ArtiGrasp: Physically Plausible Synthesis of Bi-Manual Dexterous Grasping and Articulation
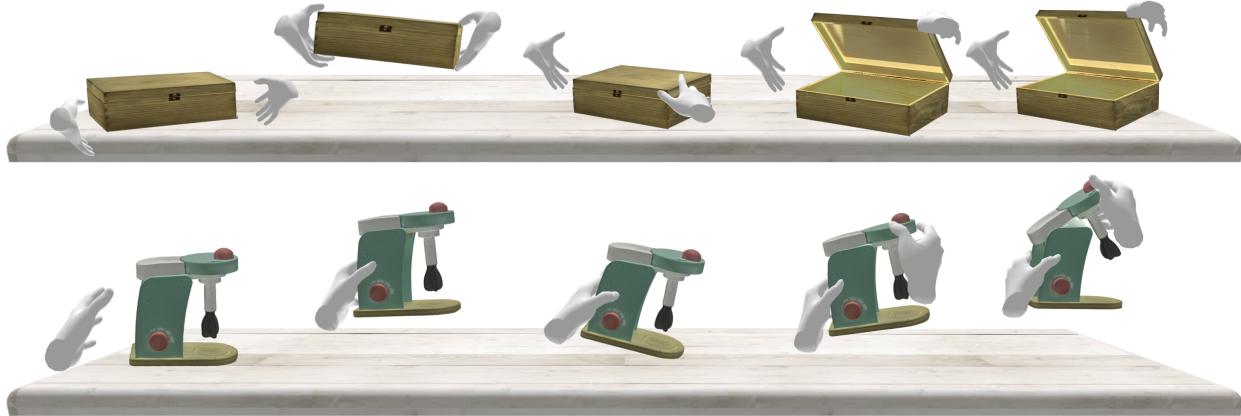
Hui Zhang[1,2*]   Sammy Christen[1*]   Zicong Fan[1,2]   Luocheng Zheng[1]
Jemin Hwangbo[3]   Jie Song[1]   Otmar Hilliges[1]

[1]ETH Zurich   [2]Max Planck Institute for Intelligent Systems
[3]Korea Advanced Institute of Science and Technology (KAIST)

Figure 1. We present a method to synthesize physically plausible bi-manual manipulation. Our method can generate motion sequences such as grasping and relocating an object with one or two hands, and opening it to a target articulation angle.

Abstract

We present ArtiGrasp, a novel method to synthesize bi-manual hand-object interactions that include grasping and articulation. This task is challenging due to the diversity of the global wrist motions and the precise finger control that are necessary to articulate objects. ArtiGrasp leverages reinforcement learning and physics simulations to train a policy that controls the global and local hand pose. Our framework unifies grasping and articulation within a single policy guided by a single hand pose reference. Moreover, to facilitate the training of the precise finger control required for articulation, we present a learning curriculum with increasing difficulty. It starts with single-hand manipulation of stationary objects and continues with multi-agent training including both hands and non-stationary objects. To evaluate our method, we introduce Dynamic Object Grasping and Articulation, a task that involves bringing an object into a target articulated pose. This task requires grasping, relocation, and articulation. We show our method's efficacy towards this task. We further demonstrate that our method can generate motions with noisy hand-object pose estimates from an off-the-shelf image-based regressor. Project page: https://eth-ait.github.io/artigrasp/.

1. Introduction

The ability to manipulate complex objects, such as operating a coffee machine, opening a laptop, or passing a box, is a fundamental part of everyday life. Providing systems with the capability to understand and perform such tasks can enable effective interactions with the physical world and provide assistance to humans in various domains. Consequently, the capacity to generate realistic hand-object interactions is paramount in fields like animation, AR/VR, human-computer interaction, and robotics. Traditional methods for capturing human motion in gaming and films, such as multi-view marker-based setups, can be costly and require substantial data cleaning for motion capture data [19, 59]. Hence, a model that can proficiently generate two-handed motions interacting with objects could reduce the costs associated with motion capture.

---

*These authors contributed equally to this work
Correspondence: huizhang@ethz.ch

Research has turned to synthesizing hand-object interactions using either data-driven [77, 81] or physics-based methods [16]. While existing data-driven methods generate hand-object motions, including object articulation [81] and two-hand manipulation [77], these methods typically depend on complete supervision from precise 3D motion data for each frame (see Table 1). Recently, physics-based methods that leverage reinforcement learning (RL) in a simulated environment have been proposed [16]. This approach reduces the data requirement for motion generation as they demand only a single hand pose reference per interaction. While physics-based approaches have primarily focused on single-hand grasping motions for rigid objects, real world hand-object interactions are often bi-manual and include articulation. However, a framework for synthesizing bi-manual grasping and articulation of objects is still missing.

Here, we go beyond single-hand grasping interaction of rigid objects and present ArtiGrasp, a novel method to synthesize dynamic bi-manual grasping and articulation of objects. We formulate this task as a reinforcement learning problem and leverage physics simulations. This allows our method to learn motions that adhere to physical plausibility, ensuring no object interpenetration and that object articulation results from stable hand-object contacts and forces. We propose a general reward function and training scheme that enables grasping and articulation of a diverse set of objects without object- or task-specific retraining.

Object articulation and bi-manual grasping present two key challenges compared to single-hand grasping. First, the articulation of different objects requires diverse wrist motions, making it challenging to define a general control strategy. For example, we show that a simple PD control scheme for relocation of objects after grasping [16] does not work well in this setting. To address this, we train an RL-based policy that learns to i) manipulate an object to a target articulation angle and ii) achieve natural interactions with the objects by utilizing only a single hand pose reference as input. The second key challenge is the precise finger control that is necessary to achieve successful articulation, where even small deviations from ideal positions on the target object impact performance. In the bi-manual manipulation setting, one hand can easily hinder the other hand from reaching its ideal position. To deal with this challenge, we introduce a learning curriculum consisting of two phases. In the first phase, we fix the object base to the surface and create separate learning environments for each hand. This allows our policies to focus on learning precise finger control for articulation. In the second phase, we fine-tune the policies using non-fixed objects in a shared physics environment, allowing the hands to cooperate.

In our experiments, we first assess both grasping and articulation separately, and then evaluate the *Dynamic Object Grasping and Articulation* task, which involves transition-

Method	Few Shot	Physics Simulation	Two-hand	Free-base Articulation
IMoS [21]	✗	✗	✓	✗
ManipNet [77]	✗	✗	✓	✗
CAMS [81]	✗	✗	✗	✓
Zhang *et al.* [80]	✗	✓	✗	✗
DexMV [50]	✗	✓	✗	✗
D-Grasp [16]	✓	✓	✗	✗
ArtiGrasp (Ours)	✓	✓	✓	✓

Table 1. **Comparison between ours and existing methods**. Ours generates two-hand manipulations using physics simulation, requires only static hand pose references (few shot), and accommodates both rigid and articulated objects with a unified policy.

ing an articulated object from its initial state into a target articulated object pose (see Fig. 1). To the best of our knowledge, there are no direct baselines for this task. Therefore, we adjust the closest related work [16] and show that simple adaptations lead to low task success rates. On the other hand, our method achieves performance gains of 5× over this baseline. We further demonstrate that our method can work with inputs from both motion capture data and noisy reconstructed poses from images with off-the-shelf hand-object pose estimation models. Lastly, we ablate the main components of our framework.

Our contributions can be summarized as follows: 1) We propose a method to achieve *Dynamic Object Grasping and Articulation* in a physically plausible manner. 2) Our method leverages RL and a general reward function to learn fine-grained wrist and finger control to grasp and articulate different objects without task- or object- specific retraining. 3) We present a learning curriculum with increasing difficulty to address the complexity of learning articulation and bi-manual manipulation. 4) We demonstrate that our method can utilize hand pose estimates from a single image as input to generate dynamic grasping and articulation.

2. Related Work

We categorize related research into human grasp generation, motion synthesis, and dexterous robotic hand manipulation. Table 1 compares different hand-object motion generation methods to ours.

2.1. Hand-object Reconstruction and Synthesis

There has been a surge of large-scale datasets introduced for modeling hand-object interaction [4, 5, 9, 24, 39, 65]. Recent datasets include interactions with articulated objects [19, 46, 69, 84]. Methods that are being developed with these datasets can be split into two categories: 1) *hand-object reconstruction*; 2) *hand grasp synthesis*.

