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)

***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 is 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 aims to bring an object into a target articulated pose. This task requires grasping, relocation, and articulation. We show our method's efficacy towards this task.**

I. 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 [2, 3]. 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 [4, 10]. Hence, a model that can proficiently generate two-handed motions interacting with objects could reduce the costs associated with motion capture.

Research has turned to synthesizing hand-object interactions using either data-driven [11, 12] or physics-based methods [1]. While existing data-driven methods generate motions including object articulation [12] and two-hand manipulation [11], these methods typically depend on complete supervision from precise 3D motion data for each frame. Recently, physics-based methods that leverage reinforcement learning (RL) in a simulated environment have been proposed [1]. 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 resulted from stable hand-object contacts and forces, ensuring physical plausibility without object interpenetration. 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 [1] 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 grasping and articulation separately, and then evaluate the *Dynamic Object Grasping and Articulation* task, which involves transitioning an articulated object from its initial state into a target articulated object pose. To the best of our knowledge, there are no direct baselines for this task. Therefore, we adjust the closest related work [1] 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 in SupMat show that our method can utilize noisy reconstructed poses

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

from images with off-the-shelf hand-object pose estimation models, and 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.

II. Method

A. Task Definition

The *Dynamic Object Grasping and Articulation* task is illustrated in Fig. 2. 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 $\overline{\mathbf{\Omega}}$, 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 [4] (see Section III and SupMat).

B. Framework Overview

We provide an overview of our policy learning framework in Fig. 1. 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 takes as input a state $\mathbf{s}$ from 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 Φ. In the physics simulation, we create controllable MANO hand models [8] with the mean shape. 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 [4]. Each mesh is split into a base and an articulation part with a single connecting joint.

C. 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:

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

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 joint local rotations and velocities, $\mathbf{f}$ are the net contact forces of each link of the hand, and $\dot{\tilde{\mathbf{T}}}$ are the wrist 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 SupMat VII-B.

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, $\tilde{\mathbf{g}}_q = \overline{\mathbf{q}} - \mathbf{q}$ is the distance between the target and the current hand joint rotations (including wrist). The term $\tilde{\mathbf{g}}_x = \overline{\mathbf{x}} - \mathbf{x}$ is the distance between the target and the current hand joint position, which can be computed from the hand pose using forward kinematics. $\mathbf{g}_c = [\overline{\mathbf{c}} | \overline{\mathbf{c}} - \mathbf{c}]$ contains the target contacts and the difference between the target and the current binary contact vector. $\mathbf{g}_a = \overline{\omega} - \omega$ 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.

D. 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:

$$r = r_{\text{im}} + r_{\text{task}}, \tag{2}$$

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

$$r_{\text{im}} = r_p + r_c + r_{\text{reg}}. \tag{3}$$

where the formulation of each reward function term is shown in SupMat. The task reward r_{task} consists of two incentives:

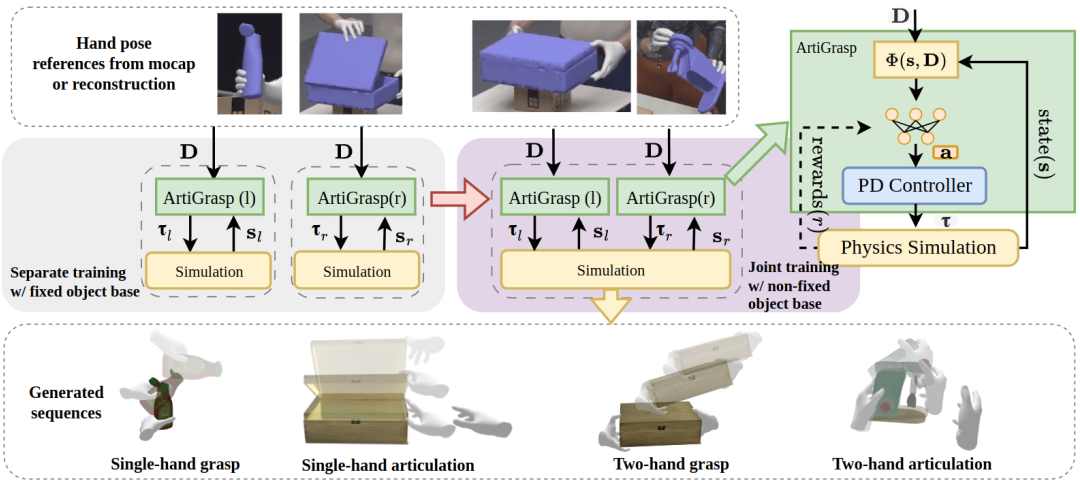


Fig. 1. **Overview of Grasping and Articulation Policy**. Our method uses static hand pose references as input (top row) and generates dynamic sequences (bottom row, where higher transparency represents further in time). We propose a curriculum that starts in a simplified setting with separate environments 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.

