GraspXL: Generating Grasping Motions for Diverse Objects at Scale

Hui Zhang[1], Sammy Christen[1], Zicong Fan[1,2], Otmar Hilliges[1], and Jie Song[1*]

[1] ETH Zürich, Switzerland
[2] Max Planck Institute for Intelligent Systems, Germany

Abstract. Human hands possess the dexterity to interact with diverse objects such as grasping specific parts of the objects and/or approaching them from desired directions. More importantly, humans can grasp objects of any shape without object-specific skills. Recent works synthesize grasping motions following single objectives such as a desired approach heading direction or a grasping area. Moreover, they usually rely on expensive 3D hand-object data during training and inference, which limits their capability to synthesize grasping motions for unseen objects at scale. In this paper, we unify the generation of hand-object grasping motions across multiple motion objectives, diverse object shapes and dexterous hand morphologies in a policy learning framework *GraspXL*. The objectives are composed of the graspable area, heading direction, wrist rotation, and hand position. Without requiring any 3D hand-object interaction data, our policy trained with 58 objects can robustly synthesize diverse grasping motions for more than **500k** unseen objects with a success rate of 82.2%. At the same time, the policy adheres to objectives, which enables the generation of diverse grasps per object. Moreover, we show that our framework can be deployed to different dexterous hands and work with reconstructed or generated objects. We quantitatively and qualitatively evaluate our method to show the efficacy of our approach.

Keywords: Motion synthesis · Hand-object interaction · Dexterous manipulation

1 Introduction

In our daily lives, we constantly engage with a wide variety of objects, from taking an apple out of a bowl and handling a knife by its grip to lifting a pair of headphones off the floor. This routine showcases the remarkable dexterity and adaptability of human hands, which effortlessly achieve complex tasks such as precisely grasping objects of different shapes in specific areas and from certain directions. Impressively, humans achieve this without needing object-specific pretraining, enabling us to manipulate items of any shape with ease. Therefore, the

[*] Now at HKUST(GZ)&HKUST

Fig. 1: Large-scale Grasping Synthesis. Our method, *GraspXL*, can be used to generate large-scale grasps with robotic hands, and the MANO hand model. Here we show large-scale generated results, better viewed when zoomed in.

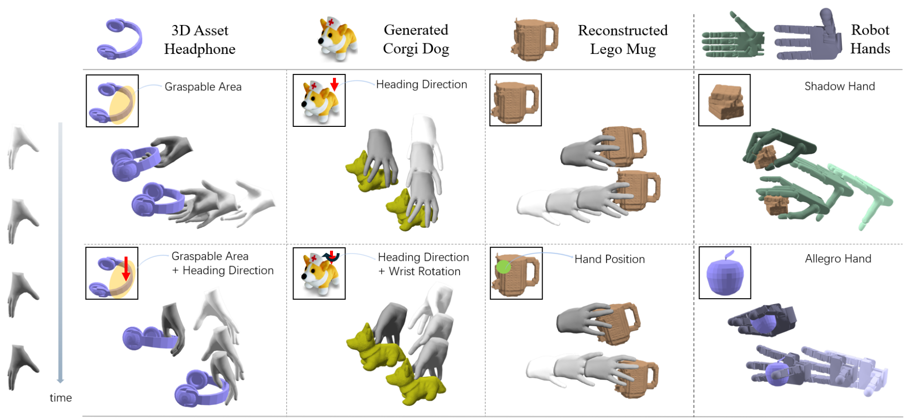


Fig. 2: *Objective-driven Grasping Motion Synthesis*. Given a hand model and an object, our goal is to synthesize grasp motions that adhere to high-level objectives, which may consist of one or multiple objectives including graspable areas (indicated by the shadow), heading directions (indicated by the red arrow), wrist rotations (indicated by the black arrow), and positions of the hand (indicated by the green dot). For each sequence, darker hand represents more recent in time.

ability to generate versatile grasping motions that adhere to certain motion objectives – like precise graspable areas and specific heading directions – holds significant benefits for fields like animation and robotic grasping [1, 7, 18].

In this paper, we present *GraspXL*, a policy learning framework capable of generating motions for a wide variety of objects, motion objectives, and hand morphologies, which is shown as Fig. 2. Our method does not rely on any 3D hand-object data to train but can robustly generalize to grasp a broad range of unseen objects. Consequently, our approach significantly scales hand-object motion generation, accommodating over **half a million** unseen objects, and we show some examples in Fig. 1. Existing methods for generating hand-object motions struggle with scalability to unseen objects due to their dependence on pose references [8, 48], their need for time-intensive optimization [7, 42], or their limitation to objects encountered during training [25]. Furthermore, these existing approaches cannot, when applied off-the-shelf, generate interacting motions that fulfill multiple motion objectives.

Scaling objective-driven grasping motion generation to a wide variety of unseen objects poses several challenges. First, there is the **generalization ability**, where the learning framework should be general enough to handle different object shapes, dexterous hand models, and motion objectives. It is necessary to design a framework that avoids specific assumptions about the object shape and hand model. Second, the model must establish **stable grasping while adhering to multiple objectives**. Given the variety of object shapes and objectives, the model needs to learn stable grasping while satisfying multiple objectives. However, these high-level goals may negatively affect each other during training as the exploration of objectives can lead the model into local optima. In such cases, the objectives are followed initially, but no stable grasp is reached due to the object movement caused by contact, making the learning difficult and increasing the requirements for control precision.

We formulate *GraspXL* in the reinforcement learning paradigm and leverage physics simulation. To allow a policy to react to varying object shapes, we capture the general shape features of diverse objects with the vectors from each finger joint to the nearest point on the object surface. To handle multiple objectives, we introduce a control scheme, dubbed *objective-driven guidance*, that guides the hand towards the desired objective(s). To achieve generalization across diverse hands, we propose a general reward function composed of a grasping reward term and an objective reward term that is agnostic to the hand morphology. Finally, to tackle the difficulty of learning stable grasping while satisfying the target objectives, we propose a learning curriculum to decompose the learning process to objective learning and grasp learning. Specifically, we first train the policy on stationary objects with a larger objective reward to learn precise finger motions for the objectives. We then fine-tune the policy on non-stationary objects with a larger grasping reward to promote stable grasping.