The goal of *hand-object reconstruction* [17, 23, 28, 29, 45, 61, 71, 82, 83, 85] is to estimate the 3D hand and object surfaces from RGB images. Methods mostly leverage deep learning models, for example, by using 3D supervision and

synthetic images [28], temporal models [61], or by integrating adversarial priors and contact constraints [17]. Others focus on the denoising of pose estimates [23, 82].

In *hand grasp synthesis*, the goal is to generate static hand grasps given an object mesh as input. Corona *et al.* [17] first generate the grasp type and then refine the grasp pose accordingly. Karunratanakul *et al.* [37] introduce a part-based implicit hand model for grasp synthesis. Hidalgo-Carvajal *et al.* [30] use 10 pre-defined poses to predict infeasible grasping areas and feasible hand poses. Some methods use contact information to predict or refine the grasp configuration [23, 35, 41, 72, 75]. Moreover, physics-based optimization can be leveraged to improve the initial grasping pose [65, 66]. Han *et al.* [25] optimize a single reference grasp pose for grasping different objects in virtual reality. Some other approaches focus on more general tasks, such as generating grasps for dexterous hands and grippers [42] or using implicit representations [36] for grasping and reconstruction.

The methods above are orthogonal to ours. They do not generate hand motion but provide useful static grasp poses that we can potentially leverage to synthesize motions involving grasping and manipulating objects.

2.2. Motion Synthesis

The synthesis of human motion is a long-standing problem in graphics and computer vision [1, 2, 22, 31, 32]. Recently, there have been methods that synthesize human body motion interacting with static scenes [7, 26, 33, 40, 58, 64, 78, 79], such as sitting on a chair or sofa. This line of work focuses on the human body motion and does not generate motions of hand-object interaction. There are methods that generate full-body grasping motion [21, 43, 60, 67], but they are purely data-driven approaches without physics. They either require a post-processing optimization for the object motion [21], or only generate the approaching motion until grasping [60, 67]. Recent work has also explored integrating physics simulations into pose reconstruction [56, 57, 76], synthesis [68], and even coarse human-object interaction pipelines [8, 27, 47]. However, these physics-based methods focus on the body motion and do not model hands to capture fine-grained hand-object interactions. In contrast, we omit the human body and focus specifically on bi-manual hand-object manipulation.

Similar to ours, some recent methods also focus on dexterous hands [63] and generate interactive sequences that include physics [6, 16], articulated objects [81], or bi-manual manipulation [77]. Wang *et al.* [63] generate multi-step human-object interactions from videos without physics. Christen *et al.* [16] generate natural grasping sequences from static grasp references in a physics simulation, but focus on rigid objects and only consider one hand. Zheng *et al.* [81] presents a data-driven method to synthesize single-hand pose sequences for articulated objects and uses post-optimization to make the sequences more physically plausible. Zhang *et al.* [77] predict finger poses for two-hand object interaction. However, they require the full trajectory of the wrist and object at inference time. In our work, we generate object and wrist motions through interaction with the objects in a physics simulation.

2.3. Dexterous Robotic Hand Manipulation

Dexterous robotic manipulation is often addressed by leveraging physics simulation and reinforcement learning with robotic hands [20, 44, 52, 55, 74]. Some common strategies to learn to grasp different objects include data augmentation [13], curriculum learning [80], or improving the movement re-targeting from human demonstrations to the robot hand in a physics simulation [14, 44, 50]. Other approaches include tele-operated sequences as training data [52] or use demonstrations as prior and predict residual actions [20]. Alternatively, a parameterized reward function based on demonstrations can be used [15]. All of these methods rely on demonstration of entire grasping sequence. In contrast, our method only requires static hand pose references, which are easier to obtain.

To alleviate the need for expert demonstrations, some methods are formulated as pure RL problems [3, 10]. Chen *et al.* [12] propose a benchmark for two-hand manipulation tasks and use standard RL algorithms. However, the hand poses remain rather unnatural. They mitigate this by fine-tuning with a human preference reward [18]. Mandikal and Grauman [48] train an affordance-aware policy for grasping. They further introduce a hand pose prior learned from YouTube videos to achieve more natural hand configurations [49]. However, they only consider a single pose per object. Qin *et al.* [51] achieve grasping objects or opening doors with noisy point clouds from a single camera. Xu *et al.* [70] and Wan *et al.* [62] generate grasp sequences on objects with point clouds. Yang *et al.* [73] focus on planning interactions with chopsticks that are already in the hand from the start. All these methods consider a single robotic hand and focus mostly on rigid objects, while our method can handle two hands for both grasping and articulation of different objects with a single policy. Chen *et al.* [11] propose a benchmark for bi-manual manipulation with different rewards and separately trained policies for each individual task and object. In contrast, our method learns a general policy across different articulated objects for approaching, grasping and articulating.

3. Task Definition

The *Dynamic Object Grasping and Articulation* task is illustrated in Fig. 3. We are given an articulated object that consist of two parts rotating about an axis $\mathbf{q}_{\text{ax}}$, an initial articulated object pose $\mathbf{\Omega}^0$, a target articulated object pose $\overline{\mathbf{\Omega}}$,
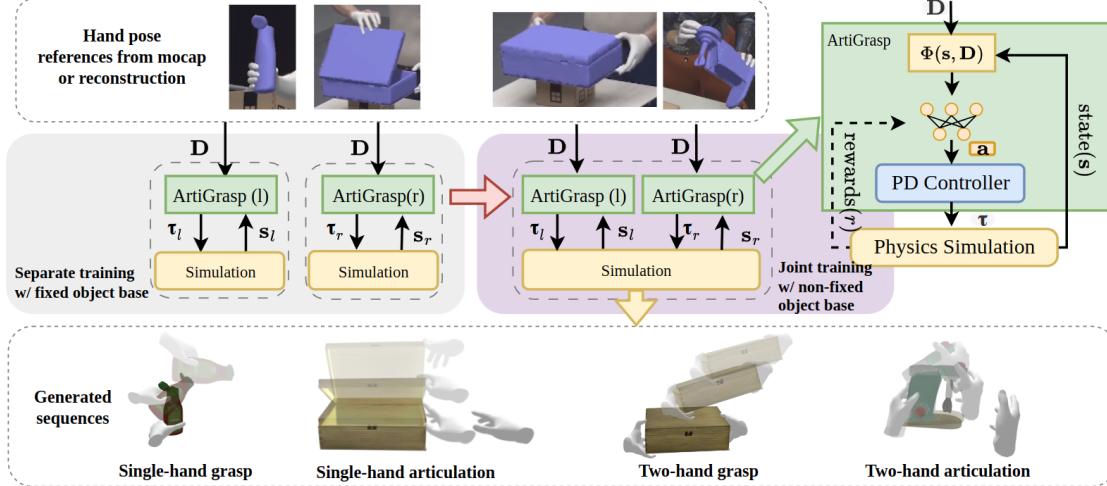
Figure 2. **Overview of Grasping and Articulation Policy**. Our method uses static hand pose references as input (top row) and generates dynamic sequences (bottom row, where higher transparency represents further in time). We propose a curriculum that starts in a simplified setting with separate environments per hand and fixed-base objects (gray solid box on the left) and continues training in a shared environment with non-fixed object base (purple solid box in the middle). Our policies are trained using reinforcement learning and a physics simulation. Rewards are only used during training. The detailed structure of our policy is shown on the right.

and two pairs of object-relative hand pose references $\mathbf{D}$ (one for grasping and one for articulation). Our goal is to generate a sequence of one or two hands interacting with the object such that the initial object pose $\mathbf{\Omega}^0$ approaches the target pose $\overline{\mathbf{\Omega}}$. An articulated object pose $\mathbf{\Omega}$ is defined by the 6 DoF global pose of the object base $\mathbf{B}$ and the 1 DoF angle ω of its articulated joint. We define the output sequence as $\{(\mathbf{q}_l^t, \mathbf{T}_l^t, \mathbf{q}_r^t, \mathbf{T}_r^t, \mathbf{\Omega}^t)\}_{t=1}^T$, where T is the number of time steps and $\mathbf{\Omega}^t$ is the articulated object pose at time step t. The hand joint rotations and the global 6D hand pose are defined by $\mathbf{q}_h^t$ and $\mathbf{T}_h^t$ where $h \in \{l, r\}$. The hand pose references $\mathbf{D} = (\overline{\mathbf{q}}_l, \overline{\mathbf{T}}_l, \overline{\mathbf{q}}_r, \overline{\mathbf{T}}_r)$ can be obtained from motion capture or grasp predictions [19] (see our experiments in Section 6).

