Dexterous Functional Pre-Grasp Manipulation with Diffusion Policy

Tianhao Wu, Yunchong Gan, Mingdong Wu, Jingbo Cheng, Yaodong Yang, Yixin Zhu, and Hao Dong

***Abstract*—In real-world scenarios, objects often require repositioning and reorientation before they can be grasped, a process known as pre-grasp manipulation. Learning universal dexterous functional pre-grasp manipulation requires precise control over the relative position, orientation, and contact between the hand and object while generalizing to diverse dynamic scenarios with varying objects and goal poses. To address this challenge, we propose a teacher-student learning approach that utilizes a novel mutual reward, incentivizing agents to optimize three key criteria jointly. Additionally, we introduce a pipeline that employs a mixture-of-experts strategy to learn diverse manipulation policies, followed by a diffusion policy to capture complex action distributions from these experts. Our method achieves a success rate of 72.6% across more than 30 object categories by leveraging extrinsic dexterity and adjusting from feedback.**

I. Introduction

Objects in human daily life serve various functions, requiring different functional grasp poses. For instance, when using a spray bottle, one typically positions fingers on the trigger, whereas when passing the bottle to another person, one typically grasps the body. Current works [1, 2] mainly focus on training models to predict the functional grasp pose or further incorporate Reinforcement Learning (RL) for grasp execution and post-grasp usage [3]. However, these works assume objects are already in highly graspable poses, overlooking the fact that objects are often not positioned with high functional graspability in the real world. For instance, a spray bottle might be lying flat on a table, making it challenging to grasp directly for its intended use. Humans typically manipulate the object into a pre-grasp pose through continuous reorientation and repositioning, a process known as pre-grasp manipulation [4, 5]. Unlike conventional pre-grasp manipulation, transitioning objects from ungraspable to graspable states, dexterous functional pre-grasp manipulation further requires both the dexterous hand and the object to satisfy a specific goal pose for subsequent functional grasping.

Dexterous functional pre-grasp manipulation of diverse objects involves intricate interactions with objects and environments, demanding closed-loop dexterous manipulation skills. Existing methods [6–8] rely on RL to train policies for general dexterous manipulation, typically focusing on satisfying the goal orientation and/or position of the objects. However, for functional use, goals must precisely align with the relative position, orientation, and contact between the dexterous hand and the object. This results in an exceedingly small solution space, making it challenging for RL agents to explore successful policies. In this scenario, conventional approaches, such as adding distance rewards [7, 9, 10], struggle. Simply adding multiple distance rewards often makes RL agents trapped in local minima, failing to devise manipulation policies that meet all criteria. It is also impossible to design specific rewards according to each object [11], since we need to generalize to diverse objects with diverse poses. Such generalization is also challenging for RL agents to learn from scratch [12–14].

To address the problem, we propose a novel mutual reward that computes a scale according to the distance of each criterion and uses the lowest scale to restrict all distance rewards, preventing the agent from getting stuck at a local minimum. Moreover, to facilitate generalization across diverse objects and functional grasp poses, we employ the teacher-student learning framework [7, 14] by training Mixture of Experts (MoE). The MoE generates diverse manipulation behavior, leading to a complex action distribution, especially for a dexterous hand with high Degrees of Freedoms (DoFs). Thus, we propose using a diffusion policy [15], which has shown great generative modeling ability to capture such complex action distributions.

Through mutual reward and mixture-of-experts training, we observe significant improvements in teacher policy learning. When distilling the teacher policy into a single-student policy using a diffusion policy, our approach achieves teacher-level performance even without object geometry. Our learned policy demonstrates adept use of extrinsic dexterity, such as leveraging tables and inertia to manipulate objects effectively, and also learns to adjust from feedback. These capabilities enhance the policy's ability to generalize across diverse objects.

In summary, our contributions are as follows: (i) We propose a novel mutual reward to address the local minimum problem, significantly improving teacher policy learning. (ii) We propose a pipeline integrating MoE and diffusion policies to learn complex and general dexterous manipulation policies. (iii) We achieve a general dexterous functional pre-grasp manipulation policy with a 72.6% success rate across 30+ object categories encompassing 1400+ objects and 10k+ goal poses.

Manuscript received xxx April 2024; accepted xx xxxxxx 2024. Date of publication xx xxxxxx 2024; date of current version xx xxxxxx 2024. This work was supported in part by the National Natural Science Foundation of China (62376006) and the National Youth Talent Support Program (8200800081). T. Wu and Y. Gan contributed equally. Corresponding author: *Hao Dong*.)

Tianhao Wu, Mingdong Wu, and Hao Dong (e-mail: hao.dong@pku.edu.cn) are with the Center on Frontiers of Computing Studies, School of Computer Science, Peking University, Beijing 100871, China, also with PKU-Agibot Lab, School of Computer Science, Peking University, Beijing 100871, China, and also with National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University, Beijing 100871, China.

Yunchong Gan and Jingbo Cheng are with the Center on Frontiers of Computing Studies, School of Computer Science, Peking University, Beijing 100871, China.

Yaodong Yang and Yixin Zhu are with the Institute for Artificial Intelligence, Peking University, Beijing 100871, China.

See additional material on https://unidexfpm.github.io

This letter has supplementary downloadable material available at https://doi.org/10.1109/LRA.2024.xxxxxx, provided by the authors.