In our experiments, we evaluate methods for multi-objective grasping motion synthesis on PartNet [27] and ShapeNet [4]. We enhance SynH2R [7], the only method offering controllability in hand heading direction, by incorporating more detailed motion objectives. Our approach, unlike SynH2R [7], achieves higher performance without time-intensive optimization for reference poses, yielding a 30% increase in success rates and reducing objective errors by 30%-50%. Additionally, we demonstrate our method's broad applicability and generalization across a diverse range of conditions: it effectively handles over half a million objects from large-scale 3D datasets [11], adapts to objects from text-to-3D generation methods [28], applies to objects from 3D reconstruction techniques [15], and operates with various robotic hands, including Shadow [34], Allegro [41], and Faive [38]. We validate our method's superiority over others through quantitative and qualitative measures and highlight its broad generalization capabilities. Additionally, we dissect our framework's critical elements and investigate the impact of various objective combinations on performance.

In summary, our contributions are: 1) *GraspXL*, a framework that synthesizes grasping motions on a large scale (500k+) of unseen objects, without relying on hand-object datasets during training. 2) A learning curriculum and objective-driven guidance to enable our method to achieve stable grasping while satisfying multiple objectives. 3) We show that our method is general enough to be deployed on reconstructed or generated objects and different dexterous hands.

2 Related Work

Table 1: Comparison with existing grasping motion synthesis methods.

Method	Multiple Objectives	Different Hands	Data-agnostic Inference	Number of Novel Test Objects
DexVIP [25]	×	×	×	0
D-Grasp [8]	×	×	×	3
UniDexGrasp [42]	×	×	✓	100
UniDexGrasp++ [40]	×	×	✓	100
SynH2R [7]	×	×	✓	1,174
GraspXL (Ours)	✓	✓	✓	**503,409**

We categorize related works into hand-object interaction synthesis and dexterous robot hand manipulation. Tab. 1 compares different grasping motion synthesis methods with ours regarding the use of objectives, different hand morphologies, whether they require datapoints at inference time, and the number of reported results on unseen objects.

2.1 Hand-object Interaction Synthesis

In the literature, hand interaction primarily focuses on hand(-object) reconstruction [3, 13, 14, 16, 17, 24, 37, 43, 50], static grasp synthesis [10, 21, 39, 45] and temporal hand-object motion synthesis [2, 8, 18, 35, 47, 49]. This work concentrates on the latter. In synthesis, some previous methods use pure data-driven approaches to synthesize human hand manipulation sequences and rely on post-processing to enhance physical plausibility [47, 49]. Data-driven methods are supervised by 3D hand-object annotated sequences during training [18, 35, 49], and some methods rely on references like wrist trajectories [47] during inference. However, acquiring accurate 3D hand-object data is infeasible to scale because of expensive capture setups [17, 36]. Due to limited training data, data-driven methods are constrained by their training distribution, making generalization challenging, especially for conditional generation tasks with motion objectives. Moreover, their dependence on data restricts their applicability across different dexterous hand platforms such as Shadow Hand [34] and Allegro Hand [41].

Some works leverage physics simulation within reinforcement-learning frameworks to alleviate the data requirement and ensure physical feasibility. Christen *et al*. [8] generate natural grasp sequences from captured or reconstructed static grasp references. Zhang *et al*. [48] achieve two-hand grasp and articulation with a single frame reference grasp. However, both methods lack extensive evaluation to demonstrate their generalization ability, and their reliance on references restricts them from scaling to more objects. Some methods [7, 42] first generate grasp reference poses and then synthesize grasping motions accordingly for thousands of objects. However, they either rely on a time-consuming optimization process [7, 42] or are limited by the diversity in the pre-collected dataset [42]. Furthermore, the pre-defined references may not be physically feasible, which introduces extra disturbances to the generated motions as the hands should adhere strictly to the imperfect references. In contrast to existing approaches, our method does not require grasping references and offers real-time inference capabilities for grasping a wide range of objects, while at the same time enabling the control over multiple motion objectives such as graspable areas, heading directions, wrist rotations, and hand positions.

2.2 Dexterous Robot Hand Manipulation

Dexterous manipulation plays a crucial role in enhancing robot capabilities [22, 44, 46]. With the motivation to learn from humans, some approaches use imitation learning. However, this requires full human demonstrations for both training and inference [6, 23, 30], which are usually expensive to collect. Some other methods utilize retargeted human demonstrations during training [44], use teleoperated sequences as training data [31], or learn a parameterized reward function based on demonstrations [9]. However, the reliance on expensive full grasping sequences during training still limits their generalization ability for out-of-distribution settings. Instead of relying on full trajectories, Xu *et al*. [42] use static pose references to guide the motion by first predicting the contact map and then optimizing the poses accordingly, with physics-based heuristics used to filter out invalid poses. This process is expensive and makes real-time inference infeasible. In contrast, our method generates motions in real-time while adhering to fine-grained motion objectives. Furthermore, without reliance on hand-object interaction data, our method can easily scale to 500k objects.

Another line of research learns grasping with RL without conditioning on reference grasps [5, 12, 29]. Wan *et al*. [40] propose a curriculum learning framework to distill a state-based policy into a vision-based policy. They generalize to many objects and achieve impressive results on vision-based grasping, whereas we focus on generalizable grasping with different objectives in the state-based setting. Similar to ours, some recent methods can achieve affordance-aware grasps. Mandikal *et al*. [26] train an affordance-aware grasp policy with RL and evaluated with only 24 novel objects, and further introduce a hand pose prior from YouTube videos for natural hand configuration [25] which is only evaluated with objects used for training. Agarwal *et al*. [1] learn a category-level policy that grasps the objects by affordance areas, but they are limited to grasp their training categories and can generalize to only 1 unseen category. Moreover, all of these methods only generate a single pose per object. In contrast, our method can generate diverse affordance-aware grasping poses that can be controlled via motion objectives, and can be deployed for over 500k unseen objects.