opening the object to a target articulation angle and avoiding the movement of the object base from its initial pose:

$$r_{\text{task}} = -w_{tq}||\overline{\omega} - \omega|| - w_{tx}||\mathbf{p}^0 - \mathbf{p}||^2, \tag{4}$$

where $\overline{\omega}$ 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.

E. Curriculum

Training our policies with non-stationary objects from scratch makes it difficult to learn the precise control necessary for fine-grained articulation. To address this, we introduce a learning curriculum with 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. 1). 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. 1). 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 SupMat VII-B, we ablate the effectiveness of our curriculum.

F. 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 II-A) by combining the different subtasks. To achieve this, we use two pairs of hand pose references $\mathbf{D}^{\text{grasp}}$ and $\mathbf{D}^{\text{art}}$. 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 $\mathbf{D}^{\text{grasp}}$ 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 following [1]. 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 $\mathbf{D}^{\text{art}}$. The pre-grasp poses correspond to $\mathbf{D}^{\text{art}}$ 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, our policy controls the hand to approach the object and articulate it to reach the target articulation angle.

III. Experiments

We conduct several experiments to evaluate our framework. We first report experimental details in Section III-A. We then conduct quantitative evaluations on grasping and articulation individually in Section III-B. We further provide quantitative evaluation on our *Dynamic Object Grasping and Articulation*task in Section III-C. We provide more experiment setting details and quantitative and qualitative results in SupMat. Experiments with imperfect hand pose references from images are shown in SupMat VII-A. And we provide ablations to show the importance of our method's components in SupMat VII-B.

A. Experimental Details

Metrics We mostly follow related work [6, 1] and define three metrics for grasping (success rate, position and angle

	Grasping			Articulation		
Model	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 I. **Quantitative comparison for grasping and articulation**

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 radians.

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 the 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 [1] is the most related work to ours, so we propose baselines following it:

D-Grasp: For grasping, we use vanilla D-Grasp and train 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.

B. Individual Grasping and Articulation

For grasping, we pre-sample 30 6D target object poses randomly (see SupMat VI-C). 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 [1] (see Section II-F). 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. I. 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. Qualitative comparisons and more examples of generated sequences (including motions generated with noisy reconstructed poses) are presented in SupMat.

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 II. **Quantitative evaluation of *Dynamic Object Grasping and Articulation*.**

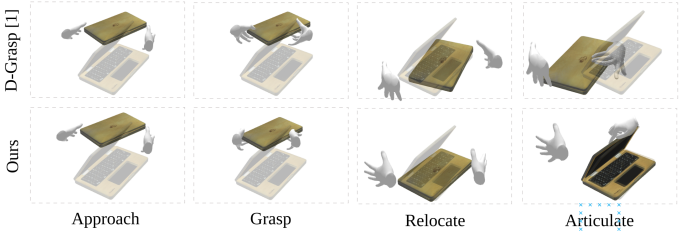


Fig. 2. **Qualitative evaluation of *Dynamic Object Grasping and Articulation*.** D-Grasp can grasp and relocate the object successfully, but fails to articulate the object. Our method is more successful at tackling this task and can articulate the object after relocation.

C. Dynamic Object Grasping and Articulation

To evaluate this task (see Section II-A), 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 VI-D). This results in roughly 7500 evaluation trials in total. We report the average pose errors and the task success rate in Tab. II. Our method outperforms D-Grasp significantly in all metrics. For example, our method achieves a success rate of $5\times$ than that of D-Grasp. This shows that while D-Grasp can perform well in grasping when decoupled, it struggles in this composed task. We provide a qualitative comparison in Fig. 2 and a longer demonstration with multiple objects in SupMat.

