M^{3}Bench: Benchmarking Whole-body Motion Generation for Mobile Manipulation in 3D Scenes

Zeyu Zhang[2*], Sixu Yan[1,2*], Muzhi Han[3], Zaijin Wang[2], Xinggang Wang[1], Song-Chun Zhu[2,4,5], Hangxin Liu[2,†]

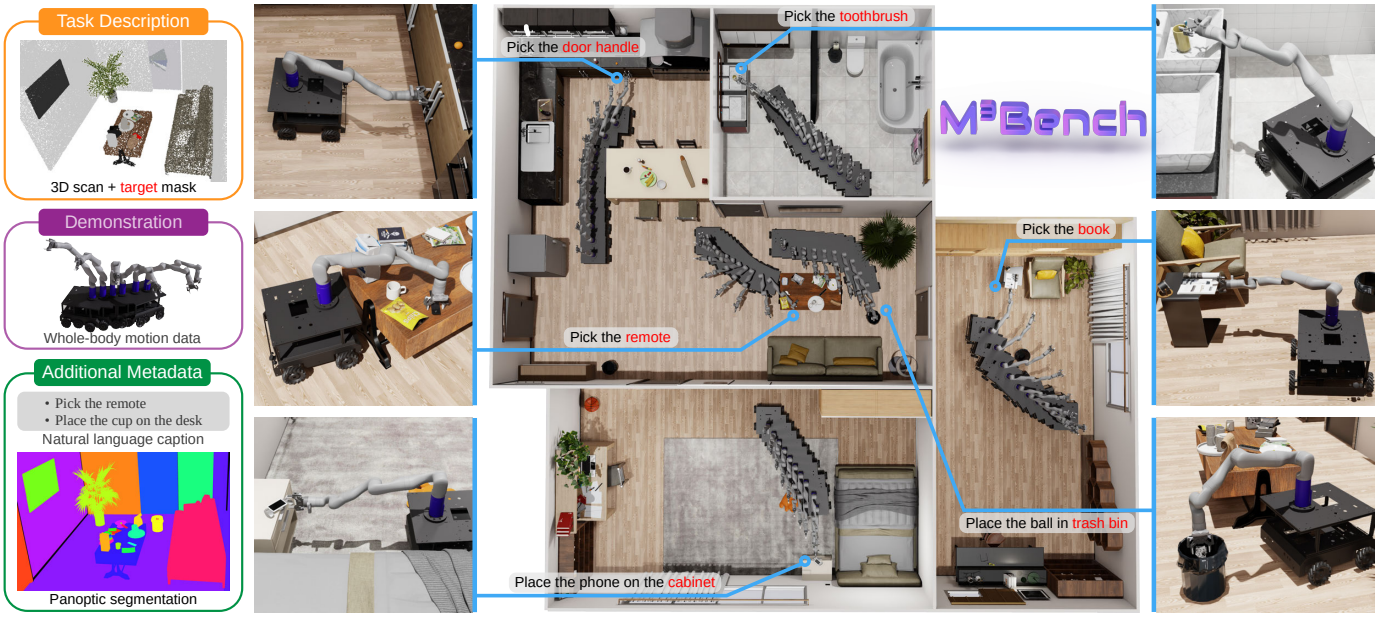


Fig. 1: **The M^{3}Bench benchmark** challenges mobile manipulators to generate whole-body motion trajectories for object manipulation in 3D scenes. Given a 3D scan, a target segmentation mask, and a task description, the robot must understand its embodiment, environment, and task objectives to produce coordinated motions for picking or placing objects.

***Abstract*—We propose M^{3}Bench, a new benchmark for whole-body motion generation in mobile manipulation tasks. Given a 3D scene context, M^{3}Bench requires an embodied agent to reason about its configuration, environmental constraints, and task objectives to generate coordinated whole-body motion trajectories for object rearrangement. M^{3}Bench features 30,000 object rearrangement tasks across 119 diverse scenes, providing expert demonstrations generated by our newly developed M^{3}BenchMaker, an automatic data generation tool that produces whole-body motion trajectories from high-level task instructions using only basic scene and robot information. Our benchmark includes various task splits to evaluate generalization across different dimensions and leverages realistic physics simulation for trajectory assessment. Extensive evaluation analysis reveals that state-of-the-art models struggle with coordinating base-arm motion while adhering to environmental and task-specific constraints, underscoring the need for new models to bridge this gap. By releasing M^{3}Bench and M^{3}BenchMaker at https://zeyuzhang.com/papers/m3bench, we aim to advance robotics research toward more adaptive and capable mobile manipulation in diverse, real-world environments.**

* Z. Zhang and S. Yan contributed equally to this work. Emails: zhangzeyu@bigai.ai, yansixu@hust.edu.cn

† Corresponding author. Emails: liuhx@bigai.ai

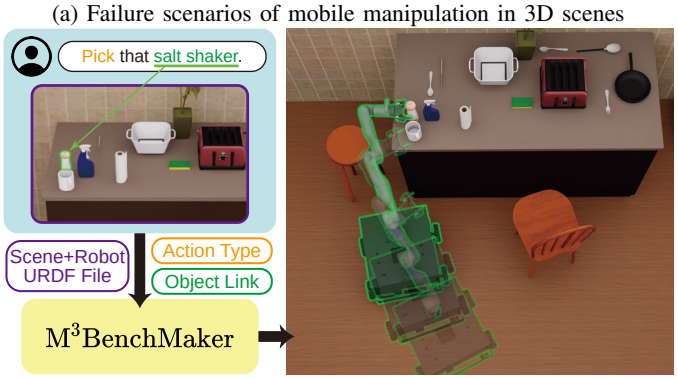
[1] School of Electronic Information and Communications, Huazhong University of Science and Technology. [2] State Key Laboratory of General Artificial Intelligence, Beijing Institute for General Artificial Intelligence (BIGAI). [3] Center for Vision, Cognition, Learning, and Autonomy (VCLA), Statistics Department, UCLA. [4] School of Intelligence Science and Technology, Peking University. [5] Institute for Artificial Intelligence, Peking University.

