2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

GAMMA: Graspability-Aware Mobile MAnipulation Policy Learning based on Online Grasping Pose Fusion

Jiazhao Zhang[1,2,*], Nandiraju Gireesh[3,*], Jilong Wang[2], Xiaomeng Fang[2], Chaoyi Xu[3], Weiguang Chen[2], Liu Dai[4], and He Wang[1,2,3,†]

***Abstract*— Mobile manipulation constitutes a fundamental task for robotic assistants and garners significant attention within the robotics community. A critical challenge inherent in mobile manipulation is the effective observation of the target while approaching it for grasping. In this work, we propose a graspability-aware mobile manipulation approach powered by an online grasping pose fusion framework that enables a temporally consistent grasping observation. Specifically, the predicted grasping poses are online organized to eliminate the redundant, outlier grasping poses, which can be encoded as a grasping pose observation state for reinforcement learning. Moreover, on-the-fly fusing the grasping poses enables a direct assessment of graspability, encompassing both the quantity and quality of grasping poses. This assessment can subsequently serve as an observe-to-grasp reward, motivating the agent to prioritize actions that yield detailed observations while approaching the target object for grasping. Through extensive experiments conducted on the Habitat and Isaac Gym simulators, we find that our method attains a good balance between observation and manipulation, yielding high performance under various grasping metrics. Furthermore, we discover that the incorporation of temporal information from grasping poses aids in mitigating the sim-to-real gap, leading to robust performance in challenging real-world experiments. Project page: `https://pku-epic.github.io/GAMMA/`**

I. INTRODUCTION

Autonomous mobile manipulation has been an essential research area in robotics [1], [2], leading to diverse applications, *e.g.*, manufacturing, warehousing, construction, and household assistance [3], [4], [5], [6]. For research on mobile manipulation, a challenging but popular task setting is to require the agent to actively observe and explore an unseen environment with the goal of manipulating a target object. Originated from the unseen nature of the environment, the agent can't directly plan a trajectory to reach and grasp the object. Instead, its has to rely on online observations and scene priors to make a success, posing many research questions to the area.

Many existing works tackle this problem via combining 3D reconstruction and scene geometry analysis with motion planning [7], [8], [9]. However, such approaches usually suffer from huge computational costs for modeling the scene geometry and entail complicated heuristic designs tailored to specific robots. Recently, reinforcement learning (RL) based approaches [10], [11], [12], [13] have gained more attention due to their simplicity and efficiency. Exemplar works [11], [3] use visibility and reachability as scene priors, which can drive the agent to observe and approach the target object. They propose to learn and use them in policy's input states as well as reward, which has significantly improved the policy performance.

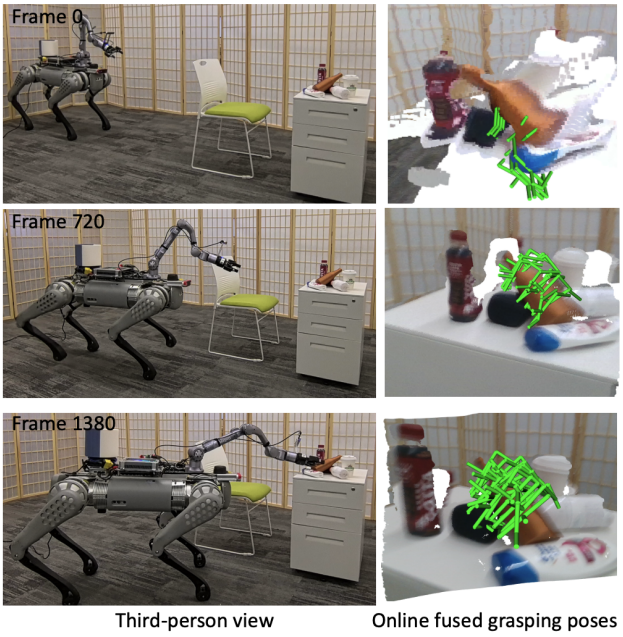

Fig. 1: We present a graspability-aware mobile manipulation approach powered by an online grasping pose fusion framework that enables a temporally consistent grasping observation and efficient grasping.

In this work, we focus on an RL-based approach and propose a novel scene prior, *graspability*, to advance mobile manipulation policy learning. We define graspability as a complete set of valid grasping poses of the target object. Compared with reachability and visibility, which provide circuitous guidance to the agent for graspings, graspability offers more direct and informative guidance for effective grasping guidance.

Note that graspability contains the full information of the valid target object graspings, online estimating graspability in an unseen environment is highly non-trivial, *e.g.*, online observations during mobile manipulation often include many occlusions as well as large overlaps, leading to noises and redundancy in grasping pose predictions [14]. We thus propose an online grasping pose fusion module, which dynamically fuses the redundant grasping poses and removes the outlier

*Joint first authors
[1]CFCS, School of Computer Science, Peking University. [2]Beijing Academy of Artificial Intelligence [3]Galbot [4]Tongji University

poses. This fusion process yields high-quality graspability estimation that achieves high precision and recall of valid grasping poses.

To facilitate agent learning, we propose the following two ways to fully utilize the estimated graspabiltity. First, we propose to encode graspabiltity into states and use it in the policy input, endowing the agent with the awareness of grasping goals. We find that our graspability-aware agent can thus learn to move its base and arm more intelligently. Second, we propose to use the number of grasps and the distance-to-grasp information in graspability as RL reward, encouraging the agent to gain more observations of valid grasping poses. We also introduce a weight schedule that combines these two rewards to balance the observation goal and the grasping goal. This reward motivates the agent to prioritize extensive observations in the initial stages, subsequently shifting its focus to grasp the target object.