Digital Object Identifier 10.1109/LRA.2024.xxxxxx

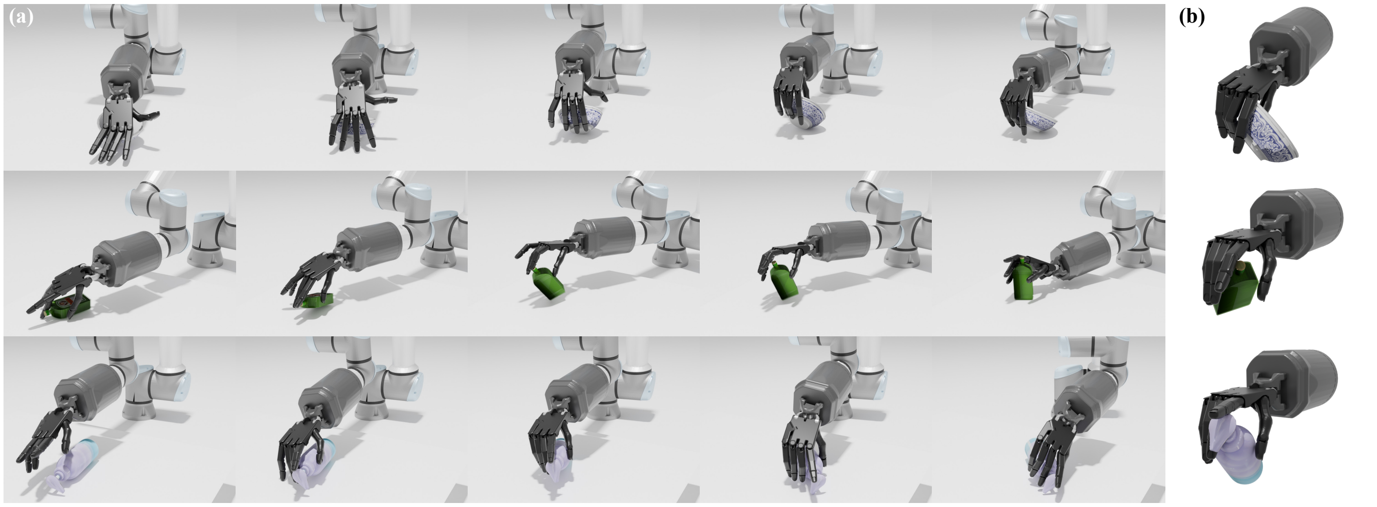

Fig. 1: **Our closed-loop manipulation policy continuously repositions and reorients diverse objects to match the functional grasp goal poses successfully.** (a) The dexterous functional pre-grasp manipulation. (b) The functional grasp goal poses.

II. Related Work

A. Dexterous Functional Grasping

Dexterous functional grasping is crucial for humans due to the diverse functionalities of objects in real-world scenarios. This encompasses both functional grasp pose generation and execution. Since the functionality of the objects is related to human design, recent frameworks [2, 16] synthesize functional grasp poses using human-labeled part-level functional information. High-quality functional grasping datasets leverage human priors [1, 17] have also been introduced for learning these poses. Additionally, functional affordance regions can be predicted using human functional grasping dataset [18, 19] or internet data [3] to indicate functionality.

To address execution, current works mainly rely on RL to learn a closed-loop policy by matching functional grasp regions [18, 19] or setting to a pre-grasp pose [3] according to functional grasp region. However, these works often assume objects are already positioned for easy functional grasping, ignoring the fact that objects are often not positioned with high functional graspability in the real world, neglecting the need for complex dexterous pre-grasp manipulation.

Our work focuses on dexterous functional pre-grasp manipulation, complementing existing works and serving as a foundation for achieving functional grasping in the real world.

B. Dexterous Manipulation

Dexterous manipulation presents a significant challenge due to the need for closed-loop policies for handling complex and discontinuous contacts, which are notoriously difficult to model accurately. Model-free RL has emerged as a popular approach for acquiring dexterous manipulation skills, as it bypasses the need for explicit contact modeling [7, 20–25].

This approach has demonstrated generalization across diverse objects and goals by shaping different distance rewards to enhance exploration. An orientation distance reward has been used for learning general in-hand reorientation [7, 8], while a position distance reward is used for learning general dynamic handover [6]. Combining both rewards has achieved general in-hand manipulation of slender cylindrical objects [23]. For general articulated object manipulation, the distance reward between the dexterous hand palm and object part has been applied to enhance reaching the goal part [10].

However, our task involves manipulating both arm and dexterous hands to achieve precise position, orientation, and contact goals, resulting in a narrow solution space. Conventional distance rewards can easily trap RL agents in local minima. Moreover, our work requires generalization to diverse objects and goals, making it difficult to design specific rewards for each object.

C. (Dexterous) Pre-grasp Manipulation

Pre-grasp manipulation has been extensively researched to enhance graspability by leveraging extrinsic dexterity. Most works focus on designing specific pre-grasp manipulation strategies to improve graspability. For parallel grippers, RL-based systems utilize external surfaces like tables and walls to transform ungraspable objects into graspable states [4]. Support surfaces and secondary arms can also be employed to achieve power grasps for objects on a table [26]. Additionally, obstacles can be adjusted to improve graspability [27].

Pre-grasp manipulation using a dexterous hand can develop more pre-grasp manipulation strategies, such as a push-and-grasp strategy where a dexterous hand pushes an ungraspable object occluded by the environment to a graspable state before grasping [28]. Another framework involves pushing, rotating, and sliding actions tailored to different objects [29]. For generalization to diverse scenes, a physics-based method has been proposed by leveraging tables and other environmental objects to transform ungraspable objects into graspable ones [5].

Unlike conventional pre-grasp manipulation, our work focuses on manipulating diverse objects to diverse goal poses for subsequent functional grasping, rather than solely achieving graspability.

D. Dexterous Diffusion Models for Grasp and Manipulation

Diffusion models have demonstrated strong generative modeling capabilities in high-dimensional spaces across various

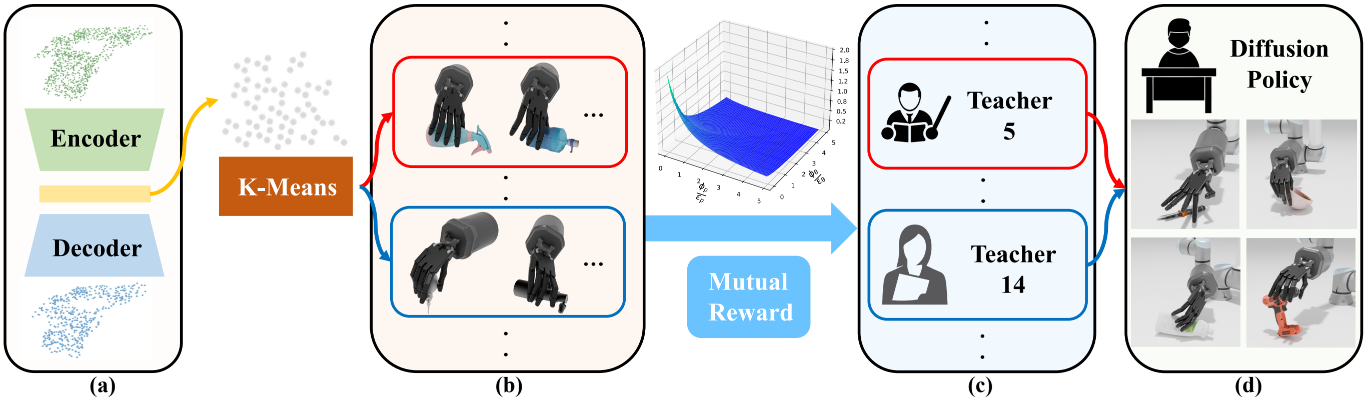

Fig. 2: **Pipeline.** (a) An Autoencoder learns latent representations based on the object-hand point cloud. (b) K-Means clusters the training set into N clusters based on the learned representations. (c) Learning an expert for each cluster based on mutual reward. (d) Distilling multi-expert knowledge into a single student using diffusion for dexterous functional pre-grasp manipulation of seen and unseen objects.

domains [30–37]. In dexterous hands, previous studies also show the potential of diffusion models to generate fine-grained, high-dimensional dexterous grasp poses, whether given full [36, 38] or partial [39] object point clouds of a single object. Moreover, diffusion models can handle the complex kinematics and dynamics involved in generating grasp poses for grasping multiple objects with one hand [40]. However, these works mainly focus on pose generation rather than learning manipulation policies.