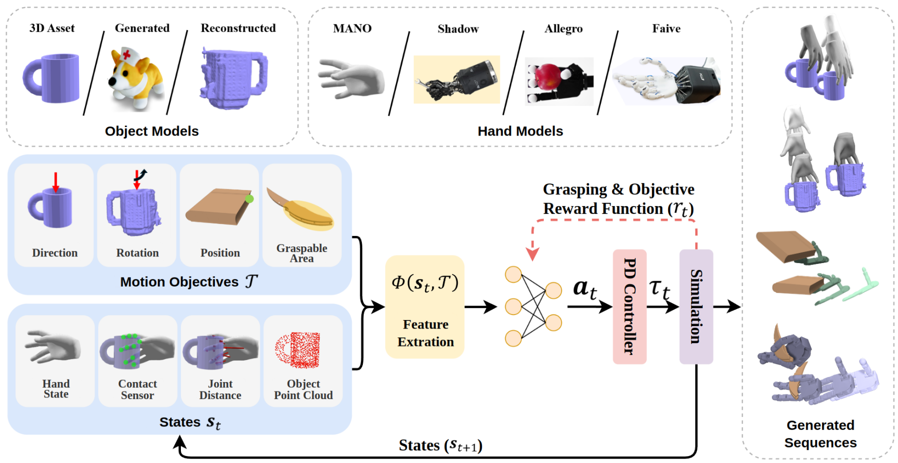
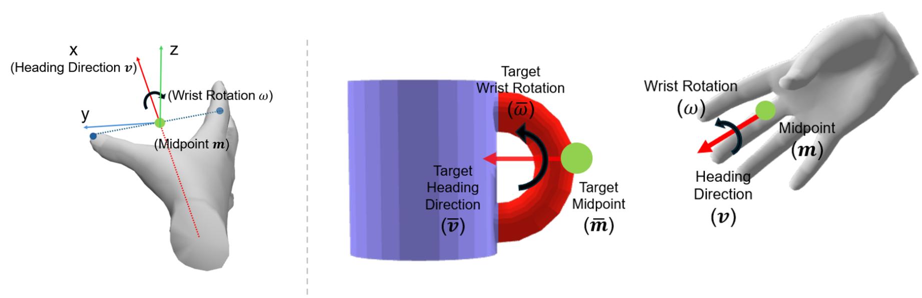


Fig. 3: Overview of *GraspXL*. As shown in the top row, our method can utilize captured, generated, or reconstructed objects, and different dexterous hand platforms such as MANO, Shadow, Allegro or Faive. With given object and hand model, the policy takes different objectives and states as inputs (on the left), and outputs dynamic grasp motions according to the specific objectives (on the right, accordingly, where darker hand represents more recent in time). The objectives can be the heading direction, wrist rotation, hand position or graspable area, and the states contain the hand state, contact, force and distance of each link with the object, and the object point cloud.

3 GraspXL

Task Definition: As illustrated in Fig. 4, we assume a hand model h with L links where $\mathbf{h}_i \in \mathbb{R}^3$ is the position of the i-th link. The hand pose consists of joint angles $\mathbf{q} \in \mathbb{R}^{L\times 3}$ and the global orientation represented by the heading direction $\mathbf{v} \in \mathbb{R}^3$ and wrist rotation $\omega \in \mathbb{R}$ about the directional vector $\mathbf{v}$. As the hand moves, it has linear and angular velocities $\mathbf{u}_h \in \mathbb{R}^6$. We define $\mathbf{m} \in \mathbb{R}^3$ as the midpoint between the thumb tip and the third joint of the middle finger. We also assume a rigid object o which has a point cloud of 3D vertices $\{\mathbf{o}_j\}$. A user may partition the point cloud into graspable/non-graspable points, specifically $\{\mathbf{o}_j\} = \{\mathbf{o}_j^+\} \cup \{\mathbf{o}_j^-\}$, where we split the points into two disjoint sets to specify whether the point should be encouraged to be in contact with the hand or not when motion is generated. As the object moves, it also has linear and angular velocities $\mathbf{u}_o \in \mathbb{R}^6$. The hand links may be in contact $\mathbf{c} \in \{0,1\}^L$ with the object using forces in magnitude $\mathbf{f} \in \mathbb{R}^{\mathrm{L}}$. In particular, the links may contact with the graspable/non-graspable area, which are denoted as $\mathbf{c}^+ \in \{0,1\}^L$ an $\mathbf{c}^- \in \{0,1\}^L$. Similarly, the force magnitude vector $\mathbf{f}$ can be decomposed into $\mathbf{f}^+ \in \mathbb{R}^{\mathrm{L}}$ and $\mathbf{f}^- \in \mathbb{R}^{\mathrm{L}}$.

A user may specify motion objectives $\mathcal{T}$ to define the target heading direction $\bar{\mathbf{v}}$, wrist rotation $\bar{\omega}$, midpoint position $\bar{\mathbf{m}}$, the partition of graspable/non-graspable object point cloud $\{\mathbf{o}_j^+\} \cup \{\mathbf{o}_j^-\}$. Given this specification, our goal is to generate a motion sequence that approaches and grasps an object without dropping whilst adhering to the objectives. Note that only the target heading direction $\bar{\mathbf{v}}$ is mandatory to specify in $\mathcal{T}$. If not specified, $\{\mathbf{o}_j\}$ will equal to $\{\mathbf{o}_j^+\}$, $\bar{\mathbf{m}}$ will be the mean of $\{\mathbf{o}_j\}$, $\bar{\omega}$ will be zero. See SupMat for more details.



(a) Hand local coordinate system (b) Notations

Fig. 4: Task definition. (a) The local coordinate system of the hand where the x-axis is the heading direction $\mathbf{v}$, the origin is the midpoint position $\mathbf{m}$ (see text for definition), the rotation about $\mathbf{v}$ is ω. (b) Given an object with user-specified graspable $\{\mathbf{o}_j^+\}$ and non-graspable vertices $\{\mathbf{o}_j^-\}$ (labelled in red and blue), the goal of the agent is to approach and grasp the object while satisfying motion objectives $\bar{\mathbf{v}}$, $\bar{\mathbf{m}}$, $\bar{\omega}$, and contact with the graspable area $\{\mathbf{o}_j^+\}$.