Through extensive experiments on two mainstream simulators, Habitat [15] and Isaac Gym [16], which include a diverse range of environments and objects, we demonstrate that our method outperforms mainstream methods on both abstract grasping metric and realistic grasping pose metric. Moreover, real-world evaluations of our approach further showcase the robustness and effectiveness of our methodology.

In summary, the contributions of our work include:

- We propose an online grasping fusion module to fuse predicted grasping poses to obtain temporally consistent grasping poses for erasability observation.
- We design an observe-to-grasp reward to effectively encourage agents to execute actions that balance both observation and grasping.
- We present a graspability-aware mobile manipulation RL system, achieving robust performance on both simulators and real-world environments.

II. Related work

Traditional mobile manipulation methods. For decades, the field of robotics has experienced significant growth in the advancement of mobile manipulation methods [1], [2]. Traditional researches [17], [7], [18] leverage scene analysis and motion planning, aiming to devise strategies for efficient task execution. However, these approaches assume access to explicit secure information regarding the environments, such as detailed maps with obstacle locations [17], [9], [19], precise object coordinates [18], [7], [8], [20].

Learning-based mobile manipulation methods. Mobile manipulation agents are trained to possess the capability to observe and interact within various scenes. One of the primary capabilities of these agents is to observe the target [21], [22], achieved by encouraging the agent to obtain multiple observations of the target object. Another crucial capability is to maneuver its arm to approach the target object, often referred to as reachability [10], [11], [23]. Regarding graspability, advanced methods do not rely on predicted grasping poses [24] or object pose estimation [25], [26], as these approaches lack temporal perception of graspability. In this paper, we introduce an online grasping pose fusion module to fuse the predicted grasping poses for encoding graspability states, enabling our method to be graspability-aware and showcasing enhanced performance.

III. Problem Statement and Method Overview

Mobile manipulation task. Given a target object location p_{goal}, the robot is tasked with navigating through an unknown environment to effectively approach and grasp the target object. We follow the mainstream setup presented in [11], [10]. The robot is equipped with a mobile base, an arm, and a parallel gripper. Two RGB-D cameras are mounted: one to the head of mobile base ($D_{\text{head}}, I_{\text{head}}$) and the other to the gripper (D_{grip},I_{grip}), where d and c represent the depth image and color image, respectively. The robot utilizes a 3-DoF configuration for its mobile base in SE(3), coupled with an $(x+1)$-DoF arm. In detail, $x = 6$ for the Spot arm and the Unitree Z1 arm, and $x = 7$ for the Fetch robots, further augmented by a 1-DoF gripper for object grasping.

Overview. Figure 2 provides an overview of our proposed graspability-aware mobile manipulation approach. To facilitate graspability, our method processes the depth image I^d_{grip} and leverages an off-the-shelf grasping module, GSNet [14], to predict grasping poses (Section IV-A). These predicted grasping poses are then fused online (Section IV-B) and encoded as the graspability state $\mathcal{S}_{\text{grasp}}$. Subsequently, our method can learn the graspability-aware mobile manipulation policy $\pi(\mathcal{A}_{\text{base}}, \mathcal{A}_{\text{arm}}, \mathcal{A}_{\text{grip}} | \mathcal{S}_{\text{grasp}}, \mathcal{S}_{\text{visual}}, \mathcal{S}_{\text{state}})$ through reinforcement learning, incorporating visual information $\mathcal{S}_{\text{visual}}$ and state information $\mathcal{S}_{\text{state}}$ (Section V-A).

In our method, the policy generates a 3-DoF SE(3) velocity for mobile base control $\mathcal{A}_{\text{base}}$, a 6-DoF residual adjustment for current arm joints $\mathcal{A}_{\text{arm}}$, and a 1-DoF switch to control the gripper $\mathcal{A}_{\text{grip}}$. During the RL training process, a composite observe-to-grasp reward is employed, incorporating both the grasping observation reward, g_{go}, and the gripper-to-grasping poses reward, g_{gg}. This reward system motivates the robot to prioritize actions based on meticulous observations, guiding it towards more optimal grasping poses (Section V-B). For ease of description, the notations used in this paper default to the world coordinate system.

IV. Graspability Estimation

To obtain accurate and complete graspability of an object, we propose to predicts grasping poses at each timesteps (see IV-A) and online fuses them together while eliminating invalid ones (see IV-B).