In closed-loop manipulation policy learning, leveraging the scalability of diffusion models to high-dimensional output space and their ability to express complex action distributions, diffusion policy has been proposed for parallel grippers to acquire dexterous manipulation skills [15, 41]. Even for complex, high-DoF dexterous hand manipulation policies, point cloud-based diffusion policy [42, 43] has been introduced, achieving impressive performance. However, these studies are focused on limited objects or a single manipulation goal.

Our focus, however, involves generalization across a wide range of objects and goals, and we leverage diffusion policy for multi-expert teacher-student learning.

III. Dexterous Functional Pre-Grasp Manipulation

We address the problem of dexterous functional pre-grasp manipulation. Given a functional grasp goal configuration, a policy is tasked to control a robotic arm and dexterous hand to manipulate the object and achieve the specified goal pose.

State and Action Spaces: We consider a tabletop scenario with a 6-DoF robotic arm $\mathbf{J}^a \in \mathbb{R}^6$ and a 24-DoF dexterous hand $\mathbf{J}^h \in \mathbb{R}^{24}$. The hand's base pose is defined as $\mathbf{b} = [\mathbf{b}_p, \mathbf{b}_q]$, where $\mathbf{b}_p \in \mathbb{R}^3$ denotes the 3D position and $\mathbf{b}_q \in \mathbb{R}^4$ the 4D quaternion. The hand joints consist of 2-DoF wrist joints $\mathbf{J}^w \in \mathbb{R}^2$, 18-DoF finger joints $\mathbf{J}^f \in \mathbb{R}^{18}$, and 4-DoF underactuated finger joints $\mathbf{J}^u \in \mathbb{R}^4$. The action space $\mathcal{A} \subseteq \mathbb{R}^{26}$ encompasses 6D relative changes for the hand base $\mathbf{a}^b$ and 20D relative changes for the actuated hand joints $\mathbf{a}^h$.

Task Simulation: For each pre-grasp manipulation trial, we sample a desired goal pose $\mathbf{g}$ from a prior goal distribution. Each $\mathbf{g}$ corresponds to a specific object O, but one O can have multiple $\mathbf{g}$. The hand's palm coordinate is denoted as P. The goal pose $\mathbf{g} = [\mathbf{g}^P_{pos}, \mathbf{g}^P_{ori}, \mathbf{g}_{fj}]$, where $\mathbf{g}^P_{pos} \in \mathbb{R}^3$ denotes the relative 3D goal position of the object's center of mass with respect to the hand's palm, $\mathbf{g}^P_{ori} \in \mathbb{R}^4$ denotes the relative 4D goal quaternion of the object's center of mass with respect to the hand's palm, and $\mathbf{g}_{fj} \in \mathbb{R}^{18}$ denotes the angle of the hand's actuated finger joints.

Observations: The policy $\pi(\mathbf{a}|\cdot)$ needs to adapt to different $\mathbf{g}$ and O. Therefore, it conditions on $\mathbf{J}^a, \mathbf{b}, \mathbf{J}^h, \mathbf{g}$, and $o^P = [o^P_p, o^P_q]$, where $o^P_p \in \mathbb{R}^3$ denotes the relative 3D position of the object's center of mass with respect to the hand's palm, and $o^P_q \in \mathbb{R}^4$ denotes the relative 4D quaternion of the object's center of mass with respect to the hand's palm.

Objective: The objective of this task is to find a policy $\pi(\mathbf{a}|\mathbf{b}, \mathbf{J}^a, \mathbf{J}^h, o^P, \mathbf{g})$ that maximizes the expected pre-grasp manipulation success rate:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\mathbf{a}_t \sim \pi(\cdot|\mathbf{b}_t, \mathbf{J}^a_t, \mathbf{J}^h_t, o^P_t, \mathbf{g})} \left[\mathbb{1}(\text{success})\right]. \quad (1)$$

The success is satisfied if $\phi_p <= \epsilon_{pos}$ and $\phi_\theta <= \epsilon_{ori}$ and $\phi_j <= \epsilon_{fj}$, where $\phi_p = |o^P_p - \mathbf{g}^P_{pos}|_2$ is the distance between o^P_p and $\mathbf{g}^P_{pos}$, $\phi_\theta = 2 \arcsin\left(\left(o^P_q \cdot (\mathbf{g}^P_{ori})^{-1}\right)_4\right)$ denotes the distance between o^P_q and $\mathbf{g}^P_{ori}$, and $\phi_j = |\mathbf{J}^f - \mathbf{g}_{fj}|_2$ denotes the distance between $\mathbf{J}^f$ and $\mathbf{g}_{fj}$. ϵ_{pos}, ϵ_{ori}, and ϵ_{fj} denote the distance threshold for position, orientation, and contact.

IV. Method

In dexterous functional pre-grasp manipulation, the high dimensionality of the dexterous hand leads to a vast policy space. Meanwhile, the task itself presents a limited solution space, as successful manipulation requires achieving precise goals that satisfy position, orientation, and contact criteria.

Despite the success of model-free RL in various manipulation tasks [6, 7], the stringent requirements in dexterous functional pre-grasp manipulation pose significant challenges to exploration, especially for agents with limited observations.

To address these challenges, we employ the teacher-student framework [7] (see Fig. 2), utilizing a pre-trained "teacher" agent with superior knowledge to guide a "student" agent during the learning process.

A. Teacher Policy Learning

Teacher policy learning aims to acquire high-performance experts without restricting access to privileged information [7]. We introduce a novel mutual reward for learning dexterous functional pre-grasp manipulation policies, followed by an MoE to enhance the overall performance of the teacher policy.