I. Introduction

HUMANS possess an innate ability to manipulate their environment with remarkable flexibility and coordination, seamlessly integrating locomotion and manipulation. In contrast, robots still struggle to achieve this level of adaptability and proficiency in mobile manipulation. Current learning-based models and motion planning methods for mobile manipulators often address individual subproblems in isolation, such as navigating to waypoints, manipulating with a fixed mobile base, or grasping objects. However, neglecting the potential of coordinated whole-body motion can lead to misalignment between module outputs and task constraints. For instance, in a typical object-fetching task, a navigable position near the target object may still be impossible for the arm to reach the object, or a feasible grasp pose may become unachievable due to collisions with surrounding objects (see Fig. 2a). These limitations underscore the necessity of coordinating whole-body motion with a comprehensive understanding of robot embodiment, environmental context, and task objectives to enable effective mobile manipulation in complex 3D scenes.

To generate whole-body motion for mobile manipulation tasks, there is an ongoing debate regarding the effectiveness and limitations of model-based motion planning methods versus data-driven learning-based models. While motion planning can produce complex whole-body mobile manipulation skills [1, 2], its effectiveness and generalizability in real-world scenarios are constrained by its reliance on perfect environ-

(a) Failure scenarios of mobile manipulation in 3D scenes



(b) An overview of whole-body motion data collection tool

Fig. 2: **Illustration of whole-body motion trajectories in 3D scenes.** (a) Treating the mobile base and arm as separate entities can lead to two typical failures: a nearby navigable position may be impractical for the arm to reach the object (red), and a feasible grasp pose may be unachievable due to the robot's embodiment and environmental constraints (orange). (b) Our tool generates feasible whole-body motion trajectories from high-level instructions, requiring only the action type, target object, and URDF files of the scene and robot. The green overlay illustrates a generated trajectory for the "pick that salt shaker" task.

mental knowledge [3–5] or predefined goal configurations (*e.g.*, grasp poses [6–8]). On the other hand, learning-based models have shown promising results in execution under perceptual and action noise, chaining primitive skills, and adapting to certain environmental variations [9, 10]. However, they have yet to demonstrate robust base-arm coordination with situated goal configurations (*e.g.*, achieving specific grasps or object placements). Learning complex mobile manipulation tasks requires datasets that capture whole-body motions in 3D scenes, yet such datasets remain scarce due to the challenges in generating whole-body motion data. Furthermore, evaluating learned models necessitates a standardized environment for fair benchmarking.

Table I presents recent state-of-the-art benchmarks for Embodied AI and robotics. Many of these benchmarks simplify actions to symbolic operations [11, 12] or navigation [13], lacking physical interaction with the environment. While more recent benchmarks enable fixed-base manipulators to interact with objects in realistic simulations [14–17] or allow mobile agents to navigate and manipulate in 3D scenes [18–20], they often overlook the necessity of coordinating base and arm motions.

To address the need for generating whole-body motions in mobile manipulation, we introduce M^{3}Bench, a comprehensive benchmark that features challenging object rearrangement tasks that require a mobile manipulator to reason about its embodiment, environmental context, and task objectives to generate coordinated motions for picking and placing objects in diverse household scenes (see Fig. 1). M^{3}Bench comprises 30,000 object rearrangement tasks involving 32 distinct object types across 119 household scenes, covering a broad spectrum of task objectives and environmental constraints relevant to embodied mobile manipulation. Additionally, it includes rich metadata, such as natural language task instructions, panoptic maps, and egocentric camera videos, making it a valuable resource for related research in Embodied AI, such as embodied instruction following and human-AI collaboration.

Leveraging M^{3}Bench, we developed M^{3}BenchMaker (see Fig. 2b), an automatic data generation tool designed to produce whole-body motion trajectories as expert demonstrations for robot learning. M^{3}BenchMaker procedurally generates coordinated trajectories from high-level task instructions, requiring only the action type, object link, and the Unified Robot Description Format (URDF) of the scene and robot. It employs an energy-based model to predict grasp pose and placement candidates [21], and it leverages an advanced virtual kinematics technique [2] to compute coordinated whole-body motion trajectories (see Sec. II for details). This tool not only addresses the scarcity of high-quality whole-body mobile manipulation data but also allows researchers to generate additional samples customized to specific robot and scene configurations for their own studies.

To enable in-depth evaluation of motion generation from 3D scans for mobile manipulation, M^{3}Bench incorporates various task splits to assess generalization across different dimensions, such as novel scenes and objects. We utilize a realistic physics simulation platform [22] to evaluate the feasibility of generated motion trajectories, ensuring that the robot can physically grasp objects and place them stably at the desired locations. Furthermore, our benchmarking reveals that sampling- and optimization-based motion planning methods [23, 24], even when augmented with affordance prediction, as well as learning-based autoregressive planning and generative AI techniques [9, 10, 25], struggle to effectively solve mobile manipulation tasks when required to account for goals such as grasp poses for picking and placement locations for placing actions. After integrating action goals into motion generation, learning-based methods outperform modularized motion planning in computational efficiency and simplicity of problem setup but still lag in motion accuracy. This underscores the importance of high-quality whole-body mobile manipulation data generated by tools like M^{3}BenchMaker and highlights the necessity of M^{3}Bench for advancing research in whole-body motion generation for mobile manipulation in 3D scenes.

Contribution: We make the following contributions:

- We introduce M^{3}Bench for benchmarking task-oriented whole-body motion generation for mobile manipulation in household environment, and we provide assets required for testing traditional planning-based methods or learning-based methods.
- We develop M^{3}BenchMaker, an automatic whole-body motion generation tool based on high-level task instruc-

TABLE I: **Relevant datasets and benchmarks in robotics.** The M^{3}Bench provides comprehensive whole-body motion demonstrations for object manipulation across 566 household scenes. **Mobile Manipulation**: Simultaneous navigation and object manipulation with foot-arm coordination. **Whole-body Demonstration**: Provides whole-body motion data. [1]Simplified cases without navigation and coordination. **Procedural Generation**: Algorithmic procedure for creating varied tasks and trajectories. **Household Scene**: Tasks performed in 3D household environments. **Language**: Natural language task descriptions. **Physical Grasp**: Realistic physics-based grasping simulation. [2]Simplified grasp (*e.g.*, attach). **Egocentric Perception**: Provides egocentric visual sensory data (*e.g.*, RGB-D images). [3]No rendered RGB images. **Flexible Material**: Customizable materials and textures for visual diversity.