4. Grasping and Articulation Policy

We provide an overview of our policy learning framework in Fig. 2. Since we formulate the problem identically for both hands, we will omit the notation "h" for simplicity in this section. ArtiGrasp is reinforcement learning based, and hence takes as input a state $\mathbf{s}$, provided by a physics simulation, and the hand pose reference $\mathbf{D}$. A feature extraction layer Φ transforms these inputs and passes them to our policy network. We train a policy $\boldsymbol{\pi}(\mathbf{a}|\Phi(\mathbf{s}, \mathbf{D}))$ for each hand. The policy predicts actions $\mathbf{a}$ as PD-control targets, from which torques $\boldsymbol{\tau}$ are computed. The torques are applied to our controllable hand model's joints in the physics simulation and the updated state is again fed to our feature extraction layer. In the physics simulation, we create controllable MANO hand models [53] with mean shape following [16].

The models have 51 DoFs each and are represented by the local hand pose $\mathbf{q} \in \mathbb{R}^{45}$ and global pose $\mathbf{T} \in \mathbb{R}^6$. The objects are represented by meshes taken from the ARCTIC dataset [19]. Each mesh is split into a base and an articulation part with a single connecting joint. We now present details about RL, the feature extraction layer, the reward function, and our learning curriculum.

4.1. RL Background

We formulate our problem as a Markov Decision Process (MDP), where the goal is to train a policy $\boldsymbol{\pi}$ to maximize the expected future reward $\mathbb{E}_{\xi \sim \boldsymbol{\pi}} \left[\sum_{t=0}^T \gamma^t r_t \right]$, where $\gamma \in [0, 1]$ is a discount factor, r_t is the reward of time step t, and $\xi = [(\mathbf{s}_0, \mathbf{a}_0), \cdots, (\mathbf{s}_T, \mathbf{a}_T)]$ a trajectory of state and action sequences generated by the policy interacting with the physics simulation. The probability distribution over all trajectories ξ is given by $p_\theta(\xi) = p(\mathbf{s}_0) \prod_{t=0}^T p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) \boldsymbol{\pi}(\mathbf{a}_t|\Phi(\mathbf{s}_t, \mathbf{D}))$, where $p(s_0)$ is the initial state distribution and $p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$ are the transitions determined by the physics simulation. From this probability we compute the expectation of the discounted future rewards. The policy $\boldsymbol{\pi}$ is represented by a neural network, whose weights are updated during training. Note that we describe the case where the two hand policies are trained in separate environments here (see Section 4.4). In the case of two hands interacting in the same environment, the update of the simulation state is influenced by the actions of both policies. To simplify notation, we omit the time indication from the equations in the following sections.

4.2. Feature Extraction

The state $\mathbf{s}$ at a time step entails the current poses of the hands and object, as well as the contacts and forces per hand joint. We convert this information into features for the policy. Since we train a left-hand and a right-hand policy, the feature space is hand-specific, however, the overall structure is identical and defined as follows:

$$\Phi(\mathbf{s}, \mathbf{D}) = (\mathcal{H}, \mathcal{O}, \mathcal{G}), \qquad (1)$$

where $\mathcal{H}$, $\mathcal{O}$, and $\mathcal{G}$ are the hand features, object features, and goal features, respectively.

The hand features $\mathcal{H}$ are defined as $\mathcal{H} = (\mathbf{q}, \dot{\mathbf{q}}, \mathbf{f}, \dot{\tilde{\mathbf{T}}})$ where $\mathbf{q}$ and $\dot{\mathbf{q}}$ are the hand joints' local rotations and velocities, $\mathbf{f}$ are the net contact forces of each link of the hand, and $\dot{\tilde{\mathbf{T}}}$ are the hand's linear and angular velocities in object-relative frame.

The object features are $\mathcal{O} = (\tilde{\mathbf{\Omega}}, \dot{\tilde{\mathbf{\Omega}}}, \mathbf{I}_{\text{art}})$. The terms $\tilde{\mathbf{\Omega}}$ and $\dot{\tilde{\mathbf{\Omega}}}$ indicate the articulated object's 7 DoF pose and velocity expressed in wrist-relative frame. We convert global information into wrist-relative features (denoted by $\tilde{\cdot}$) to make the policy independent of the global state and prevent overfitting. To provide more information about the object's state with regards to articulation to our policy, we introduce the term $\mathbf{I}_{\text{art}} = (\tilde{\mathbf{q}}_{\text{ax}}, \tilde{\mathbf{q}}_{\text{art}}, l_{\text{art}}, m_{\text{art}}, m_{\text{base}},)$, where $\tilde{\mathbf{q}}_{\text{ax}}$ and $\tilde{\mathbf{q}}_{\text{art}}$ are the direction vector of the articulation axis and the direction vector from wrist to the axis, represented in wrist-relative frame. The terms l_{art}, m_{art} and m_{base} indicate the distance from wrist to the articulation axis and the weights of the object's parts, respectively. We ablate this component $\mathbf{I}_{\text{art}}$ in Section 6.4.

The goal features $\mathcal{G}$ guide the policy towards the hand pose reference and the target articulation angle. They are defined as: $\mathcal{G} = (\tilde{\mathbf{g}}_q, \tilde{\mathbf{g}}_x, \mathbf{g}_c, \mathbf{g}_a)$. In particular, $\tilde{\mathbf{g}}_q = \bar{\mathbf{q}} - \mathbf{q}$ is the distance between the target and the current hand joint rotations (including wrist). The term $\tilde{\mathbf{g}}_x = \bar{\mathbf{x}} - \mathbf{x}$ is the distance between the target and the current hand joint position, which can be computed from the hand pose using forward kinematics. $\mathbf{g}_c = [\bar{\mathbf{c}}|\bar{\mathbf{c}} - \mathbf{c}]$ contains the target contacts and the difference between the target and the current binary contact vector. $\mathbf{g}_a = \bar{\omega} - \omega$ is the difference between the the target and the current object articulation angle. The target position, pose, and contacts are extracted from the hand pose reference. The target articulation angle is set to zero for grasping and otherwise set to a random angle during training. The goal features are expressed in either the object's base or articulation coordinate frame, depending on the part that needs to be manipulated.

4.3. Reward Function

The individual time-step reward function should guide our policy towards a solution that imitates the reference pose and fulfills the task objectives at the same time. Therefore, we define it as follows:

$$r = r_{\text{im}} + r_{\text{task}}, \qquad (2)$$

where r_{im} is the reward for imitating the reference pose and r_{task} contains the task objective. The imitation reward is defined as:

$$r_{\text{im}} = r_p + r_c + r_{\text{reg}}. \qquad (3)$$

The pose reward r_p considers both the joint position and joint angle error. The joint position error is the weighted sum of the distances between target and current positions $\bar{\mathbf{x}}$ and $\mathbf{x}$ of every joint, and the joint angle error measures the L2-norm of the differences between the target and current finger joint (and wrist) angles $\bar{\mathbf{q}}$ and $\mathbf{q}$:

$$r_p = -\sum_{i=1}^{L} w_{px}^i ||\bar{\mathbf{x}}^i - \mathbf{x}^i||^2 - w_{pq} ||\bar{\mathbf{q}} - \mathbf{q}||, \qquad (4)$$

where w_{px}^i and w_{pq} are weights for the respective terms. The contact reward r_c is composed of a relative contact term, which corresponds to the fraction of target contacts $\bar{\mathbf{c}}$ the hand has achieved, and a contact impulse term, which encourages the amount of force f applied on desired contact joints, capped by a factor proportional to the object's weight m_o:

$$r_c = w_{cc} \frac{\bar{\mathbf{c}}^T \mathbf{I}_{f>0}}{\bar{\mathbf{c}}^T \bar{\mathbf{c}}} + w_{cf} min(\bar{\mathbf{c}}^T \mathbf{f}, \lambda m_o), \qquad (5)$$

where w_{cc} and w_{cf} indicate the respective weights. The term r_{reg} regularizes the linear and angular velocities of the hand and object:

$$r_{\text{reg}} = -w_{rh} ||\dot{\mathbf{T}}||^2 - w_{ro} ||\dot{\mathbf{\Omega}}||^2. \qquad (6)$$

The task reward r_{task} consists of two incentives: opening the object to a target articulation angle and avoiding the movement of the object base from its initial pose:

$$r_{\text{task}} = -w_{tq} ||\bar{\omega} - \omega|| - w_{tx} ||\mathbf{p}^0 - \mathbf{p}||^2, \qquad (7)$$

where $\bar{\omega}$ and ω are the the target and the current articulation angle, $\mathbf{p}^0$ and $\mathbf{p}$ are the object's initial and current position. The weights w_{tq} and w_{tx} are used to balance the terms. All the weight values are reported in SupMat.