Overview: Fig. 3 outlines our RL-based method, *GraspXL*. Given the objectives $\mathcal{T}$ and the state $\mathbf{s}$ obtained from the physics simulation, we utilize a feature extraction layer Φ to derive the features. Subsequently, the policy $\boldsymbol{\pi}$ takes features as input and generates the actions $\mathbf{a}$, representing the PD-control targets used to compute the torques $\boldsymbol{\tau}$. These torques are then applied to the joints of the hand model in the physics simulation to update the state $\mathbf{s}$, which is subsequently fed back into our feature extraction layer for the next iteration.

Reinforcement Learning Background: The task is formulated as a standard Markov Decision Process (MDP). The goal is to determine the policy $\boldsymbol{\pi}$ that maximizes the expected reward $\mathbb{E}_{\xi\sim\boldsymbol{\pi}}\left[\sum_{t=0}^{T}\gamma^t r_t\right]$, where $\gamma \in [0,1]$ denotes the discount factor, r_t represents the reward at time step t, and $\xi =$

$[(\mathbf{s}_0, \mathbf{a}_0), \cdots, (\mathbf{s}_T, \mathbf{a}_T)]$ denotes a state-action pair trajectory generated by the policy $\boldsymbol{\pi}$ interacting with the physics simulation. The trajectory ξ is determined by the transition function $p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$, which is governed by the physics simulation along with an initial state distribution $p(s_0)$. The distribution of a trajectory ξ is defined as $p_\theta(\xi) = p(\mathbf{s}_0) \prod_{t=0}^{T} p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) \boldsymbol{\pi}(\mathbf{a}_t|\Phi(\mathbf{s}_t, \mathcal{T}))$. Here, $\boldsymbol{\pi}$ is represented by a neural network.

Feature Extraction: Given a hand-object state $\mathbf{s}$ and the motion objectives $\mathcal{T}$, we extract the following features with a conversion function $\Phi(\mathbf{s}, \mathcal{T})$:

$$\Phi(\mathbf{s}, \mathcal{T}) = (\mathbf{q}, \mathbf{d}, \mathbf{u}_h, \mathbf{u}_o, \mathbf{c}, \mathbf{f}, \tilde{\mathbf{v}}, \tilde{\mathbf{m}}, \tilde{\omega}, \mathbf{l}), \tag{1}$$

which includes finger joint angles $\mathbf{q}$, the finger joint tracking error (compared with the target angles for PD controller) $\mathbf{d}$, the hand and object velocities $\mathbf{u}_h$ and $\mathbf{u}_o$, the contact vector $\mathbf{c}$, the force magnitude vector $\mathbf{f}$. Finally, $\tilde{\mathbf{v}}, \tilde{\mathbf{m}}, \tilde{\omega}$ represent the differences between current and target heading directions, midpoint positions, and wrist rotation angles. To represent shape features of an object and help the hand get aware of the graspable/non-graspable areas, we construct distance features $\mathbf{l}^+ \in \mathbb{R}^{\mathrm{L}\times 3}$ where each row is the difference vector between a link position $\mathbf{h}_i$ to the closet object vertex in the graspable part $\{\mathbf{o}_j^+\}$. Similarly, we construct $\mathbf{l}^- \in \mathbb{R}^{\mathrm{L}\times 3}$ for the non-graspable part $\{\mathbf{o}_j^-\}$. We ablate this component in Sec. 4.4.

Reward Function: The reward function should guide our policy to learn a solution that can grasp the objects with the desired objectives while at the same time achieving successful grasps. Furthermore, it should be formulated without specific assumptions about the hand morphology so as to be applied on different hand models. As a result, we define our reward function as follows:

$$r = r_{\text{goal}} + r_{\text{grasp}}, \tag{2}$$

where r_{goal} is for motion objectives and r_{grasp} is for successful grasping.

In particular, the motion objective reward r_{goal} is formulated as:

$$r_{\text{goal}} = r_{\text{dis}} + r_{\mathbf{v}} + r_\omega + r_{\mathbf{m}} \tag{3}$$

where r_{dis} promotes approaching the target while avoiding non-graspable areas, and $r_{\mathbf{v}}$, r_ω, and $r_{\mathbf{m}}$ reward aligning heading direction, wrist rotation, and midpoint, respectively. Concretely, the term r_{dis} is defined as

$$r_{\text{dis}} = -\sum_{i=1}^{L} \left[w_d^+(i) ||\mathbf{h}_i - \mathbf{o}_i^+||^2 - w_d^-(i) ||\mathbf{h}_i - \mathbf{o}_i^-||^2 \right], \tag{4}$$

with the weights $w_d^+(i) \in \mathbb{R}$ and $w_d^-(i) \in \mathbb{R}$, the i-th link position ($\mathbf{h}_i$), the closest graspable/non-graspable object points to the i-th link ($\mathbf{o}_i^+$ and $\mathbf{o}_i^-$). The rewards $r_{\mathbf{v}}$, r_ω, and $r_{\mathbf{m}}$ penalize discrepancies in the heading direction $\mathbf{v}$, the wrist rotation angle ω, and midpoint position $\mathbf{m}$ from the targets:

$$r_{\mathbf{v}} = -w_{\mathbf{v}} ||\mathbf{v} - \overline{\mathbf{v}}||^2,\ r_\omega = -w_\omega ||\omega - \overline{\omega}||^2,\ r_{\mathbf{m}} = -w_{\mathbf{m}} ||\mathbf{m} - \overline{\mathbf{m}}||^2. \tag{5}$$