Mutual Reward: Reward shaping is crucial for training a proficient RL agent. In our task, even with privileged information, conventional reward shaping approaches, like adding distance rewards for each goal component [7, 9, 10], can easily trap the RL agent in a local minimum. These rewards incentivize the agent to prioritize optimizing easily achievable distance rewards, such as position distance ϕ_p and contact distance ϕ_j, by manipulating the hand base and joints. However, the agent tends to neglect the orientation distance ϕ_θ, which requires reorienting the object.

To address this, we propose a novel mutual reward. We first define a normalization function ψ to standardize different distance rewards into the range [0,1]:

$$\tilde{r} = \psi\left(\phi, \epsilon\right) = \frac{\epsilon}{\phi + \epsilon}, \tag{2}$$

where $\tilde{r}$ denotes the normalized distance reward. Given the challenge of defining the optimization order for three distance rewards, we use the minimum normalized distance reward $\tilde{r}_{\text{min}}$ as a scale to regulate all the distance rewards. Thus, the total distance reward becomes:

$$r_{\text{dist}} = \tilde{r}_{\text{min}}\left(w_p \tilde{r}_p + w_\theta \tilde{r}_\theta + w_j \tilde{r}_j\right), \tag{3}$$

where w_p, w_θ, and w_j are hyperparameters. This restriction term prevents simply minimizing ϕ_p or ϕ_j from rapidly increasing the total reward, as the typically large ϕ_θ results in a small $\tilde{r}_{\text{min}}$, as illustrated in Fig. 2. This compels the agent to jointly optimize all three distance rewards, enabling successful learning of the dexterous functional pre-grasp manipulation policy.

We also incorporate an action penalty r_{ap} to regulate arm motion:

$$r_{\text{ap}} = \|\mathbf{a}^b\|_2. \tag{4}$$

This penalty discourages excessive arm movement and encourages finger utilization for object manipulation. The success reward r_{succ} is 1 if manipulation is successful. Therefore, the total reward becomes:

$$r = r_{\text{dist}} + w_{ap} * r_{\text{ap}} + w_{succ} * r_{\text{succ}}, \tag{5}$$

where w_{ap} and w_{succ} are hyperparameters for the action penalty and success reward, respectively.

MoE: Given the goal of generalizing across diverse objects and goal poses, the manipulation process can exhibit significant diversity. This makes it challenging for RL agents to learn a good policy for all goal poses. While Unidexgrasp [13, 14] introduced a framework for learning dexterous grasping for diverse objects by starting with "GeoCurriculum," which gradually increases object instances and categories from a single object with a single pose, such a curriculum is not suitable for our task. Unlike grasping, which involves reaching and closing fingers, manipulation requires continuous repositioning and reorienting of the object. Hence, the manipulation policy for different object geometry can be different. For instance, manipulating a cylindrical bottle involves rolling it, whereas manipulating a camera requires different techniques. Thus, if the agent learns to manipulate a cylindrical bottle first, it may struggle to learn to manipulate the camera.

Although "GeoCurriculum" is not directly applicable to our task, the concept of decomposing the task space is valuable. Therefore, we initially cluster the entire task space into several clusters. Unidexgrasp++ [13] trains an autoencoder on object geometry for the reconstruction task and then uses the latent representation of each object for state-based clustering. In the case of dexterous functional pre-grasp manipulation, the task is linked to the goal pose. Given the same object with the same initial pose, the goal of grasping the handle versus grasping the body can lead to different manipulation processes. Thus, we combine the object and hand point cloud to learn a latent representation.

After clustering, we employ K-Means to partition the entire task space into N clusters. While prior work [14] suggests that a generalist can assist specialists in training dexterous grasping, in dexterous functional pre-grasp manipulation, manipulation behaviors can vary across different goals, such as manipulating a cylindrical bottle versus a camera, as described earlier. Hence, to obtain a specialized high-performance manipulation policy for each cluster, we directly train an expert for each cluster from scratch.

B. Distilling With Diffusion Policy

Once we have acquired the MoE, our objective is to distill the diverse manipulation policies into a single student policy. The student policy is constrained to only access observations available in real scenarios, as described in Sec. III. Given the complexity and diversity of the action distribution resulting from the intricate manipulation process and the MoE, coupled with the high dimensionality of the dexterous hand, we opt to utilize a diffusion policy [15] which has been shown to have the ability to learn complex high DoF dexterous hand manipulation behavior [42, 43], to model the complex action distribution of different experts. Diffusion policy formulates the robot behavior generation as a conditional denoising process.

Dataset Generation: Since the diffusion policy operates as an offline imitation learning framework, we must gather demonstrations using our teacher experts. While our teacher policy necessitates privileged information for inference, the trajectories we gather for training the diffusion policy solely comprise limited observations.

By executing the policy of our N teacher experts on the entire task space, we sample a set of trajectories $\tau_i{}_{i=1}^{M}$. However, these trajectories have different episode lengths. Following Chi *et al*. [15], for each trajectory τ_i with a step size of $\mathbf{L}_i$, we sample every sequence with a length of $\mathbf{T}_p$, where $\mathbf{T}_p$ denotes the prediction horizon. Consequently, we obtain $\mathbf{L}_i - \mathbf{T}_p + 1$ trajectory data points from τ_i. By iterating over the trajectory set $\tau_i{}_{i=1}^{M}$, we can generate the dataset $\mathbf{S}_j{}_{j=1}^{O}$ for diffusion policy training.

Diffusion Policy Training: The training process involves sampling data points from the generated dataset. For each sample $\mathbf{S}_j$, we randomly sample a time step $\mathbf{t}$, and then sample a noise $\mathbf{n}^{\mathbf{t}}$. We consider the first $\mathbf{T}_o$ steps of observations from $\mathbf{S}_j$ as the observation sequence o_j^D, and take the $\mathbf{T}_p$ steps of actions from $\mathbf{S}_j$ as the action sequence $\mathbf{A}_j^0$. We utilize o^D as a condition and define the loss function as follows:

$$\mathcal{L} = \text{MSE}(\mathbf{n}^{\mathbf{t}}, \mathbf{n}_\theta(o_j^D, \mathbf{A}_j^0 + \mathbf{n}^{\mathbf{t}}, \mathbf{t})), \tag{6}$$

where $\mathbf{n}_\theta$ is a noise prediction network.