A. Grasping pose prediction

During mobile manipulation, our graspability-aware agent constantly captures observations from RGB-D cameras and performs online predictions of grasping poses. At each time step t, the agent moves according to the policy, and then online obtains new observations of the scene. To obtain

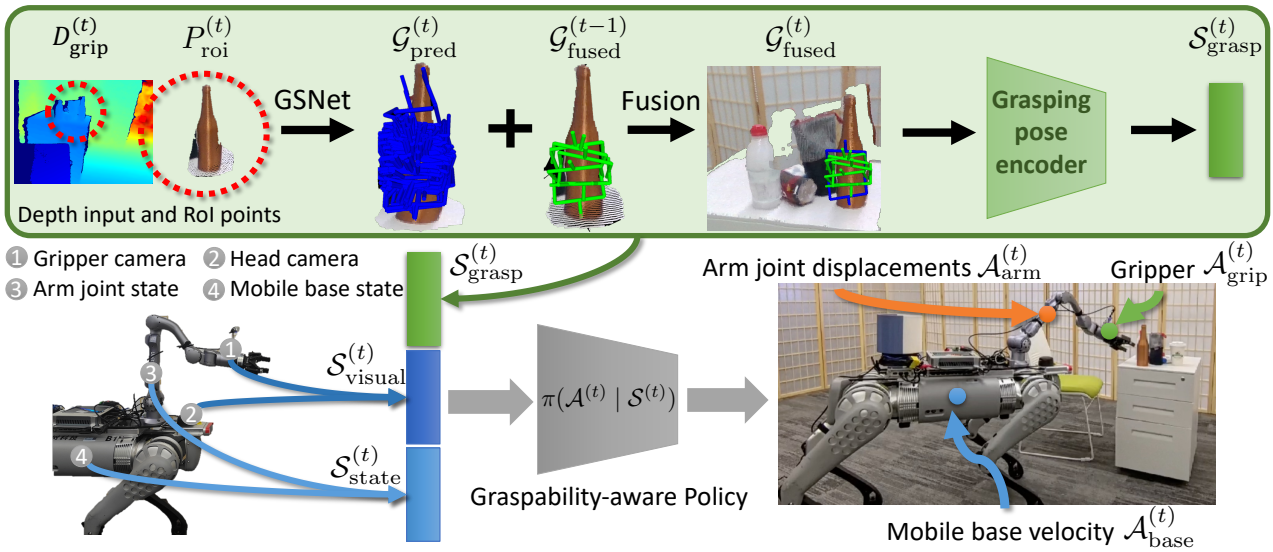

Fig. 2: Method overview. Our method processes the gripper depth map $D^{(t)}_{\text{grip}}$ to region-of-interest point cloud $P^{(t)}_{\text{roi}}$, which then be sent to GSNet for predicting grasping poses $\mathcal{G}^{(t)}_{\text{pred}}$. The $\mathcal{G}^{(t)}_{\text{pred}}$ will then be integrated into the so far fused grasping poses $\mathcal{G}^{(t-1)}_{\text{fused}}$ to obtain $\mathcal{G}^{(t)}_{\text{fused}}$. The $\mathcal{G}^{(t)}_{\text{fused}}$ will be encoded as $\mathcal{S}^{(t)}_{\text{grasp}}$, along with $\mathcal{S}^{(t)}_{\text{visual}}$ and $\mathcal{S}^{(t)}_{\text{state}}$ for learning $\pi(\mathcal{A}^{(t)} \mid \mathcal{S}^{(t)})$.

grasping poses $\mathcal{G}^{(t)}$ based on the agent's observation, we leverage GSNet [14], which is trained on a billion-scale real-world dataset [27] and demonstrates robust performance in novel scenes [28], [29]. Given the target object's location $p^{(t)}_{\text{goal}}$, we can extract a sphere-shaped region-of-interest (RoI) point cloud $P^{(t)}_{\text{roi}}$ from the 3D points $P^{(t)}_{\text{grip}}$. These points are obtained by back-projecting the depth map $D^{(t)}_{\text{grip}}$ and then be transformed to the world coordinate system. Therefore, the $P^{(t)}_{\text{roi}}$ can be formulated as follows:

$$P^{(t)}_{\text{roi}} = \{P^{(t)}_{\text{grip}} \in \mathbb{R}^3 \mid \|P^{(t)}_{\text{grip}} - p_{\text{goal}}\|_2 < \tau\}, \tag{1}$$

where τ represents the maximum distance from the target object's location. Empirically, we set $\tau = 10$ cm. Then we can utilize GSNet $G(\cdot)$ to predict the grasping poses as follows:

$$\mathcal{G}^{(t)}_{\text{pred}} = G(P^{(t)}_{\text{roi}}) = \{q^{(t)}_i, p^{(t)}_i, s^{(t)}_i\}_{i=1:n}, \tag{2}$$

q, p, and s correspond to the orientation (represented by quaternion), position, and score of the predicted poses, respectively. Additionally, n denotes the number of grasping poses, which may be zero if the quality of $P^{(t)}_{\text{grip}}$ is low.

B. Online Grasping Fusion Module.

Note that these predicted grasping poses $\mathcal{G}^{(t)}_{\text{pred}}$ from each timestep can be noisy due to occlusions and overlapping significantly with previous ones. As a result, directly combining all predicted grasping poses $\mathcal{G}^{(0)}_{\text{pred}} \cup ... \cup \mathcal{G}^{(t)}_{\text{pred}}$ to obtain graspability may lead to many errors and a high degree of redundancy, posing further challenges to policy learning. To address this, we introduce an online grasping fusion module, which is designed to maintain temporally consistent grasping pose observations. An illustration of the fusion module is presented in Figure 3.