The grasp reward r_{grasp} promotes proper contact and natural poses by combining multiple factors:

$$r_{\text{grasp}} = r_{\mathbf{c}} + r_{\mathbf{f}} + r_{\text{anatomy}} + r_{\text{reg}} \tag{6}$$

where $r_{\mathbf{c}}$ assesses the contact between fingers and the object, considering both the target and non-target areas ($r_{\mathbf{c}} = w_c^+ \left\|\mathbf{c}^+\right\|^2 - w_c^- \left\|\mathbf{c}^-\right\|^2$). The force term, $r_{\mathbf{f}}$, encourages the contact forces with the graspable area and punishes the contact forces with the non-graspable area, capped by a factor proportional to the object's weight w_o: ($r_{\mathbf{f}} = w_f^+ \mathbf{c}^+ \cdot min(\mathbf{f}^+, \lambda w_o) - w_f^- \mathbf{c}^- \cdot min(\mathbf{f}^-, \lambda w_o)$). r_{reg} penalizes excessive velocities to ensure stability ($r_{\text{reg}} = -w_h||\mathbf{u}_h||^2 - w_o||\mathbf{u}_o||^2$). If the hand model is MANO [32], we additionally apply the anatomy reward r_{anatomy} [43] on joint angles $\mathbf{q}$ for natural poses.

Curriculum: Generating grasping motion with objectives requires a policy that establishes stable grasps while achieving specific goals. This is complicated because of the potential for adverse outcomes like object flipping caused by contact when trying to accomplish the objectives. To mitigate this, we introduce a learning curriculum: we start by training the policy on static objects with increased r_{goal} to hone precise finger movements for objectives. Training progresses to moving objects with a higher r_{grasp}, enhancing wrist movements for secure, non-slip grasps. The effectiveness of this approach is discussed in Sec. 4.4.

Objective-driven hand guidance: Diverse objectives complicate policy exploration due to the need for varied wrist movements. To address this, we present a simple-yet-effective method to guide the hand during training and inference, improving exploration and control precision. Essentially, we compute the differences between the target and the current values for the heading direction $\mathbf{v}$, the wrist rotation angle ω, and the midpoint position $\mathbf{m}$. These differences are then applied directly as bias terms for the wrist's 6 degrees of freedom (DoF) PD-controller to guide the wrist toward motion objectives. This simple trick promotes quicker convergence and boosts performance, as detailed in Sec. 4.4.

4 Experiments

This section involves several experiments to evaluate our method's effectiveness and generalization capabilities. We detail the experimental setups in Sec. 4.1, compare our method's performance against others in Sec. 4.2, and examine its generalization across various unseen objects and hand models in Sec. 4.3. Lastly, we ablate our method's components and analyze the influence of different objective combinations in Sec. 4.4.

4.1 Experimental Setup

Datasets: To construct the training set, we randomly select 26 objects and 32 objects from ShapeNet [4] and PartNet [27] respectively. To demonstrate our method's generalization to large-scale object datasets, we use Objaverse [11]. Since not all objects in the three datasets are suitable for rigid-body grasping (*e.g.* a piece of paper), we filter out objects with such shapes. After filtering and preprocessing, we have 48, 3993 and 503k objects in PartNet, ShapeNet, and Objaverse for testing, respectively. The PartNet test set contains unseen objects from seen categories, while the objects of the other two test sets are completely novel. Please see SupMat for details on dataset filtering and preprocessing.

Training Details: We use PPO [33] for RL training with the RaiSim [20] physics simulator. Experiments are conducted using a single Nvidia RTX 6000 GPU with 128 CPU cores. To improve data diversity for better generalization, we sample objectives during training with the following heuristics: we randomly sample the target heading direction $\bar{\mathbf{v}} \in \mathbb{R}^3$ and wrist rotation angle $\bar{\omega} \in [0, 2\pi)$ while ensuring that the graspable part $\{\mathbf{o}_j^+\}$ of the object is narrower than 12 cm between the thumb and the other fingers (along the y axis of the hand local coordinate in Fig. 4a). As a result, the object will not be too large to be grasped. To find a midpoint, we randomly sample a point on the graspable surface $\{\mathbf{o}_j^+\}$.

Evaluation Protocol: We measure the stability of the grasps with a success metric metric. To assess the adherence to the objectives, we compute several metrics and utilize the ShapeNet and PartNet (which contains part-based objects) datasets. **Midpoint Error (Mid. Error):** The mean Euclidean distance between the final midpoint position $\mathbf{m}$ and the target $\bar{\mathbf{m}}$ measured in centimeters. **Heading Error (Head. Error):** The mean geodesic distance between the final heading direction $\mathbf{v}$ and the target $\bar{\mathbf{v}}$ measured in radian. **Wrist Rotation Error (Rot. Error):** The mean absolute error between the final wrist rotation ω and the target $\bar{\omega}$ measured in radian. **Contact Ratio:** The ratio between the number of links contacted with the graspable part $\{\mathbf{o}_j^+\}$ and the entire object $\{\mathbf{o}_j\}$. **Grasping Success Rate (Suc. Rate):** A grasp is determined as a success if the object is lifted higher than 10cm and remains stable without falling until the sequence terminates.

Baselines: We choose SynH2R as the main baseline, because it offers some degree of controllability in the motion generation, (in terms of heading direction $\mathbf{v}$). We adapt it with more fine-grained motion objectives. In particular, we have the following baselines. **SynH2R**: The original method generates grasping motion by first generating static grasping reference poses with an optimization procedure. Then it uses an RL-based policy to approach and follow the grasping reference pose. To adapt this method for objective-driven synthesis beyond heading direction $\mathbf{v}$, we add additional optimization terms to include the wrist rotation error ω and the midpoint error $\mathbf{m}$. We also encourage the contacts with the graspable part while punishing the contacts with the non-graspable part during the optimization. The RL-based policy stays the same as the original one in SynH2R, which is trained with the generated grasping references. **SynH2R-PD**: The grasping references are generated the same way as in SynH2R above. We then set the reference poses as targets for the PD controller following [8].

Table 2: Method Comparison on PartNet and ShapeNet.

