UniDexFPM: Universal Dexterous Functional Pre-grasp Manipulation Via Diffusion Policy

Tianhao Wu[1,2,3*], Yunchong Gan[1*], Mingdong Wu[1,2,3], Jingbo Cheng[1], Yaodong Yang[4,5], Yixin Zhu[4], Hao Dong[1,2,3]

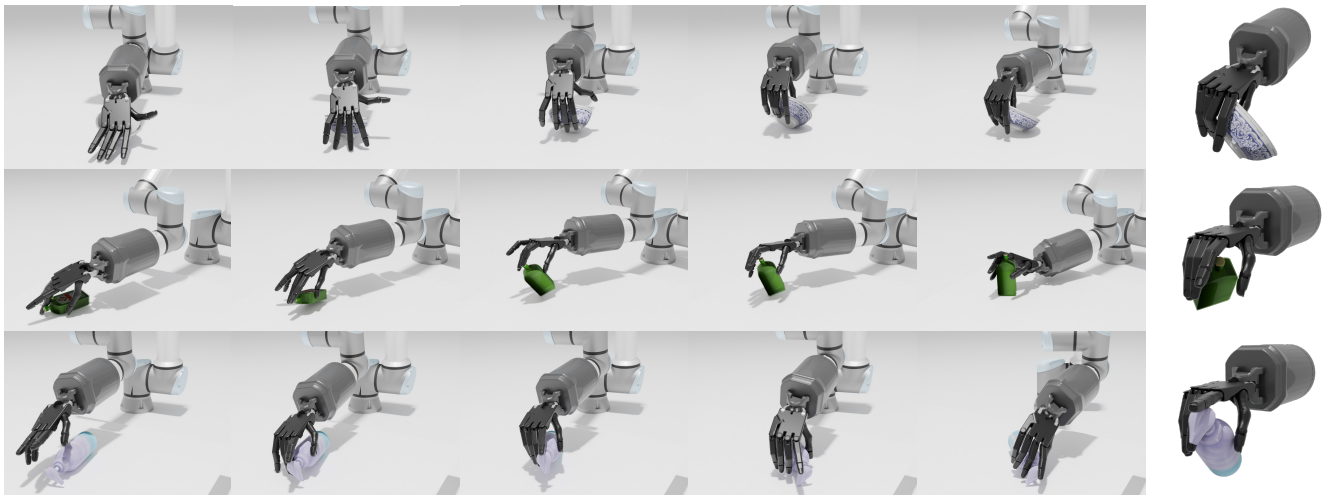

Fig. 1: Our policy utilizes extrinsic dexterity to continuously reposition and reorient diverse objects to successfully match the goal functional grasp poses. The left side depicts the dexterous functional pre-grasp manipulation process, while the right side illustrates the goal functional grasp pose that the agent needs to satisfy.

***Abstract*— Objects in the real world are often not naturally positioned for functional grasping, which usually requires repositioning and reorientation before they can be grasped, a process known as pre-grasp manipulation. However, effective learning of universal dexterous functional pre-grasp manipulation necessitates precise control over relative position, relative orientation, and contact between the hand and object, while generalizing to diverse dynamic scenarios with varying objects and goal poses. We address the challenge by using teacher-student learning. We propose a novel mutual reward that incentivizes agents to jointly optimize three key criteria. Furthermore, we introduce a pipeline that leverages 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 30+ object categories encompassing 1400+ objects and 10k+ goal poses. Notably, our method relies solely on object pose information for universal dexterous functional pre-grasp manipulation by using extrinsic dexterity and adjusting from feedback. Additional experiments under noisy object pose observation showcase the robustness of our method and its potential for real-world applications. The demonstrations can be viewed at `https://unidexfpm.github.io`.**

I. INTRODUCTION

Objects in human daily life serve various functions, which require 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 [3] incorporate with reinforcement learning (RL) for grasp execution and post grasp usage. 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, which aims to transition 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], [7], [8] rely on reinforcement learning 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. Conventional approaches, such as adding distance rewards [7], [9], [10], struggle in this scenario. Simply adding multiple distance

*: Equal contribution.
†: Corresponding emails: thwu@stu.pku.edu.cn, hao.dong@pku.edu.cn.
[1] Center on Frontiers of Computing Studies, School of Computer Science, Peking University. [2] PKU-Agibot Lab, School of Computer Science, Peking University. [3] National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University. [4] Institute for Artificial Intelligence, Peking University. [5] National Key Laboratory of General AI, Beijing Institute for General Artificial Intelligence.

rewards often leads RL agents to become trapped in local minima, failing to devise manipulation policies that meet all criteria. It is also impossible for us 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], [13], [14].