Benchmark	Mobile Manipulation	Whole-body Demonstration	Procedural Generation	Household Scene	Language	Physical Grasp	Egocentric Perception	Flexible Material
ACRV [26]	✗	✗	✗	✗	✗	✓	✓	✗
Alfred [27]	✗	✗	✓	120	✓	✗	✓	✗
ManiSkill [15, 17]	✓	✓[1]	✗	✗	✗	✓	✓	✗
Calvin [28]	✗	✗	✓	✗	✓	✓	✓	✗
Behavior [18]	✓	✗	✓	50	✗	✗	✓	✓
RLBench [14]	✗	✗	✓	✗	✗	✓[2]	✓	✗
VLMbench [29]	✗	✗	✓	✗	✓	✓[2]	✓	✗
Ravens [30]	✗	✗	✓	✗	✓	✓[2]	✓	✗
MotionBenchMaker [16]	✗	✗	✓	✗	✗	✗	✓[3]	✗
Habitat HAB [19]	✓	✗	✓	105	✗	✗	✓	✗
ARNOLD [31]	✗	✗	✗	20	✓	✓	✓	✓
Ours	✓	✓	✓	119	✓	✓	✓	✓

tions, which can be easily customized for different robot and scene configurations.

- We provide an in-depth evaluation of motion generation from 3D scans for mobile manipulation, revealing weaknesses of current arts in promoting future research in mobile manipulation across diverse 3D scenes.

Overview: The remainder of this paper is organized as follows. Sec. II describes the key components of M^{3}BenchMaker. Sec. III details the development of the environment and the benchmarking setup. We implement multiple methods for mobile manipulation and discuss their performance in Sec. IV, and we conclude the paper in Sec. V.

II. The M^{3}BenchMaker

Diverse whole-body motion trajectories for mobile manipulators in complex 3D environments is crucial for advancing embodied AI. However, collecting expert demonstrations for training models are usually time-consuming and challenging. To address this, we introduce M^{3}BenchMaker, a user-friendly tool that streamlines the generation of whole-body motion trajectories in 3D scenes, significantly reducing the time and effort required to create large-scale datasets for mobile manipulation tasks in various environments. Notably, M^{3}BenchMaker is adaptable to different robot and scene given the URDF files that describe the configurations, enabling researchers to generate customized whole-body motion trajectories for their specific research needs. Fig. 3 illustrates the architecture of the M^{3}BenchMaker.

A. Task Builder

The task builder serves as the primary user interface, allowing users to define manipulation tasks using high-level action commands such as picking, placing, and reaching. Users no longer need to manually specify grasping poses, placement locations, base positions, or create optimization programs for motion trajectories. To define a task, users simply select target object links from the scene URDF, set the robot's initial position, and specify the desired action types. The task builder then creates an instance of the data generation pipeline, integrating subsequent modules to procedually generate whole-body motion trajectories. For enhanced data diversity, the task builder supports data augmentation via the Conditional Scene Sampler (see Sec. II-B). This feature facilitates the training and evaluation of embodied AI models in complex environments by generating varied scenarios from a single task definition.

B. Conditional Scene Sampler

The conditional scene sampler generates diverse initial configurations for data augmentation by randomizing object and robot positions and orientations. It produces variations dependent on the original scene's object relations, ensuring physical feasibility and contextual consistency required by the task. For instance, in a task involving picking an object from a table, the sampler ensures the sampled objects remains on table (see Fig. 3 orange box). This is achieved by recognizing supporting planes for objects and the robot through analysis of surrounding geometries.

To identify supporting planes, we parameterize a surface plane as $\boldsymbol{\pi} = \langle \boldsymbol{n}^T, d, U \rangle$, where $\boldsymbol{n} \in \mathbb{R}^3$ is the normal vector, d is the distance to origin, and $U = \{\boldsymbol{u} | \boldsymbol{u} \in \mathbb{R}^3\}$ defines the plane's polygon outline. The most likely supporting plane $\boldsymbol{\pi}_s$ for a bottom surface $\boldsymbol{\pi}_o$ is identified by solving:

$$\underset{\boldsymbol{\pi}_s \in \prod}{\operatorname{argmax}} \ \mathrm{A}\left(U_s \cap \mathrm{proj}_{o,s}(U_o)\right) / \mathrm{A}(U_o), \tag{1}$$

$$\text{s.t.} \ \frac{1}{|U_o|} \sum_{\boldsymbol{u} \in U_o} \boldsymbol{n}_s^T \boldsymbol{u} + d_s \leqslant \theta_d, \tag{2}$$

$$\mathrm{abs}\left(\boldsymbol{n}_p^{iT} \boldsymbol{n}_c^j\right) \geqslant \theta_a. \tag{3}$$

where $\prod$ is a set of supporting plane candidates, $\mathrm{A}(\cdot)$ denotes polygon area, $\cap$ computes intersection, and $\mathrm{proj}_{o,s}(U_o) = \{\boldsymbol{u} - \left(\boldsymbol{n}_s^T \boldsymbol{u} + d_s\right) \boldsymbol{n}_s | \boldsymbol{u} \in U_o\}$ projects bottom surface points onto the supporting plane, θ_d and θ_a are distance and angle thresholds. Eq. (1) defines the contact ratio, while Eqs. (2) and (3) enforce alignment and distance constraints. The complete sampling procedure is detailed in Alg. 1. We utilize

