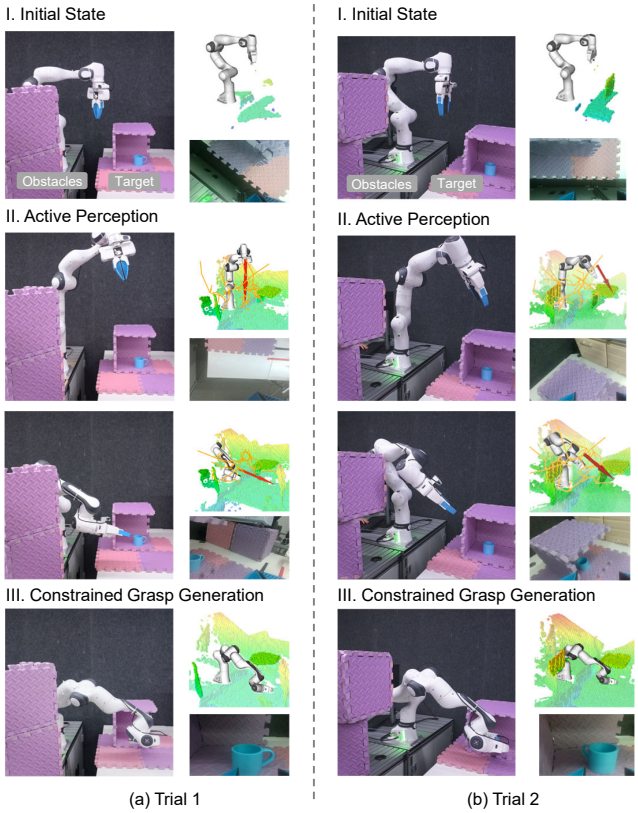

Fig. 6: Snapshots of real-world experiments.

to the final stage, computing and executing a kinematically feasible, collision-free grasp (Fig. 6 III).

Across 10 trials for each scenario, our method achieves an average Time-to-Find-Target of 159.75s and an overall success rate of 80%. This high success rate in such challenging, real-world conditions demonstrates the robustness of our integrated perception and planning pipeline. It validates that our exploration strategy can effectively perform active perception to resolve environmental uncertainty, while the subsequent grasp generation module successfully computes and executes constraint-aware grasps. The primary sources of the 20% failures were twofold: (1) motion planning failures caused by sensor noise, and (2) convergence to local optima in the exploration policy.

C. Ablation Studies

1) Near-Field Awareness for Motion Safety: To validate the effectiveness of the Near-Field Awareness stage in ensuring motion safety and kinematic feasibility, we compare our full method against a variant in which this stage was disabled. The scenes are chosen from *Level 4*. The results, presented

in Table III, show that including the near-field awareness scan significantly increased the motion planning success rate (MPSR) from 54.7% to 61.2% and the target detection success rate (DSR) from 68.3% to 94.1%, confirming its critical role in ensuring a safe exploration process.

TABLE III: Ablation study on the Near-Field Awareness

Method Variant	MPSR (%) ↑	DSR (%) ↑
COMPASS (Full Method)	**61.2**	**94.1**
w/o Near-Field Awareness	54.7	68.3

2) Constrained Grasp Pose Optimization for Grasping: To demonstrate the efficacy of our grasp constraints, we compare our full method against an unconstrained baseline in *Level 4* scenarios. As shown in Table IV, omitting these constraints leads to a sharp increase in motion collisions (MC), grasp failures (GF), and object drops (OD), confirming their necessity for safe and successful execution.

TABLE IV: Ablation study on the Constrained Grasp Optimization

Method Variant	Grasp SR (%) ↑	MC ↓	GF ↓	OD ↓
Full Method	**70.0**	**0/10**	**1/10**	**0/10**
w/o Grasp Constraints	20.0	4/10	2/10	1/10

VII. Conclusion

In this paper, we have presented COMPASS, a framework for active perception and manipulation in confined spaces. We have proposed the MUE-RRT that ensures safety and motion efficiency. Additionally, we have designed the principled and progressively challenging benchmark for confined-space manipulation tasks. We have conducted extensive simulations and real-world experiments to demonstrate the effectiveness of our method. Compared to exploration methods designed for other robots and only considering information gain, our framework increases manipulation success rate by 24.25% and reduces the average time to find target by 24.6s in simulations. In the future, we plan to extend this framework to handle dynamic environments via event-depth fusion [26] and enable open-vocabulary semantic manipulation [27].