4.4. Curriculum

Training our policies with non-stationary objects from the beginning makes it difficult to learn the precise control necessary for fine-grained articulation. To address this, we introduce a learning curriculum that consists of two phases. In the first phase, we fix the objects to the table surface and train each hand separately in its own physics environment

(grey shaded box in Fig. 2). This lets the policies learn precise finger movements and articulation. It also enables faster training, since the physics simulation speed scales roughly quadratically with the number of contacts in the environment. In the second phase, we move to the more complex setting where the object base is not fixed to the surface and the hands are both simulated in the same environment (purple shaded box in Fig. 2). In this setting, the policies need to learn to articulate the object without moving the object base or even tipping the whole object over. Additionally, the hands must collaborate, i.e., one hand should grasp the object without moving it too much, such that the other hand can successfully manipulate the object. In Section 6.4, we ablate the effectiveness of our curriculum.

5. Sequence Generation

Given the unified policy per hand that can grasp and articulate objects, we now solve the *Dynamic Object Grasping and Articulation* task (see Section 3) by combining the different subtasks. To achieve this, we use two pairs of hand pose references $\mathbf{D}^{grasp}$ and $\mathbf{D}^{art}$. In the first phase, the hand policies are executed until a stable grasp is reached. In this case, the target object articulation angle $\overline{\omega}$ is set to zero and $\mathbf{D}^{grasp}$ is used as input. To move the object to its target 6D global pose, we use the policies to keep a stable grasp on the object and employ the motion synthesis module according to D-Grasp [16]. Note that in the case where the hand pose reference contains only single hand manipulation, we simply fix the other hand. After having relocated the object, we need to transition from grasping into pre-grasp poses for articulation. This is achieved through a heuristics-based control scheme. First, we release the grasps by bringing the fingers into open hand poses and moving them away from the object following the direction that points from the object center to the wrist. Next, we linearly interpolate a trajectory between the hand poses and pre-grasp poses for articulation $\mathbf{D}^{art}$. The pre-grasp poses correspond to $\mathbf{D}^{art}$ with a linear translation in global space. They are computed by setting them at a small distance away from direction of the object center to the wrist poses of the reference. In the last phase, we use our articulation policy to approach the object and articulate it to reach the target articulation angle.

6. Experiments

We conduct several experiments to evaluate our framework. We first report experimental details in Section 6.1. We then conduct quantitative evaluations on grasping and articulation tasks in Section 6.2 and Section 6.3, including experiments with imperfect hand pose references from images. Finally, we provide ablations to show the importance of our method's components in Section 6.4. Please see our SupMat video for qualitative examples.

6.1. Experimental Details

Implementation Details We use PPO [54] for RL training and RaiSim [34] for the physics simulation. We train all policies using a single Nvidia RTX 6000 GPU and 128 CPU cores. Training our method takes roughly three days. We will release all code and data for future research.

Dataset We utilize the ARCTIC dataset [19], which contains fully annotated two-hand interaction sequences including dexterous grasping and manipulation of articulated objects. We separate all available sequences into the different interactions of grasping and articulation. For each interaction, we extract a single pair of hand pose references using heuristics (see SupMat for details). Since the ground-truth annotation for the test split is not released, we create a custom 65%/35% train/test-split over all sequences of the training and validation sets, with a total of 488 and 257 hand pose references for training and testing, respectively. For the experiments with image-based estimates, we only use the validation set consisting of 60 hand pose references, because the pose estimation model was trained on the ARCTIC training set (see Section 6.3).

Metrics We mostly follow related work [16, 35] and define three metrics for grasping (success rate, position and angle error), three metrics for articulation (success rate, simulated distance and angle error), and one additional metric for the *Dynamic Object Grasping and Articulation* task. We omit the interpenetration metric since all of our baselines include a physics simulation which exhibits no interpenetration.
Grasp Success Rate (Suc. G): A grasp is defined as success if the object is lifted higher than 0.1m and does not fall until the sequence terminates.
Position Error (PE): The mean position error between the object's final and target 3D position in meters.
Angle Error (AE): The mean angle error between the object's final and target base orientation measured as geodesic distance in radian.
Articulation Success Rate (Suc. A): An articulation is defined as success if the hand can articulate the object for more than 0.3 rad and the articulated part does not slip until sequence termination.
Articulation Angle Error (AAE): The mean error between the object final and target articulation angle in radian.
Simulated Distance (SD): As articulation should not move the object base, we report average displacement of the object base in meters.
Task Success Rate (Suc. T): We deem a task as success if the PE < 0.05m, the AE < 0.2rad, and the AAE < 0.5rad.

Baselines D-Grasp is the most related to ours (see Tab. 1) [16], so we propose baselines following it:
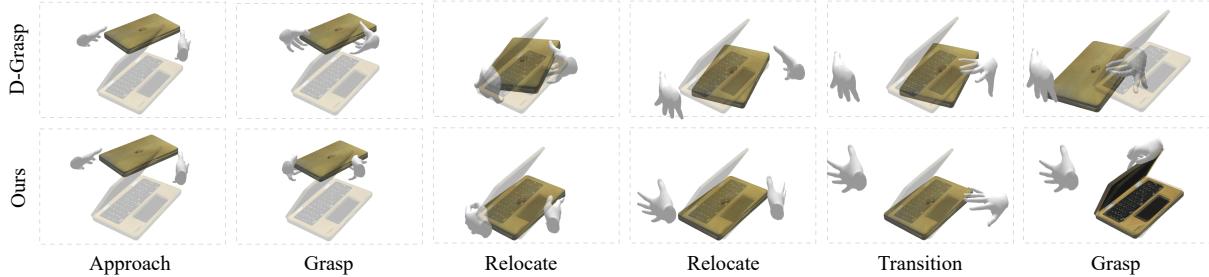**D-Grasp**: For grasping, we use vanilla D-Grasp and train

Figure 3. **Qualitative evaluation of *Dynamic Object Grasping and Articulation*.** D-Grasp can grasp and relocate the object successfully, but fails to articulate the object. Ours is more successful at tackling this task and can articulate the object after relocation.

Model	Grasping			Articulation		
	Suc. G ↑	PE ↓	AE ↓	Suc. A ↑	AAE ↓	SD ↓
PD+IK	0.13	1.20	1.50	0.28	0.80	0.39
D-Grasp	**0.72**	**0.12**	**0.62**	0.22	0.93	0.49
Ours	0.71	0.13	0.69	**0.55**	**0.57**	**0.01**

Table 2. **Quantitative comparison for grasping and articulation tasks**. When the tasks are decoupled, we find that our method outperforms the baselines on articulation and performs comparably to D-Grasp on grasping.

Models	Suc. T ↑	PE ↓	AE ↓	AAE ↓
D-Grasp	0.11	0.05	0.15	0.66
Ours	**0.50**	**0.03**	**0.10**	**0.41**

Table 3. **Evaluation for our *Dynamic Object Grasping and Articulation* task**. Our method outperforms D-Grasp on all metrics when evaluated on the task of transitioning an articulated object into a target articulated object pose.

the policies of the two hands directly with non-stationary objects. To compare D-Grasp to our method for articulation, we adjust the wrist control in D-Grasp. We first gradually increase the angle of the articulated joint and calculate the target 6D wrist pose with inverse kinematics by assuming that the wrist is fixed to the articulated part of the object. We then feed the wrist target pose to the PD controller.