To store and track the predicted grasping poses, our fusion module first partitions the 3D space into a uniform grid in the center of the target object location p_{goal}. Here, we utilize a $64 \times 64 \times 64$ cube grid (3cm voxel). For efficient memory usage, we use an off-the-shelf indexing table algorithm [30] to dynamically allocate the voxels. For each grasping pose $\mathcal{G}^{(t)}_{\text{fused}} = \{o^{(t)}_i, p^{(t)}_i, s^{(t)}_i\}_{i=1:n}$, its corresponding voxel v can be found by $v^{min}_{x/y/z} < p^{(t)}_{i,x/y/z} < v^{max}_{x/y/z}$. where $x/y/z$ represents the comparisons across the three axes. With this approach, each voxel stores a collection of grasping poses, allowing for easy identification of neighboring grasping poses within a specified 3D range.

However, these grasping poses are unorganized (lacking spatial information to one another) and redundant, requiring further refinement by merging grasping poses. Recognizing that grasping poses within a voxel are more sensitive to orientation than to translation [29], our method retains only those grasping poses that exhibit a considerable angular difference compared to other existing poses within the same voxel.

Specifically, our method iteratively calculates the angle between new grasping poses $\mathcal{G}^{(t)}_{\text{pred}}$ and fused grasping poses $\mathcal{G}^{(t-1)}_{\text{fused}}$ belonging to the same voxel. If the angle exceeds τ_{angle}, the new grasping pose will be added to the voxel grasping set. Note that, we empirically set the $\tau_{\text{angle}} = \pi/4$ through experiments. Moreover, grasping poses are mainly located within voxels that are close to the target objects, making the grasping pose query highly efficient.

Given such a fused grasping grid, we can efficiently traverse saved grasping poses to identify the $\{q^{(t-1)}_{\text{fused}}, p^{(t-1)}_{\text{fused}}, s^{(t-1)}_{\text{fused}}\}$ with the smallest angle differences. If the angle is less than τ_{angle}, we utilize a weighted average with the weight determined by the score of grasping poses,

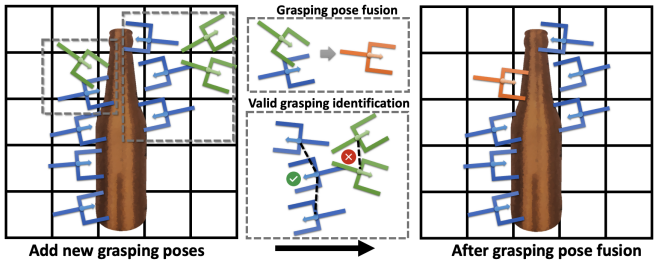

Fig. 3: An illustration of grasping pose fusion and valid grasping pose identification. New added grasps and previous fused grasping poses are indicated in green and blue, respectively.

represented as $w = s_{\text{new}}/(s_{\text{fused}} + s_{\text{new}})$:

$$\begin{aligned} p_{\text{fused}}^{(t)} &= (1-w)p_{\text{fused}}^{(t-1)} + wp_{\text{pred}}^{(t)}, \\ q_{\text{fused}}^{(t)} &= \frac{\sin((1-w)\theta)}{\sin(\theta)} q_{\text{fused}}^{(t-1)} + \frac{\sin(w\theta)}{\sin(\theta)} q_{\text{pred}}^{(t)}, \\ s_{\text{fused}}^{(t)} &= s_{\text{fused}}^{(t-1)} + s_{\text{pred}}^{(t)}, \end{aligned} \tag{3}$$

where the q_{fused} is renormalized to 1 to satisfy the quaternion constraint. Note that, the updated orientation q_{fused} may break the angle distance constraints, therefore new fused grasping pose will then be compared with other grasping poses within the same voxel until all the grasping poses satisfy the angle distance threshold. We find that the recursive grasping pose fusion is a rare occurrence due to the sparse distribution of grasping poses (typically containing approximately 4 grasping poses) in our implementation. This fusion operation leads to complete and accurate fused grasping pose results.

Valid grasping pose identification. Due to occlusions arising from observational viewpoints, the prediction outcomes may include invalid grasping poses. These grasping poses could be distant from or oriented away from the target object, as depicted in Figure 3. To remove the outlier grasping poses, we follow the basic fact that the grasping pose should be densely and tangentially distributed along the target objects. Hence, we design a grasping pose consistency verification algorithm to evaluate the density of the grasping cluster.

Specifically, our method connects neighboring grasping poses to form a 'grasping cluster' based on two criteria: (1) *Distance*: The grasping poses should locate in adjacent voxels. (2) *Orientation*: The angular difference between orientations should be less than $1.5\tau_{\text{angle}}$. Upon establishing these connections, clusters containing fewer grasping poses than τ_{count} are eliminated. These smaller clusters typically consist of outlier or error-prone poses. The resulting set preserves only the high-quality grasping poses for graspability observation.

V. Graspability-aware policy learning.

A. Graspability states

Taking the output grasping poses (a.k.a. graspablity) from the online grasping fusion module, we propose to encode them as a part of the state for RL.

Note that this encoding is highly non-trivial because the grasping poses are unordered high-dimensional vectors. As a part of the pose, the quaternion q that represents grasping pose orientation is discontinuous among SO(3) manifold [31], [32], further creating difficulties. To tackle these challenges, we first convert quaternions into continuous 6D rotation representation $F(\cdot)$ [31]; and then leverage an order-invariant neural network PointNet [33] (composed by three MLP layers M and a maxpooling layer) for state encoding:

$$S_{\text{grasp}} = maxpooling\{M(p_{\text{fused}}, F(q_{\text{fused}}), s_{\text{fused}})\}, \tag{4}$$

A corner case we need to handle: at the beginning of mobile manipulation, the camera hasn't observed the target object yet, there is no grasping pose available . In this case, we need our graspability-aware approach to degenerate into being reachability-aware, directing the agent towards the target object's location.