To tackle 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 the distance rewards to prevent the agent from quickly learning to minimize one of the distance rewards, thus prevent the agent to stuck at the local minimum. Moreover, to facilitate generalization across diverse objects and functional grasp poses, we employ the teacher-student learning framework [7], [14] by training a mixture of experts. Mixture of experts will generate diverse manipulation behavior, leading to a complex action distribution, especially for a high degree of freedom (DOF) dexterous hand. Thus, we propose to use diffusion policy [15] which has been shown to have great generative modeling ability to capture such complex action distribution.

Through mutual reward and a mixture of experts training, we observe significant improvements in teacher policy learning. When distilling the teacher policy into a single student policy using 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 to generalize across diverse objects.

In summary, our contributions are listed as follows:(1) We propose a novel mutual reward to improve the local minimum problem which greatly improves the teacher policy learning. (2) We propose a pipeline integrating a mixture of experts and diffusion policy for learning complex and general dexterous manipulation policy. (3) To the best of our knowledge, we have achieved the first general dexterous functional pre-grasp manipulation policy of 72.6% success rate across 30+ object categories encompassing 1400+ objects and 10k+ goal poses.

II. RELATED WORK

A. Dexterous Functional Grasping

Functional grasping stands as a crucial skill for humans, given the varied functionalities objects possess in real-world scenarios. Leveraging human-labeled part-level functional information, recent frameworks [16], [2] have been proposed for synthesizing functional grasp poses. Additionally, a high-quality functional grasping dataset [1] has been introduced, incorporating diverse dexterous hands. However, these efforts primarily focus on pose generation, overlooking the execution of grasps.

Utilizing the human functional grasping dataset [17], an affordance prediction model has been developed for functional region estimation, subsequently employed in RL training for functional grasping [18], [19]. To enable in-the-wild real-world functional grasping, a framework [3] using internet data for predicting functional affordance region, integrated with simulation training for real-world functional grasping. Nonetheless, these works often assume objects are already positioned in a highly graspable pose for functional grasping, thus bypassing the need for complex pre-grasp manipulation.

Our work instead focuses on dexterous functional pre-grasp manipulation, which is the complement of these existing works. Our work can serve as a foundational step towards integrating these approaches to achieve functional grasping in real-world scenarios.

B. Dexterous Manipulation

Dexterous manipulation requires closed-loop policies to handle complex and discontinuous contacts, which is challenging to model accurately. Model-free reinforcement learning, which does not require explicit modeling of such contacts, has been widely adopted for learning dexterous manipulation skills [20], [21], [22], [7], [23], [24]. This approach has also demonstrated the capability to generalize across diverse objects, either using poincloud [22], [8] or object pose information [7], [6]. However, existing tasks often focus on satisfying the goal orientation and/or position of the objects. In contrast, dexterous functional pre-grasp manipulation involves achieving precise position, orientation, and contact goals, resulting in an exceedingly narrow solution space for RL agents to explore successful policies. This challenge is further exacerbated when attempting to generalize across different objects.

C. Pre-grasp Manipulation

Pre-grasp manipulation has been widely investigated for improving the graspability, primarily focusing on leveraging extrinsic dexterity, such as utilizing tables or secondary arms, to transform ungraspable objects into graspable ones using either parallel grippers [4], [25], [26] or dexterous hands [27], [5]. Additionally, apart from manipulating target objects, obstacles can also be adjusted to improve graspability [28]. However, these studies did not consider the functionality of the graspable states or goal states. Matching position, orientation, and finger-level contact poses greater challenges for pre-grasp manipulation.

D. Diffusion Model

The diffusion model has demonstrated strong generative modeling capabilities in high dimensional space across various domains [29], [30], [31], [32]. While previous works [33], [34], [35] have primarily employed diffusion models for generating dexterous hand grasp poses, the application of diffusion policy for closed-loop manipulation policy learning has been explored more recently [15]. Specifically, diffusion policy has been proposed for parallel grippers to acquire dexterous manipulation skills within specific tasks [15] or with limited object instances [36]. However, our focus involves employing a high-degree-of-freedom (DOF) dexterous hand, which possesses a larger action space and requires generalization across a wide range of object categories and instances. Additionally, we leverage diffusion policy for multi-experts teacher-student learning.