PD+IK: We use the hand reference poses and set them as targets to the PD controller. The wrist for the articulation is controlled in the same way as in D-Grasp.

6.2. Evaluation

We first evaluate grasping and articulation tasks separately and then conduct experiments on the *Dynamic Object Grasping and Articulation* task (see Section 3).

Grasping and Articulation For grasping, we pre-sample 30 6D target object poses randomly (see SupMat for details). To control the wrist movement for relocation after grasping the object with our method and the PD+IK baseline, we adopt the same motion synthesis module as in [16] (see Section 5). For articulation, we evaluate each hand pose reference on 5 target articulation angles: $\{0.5, 0.75, 1.0, 1.25, 1.5\}$ rad. For both tasks, the initial hand poses are set at a pre-defined distance away in the direction that points from the object center to the wrist of the hand pose references, with partially opened hands. The quantitative results are shown in Tab. 2. Our method significantly outperforms the PD+IK baseline on both grasping and articulation. Our policy has considerably

better articulation performance and comparable grasping performance compared with D-Grasp. The results also show the difficulty of articulation and indicate that our learning-based wrist control is favorable for articulation compared to D-Grasp's non learning-based approach. Furthermore, as shown in the qualitative result Fig. 4, we observe some recovering behavior from failure cases, which indicates the robustness of our policy. In particular, the agent fails to grasp first but tries again to find a better grasp until it succeeds in articulating the object. Qualitative comparisons and more examples of generated sequences are presented in SupMat figures. To evaluate the generalization ability of our framework, we conduct a proof-of-concept experiment with a single left-out object. The result indicates that our framework can generalize to an unseen object with about 15% performance drop. We hypothesize that this is because the hand pose reference serves as a strong prior to the policy. However, more thorough evaluations need to be carried out once accurate 3D datasets with more articulated objects and hand-object poses become available.

Dynamic Object Grasping and Articulation To evaluate this task (see Section 3), we combine all grasping hand pose references with all articulation hand pose references per object, and sample a random target articulated object pose per trial (see SupMat for details). This results in roughly 7500 evaluation trials in total. We report the average pose errors and the task success rate in Tab. 3. Our method outperforms D-Grasp significantly in all metrics. For example, our method achieves a success rate of 5× than that of D-Grasp. This shows that while D-Grasp can perform well

Figure 4. **Qualitative articulation result.** The hand shows some recovery ability from failure cases. Zoom in for details.
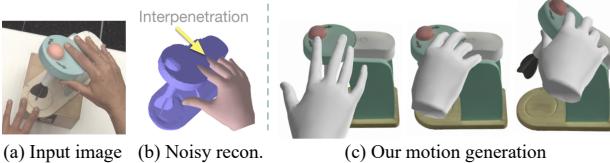

Figure 5. **Motion generation.** Our method can synthesize new motion sequences (c) with a noisy hand pose reference (b) reconstructed from a single RGB image (a).

in grasping when decoupled, it struggles in this composed task. We provide a qualitative comparison in Fig. 3 and a demonstration of a longer sequence with multiple objects in our SupMat figures. Moreover, see our SupMat video for more qualitative examples.

6.3. Generation with Reconstructed Hand Pose

We now evaluate our method with hand pose references obtained from image predictions via the off-the-shelf hand-object pose regressor from ARCTIC [19]. In particular, we estimate hand and object poses from images of the unseen validation subject in the ARCTIC and use the reconstructed results as input to our method and baselines. We separate the evaluation into grasping and articulation and present the results in Tab. 4. Despite reconstruction noise such as hand-object interpenetration, our method can retain comparable performance as in the experiment with hand pose references from motion capture. This indicates our robustness to prediction noise and its potential to synthesize new motions with hand-object pose references from single images. An example of our generated motion is shown in Fig. 5.

6.4. Ablations

We ablate the impact of the newly introduced components on our framework. To this end, we compare our full method against i) training both hands cooperatively and with a non-stationary object from the start of training (*w/o curriculum*) ii) training the hands separately and with a fixed-base object (*w/o cooperation*). Additionally, we train our method without the articulation features $\mathbf{I}_{art}$ (*w/o art. features*, cf. Section 4). The results are presented in Tab. 5. Without the curriculum, the policy achieves slightly better performance for grasping, but struggles with articulation. This is because grasping has different wrist motion with articulation which is easier to be learnt, which indicates the importance of a controlled setting to learn fine-grained articulation first. When training the hands separately without coop-

	Grasping			Articulation		
Models	Suc.G ↑	PE ↓	AE ↓	Suc.A ↑	AAE ↓	SD ↓
D-Grasp	0.60	**0.16**	**0.78**	0.20	1.07	0.63
Ours	**0.64**	**0.16**	0.80	**0.54**	**0.55**	**0.01**
Ours*	0.67	0.14	0.95	0.54	0.53	0.01

Table 4. **Results with reconstructed hand pose references.** When evaluated with predictions from images, we observe a minor drop in performance for grasping and articulation compared to mocap data. However, the overall performance shows that our method can handle noisy estimates. The asterisk (*) denotes using hand pose references from mocap.

	Grasping			Articulation		
Models	Suc. G ↑	PE ↓	AE ↓	Suc. A ↑	AAE ↓	SD ↓
w/o curriculum	**0.74**	**0.13**	**0.65**	0.36	0.77	0.02
w/o cooperation	0.21	0.32	1.43	0.48	0.65	0.02
w/o art. features	0.67	0.15	0.73	0.48	0.67	0.01
Ours	0.71	**0.13**	0.69	**0.55**	**0.57**	**0.01**

Table 5. **Ablations**. We ablate our curriculum, cooperative training, and the articulation features. All components are important aspects to achieve grasping and articulation with a single policy.

eration, grasping performance decreases because the hands cannot learn to collaborate for two-handed grasping. Lastly, the articulation features $\mathbf{I}_{art}$ improve all articulation metrics, indicating that it provides important information about the object to the policy.

7. Discussion and Conclusion

We present a method to synthesize physically plausible bi-manual grasping and articulation of objects with a single policy. We introduce an RL-based method that learns hand-object interactions in a physics simulation from static hand pose references. To address the difficulty in learning precise control for articulation, we extract articulation features and propose a curriculum with increasing task difficulty. We show our method presents a first step towards the *Dynamic Object Grasping and Articulation* task. Furthermore, we demonstrate that noisy hand-object pose estimates obtained from individual RGB images can be used as input to our method. In a proof-of-concept with a single left-out object we have shown that our policy has the potential to generalize to unseen objects, and better generalization may be achieved in the future when larger and more diverse datasets become available. A limitation of our method is that it sometimes generates unnatural poses caused by noisy hand pose references and the trade-off between the task and imitation reward, which is shown in our SupMat figures. This may be improved by integrating bio-mechanical constraints or hand pose priors obtained from data-driven methods into our framework. And generation without reference poses would be an interesting direction for the future work.

References

[1] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *International Conference on Computer Vision (ICCV)*, 2019. First two authors contributed equally. 3

[2] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. *International Conference on 3D Vision (3DV)*, 2021. 3

[3] OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Józefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020. 3

[4] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8709–8719, 2019. 2

[5] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. ContactPose: A dataset of grasps with object contact and hand pose. In *European Conference on Computer Vision (ECCV)*, pages 361–378. Springer, 2020. 2

[6] Jona Braun, Sammy Christen, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Physically plausible full-body hand-object interaction synthesis. In *International Conference on 3D Vision (3DV)*, 2024. 3

[7] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *European Conference on Computer Vision (ECCV)*, pages 387–404. Springer, 2020. 3

[8] Yu-Wei Chao, Jimei Yang, Weifeng Chen, and Jia Deng. Learning to sit: Synthesizing human-chair interactions via hierarchical control. In *AAAI Conference on Artificial Intelligence*, pages 5887–5895, 2021. 3

[9] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[10] Tao Chen, Jie Xu, and Pulkit Agrawal. A simple method for complex in-hand manipulation. In *Conference on Robot Learning (CoRL)*, 2021. 3

[11] Yuanpei Chen, Tianhao Wu, Shengjie Wang, Xidong Feng, Jiechuan Jiang, Zongqing Lu, Stephen McAleer, Hao Dong, Song-Chun Zhu, and Yaodong Yang. Towards human-level bimanual dexterous manipulation with reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5150–5163. Curran Associates, Inc., 2022. 3