We thus leverage the same encoding method of graspability states (Equation 4), and substitute the grasping pose with the target object location p_{goal}, uniform sampled orientation q_{sample} within SO(3), and a constant low score s_{reach} (set to 0.1) to form the reachability observation $S_{\text{reach}} \approx S_{\text{grasp}}$:

$$S_{\text{reach}} = maxpooling\{M(p_{\text{goal}}, F(q_{\text{sample}}^{(k)}), s_{\text{reach}})\}_{k=1:K}. \tag{5}$$

The uniformly sampled orientation $q_{\text{sample}}^{(k)}$ ($K = 128$) encourages the arm to reach the target object from any direction until the online grasping pose fusion module supplies valid grasping poses. Besides the graspability state, we also encoded the visual information $\mathcal{S}_{\text{visual}}$ from both the front camera and gripper cameras and $\mathcal{S}_{\text{state}}$ from joints state encoding. These follow the same methodology as described in [15]. Consequently, the graspability-aware policy can be expressed as $\pi(\mathcal{A}_{\text{base}}, \mathcal{A}_{\text{arm}}, \mathcal{A}_{\text{grip}} | \mathcal{S}_{\text{grasp}}, \mathcal{S}_{\text{visual}}, \mathcal{S}_{\text{state}})$.

B. Observe-to-grasp reward for RL training.

With the graspability observation, the agent is required to make a balance between observing and grasping during mobile manipulation. To this end, we design an observe-to-grasp reward mechanism, consisting of a grasping observation reward r_{go} and a gripper-to-grasping pose reward r_{gg}:

$$\begin{aligned} r_{\text{go}}^{(t)} &= \sum s_{\text{fused}}^{(t)} - \sum s_{\text{fused}}^{(t-1)}, \\ r_{\text{gg}}^{(t)} &= D_{\text{gg}}^{(t-1)} - D_{\text{gg}}^{(t)}, \end{aligned} \tag{6}$$

where $s_{\text{fused}}^{(t)} \in \mathcal{G}_{\text{fused}}^{(t)}$, and D_{gg} is the gripper to grasping pose evaluation function. For r_{go}, we leverage the online grasping fusion module, directly assessing the enhancement of graspability observation. And for the gripper-to-grasping reward r_{gg}, the gripper is encouraged to approach to fused grasping pose with a high score:

$$D_{\text{gg}} = \min\{\ e^{-s_{\text{fused}}}(\beta_1 \|p_{\text{grip}} - p_{\text{fused}}\|_2 + \beta_2 \theta(q_{\text{grip}}, q_{\text{fused}}))\}, \tag{7}$$

where $\theta(\cdot)$ compute the interval angle between two rotations (radius). And β_1 and β_2 are set to 0.3 and 0.2, respectively.

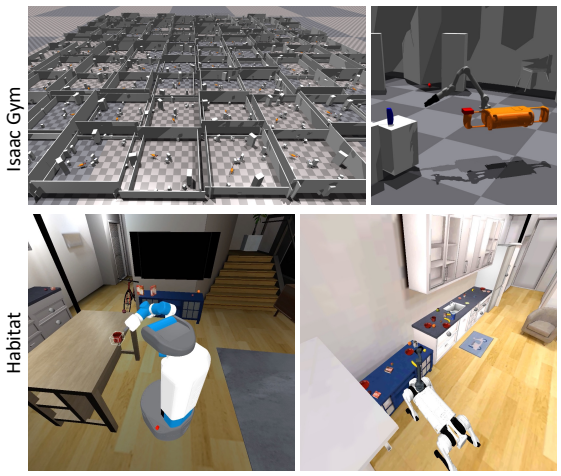

Fig. 4: Simulation setup. First row: Unitree B1 + Z1 robot dog in Isaac Gym. Second row: Fetch Robot (left) and Unitree B1 + Z1 robot dog (right) in Habitat Simulator.

Finally, we can formulate our reward as:

$$r_{\text{o2g}}^{(t)} = (1-\sigma) r_{\text{go}}^{(t)} + \sigma r_{\text{gg}}^{(t)}, \\ \sigma = \frac{1}{1+e^{0.5-t/t_{\max}}}, \tag{8}$$

and the σ is a logistic sigmoid function related to the execution steps. This approach ensures that the observe-to-grasp reward initially encourages the agent to observe, and as more steps are taken, gradually shifts to promoting grasping actions. Such a reward is dense and adaptive, guiding the agent's learning more effectively. In addition to the observe-to-grasp reward, we also leverage a sparse success reward $r_{\text{success}} = 10$ ($D_{\text{ee}} < 15$cm), a slack penalty $r_{\text{slack}} = 10^{-2}$ and a force penalty $r_{\text{force}} = 10^{-4}$. These additional rewards enhance the stability of the learning process.

Implementation details. We use Habitat 2.0 as our training simulator. Our method predicts sample actions every 20 steps and uses Proximal Policy Optimization (PPO) [34] for agent training. Given the substantial overlap between consecutive frames for increased efficiency, our method predicts grasping poses every 10 frames. We uniformly sample 128 fused grasping poses (allowing repetition if the number of fused grasping poses is fewer than 128) for $\mathcal{S}_{\text{grasp}}$ (Equ. 4). In the absence of grasping poses, we resort to using $\mathcal{S}_{\text{reach}}$ (Equ. 5). These parameters, including thresholds and size, are adopted through experiments and can be further improved with careful adjustments for specific scenes. Any parameters not detailed in this paper are adopted from [15].