III. Dexterous Functional Pre-grasp Manipulation

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

State and Action Spaces: In this task, we consider a tabletop manipulation scenario involving a 6-DOF robotic arm $\mathbf{J}^a \in \mathbb{R}^6$ and 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 3-D position and $\mathbf{b}_q \in \mathbb{R}^4$ represents the 4-D quaternion. The 24-DOF joints of the hands 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 under-actuated joints of the fingers $\mathbf{J}^u \in \mathbb{R}^4$. The action space $\mathcal{A} \subseteq \mathbb{R}^{26}$ encompasses 6-D relative changes for the hand base $\mathbf{a}^b$ and 20-D relative changes for the actuated hand joints $\mathbf{a}^h$.

Task Simulation: For each pre-grasp manipulation trail, we sample a desired goal pose $\mathbf{g} \sim p_{\mathbf{g}}(\mathbf{g})$ from a prior goal distribution. Each goal pose $\mathbf{g}$ corresponds to a specific object O. However, one object can have multiple potential goal poses $\mathbf{g}$. The goal pose $\mathbf{g}$, is represented as a combination of three elements:(1) **Relative Position:** $\mathbf{g}^P_{pos} \in \mathbb{R}^3$, representing the relative 3D position of the object's center of mass with respect to the hand's palm. (2) **Relative Orientation:** $\mathbf{g}^P_{ori} \in \mathbb{R}^4$, representing the relative 4D quaternion of the object's center of mass with respect to the hand's palm. (3) **Goal Contact:** $\mathbf{g}_{fj} \in \mathbb{R}^{18}$, representing the desired final configuration of the hand's actuated finger joints upon achieving the grasp.

Observations: This task requires the agent to adapt to different goal poses $\mathbf{g}$ and different objects O. Consequently, the policy $\pi(\mathbf{a}|\cdot)$ needs to condition on arm joint $\mathbf{J}^a$, hand base $\mathbf{b}$, hand joints $\mathbf{J}^h$, goal pose $\mathbf{g}$ and object pose o^P(with respect to hand palm). The 6-D object pose o^P is represented as $o^P = [o^P_p, o^P_q]$, where $o^P_p \in \mathbb{R}^3$ denotes the 3-D position of the object's center of mass with respect to the hand's palm. $o^P_q \in \mathbb{R}^4$ represents the 4-D 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 manipulation is successful if $\phi_p <= \epsilon_{pos}$ *and* $\phi_\theta <= \epsilon_{ori}$ *and* $\phi_j <= \epsilon_{fj}$. Where the ϕ_p represents the distance between object position o^P_p and goal position $\mathbf{g}^P_{pos}$:

$$\phi_p \left(o^P, \mathbf{g}^P_{pos}\right) = \|o^P - \mathbf{g}^P_{pos}\|_2 \quad (2)$$

ϕ_θ represents the distance between object orientation o^P_q and goal orientation $\mathbf{g}^P_{ori}$:

$$\phi_\theta \left(o^P_q, \mathbf{g}^P_{ori}\right) = 2 \arcsin \left(\left(o^P_q \cdot (\mathbf{g}^P_{ori})^{-1}\right)_4 \right) \quad (3)$$

ϕ_j represents the distance between finger joint $\mathbf{J}^f$ and goal joint $\mathbf{g}_{fj}$:

$$\phi_j \left(\mathbf{J}^f, \mathbf{g}_{fj}\right) = \|\mathbf{J}^f - \mathbf{g}_{fj}\|_2 \quad (4)$$

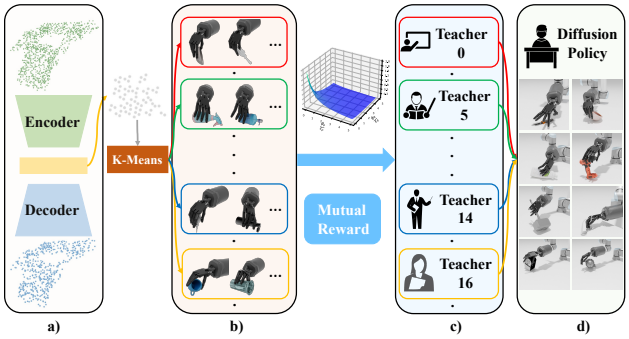

Fig. 2: The UnidexFPM Pipeline. a) Employing autoencoder to learn latent representations base on object-hand point cloud. b) Utilizing K-Means to cluster the entire training set into N clusters based on the learned representations. c) Learning an expert for each cluster based on mutual reward. d) Distilling multi-expert knowledge into a single student using diffusion policy for general dexterous functional pre-grasp manipulation of both seen and unseen objects.

ϵ_{pos}, ϵ_{ori}, and ϵ_{fj} represent the desired distance threshold for position, orientation, and contact respectively.