IV. 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.

References

[1] 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] Sammy Christen, Wei Yang, Claudia Pérez-D'Arpino, Otmar Hilliges, Dieter Fox, and Yu-Wei Chao. Learning human-to-robot handovers from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[3] Sammy Christen, Lan Feng, Wei Yang, Yu-Wei Chao, Otmar Hilliges, and Jie Song. Synh2r: Synthesizing hand-object motions for learning human-to-robot handovers. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[4] 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.

[5] 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. doi: 10.1109/LRA.2018.2792536.

[6] 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.

[7] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations (ICLR)*, 2015. URL http://arxiv.org/abs/1412.6980.

[8] 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), November 2017.

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

[10] 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. URL https://grab.is.tue.mpg.de.

[11] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. ManipNet: Neural manipulation synthesis with a hand-object spatial representation. *ACM Trans. Graph.*, 40(4), jul 2021. ISSN 0730-0301. doi: 10.1145/3450626.3459830. URL https://doi.org/10.1145/3450626.3459830.

[12] 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.

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

Supplementary Material

We describe implementation details in Section V and experimental details in Section VI. In Section VII, we provide additional qualitative and quantitative experiments.

V. Implementation Details

A. Fomulation of r_{im} Terms

Here we show the formulation of each reward function term of r_{im}, which is defined as:

$$r_{\text{im}} = r_p + r_c + r_{\text{reg}}. \tag{5}$$

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 $\mathbf{q}$:

$$r_p = -\sum_{i=1}^{L} w_{px}^i ||\bar{\mathbf{x}}^i - \mathbf{x}^i||^2 - w_{pq}||\bar{\mathbf{q}} - \mathbf{q}||, \tag{6}$$

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:

$$r_c = w_{cc}\frac{\bar{\mathbf{c}}^T \mathbf{I}_{f>0}}{\bar{\mathbf{c}}^T \bar{\mathbf{c}}} + w_{cf} min(\bar{\mathbf{c}}^T \mathbf{f}, \lambda m_o), \tag{7}$$

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:

$$r_{\text{reg}} = -w_{rh}||\dot{\mathbf{T}}||^2 - w_{ro}||\dot{\mathbf{\Omega}}||^2. \tag{8}$$

B. Training details

We use PPO [9] for RL training and RaiSim [5] 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 present an overview of the important parameters and weight values of the reward function in Tab. III.

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 [7]
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 III. **Hyperparameters of our RL algorithm and the weight values of the reward function.**.

VI. Experimental Details

A. Dataset

We utilize the ARCTIC dataset [4], which contains fully annotated two-hand interaction sequences including dexterous grasping and manipulation of articulated objects. We separate all sequences with available ground-truth annotations (training and validation sequences) into the different interactions of grasping and articulation. For each interaction, we extract a single pair of hand pose references using heuristics (See Section VI-B). 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. With all the reference poses we get, we create a custom 65%/35% train/test-split, leading to 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.

B. 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.

C. Grasping and Articulation Targets

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.

D. Dynamic Object Grasping and Articulation Targets

For the evaluation of *Dynamic Object Grasping and Articulation*, we randomly sample the target articulated object poses $\overline{\mathbf{\Omega}}$, 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 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].

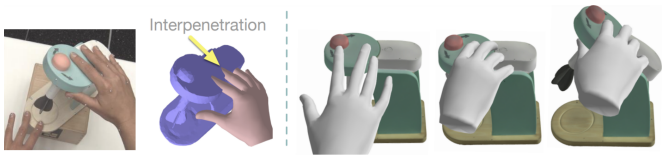
VII. Additional Experiments

A. Evaluation with Reconstructed Hand Pose References

We now evaluate our method with hand pose references obtained from image predictions via the off-the-shelf hand-object pose regressor from ARCTIC [4]. In particular, we estimate hand and object poses from images of the unseen validation subject (subject s05) of ARCTIC and use the reconstructed results as input to our method and baselines. 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 VI-B. We retrieve the images at the corresponding timesteps and pass them to the image-based prediction model. We separate the evaluation into grasping and articulation and present the results in Tab. IV. 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. 3.

	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 IV. **Results with reconstructed hand pose references**. The asterisk (*) denotes using hand pose references from mocap.



(a) Input image (b) Noisy recon. (c) Our motion generation

Fig. 3. **Motion generation.** Our method can synthesize new motion sequences (c) with a noisy hand pose reference (b) reconstructed from a single RGB image (a).

B. 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*, See Section II-C). The results are presented in Tab. V. 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 cooperation, 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.

	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 V. **Ablations**.