[12] Yuanpei Chen, Tianhao Wu, Shengjie Wang, Xidong Feng, Jiechuan Jiang, Zongqing Lu, Stephen McAleer, Hao Dong, Song-Chun Zhu, and Yaodong Yang. Towards human-level bimanual dexterous manipulation with reinforcement learning. *NeurIPS*, 35:5150–5163, 2022. 3

[13] Zoey Chen, Karl Van Wyk, Yu-Wei Chao, Wei Yang, Arsalan Mousavian, Abhishek Gupta, and Dieter Fox. Learning robust real-world dexterous grasping policies via implicit shape augmentation. In *Conference on Robot Learning (CoRL)*, 2022. 3

[14] Zoey Qiuyu Chen, Karl Van Wyk, Yu-Wei Chao, Wei Yang, Arsalan Mousavian, Abhishek Gupta, and Dieter Fox. DexTransfer: Real world multi-fingered dexterous grasping with minimal human demonstrations. *arXiv preprint arXiv:2209.14284*, 2022. 3

[15] Sammy Christen, Stefan Stevšić, and Otmar Hilliges. Guided deep reinforcement learning of control policies for dexterous human-robot interaction. In *International Conference on Robotics and Automation (ICRA)*, pages 2161–2167, 2019. 3

[16] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-Grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 4, 6, 7, 1

[17] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5031–5041, 2020. 2, 3

[18] Zihan Ding, Yuanpei Chen, Allen Z Ren, Shixiang Shane Gu, Hao Dong, and Chi Jin. Learning a universal human prior for dexterous manipulation from human preference. *arXiv preprint arXiv:2304.04602*, 2023. 3

[19] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 4, 6, 8, 3

[20] Guillermo Garcia-Hernando, Edward Johns, and Tae-Kyun Kim. Physics-based dexterous manipulations with estimated hand poses and residual reinforcement learning. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 9561–9568. IEEE, 2020. 3

[21] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. IMoS: Intent-driven full-body motion synthesis for human-object interactions. In *Eurographics*, 2023. 2, 3

[22] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. Learning human motion models for long-term predictions. In *International Conference on 3D Vision (3DV)*, pages 458–466. IEEE, 2017. 3

[23] Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. ContactOpt: Optimizing contact to improve grasps. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3

[24] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A method for 3d annotation of hand and object poses. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[25] DongHeun Han, RoUn Lee, KyeongMin Kim, and HyeongYeop Kang. VR-HandNet: A visually and physically plausible hand manipulation system in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–12, 2023. 3

[26] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. Stochastic scene-aware motion prediction. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[27] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. *arXiv preprint arXiv:2302.00883*, 2023. 3

[28] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11807–11816, 2019. 2, 3

[29] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 568–577, 2020. 2

[30] Diego Hidalgo-Carvajal, Carlos Magno C. O. Valle, Abdeldjallil Naceri, and Sami Haddadin. Object-centric grasping transferability: Linking meshes to postures. In *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*, pages 659–666, 2022. 3

[31] Daniel Holden, Jun Saito, Taku Komura, and Thomas Joyce. Learning motion manifolds with convolutional autoencoders. In *SIGGRAPH Asia 2015 Technical Briefs*, New York, NY, USA, 2015. Association for Computing Machinery. 3

[32] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Trans. Graph.*, 35(4), 2016. 3

[33] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16750–16761, 2023. 3

[34] Jemin Hwangbo, Joonho Lee, and Marco Hutter. Per-contact iteration method for solving contact dynamics. *IEEE Robotics and Automation Letters*, 3(2):895–902, 2018. 6

[35] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *International Conference on Computer Vision (ICCV)*, 2021. 3, 6

[36] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *International Conference on 3D Vision (3DV)*, pages 333–344. IEEE, 2020. 3

[37] Korrawe Karunratanakul, Adrian Spurr, Zicong Fan, Otmar Hilliges, and Siyu Tang. A skeleton-driven neural occupancy representation for articulated hands. In *International Conference on 3D Vision (3DV)*, 2021. 3

[38] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*, 2015. 1

[39] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *International Conference on Computer Vision (ICCV)*, pages 10138–10148, 2021. 2

[40] Jiye Lee and Hanbyul Joo. Locomotion-action-manipulation: Synthesizing human-scene interactions in complex 3D environments. *arXiv preprint arXiv:2301.02667*, 2023. 3

[41] Haoming Li, Xinzhuo Lin, Yang Zhou, Xiang Li, Yuchi Huo, Jiming Chen, and Qi Ye. Contact2Grasp: 3d grasp synthesis via hand-object contact constraint. *arXiv preprint arXiv: 2210.09245*, 2022. 3

[42] Puhao Li, Tengyu Liu, Yuyang Li, Yiran Geng, Yixin Zhu, Yaodong Yang, and Siyuan Huang. GenDexGrasp: Generalizable dexterous grasping. In *International Conference on Robotics and Automation (ICRA)*, pages 8068–8074. IEEE, 2023. 3

[43] Quanzhou Li, Jingbo Wang, Chen Change Loy, and Bo Dai. Task-oriented human-object interactions generation with implicit neural representations. *arXiv preprint arXiv:2303.13129*, 2023. 3

[44] Qingtao Liu, Yu Cui, Zhengnan Sun, Haoming Li, Gaofeng Li, Lin Shao, Jiming Chen, and Qi Ye. DexRepNet: Learning dexterous robotic grasping network with geometric and spatial hand-object representations. *arXiv preprint arXiv:2303.09806*, 2023. 3

[45] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[46] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. HOI4D: A 4d egocentric dataset for category-level human-object interaction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 21013–21022, 2022. 2

[47] Zhengyi Luo, Shun Iwase, Ye Yuan, and Kris Kitani. Embodied scene-aware human pose estimation. In *NeurIPS*, 2022. 3

[48] Priyanka Mandikal and Kristen Grauman. Learning dexterous grasping with object-centric visual affordances. In *International Conference on Robotics and Automation (ICRA)*, 2021. 3

[49] Priyanka Mandikal and Kristen Grauman. DexVIP: Learning dexterous grasping with human hand pose priors from video. In *Conference on Robot Learning (CoRL)*, 2021. 3

[50] Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. DexMV: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision (ECCV)*, pages 570–587. Springer, 2022. 2, 3

[51] Yuzhe Qin, Binghao Huang, Zhao-Heng Yin, Hao Su, and Xiaolong Wang. DexPoint: Generalizable point cloud reinforcement learning for sim-to-real dexterous manipulation.

In *Conference on Robot Learning (CoRL)*, pages 594–605. PMLR, 2023. 3

[52] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *arXiv preprint arXiv:1709.10087*, 2017. 3

[53] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 4

[54] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 6, 1

[55] Qijin She, Ruizhen Hu, Juzhan Xu, Min Liu, Kai Xu, and Hui Huang. Learning high-DOF reaching-and-grasping via dynamic representation of gripper-object interaction. *ACM Transactions on Graphics (SIGGRAPH 2022)*, 41(4), 2022. 3

[56] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *Transactions on Graphics (TOG)*, 39(6):1–16, 2020. 3

[57] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, Patrick Pérez, and Christian Theobalt. Neural monocular 3d human motion capture with physical awareness. *Transactions on Graphics (TOG)*, 40(4):1–15, 2021. 3

[58] Sebastian Starke, Yiwei Zhao, Taku Komura, and Kazi Zaman. Local motion phases for learning multi-contact character movements. *Transactions on Graphics (TOG)*, 39(4): 54–1, 2020. 3

[59] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[60] Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[61] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+O: Unified egocentric recognition of 3D hand-object poses and interactions. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4511–4520, 2019. 2, 3

[62] Weikang Wan, Haoran Geng, Yun Liu, Zikang Shan, Yaodong Yang, Li Yi, and He Wang. UniDexGrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. *arXiv preprint arXiv:2304.00464*, 2023. 3

[63] He Wang, Sören Pirk, Ersin Yumer, Vladimir G Kim, Ozan Sener, Srinath Sridhar, and Leonidas J Guibas. Learning a generative model for multi-step human-object interactions from videos. In *Computer Graphics Forum*, pages 367–378. Wiley Online Library, 2019. 3