IV. Method

In dexterous functional pre-grasp manipulation, the high dimensionality of the dexterous hand results in a vast policy space. However, the task itself encompasses an exceedingly limited solution space, as successful pre-grasp manipulation requires achieving precise goals that simultaneously fulfill position, orientation, and contact requirements.

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

To tackle these challenges, we employ the teacher-student framework [7], as shown in Figure 2. This framework utilizes 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 get high-performance experts without constraining access to privileged information [7]. We introduce a novel mutual reward to enable the learning of dexterous functional pre-grasp manipulation policies, followed by the utilization of a mixture of experts to enhance the overall performance of the teacher policy.

1) Mutual Reward: Reward shaping is a crucial aspect of training a proficient RL agent. In our task, even when provided with privileged information, conventional reward shaping approaches, such as adding distance rewards for each goal component [7], [9], [10], can readily cause the RL agent to become trapped at a local minimum, as shown in Figure 3a). This type of reward incentivizes the RL agent to prioritize optimizing distance rewards that are easy to achieve, such as position distance ϕ_p and contact distance ϕ_j, which can be easily adjusted by manipulating the hand base and hand joint. However, the RL agent tends to disregard the orientation distance ϕ_θ, which requires reorienting the object to minimize.

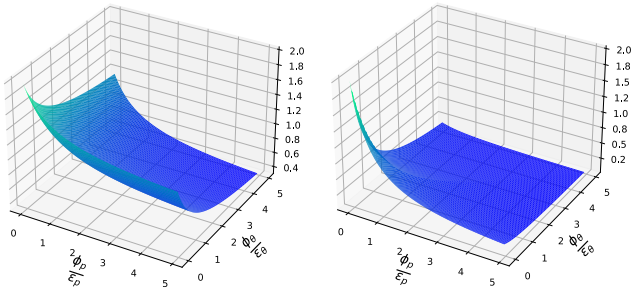

Fig. 3: Compasion of Rewards. To compare different rewards, we select position and orientation distance, setting $w_p = w_\theta = 1$ for simplicity. We visualize the trend of reward with the changing of $\frac{\phi}{\epsilon}$. a) Sum reward: Optimizing towards either distance reward leads to an increase in the total reward. b) Mutual reward: The total reward increases only when all distance rewards are jointly optimized.

To tackle the problem, we propose a novel mutual reward. More specially, we first define a normalization function ψ, to standardize different distance rewards into range [0,1]:

$$\tilde{r} = \psi(\phi, \epsilon) = \frac{\epsilon}{\phi + \epsilon} \tag{5}$$

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

$$r_{\text{dist}} = \tilde{r}_{\text{min}} \left(w_p \tilde{r}_p + w_\theta \tilde{r}_\theta + w_j \tilde{r}_j \right) \tag{6}$$

Here, w_p, w_θ, and w_j are hyperparameters. By incorporating this restriction term, simply minimizing the position distance ϕ_p or contact distance ϕ_j will not lead to a rapid increase in the total distance reward, as the orientation distance ϕ_θ is typically large, resulting in a small value for $\tilde{r}_{\text{min}}$, as illustrated in Figure 3 b). This reward mechanism compels the RL agent to jointly optimize all three distance rewards. This enables the RL agent to successfully learn the dexterous functional pre-grasp manipulation policy.

In addition to the mutual reward, we incorporate an action penalty, denoted as r_{ap}, to regulate arm motion:

$$r_{\text{ap}} = \|\mathbf{a}^b\|_2 \tag{7}$$

This penalty aims to discourage excessive arm motion and encourage the agent to utilize fingers for object manipulation. The success reward, r_{succ}, is assigned a value of 1 if the manipulation is successful. Therefore, the total reward becomes:

$$r = r_{\text{dist}} + w_{ap} * r_{\text{ap}} + w_{succ} * r_{\text{succ}} \tag{8}$$

Where w_{ap} represents the hyperparameter for the action penalty, and w_{succ} represents the hyperparameter for the success reward.