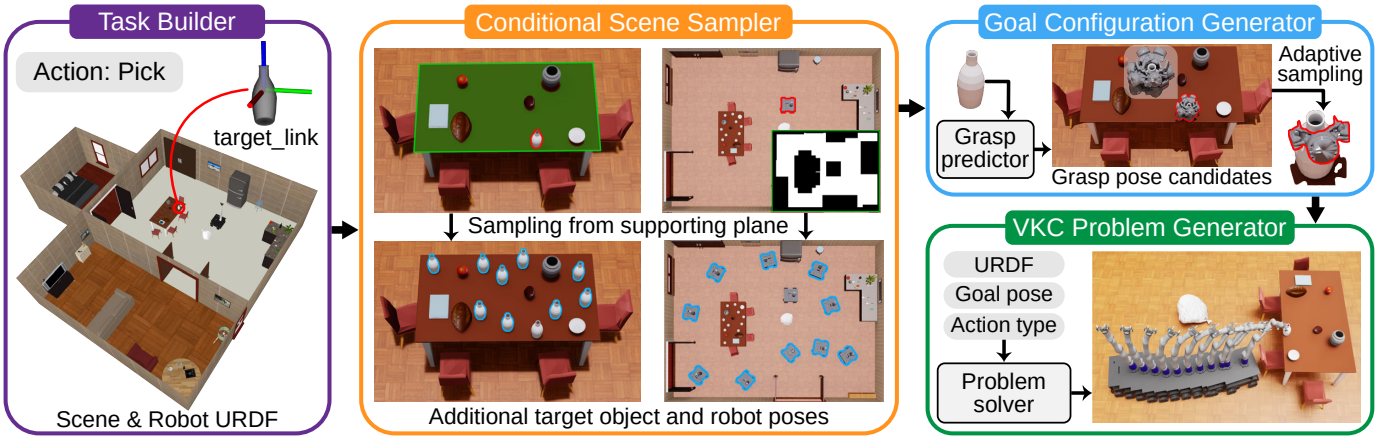


Fig. 3: **Overview of the M^{3}BenchMaker.** The **Task Builder** allows users to specify manipulation tasks via high-level definitions using URDF, target object link, and action. The **Conditional Scene Sampler** augments data by generating object and robot poses (blue outline) in terms of their supporting planes (green outline) of target objects (red outline). The **Goal Configuration Generator** produces task-specific goal poses using a pre-trained model for grasp/placement candidates. The **VKC Problem Generator** constructs optimization programs for computing whole-body motion trajectories that satisfy task objectives and constraints via Virtual Kinematic Chain (VKC) [2].

Algorithm 1: Conditional Scene Sampler

Input : Target object $\boldsymbol{\pi}_o$, candidate planes $\prod$, thresholds θ_d, θ_a, number of samples N
Output: A set of feasible object poses $\mathcal{S}$

$\mathcal{S} \leftarrow \varnothing$;
// Filter candidates in terms of Eqs. (2) and (3)
$\prod_c \leftarrow \texttt{FilterSupportPlane}(\boldsymbol{\pi}_o, \prod, \theta_d, \theta_a)$
// Determine supporting plane according to Eq. (1)
$\boldsymbol{\pi}_s \leftarrow \texttt{CalcSupportPlane}(\boldsymbol{\pi}_o, \prod_c)$
while $\mathcal{S}.size() < N$ **do**
  $p \leftarrow \texttt{samplePoseOnPolygon}(\boldsymbol{\pi}_s)$
  // Check if sampled pose is within plane $\boldsymbol{\pi}_s$
  if $withinPolygon(\boldsymbol{\pi}_o, p, \boldsymbol{\pi}_s)$ **then**
    $\mathcal{S}.\texttt{add}(p)$ // add sampled pose to $\mathcal{S}$
return $\mathcal{S}$

Algorithm 2: Adaptive Goal Sampling

Input : candidate set $\mathcal{C}$
Output: goal configuration g

$T \leftarrow \texttt{KDTree}(\mathcal{C})$
$scores \leftarrow \texttt{initFeasibilityScore}(\mathcal{C})$
while *feasible goal not found* **do**
  // calculate probability for each candidate
  $probs \leftarrow \texttt{calcSamplingProb}(\mathcal{C}, scores)$
  // draw a single condidate index from distribution
  $i \leftarrow \texttt{drawSample}(\mathcal{C}, probs)$
  // Check feasibility of sampled candidate
  if $checkFeasibility(\mathcal{C}[i])$ **then**
    // found feasible configuration
    $g \leftarrow \mathcal{C}[i]$
    break
  // Update feasibility scores in neighbors
  $neighbors \leftarrow T.\texttt{GetNeighbors}(\mathcal{C}[i])$
  $scores[neighbors] \leftarrow scores[neighbors] \times 0.5$
  // update K-D tree and remove checked candidate
  $T \leftarrow \texttt{UpdateKDTree}(T, \mathcal{C}[i])$
  $\mathcal{C}.\texttt{remove}(i)$
return g

the method in [4] to extract surface planes and solve the optimization problem by iteratively identifying the plane that maximizes Eq. (1) while satisfying the constraints.

C. Goal Configuration Generator

This module efficiently generates 6D end-effector poses for grasping or placing target objects, serving as optimization objectives for motion planning. We employ an energy-based model to predict candidate goal configurations based on target object geometry [21]. However, this object-centric approach, which considers only object geometry without accounting for the robot's kinematic constraints or environmental contexts, results in only a small subset of candidates being feasible for the task. To address the computational expense of evaluating all candidates through motion planning, we developed an adaptive sampling algorithm that efficiently draws samples from the candidate set, significantly accelerating the motion generation process.

Detailed in Alg. 2, our algorithm iteratively selects and updates the sampling probability of candidates based on their feasibility scores. It utilizes a K-D tree for efficient neighbor search and initializes feasibility scores using the candidates' energy values. When a candidate fails the feasibility check, the feasibility scor es of its neighbors, identified via the K-D tree within a specified distance, are halved during the update. By concentrating sampling in promising regions of the goal configuration space while maintaining exploration, the algorithm significantly reduces the number of expensive feasibility checks required to identify viable goal configurations.

D. VKC Problem Generator

The VKC problem generator automates the construction of motion planning programs, formulating comprehensive optimization problems that encapsulate all necessary constraints and objectives for computing whole-body motion trajectories,

utilizing task specifications, URDF, and goal configurations from preceding modules. We employ the VKC approach [2] to solve for whole-body motion of mobile manipulators, modeling the mobile base, robot arm, and manipulated object as a unified system, achieving superior foot-arm coordination through simultaneous optimization and surpassing traditional methods that separate base and arm planning.

Our implementation follows TrajOpt and ROS-Industrial Tesseract conventions [32], effectively incorporating kinematic constraints while avoiding large-space searches. The trajectory optimization minimizes joint travel distances and overall smoothness, with inequality constraints for joint limits, collision avoidance, and end-effector pose reaching. We adopt a sequential convex optimization method [24] to solve the resulting problem, yielding feasible, coordinated whole-body motion trajectories for diverse mobile manipulation tasks without manual task-specific planner programming.

By automating these processes, M^{3}BenchMaker empowers researchers to efficiently collect tailored whole-body motion trajectories, significantly advancing embodied AI in complex 3D environments.