[64] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *Computer Vision and Pattern Recognition (CVPR)*, pages 12206–12215, 2021. 3

[65] Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. DexGraspNet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. In *International Conference on Robotics and Automation (ICRA)*, pages 11359–11366. IEEE, 2023. 2, 3

[66] Albert Wu, Michelle Guo, and Karen Liu. Learning diverse and physically feasible dexterous grasps with generative model and bilevel optimization. In *Conference on Robot Learning (CoRL)*, pages 1938–1948. PMLR, 2022. 3

[67] Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. SAGA: Stochastic whole-body grasping with contact. In *European Conference on Computer Vision (ECCV)*, pages 257–274. Springer, 2022. 3

[68] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based human motion estimation and synthesis from videos. In *International Conference on Computer Vision (ICCV)*, pages 11532–11541, 2021. 3

[69] Xiang Xu, Hanbyul Joo, Greg Mori, and Manolis Savva. D3D-HOI: Dynamic 3d human-object interactions from videos. *arXiv preprint arXiv:2108.08420*, 2021. 2

[70] Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruicheng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, et al. UniDexGrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4737–4746, 2023. 3

[71] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. CPF: Learning a contact potential field to model the hand-object interaction. In *International Conference on Computer Vision (ICCV)*, pages 11097–11106, 2021. 2

[72] Lixin Yang, Kailin Li, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu Lu. ArtiBoost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[73] Zeshi Yang, Kangkang Yin, and Libin Liu. Learning to use chopsticks in diverse gripping styles. *ACM Trans. Graph.*, 41(4), 2022. 3

[74] Jianglong Ye, Jiashun Wang, Binghao Huang, Yuzhe Qin, and Xiaolong Wang. Learning continuous grasping function with a dexterous hand from human demonstrations. *IEEE Robotics and Automation Letters*, 8(5):2882–2889, 2023. 3

[75] Yuting Ye and C Karen Liu. Synthesis of detailed hand manipulations using contact sampling. *Transactions on Graphics (TOG)*, 31(4):1–10, 2012. 3

[76] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. SimPoE: Simulated character control for 3d human pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7159–7169, 2021. 3

[77] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. ManipNet: Neural manipulation synthesis with a hand-object spatial representation. *ACM Trans. Graph.*, 40 (4), 2021. 2, 3

[78] Wanyue Zhang, Rishabh Dabral, Thomas Leimkühler, Vladislav Golyanik, Marc Habermann, and Christian Theobalt. Roam: Robust and object-aware motion generation using neural pose descriptors. *arXiv preprint arXiv:2308.12969*, 2023. 3

[79] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. COUCH: Towards controllable human-chair interactions. In *European Conference on Computer Vision*, pages 518–535. Springer, 2022. 3

[80] Yunbo Zhang, Alexander Clegg, Sehoon Ha, Greg Turk, and Yuting Ye. Learning to transfer in-hand manipulations using a greedy shape curriculum. In *Computer Graphics Forum*, pages 25–36. Wiley Online Library, 2023. 2, 3

[81] Juntian Zheng, Qingyuan Zheng, Lixing Fang, Yun Liu, and Li Yi. CAMS: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 585–594, 2023. 2, 3

[82] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object correspondence to hand for motion refinement. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 2, 3

[83] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5345–5354, 2020. 2

[84] Zehao Zhu, Jiashun Wang, Yuzhe Qin, Deqing Sun, Varun Jampani, and Xiaolong Wang. ContactArt: Learning 3d interaction priors for category-level articulated object and hand poses estimation. *arXiv preprint arXiv:2305.01618*, 2023. 2

[85] Andrea Ziani, Zicong Fan, Muhammed Kocabas, Sammy Christen, and Otmar Hilliges. Tempclr: Reconstructing hands via time-coherent contrastive learning. In *International Conference on 3D Vision (3DV)*, 2022. 2

ArtiGrasp: Physically Plausible Synthesis of Bi-Manual Dexterous Grasping and Articulation

Supplementary Material

The supplementary material contains this document and a video. We describe implementation details in Section A and experimental details in Section B. In Section C, we provide additional experiments. We will release all code and pre-trained models.

A. Implementation Details

For training, we use PPO [54] and follow the implementation provided in [16]. We present an overview of the important parameters and weight values of the reward function in Tab. 6.

Hyperparameters PPO	Value
Epochs	1e4
Steps per epoch	6e5
Environment steps per episode	300
Batch size	2000
Updates per epoch	20
Simulation timestep	2.5e-3s
Simulation steps per action	4
Discount factor γ	0.996
GAE parameter λ	0.95
Clipping parameter	0.2
Max. gradient norm	0.5
Value loss coefficient	0.5
Entropy coefficient	0.0
Optimizer	Adam [38]
Learning rate	5e-4
Hidden units	128
Hidden layers	2

Weight Parameters	Value
w_{px}	3.0
$w_{px,\text{fingertip}}$	12.0
w_{pq}	0.2
w_{cc}	1.5
w_{cf}	1.5
w_{rh}	0.5
w_{ro}	0.2
w_{tq}	1.5
w_{tx}	0.2
λ	5.0

Table 6. Hyperparameters of our RL algorithm and the weight values of the reward function.

B. Experimental Details

B.1. Dataset

In our experiments, we use the ARCTIC dataset [19] and include sequences from all training subjects and the recently released data from the validation subject s05. We exclude the three objects "scissors", "capsule machine", and "phone" from our experiments. "Scissors" is different from all other objects as it cannot be split into a clear base and articulation part and requires in-hand manipulation. "Phone" and "capsule machine" have very small and thin articulation parts which cannot be modeled with our method currently. We then extract hand pose references according to Section B.2. From these references, we create a 65%/35% train/test-split. In total, we generate 745 hand pose references for 8 objects from the dataset, with a train/test split of 488/257 hand pose references.

B.2. Hand Pose Reference Generation

We now describe the procedure of retrieving hand pose references from motion capture sequences. Since the sequences contain several different interactions of object manipulation, from which a lot of hand pose references could be extracted, we devise heuristics to obtain diverse frames and avoid redundancy. We distinguish between two types of manipulation in this paper: grasping and articulation.

For each sequence, we first remove all frames where none of the hands is in contact with an object. Next, we filter all remaining frames for grasping and articulation. An interaction is determined as grasping if an object is moved from its underlying surface, i.e., if the velocity of the object base $\dot{\mathbf{B}}$ is higher than a threshold ϵ_v. On the other hand, if the articulation angle ω is changed, we deem an interaction as articulation. To avoid redundancy in hand pose reference frames, we make the assumption that the hand pose does not drastically change during one interaction. Hence, we choose one frame per interaction subsequence.

B.3. Grasping and Articulation

To evaluate grasping, we generate 30 target 6D object poses for each hand-pose reference. The target positions are sampled within a range of [-0.15m, 0.15m] in x and y directions and [0.15m, 0.45m] for the z direction. The target object orientation is the initial object orientation disturbed with noise in the range of [-0.3rad, 0.3rad] for all rotation axes. To evaluate articulation, we set 5 target joint angles per trial: 0.5rad, 0.75rad, 1.0rad, 1.25rad and 1.5rad.
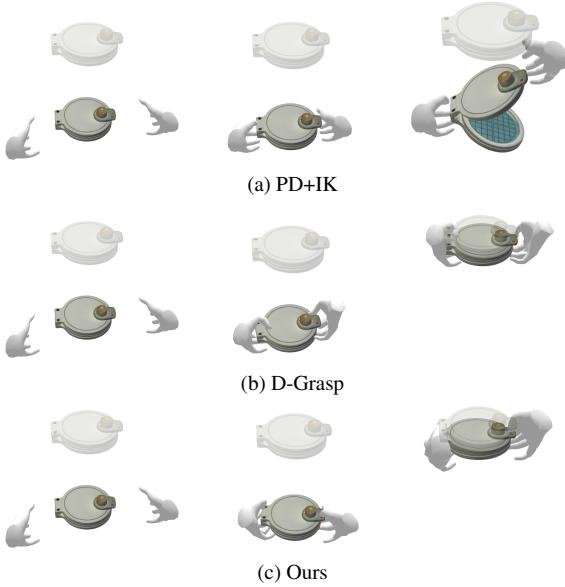
(a) PD+IK