2) Mixture of Experts: Given the need for our task to generalize across diverse objects and goal poses, the manipulation process can exhibit considerable diversity. Consequently, it is challenging for RL agents to learn a good policy for all goals. While Unidexgrasp [13], [14] introduced a framework for learning dexterous grasping for diverse objects by starting with "GeoCurriculum", which gradually increases the object instances and categories from a single object with a single pose. However, 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] train an autoencoder on object geometry for the reconstruction task and then use the latent representation of each object for state-based clustering. In the case of 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 mixture of experts, 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 Section III. Given the complexity and diversity of the action distribution resulting from the intricate manipulation process and the mixture of experts, coupled with the high dimensionality of the dexterous hand, we opt to utilize a diffusion policy [15] to model the 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 [15], for each trajectory τ_i with a stepsize of $\mathbf{L}_i$, we sample every sequence with 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 $\mathbf{n^t}$. 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:

$$\mathcal{L} = \mathrm{MSE}(\mathbf{n^t}, \mathbf{n}_\theta(o_j^D, \mathbf{A}_j^0 + \mathbf{n^t}, \mathbf{t})) \tag{9}$$

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 [37] 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:

$$\mathbf{A}_{s_i}^{\mathbf{t}-1} = \alpha(\mathbf{A}_{s_i}^{\mathbf{t}} - \gamma \mathbf{n}_\theta(o_{s_i}^D, \mathbf{A}_{s_i}^{\mathbf{t}}, \mathbf{t}) + \mathcal{N}(0, \sigma^2 I)) \tag{10}$$

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 [38], 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.

Variable	Dimension	Description	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)
$\boldsymbol{v}_o$	(3,)	object linear velocity	$\boldsymbol{w}_o$	(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

TABLE I: Teacher Observation. The superscript P represents the variable is with respect to the hand-palm coordinate.

Mixture of Experts: For each goal pose $\mathbf{g}_k$ in the set $\{\mathbf{g}_k\}_{k=1}^{O}$, we sample 1024 points from the corresponding object mesh and hand mesh. These point clouds are encoded using PointNet++ [39], and the reconstruction loss is computed with Chamfer Distance. The entire task space is divided into 20 clusters.

Diffusion Policy: We configure $\mathbf{T}_p = 4$, $\mathbf{T}_o = 2$, and $\mathbf{T}_a = 1$. Because we use the relative action for policy learning, we use the transformer backbone [40] for handling quick and sharp changes in action sequence [15].

V. EXPERIMENT SETUPS

A. Task Simulation

Environment setup: We created a simulation environment based on Isaac Gym [41] using ShadowHand and UR10e robots. Each environment consists of an object randomly placed on a table, the object's mass is randomized from 0.01kg to 0.5kg due to the diversity of object categories we have. A UR10e robot is positioned outside the table with the ShadowHand mounted on the end of the arm, as shown in Figure 1. The max episode length is 300 steps. Episodes terminate if reach the goal pose, or prematurely if the object falls off the table or the maximum steps are reached.

Goal pose generation: Currently, there exists no publicly available functional grasp pose dataset. We utilize the Oakink dataset [42], which covers diverse functional intents for a wide range of objects. However, since this dataset is based on 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 [43] for optimization. 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. 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 teacher policy, we compare our method with the following methods: 1) **PPO-Sum**: In this baseline, we adopt a sum reward approach, combining three distance rewards for RL training, while keeping other rewards the same as *Ours*. Based on our proposed reward, we further conduct experiments on 2) **Ours-SE**: Here, we only train a single expert for the entire training set. 3) **Ours-MoEF**: In this comparison, we utilize a mixture of experts. However, rather than training them from scratch, we fine-tune them from the *Ours-SE*. 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 1) **PPO-OS**: This baseline employs PPO [38] as a one-stage method. It uses the same mutual reward as *Ours* but without teacher-student learning.

2) **BC**: Behavior Cloning servers as an offline imitation learning framework, learning directly from expert demonstrations via supervised learning. This baseline employs the same settings and teacher policy as *Ours*. 3) **Dagger**: Dagger [44] is an online imitation learning framework, that tackles the covariate shift problem through iterative sampling with a learned policy via online interaction.

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

VI. RESULTS

A. Teacher Policy Comparison

As depicted in Table II, when lacking mutual reward, the RL agent fails to explore a successful manipulation policy.

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 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.

By observing the learning policy, we found the agent rapidly learns to align positions and contacts. However, the agent is stuck at this local minima and fails to align orientations. In contrast, our mutual reward prevents the agent from prematurely optimizing towards part of distance rewards. Instead, it encourages the agent to simultaneously optimize each distance reward, leading to a significant improvement in success rate from 0.0% to 58.0%.

Compared to a single expert, utilizing multiple experts improves the success rate from 58.0% to 75.0%. Additionally, we conducted an experiment to show that training from scratch is better than fine-tuning from a generalist. We sampled five clusters with varying learning difficulty from the entire training set due to computational cost. "Ours (sub)" was trained from scratch on each cluster, while "Ours-MoEF (sub)" was fine-tuned from the pre-trained single expert "Ours-SE (sub)". As shown in Table II, training from scratch performs better overall. This is due to the diversity of objects and poses in our dataset and the complexity of manipulation, making it challenging to transfer a general policy to objects and poses with significant variability.