Method	PartNet Test Set					ShapeNet Test Set			
	Suc. Rate [%] ↑	Mid. Error [cm] ↓	Head. Error [rad] ↓	Rot. Error [rad] ↓	Contact Ratio [%] ↑	Suc. Rate [%] ↑	Mid. Error [cm] ↓	Head. Error [rad] ↓	Rot. Error [rad] ↓
PD	26.5	4.30	0.767	0.857	13.0	21.9	4.60	0.850	0.964
SynH2R	82.3	4.06	0.522	0.568	53.4	65.8	4.49	0.642	0.688
Ours	**95.0**	**2.85**	**0.270**	**0.306**	**86.7**	**81.0**	**3.22**	**0.292**	**0.338**

4.2 Method Comparison

Following our evaluation protocol in Sec. 4, we compare with SynH2R and SynH2R-PD for objective-driven motion synthesis. All methods use the MANO hand model and the same pre-sampled objectives. We initialize the hand state in the same way for all methods. In particular, the starting hand pose is set to the mean pose of MANO with an open thumb (see Fig. 4a), and the wrist is positioned $30cm$ away from the target midpoint position $\bar{\mathbf{m}}$ along the target heading direction $\bar{\mathbf{v}}$. For each sequence, we control the hand to grasp the object, and then apply torques to the wrist to lift the object. It's noteworthy that our method can directly infer in simulation whereas the baselines rely on pre-generated hand pose references through a time-consuming optimization procedure.

PartNet Evaluation: We use the PartNet test set to evaluate all methods with the given objectives. For each object, we calculate the average performance among 25 randomly sampled heading directions $\bar{\mathbf{v}}$, wrist rotations $\bar{\omega}$ and midpoints $\bar{\mathbf{m}}$. Tab. 2 shows that our method significantly outperforms the baselines across all metrics. In particular, our method more closely follows the objectives in terms of hand grasping position (Midpoint Error), heading direction (Heading Error), wrist rotation (Wrist Rotation Error), and contact points with the graspable/non-graspable parts (Contact Ratio). At the same time, we achieve the most stable grasps (Grasping Success Rate).

ShapeNet Evaluation: While PartNet specializes in part-based objects at a smaller scale, we extend our evaluation to ShapeNet to test on more diverse and unseen objects. We measure average performance across 5 randomly sampled heading directions $\bar{\mathbf{v}}$, wrist rotations $\bar{\omega}$ and midpoints $\bar{\mathbf{m}}$ per object. As shown in Tab. 2, our method outperforms all baselines, demonstrating less decrease in performance when comparing PartNet and ShapeNet, thereby proving superior generalization. The Grasping Success Rate has a drop for all the methods (even for the non-learning-based SynH2R-PD baseline), which we claim is caused by some objects that are too large or too heavy to be grasped. It is worth noting that generating grasp pose references for the ShapeNet test set using the baselines requires approximately a week, whereas our method can directly perform real-time inference in simulation, showcasing its efficiency. Please refer to SupMat for further qualitative comparisons.

Table 3: Generalization to the Large-scale Objaverse Dataset.

Objects	Suc. Rate [%] ↑	Mid. Error [cm] ↓	Head. Error [rad] ↓	Rot. Error [rad] ↓
Small	**85.9**	3.20	0.311	0.362
Medium	84.5	**3.16**	0.274	0.315
Large	79.0	3.50	**0.271**	**0.306**
Average	82.2	3.32	0.279	0.319

Table 4: Generalization to Generated and Reconstructed Objects.

Objects	Suc. Rate [%] ↑	Mid. Error [cm] ↓	Head. Error [rad] ↓	Rot. Error [rad] ↓
Generated Obj. (DreamFusion)	88.4	2.85	0.310	0.361
Reconstructed in-the-wild Obj. (HOLD)	77.5	3.68	0.222	0.281
Reconstructed HO3D Obj. (HOLD)	74.0	3.84	0.207	0.247
Ground-truth HO3D Obj. (HOLD)	73.7	3.63	0.259	0.301

4.3 Generalization

Generalization to Large-scale Object Dataset: Our method's scalability was tested using the large Objaverse [11] dataset, resulting in a test set of over half a million objects of three different sizes: small, medium and large (refer to SupMat for preprocessing details). Performance on this set, as detailed in Tab. 3, is comparable to our ShapeNet results, underscoring our method's ability to scale. Specifically, medium-sized objects offer optimal grasping success. Smaller objects enable higher success rates and more accurate grasp positioning but suffer from larger heading direction and wrist rotation errors, as smaller objects are easier to grasp but also easier to rotate in hand. Overall, our method consistently shows excellent performance across different object sizes, confirming its effectiveness for varying scales. The generated grasping motions of different hands are released for further research.
Generalization to Reconstructed and Generated Objects: Our method not only synthesizes grasping motions for standard 3D assets [4, 11, 27] but also effectively handles reconstructed and generated objects, expanding its usability. We evaluated its performance on objects reconstructed via HOLD [15] and objects generated by DreamFusion [28]. We analyzed the average performance for each object across 25 random sets of motion objectives, with findings detailed in Tab. 4. Despite the objects being novel and containing severe artifacts (see the Rubik's Cube in Fig. 2), our method maintains similar performance across all metrics compared to other experiments. Notably, the performance of objects reconstructed from HO3D [19] is similar with the performance of ground-truth objects. This shows our method's robustness towards reconstruction noise in object meshes. Please refer to the SupMat for a more detailed setting explanation.
Generalization on Different Robotic Hands: We assess our framework's generalization capabilities on various dexterous robotic hands, including Shadow Hand [34], Allegro Hand [41], and Faive Hand [38], each differing in size and joint structure (Faive Hand has 30, Shadow Hand 22, Allegro Hand 16, and MANO 45 finger joints). Our method adapts seamlessly to different models by adjusting

Table 5: Generalization to Different Hand Models.