Action Generation with Diffusion Policy: Upon training the noise prediction network $\mathbf{n}_\theta$, for each simulation step s_i, the DDPM [44] performs $\mathbf{t}$ steps denoising from the noise action sequence $\mathbf{A}_{s_i}^{\mathbf{t}}$ sampled from Gaussian noise, until obtaining the noise-free action sequence $\mathbf{A}_{s_i}^0$. Following equation:

$$\mathbf{A}_{s_i}^{\mathbf{t}-1} = \alpha(\mathbf{A}_{s_i}^{\mathbf{t}} - \gamma \mathbf{n}_\theta(o_{s_i}^D, \mathbf{A}_{s_i}^{\mathbf{t}}, \mathbf{t}) + \mathcal{N}(0, \sigma^2 I)), \tag{7}$$

where α, γ, and σ are parameters for noise scheduler. We then execute $\mathbf{T}_a$ steps of the denoised action sequence $\mathbf{A}_{s_i}^0$.

C. Implementation Details

Teacher Policy: Our RL backbone is PPO [45]. We configure hyperparameters with $w_p = w_\theta = w_j = 3$, $w_{ap} = -0.01$, and $w_{succ} = 800$. Privileged information details for teacher policy training are provided in Table I.

TABLE I: **Teacher Observation.** The superscript P represents the variable is w.r.t. the hand-palm coordinate.

Variable	Dimension	Description
$\mathbf{b}_p$	(3,)	hand base positions
$\mathbf{b}_q$	(4,)	hand base orientations
$\mathbf{J}^a$	(6,)	arm joint angles
$\dot{\mathbf{J}}^a$	(6,)	arm joint velocities
$\mathbf{J}^h$	(24,)	hand joint angle
$\dot{\mathbf{J}}^h$	(24,)	hand joint velocities
${f_p}^P$	(5, 3)	fingertip positions (to Palm)
${f_q}^P$	(5, 4)	fingertip orientations (to Palm)
$\boldsymbol{v}_f$	(5, 3)	fingertip linear velocities
$\boldsymbol{w}_f$	(5, 3)	fingertip angular velocities
o_p	(3,)	object position
o_q	(4,)	object orientation
o_p^P	(3,)	object position (to Palm)
o_q^P	(4,)	object orientation (to Palm)
$\boldsymbol{v}_{\text{o}}$	(3,)	object linear velocity
$\boldsymbol{w}_{\text{o}}$	(3,)	object angular velocity
$\boldsymbol{bbox}_{\text{object}}$	(2, 3)	object boundingbox
$\mathbf{g}_{pos}^P$	(3,)	target object position (to Palm)
ϕ_p	(3,)	position distance
$\mathbf{g}_{ori}^P$	(4,)	target object orientation (to Palm)
ϕ_θ	(4,)	orientation distance
$\mathbf{g}_{fj}$	(18,)	target hand joint angles
ϕ_j	(18,)	joint distance

MoE: For each goal pose $\mathbf{g}_k$ in the set $\mathbf{g}_k{}_{k=1}^O$, we sample 512 points from the corresponding object mesh and 512 points from the corresponding hand mesh. These point clouds are concatenated and then encoded using PointNet++ [46], and the reconstruction loss is computed with Chamfer Distance. The entire task space is divided into 20 clusters. For each expert, we train on 12000 parallel environments until convergence.

Diffusion Policy: We configure $\mathbf{T}_p = 4$, $\mathbf{T}_o = 2$, and $\mathbf{T}_a = 1$, while keeping other parameters the same as [15]. Because we use the relative action for policy learning, we use the transformer backbone [47] for handling quick and sharp changes in action sequence [15].

V. Experimental Setups

A. Task Simulation

Environment Setup: We created a simulation environment based on Isaac Gym [48] using ShadowHand and UR10e robots. Each environment consists of an object randomly placed on a table, with its mass randomized from 0.01kg to 0.5kg due to the diversity of object categories. A UR10e robot is positioned outside the table with the ShadowHand mounted on the end of the arm, as shown in Fig. 1. The maximum episode length is 300 steps. Episodes terminate upon reaching the goal pose or prematurely if the object falls off the table or the maximum steps are reached.

Goal Pose Generation: We utilize the Oakink dataset [49], which covers diverse functional intents for a wide range of objects, to generate a dexterous hand functional grasp pose dataset. Since the Oakink dataset is based on the human hand, it differs in structural and shape characteristics from robotic hands. To adapt the hand poses, we employ a retargeting algorithm [22] based on task space vectors to map the mano hand pose to the ShadowHand pose. Next, to refine poses prone to collision and non-force closure grasp, we utilize Dexgraspnet [50] for optimization. To enhance grasping, we optimize the joints by calculating the gradient corresponding to the movement of the contact point along the normal direction. Finally, all refined poses undergo validation in a simulated environment to eliminate those unstable under the influence of gravity.

Due to the uneven distribution of object instances within each category in the Oakink dataset, we implement a stratified splitting approach for training and testing sets. Categories with a larger number of instances are randomly divided into a 70% training set and a 30% testing set, while categories with fewer instances are split evenly into 50% for training and testing. Overall, our training set comprises 1026 object instances with a total of 6968 goal poses, while the testing set consists of 443 object instances with a total of 3034 goal poses.

B. Baselines and Metrics

For the teacher policy, we compare our method with the following: (i) **PPO-Sum**: This baseline adopts a sum reward approach, combining three distance rewards for RL training, while keeping other rewards the same as *Ours*. (ii) **Ours-SE**: Based on our proposed reward, we train a single expert for the entire training set. (iii) **Ours-MoEF**: This comparison utilizes a MoE, but instead of training them from scratch, we fine-tune them from the *Ours-SE* model. Due to computational cost, this comparison is conducted on a subset of our training data.

For comparison based on student observations, we evaluate our method with: (i) **PPO-OS** employs PPO as a one-stage method, using the same mutual reward as *Ours* but without teacher-student learning. (ii) **Behavior Cloning (BC)** serves as an offline imitation learning framework, learning directly from expert demonstrations via supervised learning. (iii) **Dagger** [51] is an online imitation learning framework that tackles the covariate shift problem through iterative sampling with a learned policy via online interaction.

We employ the success rate as the metric for all comparisons. Our task employs stringent criteria, with $\epsilon_{pos} = 1cm$, $\epsilon_{ori} = 0.1rad$, and $\epsilon_{fj} = 0.2rad$, which are challenging thresholds to meet.

VI. Results

We conducted extensive experiments to validate our approach. Section VI-A compares our proposed reward with a baseline using privileged information. Building upon our reward, Sec. VI-B explores the challenges of learning general dexterous functional pre-grasp manipulation without a teacher-student framework. Within this framework, Secs. VI-C and VI-D evaluate the effectiveness of using MoE for teacher policy training and a diffusion policy for distillation, respectively. To assess our reliance on object pose observation, Secs. VI-E and VI-F present results concerning different geometry types and robustness. Finally, Sec. VI-G analyzes performance across object categories.