B. Student observation based Comparsion

Method	Teacher	Succ (Train)	Succ (Test)
PPO-OS	-	6.5%	6.1%
Dagger	SE	52.2%	52.3%
Dagger	MoE	17.5%	17.3%
Ours	MoE	73.7%	70.1%

TABLE III: Success Rate of Student Observation-Based Policy. In here, we present the results of the methods without requiring demonstrations for training. "SE": single expert; "MoE": mixture of experts; "Succ (Train)": success rate on the training set; "Succ (Test)": success rate on the testing set.

As indicated in Table III, the *PPO-OS* baseline, which undergoes end-to-end training, exhibits a very low success rate even after interacting with the environment for 5.76 billion steps. Regarding Dagger, while it can achieve performance comparable to the single-expert policy, we observed it struggles to learn an effective policy under the mixture of experts. We suspect this challenge arises from the high diversity and complexity of the expert policies, leading to continuous changes in the action distribution. Consequently, the agent struggles to learn a robust policy in such a non-stationary scenario.

Compared to methods requiring interaction with the environment, offline imitation learning methods demonstrate superior results. Due to the critical role of data quantity in imitation learning, we conduct comparisons between *Ours* and *BC* across various demonstration numbers. As shown in Figure 4, *Ours* is consistently better than *BC* on both the training and testing sets and outperforms *BC* when have limited demonstration numbers. Notably, using only half the number of demonstrations required by *BC*, *Ours* can still achieve comparable performance. With a large number of demonstrations, *Ours* can approach teacher-level performance.

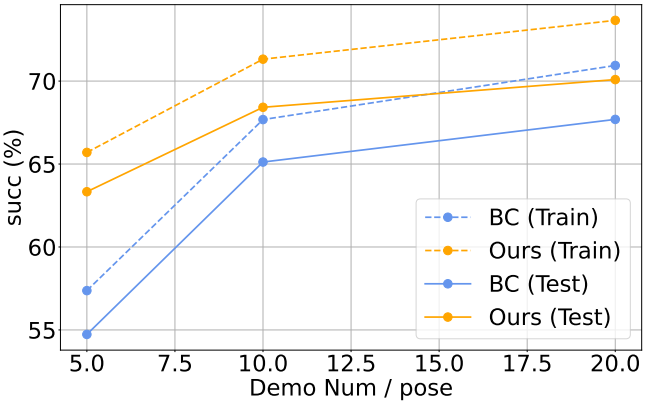

Fig. 4: Success Rate of Ours and BC under different demonstration numbers. "Succ": success rate; "Demo Num / Pose": the number of demonstrations collected for each pose in the training set, used for distilling the student policy.

C. Difficulity of general dexterous functional pre-grasp manipulation

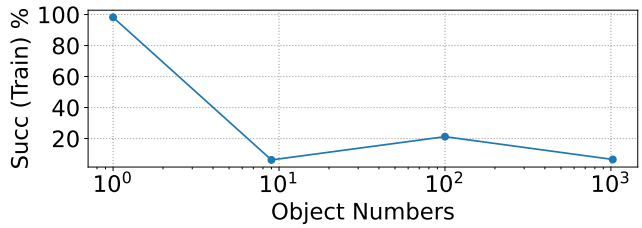

Fig. 5: Success Rate of One-stage PPO under Different Sizes of Training Set. "Succ (Train)": success rate on the training set. As the number of objects increases, finding a general manipulation policy across diverse objects becomes increasingly challenging for one-stage PPO.

To demonstrate the difficulty of learning general dexterous functional pre-grasp manipulation, we conducted experiments using one-stage PPO, incorporating our mutual reward. We trained PPO across varying numbers of objects, for each PPO model, we trained until convergence or until the maximum interaction steps (5.76 billion) was reached.

As depicted in Figure 5, 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 a mixture of experts.

D. Ablation on geometry type

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 IV: 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, we conduct this experiment using 5 demonstrations per pose.

From a common sense perspective, having information about an object's geometry is crucial for manipulating various objects. However, as shown in Table IV, while providing more detailed geometry information can lead to a better student policy, it does not significantly affect performance. By observing the learned policy, we discovered that our policy utilizes extrinsic dexterity, such as using the table to roll objects or leveraging inertia to aid in manipulating objects, as shown in Figure 1. Moreover, our policy learns to adjust based on feedback, as depicted in Figure 6. These capabilities enhance the agent's ability to generalize to different objects and goal poses.