III. The M^{3}Bench

The M^{3}Bench aims to advance robot capabilities in coordinating whole-body movements within complex environments, inspired by human ability to seamlessly perform such tasks. It challenges mobile manipulators to generate coordinated whole-body motion trajectories for picking or placing everyday objects in 3D scenes, requiring agents to jointly understand their embodiment, environmental contexts, and task objectives from 3D scans.

A. Simulation Environment

Simulation Platform. Our benchmark, built on Isaac Sim [22], provides a high-fidelity physics simulation that meticulously models real-world properties and interactions. This platform enables precise evaluation of motion trajectory feasibility, grasping abilities, and the complex interplay between mobility and manipulation. Additionally, it could generate rich perceptual data (*e.g.*, RGB-D image) that closely mimics the sensory input available to real-world robots.

Scene and Robot Configuration. The benchmark comprises 119 diverse household scenes containing 32 types of objects, curated from PhyScene [33]. These interactive 3D scenes are enhanced with physical properties and rich materials for photo-realistic and physics-realistic simulation. For the robot, we employ a common mobile manipulator configuration: a 7-DoF Kinova Gen3 robotic arm with a parallel gripper, mounted on an omnidirectional mobile base. This setup facilitates complex manipulations requiring coordinated base and arm movements.

B. Task Design and Variations

M^{3}Bench focuses on two primary object rearrangement tasks: picking and placing. Given a 3D point cloud of the scene, a mask of the target object, and its initial configuration, the robot must generate whole-body motion trajectories to manipulate the object. The tasks are defined as: (i) *Pick tasks*: Navigate to, reach, and grasp a specified object from its initial location; (ii) *Place tasks*: Transport a held object to a designated location and place it stably. Success in both tasks requires avoiding collisions along the trajectory and maintaining the desired goal state for 2 seconds.

TABLE II: Number of pick/place task samples in each data split.

Split	*Pick*	*Place*
Train	14,793	7,478
Val	948	479
Test	3,225	1,630
Novel Object	688	397
Novel Scene	762	369
Novel Scenario	204	77
Total	20,620	10,430

TABLE III: Number of rooms and target objects in M^{3}Bench

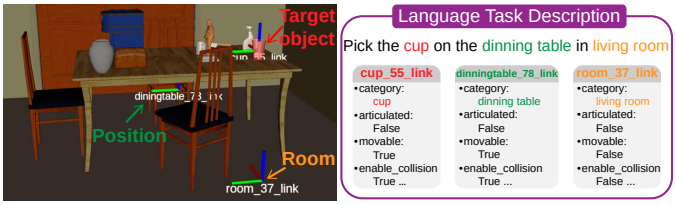
Statistics	Value
Bathroom	132
Bedroom	198
Kitchen	97
Living room	129
Total scenes	119
Object types	32
Total objects	588

The task pool encompasses a wide range of mobile manipulation scenarios, featuring 32 object types with varying properties across 119 diverse household scenes. Each scene presents unique layouts, furniture arrangements, and obstacle configurations. Tasks are generated by selecting appropriate objects and placement locations based on scene categories. We employ the conditional scene sampler (Sec. II-B) to generate various initial configurations, further challenging the robot to generate coordinated whole-body motions while adapting to environmental constraints and task objectives.

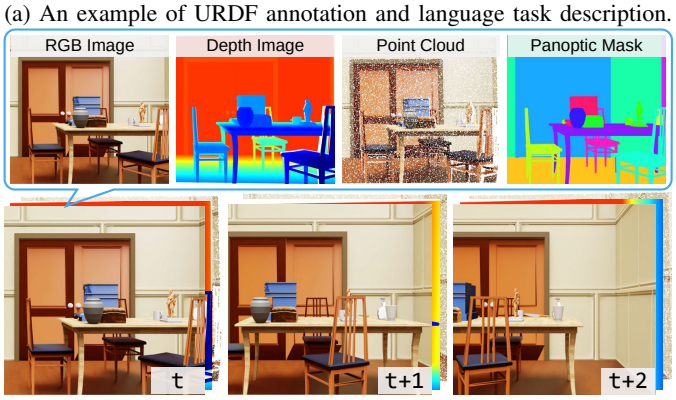
C. Data Collection

Demonstration Generation. We utilized our developed M^{3}BenchMaker to generate demonstrations for each task. The tool takes as input the scene and robot URDF, target object link, and task type (pick or place), then generates a whole-body motion trajectory for the robot. The optimization program in M^{3}BenchMaker ensures these trajectories are collision-free and kinematically feasible. Each trajectory is then verified for physical feasibility in Isaac Sim, with only valid demonstrations and their corresponding tasks included in the benchmark. In total, we collected 30k valid demonstrations, each containing 30 waypoints.

Additional Metadata. To facilitate embodied AI research, we provide comprehensive metadata for each task (see Fig. 4). This includes annotations for all links in the scene URDF, covering object categories and simulation properties. We employ a template-based approach with lexicalized phrase candidates to generate language instructions for each task. For example, the template "Pick [object] in [room] on [position]" might be realized as "Pick the cup in the living room on the dining table". During task execution, Isaac Sim's built-in rendering capabilities, combined with annotated information, generate pixel-accurate semantic and instance segmentations along with egocentric camera views. This rich combination of annotations, trajectory data, and language instructions creates a comprehensive resource for exploring various aspects of embodied intelligence.



(a) An example of URDF annotation and language task description.



(b) An example of sequential egocentric views.

Fig. 4: **An illustration of metadata.**

D. Benchmark

Data Split and Statistics. The tasks in M^{3}Bench are carefully divided into several splits to assess different aspects of generalization capabilities. Objects and scenes are randomly categorized into seen and unseen subsets. The primary evaluation set, the *Base* split, encompasses all seen objects and scenes, divided into *Train* (75%), *Val* (5%), and *Test* (20%) sets. Three additional splits challenge model generalization: *Novel Object* (unseen objects in seen scenes), *Novel Scene* (seen objects in unseen scenes), and *Novel Scenario* (unseen objects in unseen scenes). Tables II and III present detailed statistics of these splits and task configurations, enabling systematic evaluation of model generalization across various dimensions of mobile manipulation in 3D scenes.

Metrics. We employ a multi-faceted approach to evaluate motion generation models. Task success rate serves as the primary metric, determined by the robot's ability to complete specified tasks and maintain the desired state for 2 seconds, as verified by the Isaac Sim physics engine. We also measure the closest distance from the end-effector to the target as an auxiliary metric, reflecting the trajectory's effectiveness in reaching the object or placement location. To assess trajectory quality, we utilize several quantitative measures: environment collision, self-collision, joint limit violation, and trajectory solving time. This comprehensive set of metrics evaluates models' capabilities in generating effective and efficient motion trajectories for mobile manipulation in 3D scenes.