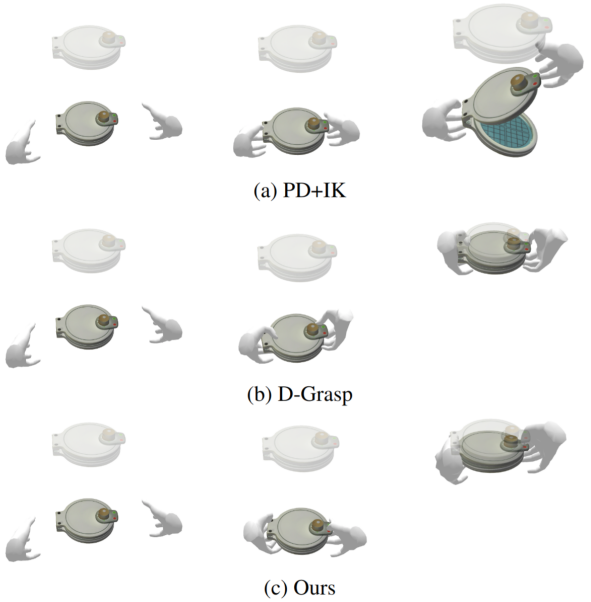


Fig. 4. **Qualitative evaluation of grasping.** When evaluated only on grasping, PD+IK often fails to successfully grasp the object. On the other hand, D-Grasp and ours succeed at the task.

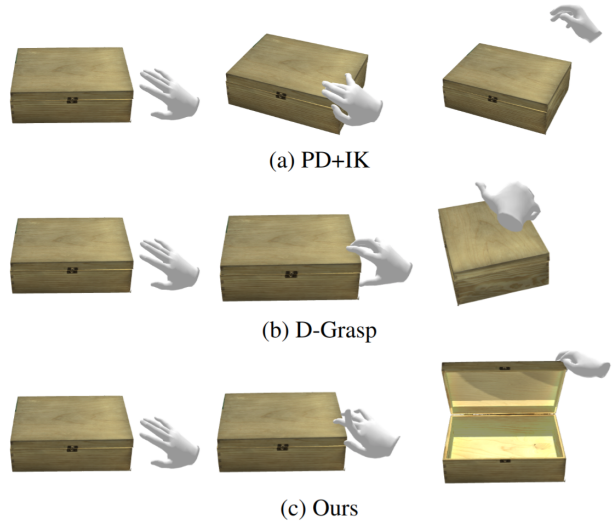


Fig. 5. **Qualitative evaluation of articulation.** When evaluated only on articulation, both PD+IK and D-Grasp often fail at the task. On the other hand, our method can articulate the object successfully.

C. Qualitative Evaluation

We provide additional qualitative comparisons of our method with baselines for grasping and articulation in Fig. 4 and Fig. 5, respectively. Given a pair of policies (one per hand), our method can generate diverse grasping and articulation sequences across different objects as shown in Fig. 6, and complete the task 'task as shown in Fig. 7. Furthermore, as shown in the qualitative result Fig. 9, 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.

D. Unnatural Poses

As shown in Fig. 10, our method can generate unnatural poses, which we argue occurs because of noisy pose references from ARCTIC [4]. 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.

E. Long Sequence with Multiple Objects

Our method can generate long motion sequences in environments with multiple objects, which is shown in Fig. 8. We use a heuristics-based planner to compose the sequences. 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 object can first be articulated, and then be moved to a different location.

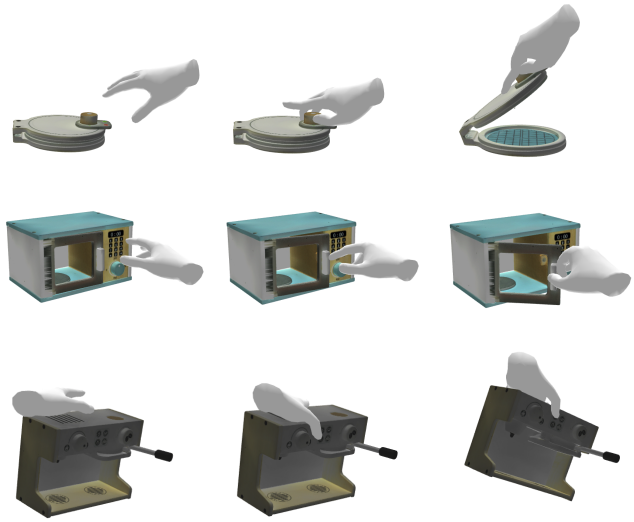


Fig. 6. **Qualitative outputs of our method.** We provide more sequences for grasping and articulation, which are generated by our method with a single pair of hand pose reference label per interaction. Each sequence is shown from left to right.