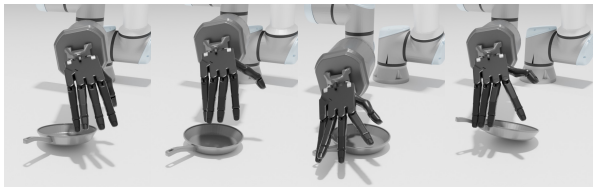

Fig. 6: Adjustment of Our Learned Policy. Although our policy failed to pull up the pan initially, it adjusted by lowering the arm on the second attempt, successfully pulling up the pan.

However, these capabilities also have drawbacks. We observed examples 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.

E. Robustness under noisy object pose observation

	1 ϵ	1.5 ϵ	2 ϵ
0°, 0cm	70.1%	77.8%	81.2%
2°, 2cm	38.1%	67.7%	75.2%
5°, 5cm	0.0%	6.5%	8.71%

TABLE V: Success Rate under Different Levels of Object Pose Estimation Noise and Success Threshold. The Gaussian noise is determined by the standard deviation of the specified threshold and added separately to observations of wrist position and orientation. Any noise exceeding these bounds will be clipped.

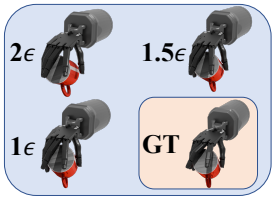

Fig. 7: Visualization of Achieved Functional Pose under Different Success Thresholds. Even when the threshold is doubled, the achieved functional pose remains meaningful and comparable to the GT pose.

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 experiment under varying levels of noisy object pose observations [45]. As depicted in Table V, 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 is also a reasonable threshold, as shown in Figure 7. By slightly adjusting the threshold, we found that our method can still achieve a high success rate. This underscores the robustness of our approach and the potential for real-world applications.

F. Performance under different object categories

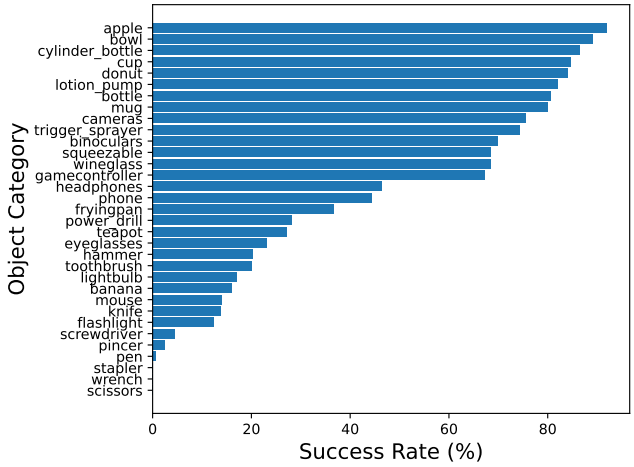

Fig. 8: Success Rate of Different Object Categories.

As depicted in Figure 8, 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 .

VII. CONCLUSIONS

In this work, we focus on general dexterous functional pre-grasp manipulation. This entails 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 to prevent the RL agent from getting stuck in local minima, greatly enhancing teacher policy learning. Furthermore, we propose employing a mixture of experts and distillation with a diffusion policy to facilitate learning diverse and complex manipulation behavior. Our experiments showcase the effectiveness and robustness of our approach, revealing its 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 the 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 sim2real gap.

Acknowledgments: We thank Jiyao Zhang (PKU), Haoran Geng (PKU), Zeyuan Chen (PKU), Tianyu Wang (PKU) for their helpful assistance.

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," *TPAMI*, 2023.

[3] A. Agarwal, S. Uppal, K. Shaw, and D. Pathak, "Dexterous functional grasping," *arXiv preprint arXiv:2312.02975*, 2023.

[4] W. Zhou and D. Held, "Learning to grasp the ungraspable with emergent extrinsic dexterity," in *CoRL*. PMLR, 2023, pp. 150–160.

[5] S. Chen, A. Wu, and C. K. Liu, "Synthesizing dexterous nonprehensile pregrasp for ungraspable objects," in *SIGGRAPH*, 2023, pp. 1–10.

[6] B. Huang, Y. Chen, T. Wang, Y. Qin, Y. Yang, N. Atanasov, and X. Wang, "Dynamic handover: Throw and catch with bimanual hands," *arXiv preprint arXiv:2309.05655*, 2023.

[7] T. Chen, J. Xu, and P. Agrawal, "A system for general in-hand object re-orientation," in *CoRL*. PMLR, 2022, pp. 297–307.