VI. Experiments

A. Experimental setup

Synthetic environment setup. We evaluate our method on the Habitat 2.0 simulator [15] and Isaac Gym [16]. The Habitat 2.0 features photo-realistic reconstructions of apartment scenes from ReplicaCAD [15], stuffed by objects from YCB dataset [36]. We use both two datasets on Habitat, including 1000 Habitat episodes and a self-build 1,000 challenging episode dataset with cluttered object layouts (approximately 10 objects on each receptacle). For Isaac Gym, we create a mobile manipulation environment similar to [11]. *The episode data will be released to the public.*

Real-world environment setup. We deploy the B1+Z1 robot in the real world to grasp the target object from a cluttered receptacle (with 4-6 objects) while performing obstacle avodiance. In detail, We mount an Azure Kinect DK on the head of the B1 robot dog and a Realsense D415 on the gripper of the Z1 arm. During the experiments, we make use of ORB-SLAM3 [37] to obtain 6D grasping poses based on the gripper camera observations. Note that the extrinsic parameters between the cameras and the arm, as well as the robot base, are pre-calibrated.

Baselines. Given the intricacy of the task, ensuring a fair comparison among all mainstream methods is a formidable challenge. Hence, we focus on comparing with other methods that are most pertinent to our approach and have been assessed within the same simulator environment. Specifically, we consider:

1) Multi-skill Mobile Manipulation **(M3)** [35]: A modular method which incorporates mobility for enhanced flexibility in object interactions.
2) Habitat-Baselines **(HB)**: A standard baseline method provided by Habitat 2.0.
3) Reachability-aware policy **(ReachMM)**: An approach that leverages the reachability encoded state $\mathcal{S}_{\text{reach}}$ as defined in Equ. 5.
4) Non-fusion graspability-aware policy **(GAMMA without fusion)**: A method which doesn't make use of our proposed OGFM module and predicts grasp poses for every frame.

Metrics. We measure the episode when the grasp action is called. In simulator, we consider two following metrics: (1) *Gaze Success Rate* (GazeSR), an episode is deemed successful if the distance between the arm camera position and the target object position is within the 15 cm and the angle between the camera ray and the object-to-camera ray is less than 10°. (2) *Grasp Success Rate* (GraspSR), Success is achieved if the gripper's pose closely matches any densely annotated grasping pose, with deviations less than 10 cm in distance and 10° in angle.

B. Results

Comparsion on Habitat simulator. To comprehensively evaluate our method, we conduct extensive experiments in the Habitat simulator. These results are demonstrated in Table I. Here, we find that our method achieves state-of-the-art performance in all challenging settings. Moreover, we find other methods have an apparent performance drop from non-cluttered episodes to cluttered episodes because the cluttered scenes required more accurate gripper poses to avoid mistakingly grasping other objects. Our methods benefit from temporally consistent grasping poses and directly learn how to drive the gripper to the grasping pose, showing only a

TABLE I: Quantitative comparison and ablation study on the Habitat simulator on both non-cluttered and cluttered episodes.

Methods	Fetch Robot				Unitree B1+Z1 arm			
	Non-cluttered env.		Cluttered env.		Non-cluttered env.		Cluttered env.	
	GazeSR	GraspSR	GazeSR	GraspSR	GazeSR	GraspSR	GazeSR	GraspSR
HB[15]	45.8	-	22.0	-	57.0	-	31.9	-
M3[35]	55.2	49.4	43.6	39.5	-	-	-	-
ReachMM	36.5	29.1	18.2	11.3	39.1	29.4	21.5	12.7
GAMMA without fusion	10.3	7.2	8.1	3.2	10.9	7.3	8.1	2.1
GAMMA (Ours)	**67.3**	**62.4**	**60.7**	**57.1**	**71.0**	**69.2**	**66.3**	**64.5**

TABLE II: Comparisons on Isaac Gym simulator and real-world experiments under challenging conditions (with obstacles and cluttered objects).

Methods	Isaac Gym Simulator		Real-world Env.
	GazeSR	GraspSR	SR
GAMMA without fusion	89.0%	29.6%	53.3%
GAMMA (Ours)	**96.6%**	**86.6%**	**73.3%**

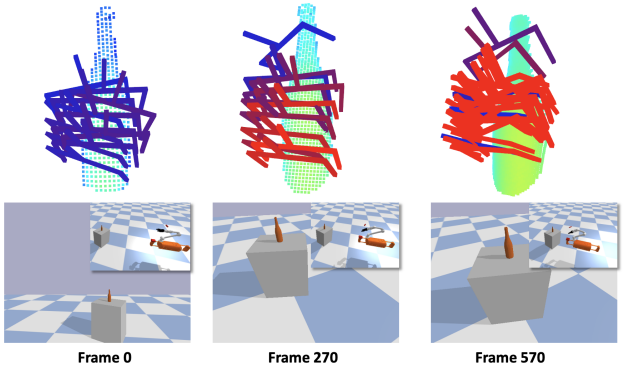

Fig. 5: Visualization of the quality of the fused grasping poses during mobile manipulation. The grasping are color-coded based on their graspability feature (to red the better).

small performance drop. Another interesting finding is that the GazeSR and GraspSR have large performance gap in many methods that use abstract grasp (M3 [35], HB [15], ReachMM), which proved that the grasping poses perform better to motivate grasping strategy.