A. Reward Comparison

As shown in Tab. II, without a mutual reward, the *PPO-Sum* agent fails to learn a successful manipulation policy. We observed that the agent quickly learns to align positions and contacts but is stuck at a local minimum, failing to align orientations. In contrast, *Ours-SE* with a mutual reward prevents premature optimization, encouraging simultaneous optimization of each reward. This leads to a significant improvement in the success rate from 0.0% to 58.0%.

B. Challenges in Functional Pre-grasp Manipulation

To demonstrate the difficulty of learning general dexterous functional pre-grasp manipulation, we conducted experiments using *PPO-OS* based on student observation, incorporating our mutual reward. We trained PPO across varying numbers of objects, with each PPO model trained until convergence or reaching the maximum interaction steps (5.76 billion).

As depicted in Tab. III, when trained on a single object, the RL agent rapidly learns a policy with a nearly 100% success rate. However, as the number of objects increases, the success rate declines steeply, highlighting the difficulty of **general dexterous functional pre-grasp manipulation**. Interestingly, the success rate for 9 objects is lower than for 100 objects. This is because within the set of 9 objects, the presence of challenging objects, such as knives, is proportionately higher, hindering exploration. This underscores the necessity of employing an MoE.

TABLE II: **Success Rate of Teacher Policy.** "All": trained on the entire training set; "Sub": trained on a subset of the training set; "SE": single expert; "MoE": mixture of experts; "Succ (Train)": success rate on the training set.

Method	Training set	Reward	Teacher	Succ (Train)
PPO-Sum	All	Sum	SE	0.0%
Ours-SE	All	Mutual	SE	58.0%
Ours-SE (sub)	Sub	Mutual	SE	55.2%
Ours-MoEF (sub)	Sub	Mutual	MoE	63.9%
Ours (sub)	Sub	Mutual	MoE	67.4%
Ours	All	Mutual	MoE	75.0%

TABLE III: **Success Rate of One-stage PPO under Different Sizes of Training Set.** "Succ (Train)": success rate on the training set. As the number of objects increases, finding a general manipulation policy across diverse objects becomes increasingly challenging for one-stage PPO.

Obj Num	**1**	**9**	**100**	**1026 (All)**
Succ (Train)	98.2%	6.2%	21.2%	6.5%

C. Teacher Policy Comparison

Table II demonstrates that employing multiple experts leads to a further increase in the success rate from 58.0% to 75.0% compared to a single expert. This is because multiple experts allow the agent to learn more tailored policies for each cluster. Additionally, our experiment shows that training from scratch outperforms fine-tuning from a generalist. We sampled five clusters with varying learning difficulty and trained *Ours (sub)* from scratch on each cluster, while *Ours-MoEF (sub)* was fine-tuned from the pre-trained single expert *Ours-SE (sub)*. As shown in Tab. II, training from scratch achieves better overall performance due to the diversity of objects and poses and the complexity of manipulation, making it challenging to transfer a general policy to objects and poses with significant variability.

D. Student Policy Comparison

All methods utilizing teacher-student learning outperform *PPO-OS*, which undergoes end-to-end training. As indicated in Tab. IV, Dagger can achieve performance comparable to the single-expert teacher policy but struggles to learn an effective policy under the mixture of expert teacher policy.

Offline imitation learning methods demonstrate superior results compared to those requiring environment interaction. Due to the critical role of data quantity, we compare *Ours* and *BC* across various demonstration numbers. Table IV shows that *Ours* consistently outperforms *BC* on both training and testing sets, especially with limited demonstrations. Notably, *Ours* can achieve comparable performance with only half the number

TABLE IV: **Success Rate of Student Policy.** This table presents the results of methods that require teachers for training. "SE": single expert; "MoE": mixture of experts; "Demo Num": the maximum number of demonstrations collected for each pose in the training set, used for distilling the student policy; "Succ (Train)": success rate on the training set; "Succ (Test)": success rate on the testing set.

Method	Teacher	Demo Num	Succ (Train)	Succ (Test)
Dagger	SE	-	52.2%	52.3%
Dagger	MoE	-	17.5%	17.3%
BC	MoE	5	57.4%	54.7%
BC	MoE	10	67.7%	65.1%
BC	MoE	20	70.9%	67.7%
Ours	MoE	5	65.7%	63.3%
Ours	MoE	10	71.3%	68.4%
Ours	MoE	20	73.7%	70.1%

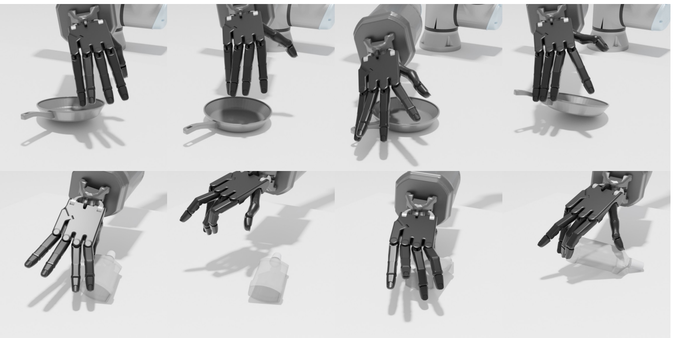

Fig. 3: **Adaptability of Our Learned Policy.** Although our agent may initially fail to manipulate objects, it adjusts its policy on the second attempt, successfully manipulating them. This capability helps the agent handle diverse dynamics.

of demonstrations required by *BC*. With a large number of demonstrations, *Ours* approaches teacher-level performance.

E. Ablation on Geometry Type

While common sense suggests that object geometry information is crucial for manipulation, Tab. V shows that providing more detailed geometry information does not significantly impact performance, although it can lead to a better student policy. Observing the learned policy, we discovered that our policy utilizes extrinsic dexterity, such as using the table to roll objects or leveraging inertia, as shown in Fig. 1. Additionally, our policy learns to adjust based on feedback (Fig. 3). These capabilities enhance the agent's ability to generalize to different objects and goal poses.

However, these capabilities also have drawbacks. We observed instances where the agent pushes objects down to better utilize extrinsic dexterity, which may need improvement in the future through the design of new reward mechanisms.

F. Robustness under Noisy Object Pose

As we solely depend on object pose for dexterous functional pre-grasp manipulation, and object pose is actually hard to be accurate in the real world due to sensor noise and occlusion, we further conduct experiments under varying levels of noisy object pose observations [52].