[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 *CoRL*. PMLR, 2023, pp. 594–605.

[10] C. Bao, H. Xu, Y. Qin, and X. Wang, "Dexart: Benchmarking generalizable dexterous manipulation with articulated objects," in *CVPR*, 2023, pp. 21 190–21 200.

[11] Y. Chen, T. Wu, S. Wang, X. Feng, J. Jiang, Z. Lu, S. McAleer, H. Dong, S.-C. Zhu, and Y. Yang, "Towards human-level bimanual dexterous manipulation with reinforcement learning," *NeurIPS*, vol. 35, pp. 5150–5163, 2022.

[12] T. Wu, M. Wu, J. Zhang, Y. Gan, and H. Dong, "Learning score-based grasping primitive for human-assisting dexterous grasping," *NeurIPS*, vol. 36, 2024.

[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 *CVPR*, 2023, pp. 4737–4746.

[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," *arXiv preprint arXiv:2304.00464*, 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," *arXiv preprint arXiv:2303.04137*, 2023.

[16] T. Zhu, R. Wu, X. Lin, and Y. Sun, "Toward human-like grasp: Dexterous grasping via semantic representation of object-hand," in *ICCV*, 2021, pp. 15 741–15 751.

[17] S. Brahmbhatt, C. Ham, C. C. Kemp, and J. Hays, "Contactdb: Analyzing and predicting grasp contact via thermal imaging," in *CVPR*, 2019, pp. 8709–8719.

[18] P. Mandikal and K. Grauman, "Learning dexterous grasping with object-centric visual affordances," in *ICRA*. IEEE, 2021, pp. 6169–6176.

[19] ——, "Dexvip: Learning dexterous grasping with human hand pose priors from video," in *5th Annual Conference on Robot Learning*, 2021.

[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," *arXiv preprint arXiv:1709.10087*, 2017.

[21] D. Jain, A. Li, S. Singhal, A. Rajeswaran, V. Kumar, and E. Todorov, "Learning deep visuomotor policies for dexterous hand manipulation," in *ICRA*. IEEE, 2019, pp. 3636–3643.

[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 *ECCV*. Springer, 2022, pp. 570–587.

[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 C. K. Liu, "Sequential dexterity: Chaining dexterous policies for long-horizon manipulation," *arXiv preprint arXiv:2309.00987*, 2023.

[25] M. R. Dogar and S. S. Srinivasa, "Push-grasping with dexterous hands: Mechanics and a method," in *IROS*. IEEE, 2010, pp. 2123–2130.

[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] D. Kappler, L. Chang, M. Przybylski, N. Pollard, T. Asfour, and R. Dillmann, "Representation of pre-grasp strategies for object manipulation," in *Humanoids*. IEEE, 2010, pp. 617–624.

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

[29] M. Wu, F. Zhong, Y. Xia, and H. Dong, "TarGF: Learning target gradient field for object rearrangement," *arXiv preprint arXiv:2209.00853*, 2022.

[30] Y. Song, L. Shen, L. Xing, and S. Ermon, "Solving inverse problems in medical imaging with score-based generative models," *arXiv preprint arXiv:2111.08005*, 2021.

[31] H. Ci, M. Wu, W. Zhu, X. Ma, H. Dong, F. Zhong, and Y. Wang, "Gfpose: Learning 3d human pose prior with gradient fields," *arXiv preprint arXiv:2212.08641*, 2022.

[32] R. Cai, G. Yang, H. Averbuch-Elor, Z. Hao, S. Belongie, N. Snavely, and B. Hariharan, "Learning gradient fields for shape generation," in *ECCV*. Springer, 2020, pp. 364–381.

[33] 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 *CVPR*, June 2023, pp. 16 750–16 761.

[34] Y. Li, B. Liu, Y. Geng, P. Li, Y. Yang, Y. Zhu, T. Liu, and S. Huang, "Grasp multiple objects with one hand," *RA-L*, 2024.

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

[36] H. Ha, P. Florence, and S. Song, "Scaling up and distilling down: Language-guided robot skill acquisition," in *CoRL*. PMLR, 2023, pp. 3766–3777.

[37] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, vol. 33, pp. 6840–6851, 2020.

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

[39] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *NeurIPS*, vol. 30, 2017.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[41] 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.

[42] 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 *CVPR*, 2022, pp. 20 953–20 962.

[43] 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 *ICRA*. IEEE, 2023, pp. 11 359–11 366.

[44] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *AISTATS*. JMLR Workshop and Conference Proceedings, 2011, pp. 627–635.

[45] 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 *CVPR*, 2022, pp. 2781–2790.