Comparsion in Isaac Gym and real-world environments. We deploy our policy, trained on the Isaac Gym simulator, to real-world environments (Unitree B1 + Z1 arm). The results on both Isaac Gym and the real-world environment are reported in Table II. We evaluate two widely-demanded skills grasping within cluttered objects and avoiding obstacles simultaneously. Our findings indicate that directly deploying our method yields robust performance in real-world settings. Furthermore, when comparing with the GAMMA without fusion, we observed a performance drop from GazeSR to GraspSR, a trend also identified in the Habitat [15] environment(Table I). This supports the importance of complete and accurate graspability in RL training.

Mobile manipulation process analysis. We plot two main observations related to graspability-aware mobile manipulation: the number of grasping poses and the distance from the gripper to these poses (Equ. 7). The results, showcased in Fig.6, originate from the Habitat environment. These data

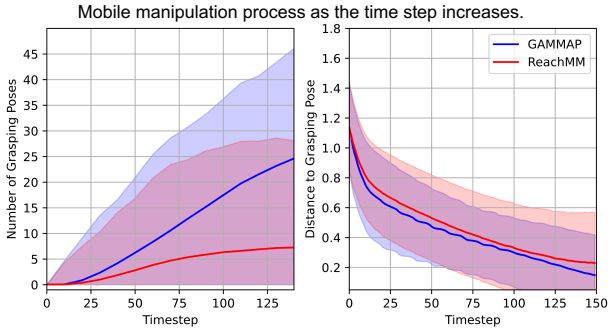

Fig. 6: Compared with ReachMM in terms of (1) the number of grasping poses and (2) the distance to the grasping poses.

compellingly indicate that our policy adeptly directs the agent to identify an increased number of grasping poses while simultaneously nearing those specified positions effectively. Additionally, to elucidate the significance of the graspability state, we color-code the integrated grasping pose and employ a fine voxel size of 1cm to ensure a detailed visualization. Throughout the mobile manipulation process, it's evident that the agent gains confidence while observing and approaching the grasping poses.

Runtime and memory analysis. The online grasping fusion module is highly efficient, both in terms of memory usage and execution speed. It demands less than 0.1 GB of memory for a single scene and operates at real-time frame rates. For training, our method necessitates 36 GPU hours on an A100 to attain state-of-the-art performance.

VII. CONCLUSIONS

We introduce a graspability-aware mobile manipulation policy, enabling agents to achieve robust and accurate mobile grasping. This capability is powered by an online grasping pose fusion module, which fuses online predicted grasping poses. This leads to temporally consistent grasping pose observations, facilitating learning graspability. Our method demonstrates superior performance on Habitat. We also deploy our approach on the Unitree B1 with Z1 arm for real-world experiments, further showcasing the robustness of our methodology. In the future, we would like to explore the potential of this framework in highly dynamic environments or for long-horizon tasks.

VIII. ACKNOWLEDGMENT

We thank all reviewers for their insightful comments and valuable suggestions. This project is supported by the National Natural Science Foundation of China (No. 62306016) and Beijing Academy of Artificial Intelligence (BAAI).

REFERENCES

[1] O. Brock, J. Park, and M. Toussaint, "Mobility and manipulation," *Springer Handbook of Robotics*, pp. 1007–1036, 2016.

[2] P. Hebert, M. Bajracharya, J. Ma, N. Hudson, A. Aydemir, J. Reid, C. Bergh, J. Borders, M. Frost, M. Hagman *et al.*, "Mobile manipulation and mobility as manipulation—design and algorithms of robosimian," *Journal of Field Robotics*, vol. 32, no. 2, pp. 255–274, 2015.

[3] D. Watkins-Valls, P. K. Allen, H. Maia, M. Seshadri, J. Sanabria, N. Waytowich, and J. Varley, "Mobile manipulation leveraging multiple views," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 4585–4592.

[4] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke *et al.*, "Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation," *arXiv preprint arXiv:1806.10293*, 2018.

[5] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg, "Learning ambidextrous robot grasping policies," *Science Robotics*, vol. 4, no. 26, p. eaau4984, 2019.

[6] K. C. Vivaldini, J. P. Galdames, T. S. Bueno, R. C. Araújo, R. M. Sobral, M. Becker, and G. A. Caurin, "Robotic forklifts for intelligent warehouses: Routing, path planning, and auto-localization," in *2010 IEEE International Conference on Industrial Technology*. IEEE, 2010, pp. 1463–1468.

[7] F. Sun, Y. Chen, Y. Wu, L. Li, and X. Ren, "Motion planning and cooperative manipulation for mobile robots with dual arms," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, no. 6, pp. 1345–1356, 2022.

[8] H. Chen, X. Zang, Y. Liu, X. Zhang, and J. Zhao, "A hierarchical motion planning method for mobile manipulator," *Sensors*, vol. 23, no. 15, p. 6952, 2023.

[9] S. Patki, E. Fahnestock, T. M. Howard, and M. R. Walter, "Language-guided semantic mapping and mobile manipulation in partially observable environments," in *Conference on Robot Learning*. PMLR, 2020, pp. 1201–1210.