IV. Experiments

A. Experimental Setup

Models for M^{3}Bench. Due to the lack of existing models for whole-body motion generation in mobile manipulation within 3D scenes, we adapt five state-of-the-art approaches to our benchmark:

- ModMP[O]: Integrates a VKC motion planner [2] with grasp pose predictor [21] and heuristic placement.
- ModMP[S]: Similar to ModMP[O], we replace the planer with a sampling-based planner RRT-Connect [34].
- MπNet [25]: Extended from stationary to mobile manipulation by incorporating whole-body joint generation and Signed Distance Function (SDF) [35] for collision loss computation in complex 3D scans.
- MπFormer: A skill transformer [36] variant using PointNet++ [37] for 3D scan processing and decision transformer architecture [38] for enhanced sequence modeling.
- MDiffusion: Utilizes a conditional diffusion model [39], encoding 3D scans with a Point Transformer [40] and employing a cross-attention module to predict denoising scores conditioned on 3D features.

Implementation Details. For MπNet, MDiffusion and MπFormer, we generate 3D scans from scene URDF. To enhance learning tractability, we apply a perception bounding box around the robot and target object to crop the scans, focusing the model's attention on relevant spatial information. We train MπNet, MDiffusion and MπFormer on the *Train* split and perform model selection on *Val*. In constrast, as ModMP[O]and ModMP[S] does not involve learning procedure, we evaluate it directly on the *Test* and *Novel* splits. To simplify the optimization problem in ModMP[O] and ModMP[S], we ignore collisions between the end-effector and target object during motion planning, as considering these collisions would frequently result in infeasible trajectories.

B. Experimental Results

The experimental results are summarized in Tab. IV. Trajectories are evaluated in Isaac Sim using metrics described in Sec. III-D. Particularly, for the ModMP[O] and ModMP[S] model, when motion planning fails to solve the problem (*i.e.*, optimization does not converge), we consider it as a failure instance.

Across Models. ModMP[O] consistently outperforms other approaches in most pick-and-place tasks, achieving higher success rates and closer distances to the goal. This supports our hypothesis that integrating conventional motion planning with affordance prediction could generalize across diverse 3D scenes. While the performance of the sampling-based planner ModMP[S] is comparable to that of the optimization-based planner ModMP[O], it requires significantly more time to find solutions. However, ModMP[O]'s superior performance also comes at the cost of increased computation time, driven by the optimization complexity in large-scale 3D environments. Its effectiveness also heavily depends on the quality of predicted grasp and placement poses; inaccurate predictions can result in optimization failures or environmental collisions (see Fig. 2a). Although ModMP[O] shows promising results, the overall low success rates suggest that integrating conventional motion planning with affordance prediction alone is insufficient for robust performance.

In contrast, learning-based models are more time-efficient but often fail to generate feasible solutions in unseen scenarios. Specifically, while MDiffusion achieves comparable perfor-

TABLE IV: Quantitative results on M^{3}Bench, measured by success rate (*Succ*), distance to goal (*Dist*), joint violation rate (*J.Vio*), environment collision rate (*E.Coll*), self-collision rate (*S.Coll*), and execution time (*Time*). Best performance is shown in bold.

Test Split	Method	Pick Task						Place Task					
		Succ(%)↑	Dist(m)↓	JVio(%)↓	EnvColl(%)↓	SelfColl(%)↓	Time(s)↓	Succ(%)↑	Dist(m)↓	JVio(%)↓	EnvColl(%)↓	SelfColl(%)↓	Time(s)↓
Base Test	MπNet	0.07	0.34	20.79	16.53	0.36	**0.48**	0.80	1.68	34.67	42.75	1.24	0.59
	MπFormer	0.00	1.36	**0.00**	44.58	**0.00**	0.93	0.15	0.92	0.15	23.38	**0.00**	1.16
	MDiffusion	18.12	0.04	0.59	19.09	0.53	**0.48**	**5.83**	0.04	0.36	39.67	0.32	**0.47**
	ModMP[S]	16.90	0.03	**0.00**	12.13	**0.00**	87.53	1.98	0.31	**0.00**	3.58	**0.00**	89.74
	ModMP[O]	**20.13**	**0.01**	**0.00**	**9.70**	**0.00**	19.63	2.76	**0.29**	**0.00**	**2.65**	**0.00**	28.58
Novel Object	MπNet	0.15	0.34	29.07	22.38	0.44	0.47	0.76	1.55	35.26	45.84	**0.00**	0.59
	MπFormer	0.44	1.39	**0.00**	53.49	**0.00**	0.94	0.25	0.70	**0.00**	31.74	**0.00**	1.16
	MDiffusion	9.30	0.05	0.15	37.24	**0.00**	**0.45**	1.26	0.06	0.50	35.32	**0.00**	**0.44**
	ModMP[S]	18.39	0.01	**0.00**	19.75	**0.00**	88.65	3.41	0.14	**0.00**	4.53	**0.00**	88.31
	ModMP[O]	**21.80**	**0.00**	**0.00**	**13.15**	**0.00**	18.74	**5.10**	**0.12**	**0.00**	**0.00**	**0.00**	29.89
Novel Scene	MπNet	0.00	0.42	13.73	43.88	0.13	**0.48**	0.84	2.31	41.78	45.96	4.18	0.59
	MπFormer	0.00	2.06	**0.00**	60.13	**0.00**	0.93	0.00	1.04	**0.00**	13.65	**0.00**	1.17
	MDiffusion	7.25	0.04	0.13	38.53	0.13	**0.48**	1.95	0.07	0.28	45.13	**0.00**	**0.44**
	ModMP[S]	19.20	0.02	**0.00**	13.27	**0.00**	89.10	7.80	0.23	**0.00**	3.49	**0.00**	88.93
	ModMP[O]	**25.59**	**0.00**	**0.00**	**10.82**	**0.00**	20.13	**9.76**	**0.18**	**0.00**	**1.10**	**0.00**	27.39
Novel Scenario	MπNet	0.00	0.61	16.67	25.49	**0.00**	0.47	0.00	2.74	16.88	9.09	1.30	0.59
	MπFormer	0.00	2.58	**0.00**	70.59	**0.00**	0.92	0.00	1.68	9.09	12.99	**0.00**	1.17
	MDiffusion	5.88	0.04	**0.00**	26.76	**0.00**	**0.46**	2.60	0.04	**0.00**	7.49	**0.00**	**0.45**
	ModMP[S]	20.12	0.02	**0.00**	14.97	**0.00**	89.32	4.31	0.38	**0.00**	2.67	**0.00**	89.19
	ModMP[O]	**23.94**	**0.00**	**0.00**	**11.81**	**0.00**	19.49	**6.52**	**0.25**	**0.00**	**0.00**	**0.00**	28.31