References

[1] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, p. 02783649241273668, 2023.

[2] M. Pan, J. Zhang, T. Wu, Y. Zhao, W. Gao, and H. Dong, "Omnimanip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 17 359–17 369.

[3] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[4] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," *arXiv preprint arXiv:2403.03954*, 2024.

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

[6] B. Zhao, Z. Wang, J. Fang, C. Gao, F. Man, J. Cui, X. Wang, X. Chen, Y. Li, and W. Zhu, "Embodied-r: Collaborative framework for activating embodied spatial reasoning in foundation models via reinforcement learning," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 11 071–11 080.

[7] M. Dharmadhikari, T. Dang, L. Solanka, J. Loje, H. Nguyen, N. Khedekar, and K. Alexis, "Motion primitives-based path planning for fast and agile exploration using aerial robots," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 179–185.

[8] D. Shah, B. Eysenbach, G. Kahn, N. Rhinehart, and S. Levine, "Rapid exploration for open-world navigation with latent goal models," *arXiv preprint arXiv:2104.05859*, 2021.

[9] H. Zhu, C. Cao, Y. Xia, S. Scherer, J. Zhang, and W. Wang, "Dsvp: Dual-stage viewpoint planner for rapid exploration by dynamic expansion," in *2021 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2021, pp. 7623–7630.

[10] C. Cao, H. Zhu, H. Choset, and J. Zhang, "Tare: A hierarchical framework for efficiently exploring complex 3d environments." in *Robotics: Science and Systems*, vol. 5, 2021, p. 2.

[11] 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.

[12] M. Naazare, F. G. Rosas, and D. Schulz, "Online next-best-view planner for 3d-exploration and inspection with a mobile manipulator robot," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3779–3786, 2022.

[13] H. Ren and A. H. Qureshi, "Robot active neural sensing and planning in unknown cluttered environments," *IEEE Transactions on Robotics*, vol. 39, no. 4, pp. 2738–2750, 2023.

[14] 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*, 2020, pp. 11 444–11 453.

[15] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3929–3945, 2023.

[16] M. Kang, H. Kee, J. Kim, and S. Oh, "Grasp planning for occluded objects in a confined space with lateral view using monte carlo tree search," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 10 921–10 926.

[17] S. Elliott, M. Valente, and M. Cakmak, "Making objects graspable in confined environments through push and pull manipulation with a tool," in *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 4851–4858.

[18] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "Octomap: An efficient probabilistic 3d mapping framework based on octrees," *Autonomous robots*, vol. 34, no. 3, pp. 189–206, 2013.

[19] A. Orthey, C. Chamzas, and L. E. Kavraki, "Sampling-based motion planning: A comparative review," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 7, 2023.

[20] T. Dang, M. Tranzatto, S. Khattak, F. Mascarich, K. Alexis, and M. Hutter, "Graph-based subterranean exploration path planning using aerial and legged robots," *Journal of Field Robotics*, vol. 37, no. 8, pp. 1363–1388, 2020.

[21] T. Yoshikawa, "Manipulability of robotic mechanisms," *The international journal of Robotics Research*, vol. 4, no. 2, pp. 3–9, 1985.

[22] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "Yolo-world: Real-time open-vocabulary object detection," *arXiv preprint arXiv:2401.17270*, 2024.

[23] NVIDIA Corporation, "NVIDIA Isaac Sim," https://developer.nvidia.com/isaac/sim.

[24] S. Chitta, I. Sucan, and S. Cousins, "Moveit![ros topics]," *IEEE robotics & automation magazine*, vol. 19, no. 1, pp. 18–19, 2012.

[25] S. Isler, R. Sabzevari, J. Delmerico, and D. Scaramuzza, "An information gain formulation for active volumetric 3d reconstruction," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 3477–3484.

[26] X. Luo, H. Wang, C. Ruan, C. Liang, J. Xu, and X. Chen, "Eventtracker: 3d localization and tracking of high-speed object with event and depth fusion," in *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, 2024, pp. 1974–1979.

[27] Q. Gu, Z. Ye, J. Yu, J. Tang, T. Yi, Y. Dong, J. Wang, J. Cui, X. Chen, and Y. Wang, "Mr-cographs: Communication-efficient multi-robot open-vocabulary mapping system via 3d scene graphs," *IEEE Robotics and Automation Letters*, 2025.