[10] N. Yokoyama, A. W. Clegg, E. Undersander, S. Ha, D. Batra, and A. Rai, "Adaptive skill coordination for robotic mobile manipulation," *arXiv preprint arXiv:2304.00410*, 2023.

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

[12] C. Wang, Q. Zhang, Q. Tian, S. Li, X. Wang, D. Lane, Y. Petillot, and S. Wang, "Learning mobile manipulation through deep reinforcement learning," *Sensors*, vol. 20, no. 3, p. 939, 2020.

[13] C. Sun, J. Orbik, C. M. Devin, B. H. Yang, A. Gupta, G. Berseth, and S. Levine, "Fully autonomous real-world reinforcement learning with applications to mobile manipulation," in *Conference on Robot Learning*. PMLR, 2022, pp. 308–319.

[14] C. Wang, H. Fang, M. Gou, H. Fang, J. Gao, C. Lu, and S. J. Tong, "Graspness discovery in clutters for fast and accurate grasp detection," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15 944–15 953, 2021.

[15] A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. S. Chaplot, O. Maksymets, A. Gokaslan, V. Vondrus, S. Dharur, F. Meier, W. Galuba, A. X. Chang, Z. Kira, V. Koltun, J. Malik, M. Savva, and D. Batra, "Habitat 2.0: Training home assistants to rearrange their habitat," *ArXiv*, vol. abs/2106.14405, 2021.

[16] 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," *arXiv preprint arXiv:2108.10470*, 2021.

[17] A. Heins, M. Jakob, and A. P. Schoellig, "Mobile manipulation in unknown environments with differential inverse kinematics control," in *2021 18th Conference on Robots and Vision (CRV)*, 2021, pp. 64–71.

[18] C. R. Garrett, C. Paxton, T. Lozano-Pérez, L. P. Kaelbling, and D. Fox, "Online replanning in belief space for partially observable task and motion problems," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 5678–5684.

[19] L. Zheng, C. Zhu, J. Zhang, H. Zhao, H. Huang, M. Niessner, and K. Xu, "Active scene understanding via online semantic reconstruction," in *Computer Graphics Forum*, vol. 38, no. 7. Wiley Online Library, 2019, pp. 103–114.

[20] J. Zhang, L. Dai, F. Meng, Q. Fan, X. Chen, K. Xu, and H. Wang, "3d-aware object goal navigation via simultaneous exploration and identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6672–6682.

[21] K. He, R. Newbury, T. Tran, J. Haviland, B. Burgess-Limerick, D. Kulić, P. Corke, and A. Cosgun, "Visibility maximization controller for robotic manipulation," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 8479–8486, 2022.

[22] K. Xu, Y. Shi, L. Zheng, J. Zhang, M. Liu, H. Huang, H. Su, D. Cohen-Or, and B. Chen, "3d attention-driven depth acquisition for object identification," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, pp. 1–14, 2016.

[23] Q. Dai, Y. Zhu, Y. Geng, C. Ruan, J. Zhang, and H. Wang, "Graspnerf: multiview-based 6-dof grasp detection for transparent and specular objects using generalizable nerf," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 1757–1763.

[24] M. Bajracharya, J. Borders, R. Cheng, D. M. Helmick, L. Kaul, D. Kruse, J. Leichty, J. Ma, C. Matl, F. Michel, C. Papazov, J. Petersen, K. Shankar, and M. Tjersland, "Demonstrating mobile manipulation in the wild: A metrics-driven approach," *Robotics: Science and Systems XIX*, 2023.

[25] K. Yamazaki, S. Suzuki, and Y. Kuribayashi, "Approaching motion planning for mobile manipulators considering the uncertainty of self-positioning and object's pose estimation," *Robotics and Autonomous Systems*, vol. 158, p. 104232, 2022.

[26] A. B. Chowdhury, J. Li, and D. J. Cappelleri, "Neural network-based pose estimation approaches for mobile manipulation," *Journal of Mechanisms and Robotics*, vol. 15, no. 1, p. 011009, 2023.

[27] H. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11 441–11 450, 2020.

[28] M. Gou, H. Pan, H. Fang, Z. Liu, C. Lu, and P. Tan, "Unseen object 6d pose estimation: A benchmark and baselines," *ArXiv*, vol. abs/2206.11808, 2022.

[29] H. 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," *ArXiv*, vol. abs/2212.08333, 2022.

[30] J. Zhang, C. Zhu, L. Zheng, and K. Xu, "Fusion-aware point convolution for online semantic 3d scene segmentation," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4533–4542, 2020.

[31] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5738–5746, 2018.

[32] J. Chen, Y. Yin, T. Birdal, B. Chen, L. J. Guibas, and H. Wang, "Projective manifold gradient layer for deep rotation regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6646–6655.

[33] C. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 77–85, 2016.

[34] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *ArXiv*, vol. abs/1707.06347, 2017.

[35] J. Gu, D. S. Chaplot, H. Su, and J. Malik, "Multi-skill mobile manipulation for object rearrangement," *ArXiv*, vol. abs/2209.02778, 2022.

[36] B. Çalli, A. Singh, A. Walsman, S. S. Srinivasa, P. Abbeel, and A. M. Dollar, "The ycb object and model set: Towards common benchmarks for manipulation research," *2015 International Conference on Advanced Robotics (ICAR)*, pp. 510–517, 2015.

[37] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.