	PartNet Test Set					ShapeNet Test Set			
Hand Model	Suc. Rate [%] ↑	Mid. Error [cm] ↓	Head. Error [rad] ↓	Rot. Error [rad] ↓	Contact Ratio [%] ↑	Suc. Rate [%] ↑	Mid. Error [cm] ↓	Head. Error [rad] ↓	Rot. Error [rad] ↓
Allegro	95.3	4.38	0.291	0.300	81.1	83.4	4.45	0.271	0.292
Shadow	94.0	3.57	0.317	0.320	83.4	83.2	3.67	0.363	0.381
Faive	95.8	2.85	0.228	0.243	88.7	82.4	3.44	0.250	0.262
MANO	95.0	2.85	0.270	0.306	86.7	81.0	3.22	0.292	0.338

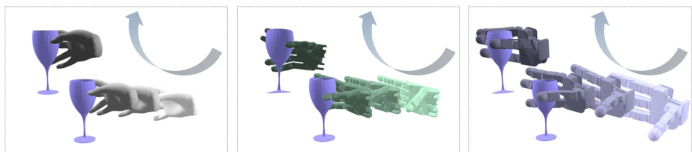


Fig. 5: Generated Motions of Different Hands with the Same Objectives. We require the hands to approach from the right and grasp the upper part of the glass.

hand state and action space dimensions. According to Tab. 5, all tested hands showed similar success rates. However, the Allegro Hand has greater Mid. Errors and lower Contact Ratios due to its larger size, which makes it difficult to achieve precise position control. The Shadow Hand shows marginally higher objective errors due to its broad, flat palm which limits its dexterity. Despite these structural differences affecting performance metrics, all hands achieved commendable results, highlighting our method's ability to generalize. This suggests our framework's capacity to deal with hand models with different morphologies. Fig. 5 illustrates the variety of grasping motions generated for different hands under the same motion objectives.

4.4 Ablation and Analysis

Ablation: We conduct an ablation study to evaluate the impact of different components in our method. Specifically, we assessed variations without the hand guidance technique (w/o Guidance), the joint distance features (w/o Distance), and the learning curriculum (w/o Curriculum). Results are detailed in Tab. 6. The version without the curriculum slightly outperforms in contact ratio on the PartNet set and maintains similar success rates across tests but suffers significantly in objective-related metrics. We also notice the absence of the curriculum makes the results more sensitive to the exact reward coefficients, leading to more grasp failure with slightly higher objective rewards or poor objective precision with higher grasping rewards. These underscore the curriculum's role in decoupling the learning of stable grasping and objective fulfillment, which can help the policy avoid getting stuck in local optima caused by their influence on each other during training. The model without the joint distance features gets a slightly larger midpoint error and significantly underperforms in all other aspects, highlighting the distance features' value in understanding object shapes

Table 6: Effects of Different Components in *GraspXL*.

Model	PartNet Test Set Suc. Rate [%] ↑	Mid. Error [cm] ↓	Head. Error [rad] ↓	Rot. Error [rad] ↓	Contact Ratio [%] ↑	ShapeNet Test Set Suc. Rate [%] ↑	Mid. Error [cm] ↓	Head. Error [rad] ↓	Rot. Error [rad] ↓
w/o Guidance	90.0	3.22	0.394	0.425	82.2	68.5	3.74	0.455	0.528
w/o Distance	81.6	2.90	0.419	0.475	84.2	70.7	3.34	0.467	0.510
w/o Curriculum	**96.2**	4.12	0.381	0.462	**88.8**	79.6	4.60	0.396	0.461
Ours	95.0	**2.85**	**0.270**	**0.306**	86.7	**81.0**	**3.22**	**0.292**	**0.338**

Table 7: Generation Performance with Different Objective Combinations.

Combination	PartNet Test Set Suc. Rate [%] ↑	Mid. Error [cm] ↓	Head. Error [rad] ↓	Rot. Error [rad] ↓	Contact Ratio [%] ↑	ShapeNet Test Set Suc. Rate [%] ↑	Mid. Error [cm] ↓	Head. Error [rad] ↓	Rot. Error [rad] ↓
Direction	**96.1**	-	**0.256**	-	**90.2**	**82.5**	-	**0.268**	-
Direction+Rotation	95.1	-	0.263	**0.303**	87.9	81.8	-	0.276	**0.320**
Direction+Midpoint	95.2	**2.84**	0.268	-	88.7	81.4	**3.15**	0.284	-
Dir.+Rot.+Mid.	95.0	2.85	0.270	0.306	86.7	81.0	3.22	0.292	0.338

and adjusting to contact-induced movements. Lastly, excluding hand guidance worsens all performance metrics, proving its efficacy in directing the hand to achieve the desired grasp.

Evaluation on Different Objective Combinations: Our method's adaptability was tested across various objective combinations to gauge performance impacts. Evaluations were conducted for different scenarios: solely controlling heading direction ($\bar{\mathbf{m}}$), adding wrist rotation to heading direction ($\bar{\mathbf{v}} + \bar{\omega}$), combining heading direction with midpoint position ($\bar{\mathbf{v}} + \bar{\mathbf{m}}$), and integrating all three objectives ($\bar{\mathbf{v}} + \bar{\omega} + \bar{\mathbf{m}}$). This variety helps to understand our approach's efficacy in handling multiple, concurrent objectives and understanding the complexity levels our method can effectively manage. For each objective set, we sample 25 sets of motion objectives for PartNet objects and 5 for ShapeNet objects. Results in Tab. 7 show slightly better performance with fewer motion objectives, highlighting the increased control challenge with multiple objectives. This demonstrates our method's capability to handle varying difficulty levels and its robustness, as indicated by the minimal performance decline.

5 Conclusion

We presented *GraspXL*, an RL-based method that learns to synthesize grasping motions on large-scale while satisfying one or multiple motion objectives. We introduced a learning curriculum to deal with the control complexity and an objective-driven guidance technique to accelerate exploration during training. To improve the generalization ability, we adopted a joint distance sensor to capture object local shape features. Notably, we demonstrated that our method exhibits a high success rate of 82.2% on a test set with more than 500k unseen objects. Moreover, we showed that our approach works well even when applied to different dexterous hand platforms and with reconstructed or generated objects.