mance to planning-based methods, its success drops significantly in the *Novel* splits. Additionally, unlike planning-based methods that enforce hard constraints, learning-based approaches frequently produce trajectories that violate joint limitations and are more prone to collisions. Moreover, MπNet and MπFormer barely achieve any success in the test splits. This result suggests that directly adapting stationary manipulation models to mobile manipulation tasks in 3D scenes is infeasible unless the underlying model is powerful enough to capture the complexity of the 3D environment and associated tasks. These findings highlight the persistent challenge of efficiently generating whole-body motion trajectories in complex 3D environments and emphasize the need for further research to develop more sophisticated models for mobile manipulation tasks.

Across Tasks. The experiment results reveal distinct performance patterns between pick and place tasks. While ModMP[O] maintains better performance in both tasks, its success rates significantly drop in place tasks, and all models require more time to generate trajectories for the *place tasks*. This discrepancy suggests that generating coordinated whole-body motion trajectories for placing objects is more challenging than for picking objects, as it involves additional constraints such as stable placement locations, appropriate object orientation, and reachable motion trajectory. The increased complexity of place tasks explains the lower success rates and longer execution times observed across all models.

On Generalization. While MDiffusion achieves comparable performance to planning-based methods in the *Base* split, its performance deteriorates in the *Novel* splits,indicating persistent generalization challenges for learning-based models. Particularly, the distance to goal in unfamiliar scenes (*Novel Scene* and *Novel Scenario* splits) exceeds that of the *Novel Object* split, suggesting that the impact of novel scenes is more significant than novel objects. The conventional planning-based method, in contrast, maintains consistent performance across all splits, though its relatively low success rates across the board underscore the inherent complexity of mobile manipulation tasks in diverse household environments.

Remarks. Our experiments reveal two crucial insights:

- Although combining motion planning with affordance prediction demonstrates consistent performance across all splits, its overall success rates remain low. This highlights the limitations of even advanced hybrid approaches in addressing the challenges of mobile manipulation in diverse 3D scenes, emphasizing the need for models capable of holistically solving such complex tasks.
- Mobile manipulation tasks demand learning-based models with significantly greater expressiveness and generalization than stationary manipulation tasks. Their poor performance in unseen scenarios (*i.e.*, the *Novel* splits) underscores the need for advancements in two key areas: (a) developing fine-grained representations of perceptual inputs to better capture the complexity of 3D environments, and (b) designing more sophisticated models for continuous whole-body motions to generate feasible trajectories in challenging scenarios.

V. Conclusion

We introduced M^{3}Bench, a comprehensive benchmark for whole-body motion generation in mobile manipulation tasks across diverse 3D environments, featuring 30k object rearrangement tasks in 119 household scenes. M^{3}Bench provides a standardized platform for both planning and learning communities. Through comprehensive evaluations of state-of-the-art models, we highlighted the persistent challenges in generating coordinated base-arm motion trajectories that satisfy both environmental constraints and task objectives. Furthermore, we developed M^{3}BenchMaker, a tool designed to efficiently generate whole-body motion trajectories from high-level instructions, which can serve as a valuable resource for researchers in their own studies. We hope M^{3}Bench opens new opportunities for robotics research and catalyzes progress toward developing more adaptive and capable embodied agents.

Acknowledgement: This work was supported in part by the National Natural Science Foundation of China (No. 62403064, 62403063). We thank the all the colleagues in Robotics Lab from BIGAI for fruitful discussions.

REFERENCES

[1] A. Jain and C. C. Kemp, "Pulling open doors and drawers: Coordinating an omni-directional base and a compliant arm with equilibrium point control," in *International Conference on Robotics and Automation (ICRA)*, 2010.

[2] Z. Jiao, Z. Zhang, X. Jiang, D. Han, S.-C. Zhu, Y. Zhu, and H. Liu, "Consolidating kinematic models to promote coordinated mobile manipulations," in *International Conference on Intelligent Robots and Systems (IROS)*, 2021.

[3] K. Mo, Y. Qin, F. Xiang, H. Su, and L. Guibas, "O2O-Afford: Annotation-free large-scale object-object affordance learning," in *Conference on Robot Learning (CoRL)*, 2021.

[4] M. Han, Z. Zhang, Z. Jiao, X. Xie, Y. Zhu, S.-C. Zhu, and H. Liu, "Scene reconstruction with functional objects for robot autonomy," *International Journal of Computer Vision (IJCV)*, 2022.

[5] Z. Zhang, L. Zhang, Z. Wang, Z. Jiao, M. Han, Y. Zhu, S.-C. Zhu, and H. Liu, "Part-level scene reconstruction affords robot interaction," in *International Conference on Intelligent Robots and Systems (IROS)*, 2023.

[6] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," in *International Conference on Robotics and Automation (ICRA)*, 2021.

[7] J. Gu, D. S. Chaplot, H. Su, and J. Malik, "Multi-skill mobile manipulation for object rearrangement," in *International Conference on Learning Representations (ICLR)*, 2022.

[8] P. Li, T. Liu, Y. Li, Y. Geng, Y. Zhu, Y. Yang, and S. Huang, "Gendexgrasp: Generalizable dexterous grasping," in *International Conference on Robotics and Automation (ICRA)*, IEEE, 2023.

[9] N. Hansen, Y. Lin, H. Su, X. Wang, V. Kumar, and A. Rajeswaran, "MoDem: Accelerating Visual Model-Based Reinforcement Learning with Demonstrations," in *International Conference on Learning Representations (ICLR)*, 2022.