Table VI shows that injecting 2°, 2cm noise results in a decrease in success rate. However, given our stringent criteria, we also tested with a larger success threshold, which remains reasonable. Even when doubled, the achieved functional pose remains meaningful and comparable. By slightly relaxing the threshold, our method still achieves a high success rate, underscoring its robustness and potential for real-world applications.

TABLE V: **Success Rate of Different Geometries.** "Succ (Train)": success rate on the training set; "Succ (Test)": success rate on the testing set. Due to computational cost, this experiment was conducted using 5 demonstrations per pose.

Geometry Type	Succ (Train)	Succ (Test)
Pose + Point Cloud	66.5%	63.3%
Pose + Bounding Box	65.9%	62.8%
Pose	65.7%	63.3%

TABLE VI: **Success Rate under Different Levels of Object Pose Estimation Noise and Success Threshold.** "Succ (Test)": success rate on the testing set. The noise level is determined by the standard deviation of the specified noise.

Noise level	**Threshold**		
	1ϵ	1.5ϵ	2ϵ
	Succ (Test)		
0°, 0cm	70.1%	77.8%	81.2%
2°, 2cm	38.1%	67.7%	75.2%
5°, 5cm	0.0%	6.5%	8.7%

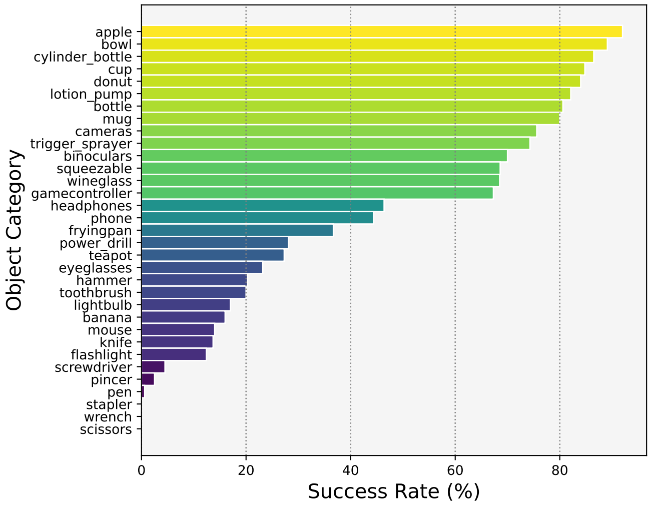

Fig. 4: **Success Rate of Different Object Categories.**

G. Performance across Object Categories

Figure 4 shows that while our method achieves a high success rate across the entire dataset, it still struggles with irregularly shaped objects, particularly thin and slender ones like knives and pens. Even when trained from scratch, the experts fail to perform well on these objects, indicating a need for specific design considerations.

VII. Conclusion

This work focuses on general dexterous functional pre-grasp manipulation, repositioning and reorienting various objects to precisely match diverse functional grasp poses, crucial for real-world functional grasping. We adopt a teacher-student learning framework, introducing a novel mutual reward that greatly enhances teacher policy learning. Furthermore, we propose employing a MoE and distillation with a diffusion policy for learning diverse manipulation behavior. Our experiments showcase efficacy and robustness, revealing the potential for real-world applications.

Limitations and Future Works: Although our teacher policy shows promising results, it still struggles with objects of irregular shapes. Integrating human demonstrations could potentially improve performance. Additionally, our current focus is solely on pre-grasp manipulation. To achieve functional grasping in real-world scenarios, it is essential to integrate pre-grasp manipulation with functional grasp pose generation and grasping, alongside addressing the sim2real gap.

Acknowledgments: We thank Jiyao Zhang, Haoran Geng, Zeyuan Chen, Tianyu Wang, Xiangyu Huang, and Jinghui Zhuang for their insightful discussions.

References

[1] W. Wei, P. Wang, and S. Wang, "Generalized anthropomorphic functional grasping with minimal demonstrations," *arXiv preprint arXiv:2303.17808*, 2023.

[2] T. Zhu, R. Wu, J. Hang, X. Lin, and Y. Sun, "Toward human-like grasp: Functional grasp by dexterous robotic hand via object-hand semantic representation," *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 45, no. 10, pp. 12 521–12 534, 2023.

[3] A. Agarwal, S. Uppal, K. Shaw, and D. Pathak, "Dexterous functional grasping," in *Conference on Robot Learning (CoRL)*, 2023.

[4] W. Zhou and D. Held, "Learning to grasp the ungraspable with emergent extrinsic dexterity," in *Conference on Robot Learning (CoRL)*, 2023.

[5] S. Chen, A. Wu, and C. K. Liu, "Synthesizing dexterous nonprehensile pregrasp for ungraspable objects," in *ACM SIGGRAPH Conference Proceedings*, 2023.

[6] B. Huang, Y. Chen, T. Wang, Y. Qin, Y. Yang, N. Atanasov, and X. Wang, "Dynamic handover: Throw and catch with bimanual hands," in *Conference on Robot Learning (CoRL)*, 2023.

[7] T. Chen, J. Xu, and P. Agrawal, "A system for general in-hand object re-orientation," in *Conference on Robot Learning (CoRL)*, 2022.

[8] T. Chen, M. Tippur, S. Wu, V. Kumar, E. Adelson, and P. Agrawal, "Visual dexterity: In-hand reorientation of novel and complex object shapes," *Science Robotics*, vol. 8, no. 84, p. eadc9244, 2023.

[9] Y. Qin, B. Huang, Z.-H. Yin, H. Su, and X. Wang, "Dexpoint: Generalizable point cloud reinforcement learning for sim-to-real dexterous manipulation," in *Conference on Robot Learning (CoRL)*, 2023.

[10] C. Bao, H. Xu, Y. Qin, and X. Wang, "Dexart: Benchmarking generalizable dexterous manipulation with articulated objects," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[11] Y. Chen, T. Wu, S. Wang, X. Feng, J. Jiang, Z. Lu, S. M. McAleer, H. Dong, S.-C. Zhu, and Y. Yang, "Towards human-level bimanual dexterous manipulation with reinforcement learning," in *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022.

[12] T. Wu, M. Wu, J. Zhang, Y. Gan, and H. Dong, "Learning score-based grasping primitive for human-assisting dexterous grasping," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[13] Y. Xu, W. Wan, J. Zhang, H. Liu, Z. Shan, H. Shen, R. Wang, H. Geng, Y. Weng, J. Chen *et al.*, "Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[14] W. Wan, H. Geng, Y. Liu, Z. Shan, Y. Yang, L. Yi, and H. Wang, "Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning," in *International Conference on Computer Vision (ICCV)*, 2023.