(b) D-Grasp

(c) Ours

Figure 6. **Qualitative evaluation of grasping.** When evaluated only on grasping, PD+IK often fails to successfully grasp the object. On the other hand, D-Grasp and ours succeed at the task.
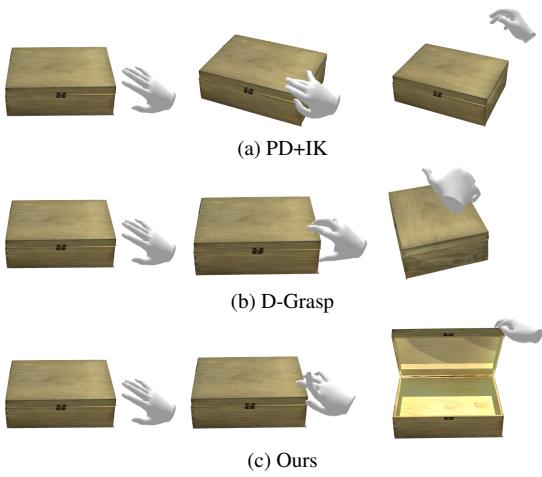

(a) PD+IK

(b) D-Grasp

(c) Ours

Figure 7. **Qualitative evaluation of articulation.** When evaluated only on articulation, both PD+IK and D-Grasp often fail at the task. On the other hand, our method can articulate the object successfully.

B.4. Dynamic Object Grasping and Articulation

For the evaluation of *Dynamic Object Grasping and Articulation*, we randomly sample the target articulated object poses $\overline{\Omega}$, which consists of the target 6D base pose and the target object joint articulation angle. The target base position is sampled within a range of [-0.1m, -0.05m] in x and y directions and 0m in z direction, since the objects should be relocated back onto the table. The target base orientation is the initial object orientation disturbed with noise in the
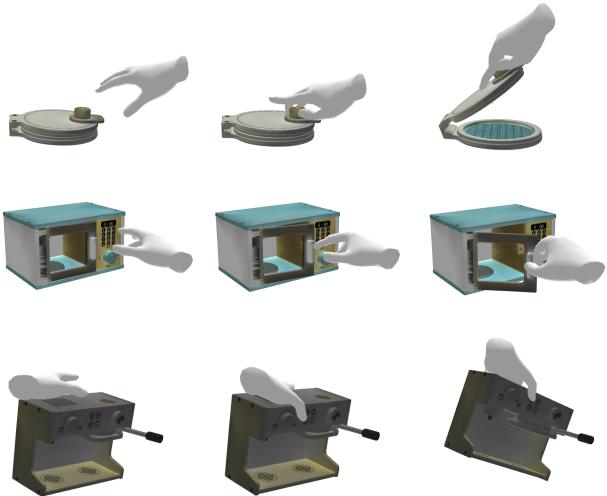

Figure 8. **Qualitative outputs of our method.** We provide more sequences for grasping and articulation, which are generated by our method with a single pair of hand pose reference label per interaction. Each sequence is shown from left to right.

range of [-0.4rad, 0.4rad] for the yaw axis. The target articulation angle is randomly sampled in the range of [0.5rad, 0.6rad].

B.5. Hand Pose Reconstruction

In this experiment, we use the pretrained image-based reconstruction model of ARCTIC [19] to predict hand-object poses from single images. Since their model is trained on the full training set and the test data is not released, we use the validation set (subject s05) to evaluate this experiment. This allows us to do a direct comparison of individual hand pose references from motion capture and image-based predictions. Our evaluation is conducted on 60 hand pose references selected with the heuristics explained in Section B.2. We retrieve the images at the corresponding timesteps and pass them to the image-based prediction model.

C. Additional Experiments

We provide additional qualitative comparisons of our method with baselines for grasping and articulation in Fig. 6 and Fig. 7, respectively. Given a pair of policies (one per hand), our method can generate diverse grasping and articulation sequences across different objects, which is shown in Fig. 8. Please see our SupMat video for more qualitative examples.

C.1. Long Sequence with Multiple Objects

Our method can generate long motion sequences in environments with multiple objects, which is shown in Fig. 9. We use a heuristics-based planner to compose the sequences.
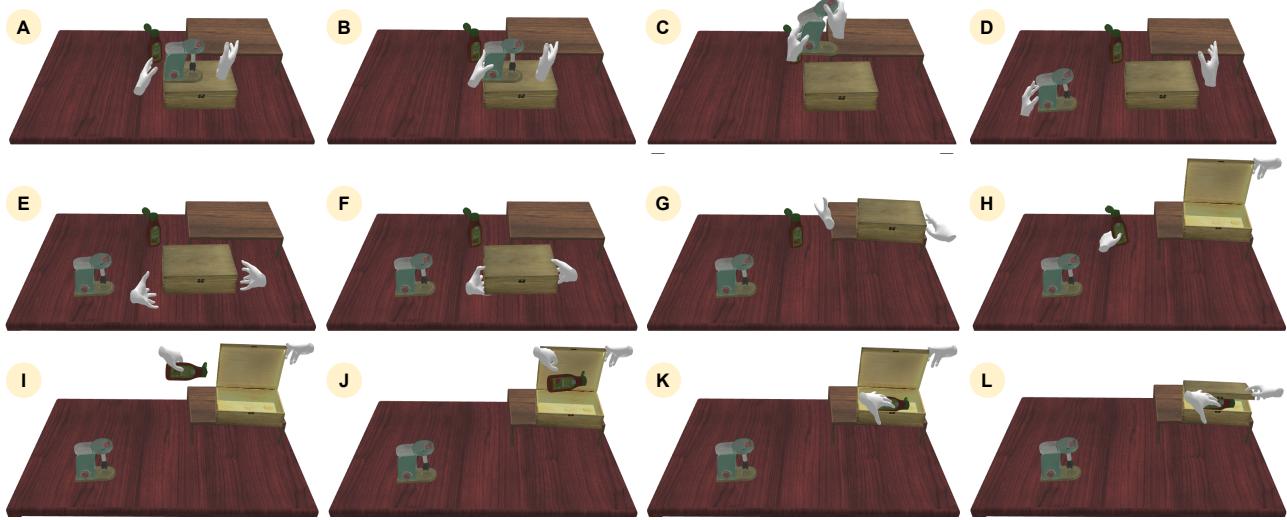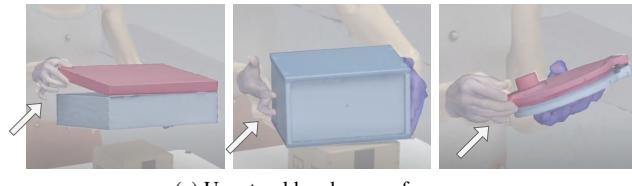
Figure 9. **Long sequence with multiple objects.** We show that our method can generate sequences of manipulating multiple objects. (A) Approaching the mixer with the left hand. (B) Grasping the mixer with the left hand. (C) Articulating the mixer with the right hand while the left hand is holding it. (D) Putting the mixer down on the table. (E) Approaching the box with both hands. (F) Grasping the box with both hands. (G) Relocating the box on the table and moving the left hand to the ketchup bottle. (H) Grasping the ketchup bottle with the left hand and opening the box with the right hand. (I) Relocating the ketchup bottle while the box is being held open. (J) Dropping the ketchup bottle into the box. (K) Moving the left hand away from the box. (L) Closing the box with the right hand.

(a) Unnatural hand pose references

(b) Unnatural generated hand poses

Figure 10. **Unnatural hand poses** (a) Some of the hand pose references we extract from the ARCTIC dataset contain unnatural hand poses. (b) Our method can output some unnatural hand poses, which can be due to noise in the hand pose references or because of the trade-off in the task objective.

object can first be articulated, and then be moved to a different location.

C.2. Unnatural Poses

As shown in Fig. 10b, our method can generate unnatural poses, which we argue occurs because of noisy pose references from ARCTIC [19] as seen in Fig. 10a. We find that especially the index finger is often poorly labeled in the data, which translates to our policies. Developing hand pose priors to incentivize natural poses could be one way to mitigate this issue.

Learning a high-level planning module to couple the different phases is an interesting direction to explore in the future. Note that while we propose a controlled setting to evaluate the *Dynamic Object Grasping and Articulation* task, the order of manipulations can also be reversed. For example, an