[10] P. Xie, R. Chen, S. Chen, Y. Qin, F. Xiang, T. Sun, J. Xu, G. Wang, and H. Su, "Part-Guided 3D RL for Sim2Real Articulated Object Manipulation," in *IEEE Robotics and Automation Letters (RA-L)*, 2023.

[11] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, M. Deitke, K. Ehsani, D. Gordon, Y. Zhu, *et al.*, "Ai2-thor: An interactive 3d environment for visual ai," *arXiv preprint arXiv:1712.05474*, 2017.

[12] C. H. Song, J. Wu, C. Washington, B. M. Sadler, W.-L. Chao, and Y. Su, "Llm-planner: Few-shot grounded planning for embodied agents with large language models," in *International Conference on Computer Vision (ICCV)*, October 2023.

[13] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3674–3683, 2018.

[14] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "Rlbench: The robot learning benchmark & learning environment," *IEEE Robotics and Automation Letters (RA-L)*, vol. 5, no. 2, pp. 3019–3026, 2020.

[15] T. Mu, Z. Ling, F. Xiang, D. Yang, X. Li, X. Li, S. Tao, Z. Huang, Z. Jia, and H. Su, "Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations," in *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks* (J. Vanschoren and S. Yeung, eds.), 2021.

[16] C. Chamzas, C. Quintero-Pena, Z. Kingston, A. Orthey, D. Rakita, M. Gleicher, M. Toussaint, and L. E. Kavraki, "Motionbenchmaker: A tool to generate and benchmark motion planning datasets," *IEEE Robotics and Automation Letters (RA-L)*, vol. 7, no. 2, pp. 882–889, 2021.

[17] J. Gu, F. Xiang, X. Li, Z. Ling, X. Liu, T. Mu, Y. Tang, S. Tao, X. Wei, Y. Yao, *et al.*, "Maniskill2: A unified benchmark for generalizable manipulation skills," in *International Conference on Learning Representations (ICLR)*, 2022.

[18] C. Li, R. Zhang, J. Wong, C. Gokmen, S. Srivastava, R. Martín-Martín, C. Wang, G. Levine, M. Lingelbach, J. Sun, M. Anvari, M. Hwang, M. Sharma, A. Aydin, D. Bansal, S. Hunter, K.-Y. Kim, A. Lou, C. R. Matthews, I. Villa-Renteria, J. H. Tang, C. Tang, F. Xia, S. Savarese, H. Gweon, K. Liu, J. Wu, and L. Fei-Fei, "Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation," in *Conference on Robot Learning (CoRL)*, 2023.

[19] A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. S. Chaplot, O. Maksymets, *et al.*, "Habitat 2.0: Training home assistants to rearrange their habitat," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[20] J. Wu, R. Antonova, A. Kan, M. Lepert, A. Zeng, S. Song, J. Bohg, S. Rusinkiewicz, and T. Funkhouser, "Tidybot: Personalized robot assistance with large language models," *Autonomous Robots*, vol. 47, no. 8, pp. 1087–1102, 2023.

[21] J. Urain, N. Funk, J. Peters, and G. Chalvatzaki, "Se(3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion," in *International Conference on Robotics and Automation (ICRA)*, 2023.

[22] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, *et al.*, "Isaac gym: High performance gpu based physics simulation for robot learning," in *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2021.

[23] J. Schulman, J. Ho, A. X. Lee, I. Awwal, H. Bradlow, and P. Abbeel, "Finding locally optimal, collision-free trajectories with sequential convex optimization.," in *Robotics: Science and Systems (RSS)*, 2013.

[24] J. Schulman, Y. Duan, J. Ho, A. Lee, I. Awwal, H. Bradlow, J. Pan, S. Patil, K. Goldberg, and P. Abbeel, "Motion planning with sequential convex optimization and convex collision checking," *International Journal of Robotics Research (IJRR)*, vol. 33, no. 9, pp. 1251–1270, 2014.

[25] A. Fishman, A. Murali, C. Eppner, B. Peele, B. Boots, and D. Fox, "Motion policy networks," in *Conference on Robot Learning (CoRL)*, 2023.

[26] J. Leitner, A. W. Tow, N. Sünderhauf, J. E. Dean, J. W. Durham, M. Cooper, M. Eich, C. Lehnert, R. Mangels, C. McCool, *et al.*, "The acrv picking benchmark: A robotic shelf picking benchmark to foster reproducible research," in *International Conference on Robotics and Automation (ICRA)*, 2017.

[27] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, and D. Fox, "ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[28] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard, "Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks," *IEEE Robotics and Automation Letters (RA-L)*, vol. 7, no. 3, pp. 7327–7334, 2022.

[29] K. Zheng, X. Chen, O. C. Jenkins, and X. Wang, "Vlmbench: A compositional benchmark for vision-and-language manipulation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[30] A. Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, V. Sindhwani, *et al.*, "Transporter networks: Rearranging the visual world for robotic manipulation," in *Conference on Robot Learning (CoRL)*, 2021.

[31] R. Gong, J. Huang, Y. Zhao, H. Geng, X. Gao, Q. Wu, W. Ai, Z. Zhou, D. Terzopoulos, S.-C. Zhu, *et al.*, "Arnold: A benchmark for language-grounded task learning with continuous states in realistic 3d scenes," in *International Conference on Computer Vision (ICCV)*, 2023.

[32] L. Armstrong, "Optimization motion planning with tesseract and trajopt for industrial applications - ros-industrial."

[33] Y. Yang, B. Jia, P. Zhi, and S. Huang, "Physcene: Physically interactable 3d scene synthesis for embodied ai," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[34] J. J. Kuffner and S. M. LaValle, "Rrt-connect: An efficient approach to single-query path planning," in *International Conference on Robotics and Automation (ICRA)*, 2000.

[35] P.-S. Wang, Y. Liu, and X. Tong, "Dual octree graph networks for learning adaptive volumetric shape representations," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–15, 2022.

[36] X. Huang, D. Batra, A. Rai, and A. Szot, "Skill transformer: A monolithic policy for mobile manipulation," in *International Conference on Computer Vision (ICCV)*, 2023.

[37] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[38] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, "Decision transformer: Reinforcement learning via sequence modeling," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[39] S. Huang, Z. Wang, P. Li, B. Jia, T. Liu, Y. Zhu, W. Liang, and S.-C. Zhu, "Diffusion-based generation, optimization, and planning in 3d scenes," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[40] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *International Conference on Computer Vision (ICCV)*, 2021.