[15] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Robotics: Science and Systems (RSS)*, 2023.

[16] T. Zhu, R. Wu, X. Lin, and Y. Sun, "Toward human-like grasp: Dexterous grasping via semantic representation of object-hand," in *International Conference on Computer Vision (ICCV)*, 2021.

[17] J. Hang, X. Lin, T. Zhu, X. Li, R. Wu, X. Ma, and Y. Sun, "Dexfuncgrasp: A robotic dexterous functional grasp dataset constructed from a cost-effective real-simulation annotation system," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

[18] P. Mandikal and K. Grauman, "Learning dexterous grasping with object-centric visual affordances," in *International Conference on Robotics and Automation (ICRA)*, 2021.

[19] ——, "Dexvip: Learning dexterous grasping with human hand pose priors from video," in *Conference on Robot Learning (CoRL)*, 2022.

[20] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine, "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," in *Robotics: Science and Systems (RSS)*, 2018.

[21] D. Jain, A. Li, S. Singhal, A. Rajeswaran, V. Kumar, and E. Todorov, "Learning deep visuomotor policies for dexterous hand manipulation," in *International Conference on Robotics and Automation (ICRA)*, 2019.

[22] Y. Qin, Y.-H. Wu, S. Liu, H. Jiang, R. Yang, Y. Fu, and X. Wang, "Dexmv: Imitation learning for dexterous manipulation from human videos," in *European Conference on Computer Vision (ECCV)*, 2022.

[23] W. Hu, B. Huang, W. W. Lee, S. Yang, Y. Zheng, and Z. Li, "Dexterous in-hand manipulation of slender cylindrical objects through deep reinforcement learning with tactile sensing," *arXiv preprint arXiv:2304.05141*, 2023.

[24] Y. Chen, C. Wang, L. Fei-Fei, and K. Liu, "Sequential dexterity: Chaining dexterous policies for long-horizon manipulation," in *Conference on Robot Learning (CoRL)*, 2023.

[25] Z.-H. Yin, B. Huang, Y. Qin, Q. Chen, and X. Wang, "Rotating without seeing: Towards in-hand dexterity through touch," in *Robotics: Science and Systems (RSS)*, 2023.

[26] I. Baek, K. Shin, H. Kim, S. Hwang, E. Demeester, and M.-S. Kang, "Pre-grasp manipulation planning to secure space for power grasping," *Ieee Access*, vol. 9, pp. 157 715–157 726, 2021.

[27] M. Moll, L. Kavraki, J. Rosell *et al.*, "Randomized physics-based motion planning for grasping in cluttered and uncertain environments," *IEEE Robotics and Automation Letters (RA-L)*, vol. 3, no. 2, pp. 712–719, 2017.

[28] M. R. Dogar and S. S. Srinivasa, "Push-grasping with dexterous hands: Mechanics and a method," in *International Conference on Intelligent Robots and Systems (IROS)*, 2010.

[29] D. Kappler, L. Chang, M. Przybylski, N. Pollard, T. Asfour, and R. Dillmann, "Representation of pre-grasp strategies for object manipulation," in *International Conference on Humanoid Robots (Humanoids)*, 2010.

[30] R. Cai, G. Yang, H. Averbuch-Elor, Z. Hao, S. Belongie, N. Snavely, and B. Hariharan, "Learning gradient fields for shape generation," in *European Conference on Computer Vision (ECCV)*, 2020.

[31] Y. Song, L. Shen, L. Xing, and S. Ermon, "Solving inverse problems in medical imaging with score-based generative models," in *International Conference on Learning Representations (ICLR)*, 2021.

[32] P. Yu, S. Xie, X. Ma, B. Jia, P. Bo, R. Gao, Y. Zhu, Y. Wu, and S.-C. Zhu, "Unsupervised foreground extraction via deep region competition," in *International Conference on Machine Learning (ICML)*, 2022.

[33] M. Wu, Y. Xia, H. Dong *et al.*, "Targf: Learning target gradient field to rearrange objects without explicit goal specification," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[34] Z. Wang, Y. Chen, T. Liu, Y. Zhu, W. Liang, and S. Huang, "Humanise: Language-conditioned human motion generation in 3d scenes," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[35] H. Ci, M. Wu, W. Zhu, X. Ma, H. Dong, F. Zhong, and Y. Wang, "Gfpose: Learning 3d human pose prior with gradient fields," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[36] 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.

[37] Z. Wang, Y. Chen, B. Jia, P. Li, J. Zhang, J. Zhang, T. Liu, Y. Zhu, W. Liang, and S. Huang, "Move as you say, interact as you can: Language-guided human motion generation with scene affordance," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[38] J. Lu, H. Kang, H. Li, B. Liu, Y. Yang, Q. Huang, and G. Hua, "Ugg: Unified generative grasping," *arXiv preprint arXiv:2311.16917*, 2023.

[39] Z. Weng, H. Lu, D. Kragic, and J. Lundell, "Dexdiffuser: Generating dexterous grasps with diffusion models," *arXiv preprint arXiv:2402.02989*, 2024.

[40] Y. Li, B. Liu, Y. Geng, P. Li, Y. Yang, Y. Zhu, T. Liu, and S. Huang, "Grasp multiple objects with one hand," *IEEE Robotics and Automation Letters (RA-L)*, 2024.

[41] H. Ha, P. Florence, and S. Song, "Scaling up and distilling down: Language-guided robot skill acquisition," in *Conference on Robot Learning (CoRL)*, 2023.

[42] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy," *arXiv preprint arXiv:2403.03954*, 2024.

[43] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. K. Liu, "Dexcap: Scalable and portable mocap data collection system for dexterous manipulation," *arXiv preprint arXiv:2403.07788*, 2024.

[44] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[45] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[46] 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.

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[48] 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.

[49] L. Yang, K. Li, X. Zhan, F. Wu, A. Xu, L. Liu, and C. Lu, "Oakink: A large-scale knowledge repository for understanding hand-object interaction," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[50] R. Wang, J. Zhang, J. Chen, Y. Xu, P. Li, T. Liu, and H. Wang, "Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation," in *International Conference on Robotics and Automation (ICRA)*, 2023.

[51] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

[52] H. Chen, P. Wang, F. Wang, W. Tian, L. Xiong, and H. Li, "Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.