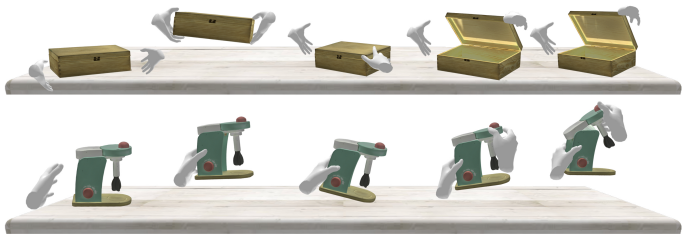


Fig. 7. Our method can synthesize physically plausible motions to grasp and relocate an object with one or two hands, and open it to a target articulation angle.

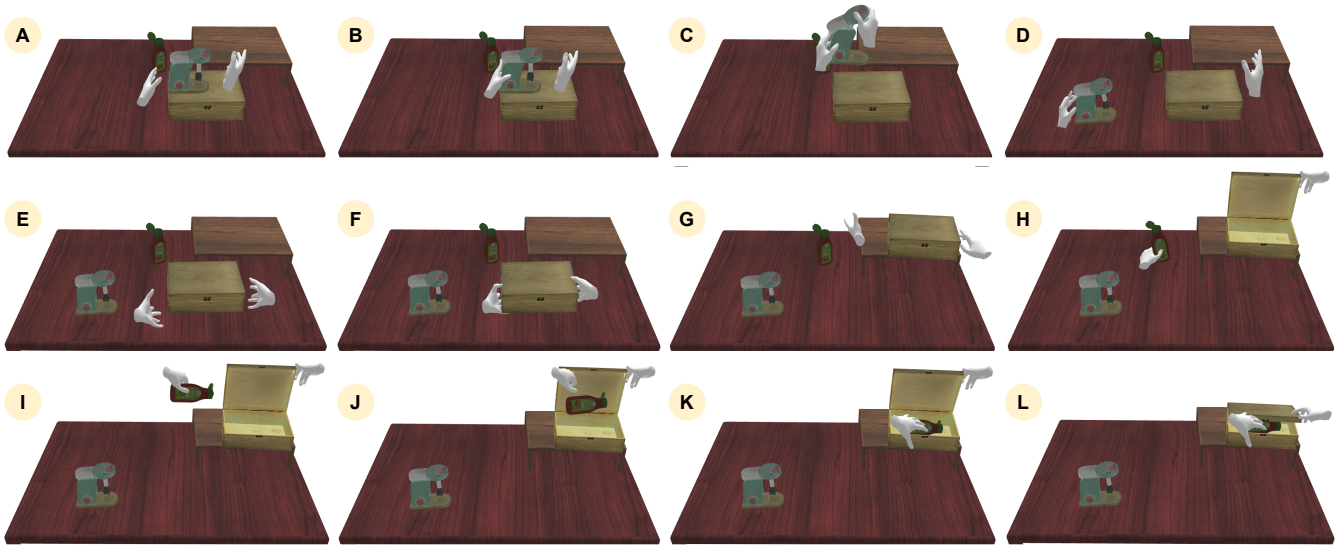
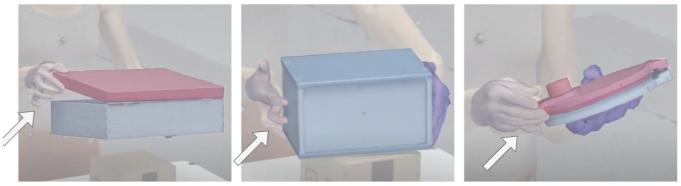


Fig. 8. **Long sequence with multiple objects.** We show that our method can generate sequences of manipulating multiple objects. (A) Approaching the mixer with the left hand. (B) Grasping the mixer with the left hand. (C) Articulating the mixer with the right hand while the left hand is holding it. (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.

Fig. 9. **Qualitative articulation result.** The hand shows some recovery ability from failure cases. Zoom in for details.



(a) Unnatural hand pose references

(b) Unnatural generated hand poses

Fig. 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.