References

1. Agarwal, A., Uppal, S., Shaw, K., Pathak, D.: Dexterous functional grasping. In: Conference on Robot Learning (CoRL) (2023)
2. Braun, J., Christen, S., Kocabas, M., Aksan, E., Hilliges, O.: Physically plausible full-body hand-object interaction synthesis. In: International Conference on 3D Vision (3DV) (2024)
3. Cao, Z., Radosavovic, I., Kanazawa, A., Malik, J.: Reconstructing hand-object interactions in the wild. In: International Conference on Computer Vision (ICCV). pp. 12417–12426 (2021)
4. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: ShapeNet: An Information-Rich 3D Model Repository. Tech. Rep. arXiv:1512.03012 (2015)
5. Chen, Y., Wu, T., Wang, S., Feng, X., Jiang, J., Lu, Z., McAleer, S., Dong, H., Zhu, S.C., Yang, Y.: Towards human-level bimanual dexterous manipulation with reinforcement learning. Adv. Neural Inform. Process. Syst. (2022)
6. Chen, Z.Q., Van Wyk, K., Chao, Y.W., Yang, W., Mousavian, A., Gupta, A., Fox, D.: DexTransfer: Real world multi-fingered dexterous grasping with minimal human demonstrations. arXiv:2209.14284 (2022)
7. Christen, S., Feng, L., Yang, W., Chao, Y.W., Hilliges, O., Song, J.: Synh2r: Synthesizing hand-object motions for learning human-to-robot handovers. In: IEEE International Conference on Robotics and Automation (ICRA) (2024)
8. Christen, S., Kocabas, M., Aksan, E., Hwangbo, J., Song, J., Hilliges, O.: D-Grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In: Computer Vision and Pattern Recognition (CVPR) (2022)
9. Christen, S., Stevšić, S., Hilliges, O.: Demonstration-guided deep reinforcement learning of control policies for dexterous human-robot interaction. In: International Conference on Robotics and Automation (ICRA) (2019)
10. Corona, E., Pumarola, A., Alenyà, G., Moreno-Noguer, F., Rogez, G.: GanHand: Predicting human grasp affordances in multi-object scenes. In: Computer Vision and Pattern Recognition (CVPR). pp. 5030–5040 (2020)
11. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. arXiv preprint arXiv:2212.08051 (2022)
12. Ding, Z., Chen, Y., Ren, A.Z., Gu, S.S., Dong, H., Jin, C.: Learning a universal human prior for dexterous manipulation from human preference. arXiv:2304.04602 (2023)
13. Duran, E., Kocabas, M., Choutas, V., Fan, Z., Black, M.J.: HMP: Hand motion priors for pose and shape estimation from video. In: Winter Conference on Applications of Computer Vision (WACV). pp. 6353–6363 (January 2024)
14. Fan, Z., Ohkawa, T., Yang, L., Lin, N., Zhou, Z., Zhou, S., Liang, J., Gao, Z., Zhang, X., Zhang, X., Li, F., Zheng, L., Lu, F., Zeid, K.A., Leibe, B., On, J., Baek, S., Prakash, A., Gupta, S., He, K., Sato, Y., Hilliges, O., Chang, H.J., Yao, A.: Benchmarks and challenges in pose estimation for egocentric hand interactions with objects. In: European Conference on Computer Vision (ECCV) (2024)
15. Fan, Z., Parelli, M., Kadoglou, M.E., Kocabas, M., Chen, X., Black, M.J., Hilliges, O.: HOLD: Category-agnostic 3d reconstruction of interacting hands and objects from video (2024)
16. Fan, Z., Spurr, A., Kocabas, M., Tang, S., Black, M., Hilliges, O.: Learning to disambiguate strongly interacting hands via probabilistic per-pixel part segmentation. In: International Conference on 3D Vision (3DV) (2021)

17. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
18. Ghosh, A., Dabral, R., Golyanik, V., Theobalt, C., Slusallek, P.: IMoS: Intent-driven full-body motion synthesis for human-object interactions. In: Eurographics (2023)
19. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: HOnnotate: A method for 3d annotation of hand and object poses. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
20. Hwangbo, J., Lee, J., Hutter, M.: Per-contact iteration method for solving contact dynamics. Robotics and Automation Letters (RA-L) (2018)
21. Jiang, H., Liu, S., Wang, J., Wang, X.: Hand-object contact consistency reasoning for human grasps generation. In: International Conference on Computer Vision (ICCV) (2021)
22. Li, S., Jiang, J., Ruppel, P., Liang, H., Ma, X., Hendrich, N., Sun, F., Zhang, J.: A mobile robot hand-arm teleoperation system by vision and imu. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 10900–10906. IEEE (2020)
23. Liu, Q., Cui, Y., Ye, Q., Sun, Z., Li, H., Li, G., Shao, L., Chen, J.: Dexrepnet: Learning dexterous robotic grasping network with geometric and spatial hand-object representations. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3153–3160. IEEE (2023)
24. Liu, S., Jiang, H., Xu, J., Liu, S., Wang, X.: Semi-supervised 3D hand-object poses estimation with interactions in time. In: Computer Vision and Pattern Recognition (CVPR). pp. 14687–14697 (2021)
25. Mandikal, P., Grauman, K.: DexVIP: Learning dexterous grasping with human hand pose priors from video. In: Conference on Robot Learning (CoRL) (2021)
26. Mandikal, P., Grauman, K.: Learning dexterous grasping with object-centric visual affordances. In: International Conference on Robotics and Automation (ICRA) (2021)
27. Mo, K., Zhu, S., Chang, A.X., Yi, L., Tripathi, S., Guibas, L.J., Su, H.: PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In: Computer Vision and Pattern Recognition (CVPR)
28. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv (2022)
29. Qin, Y., Huang, B., Yin, Z.H., Su, H., Wang, X.: DexPoint: Generalizable point cloud reinforcement learning for sim-to-real dexterous manipulation. In: Conference on Robot Learning (CoRL) (2023)
30. Qin, Y., Wu, Y.H., Liu, S., Jiang, H., Yang, R., Fu, Y., Wang, X.: DexMV: Imitation learning for dexterous manipulation from human videos. In: European Conference on Computer Vision (ECCV) (2022)
31. Rajeswaran, A., Kumar, V., Gupta, A., Vezzani, G., Schulman, J., Todorov, E., Levine, S.: Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. In: Robotics: Science and Systems (RSS) (2018)
32. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) (2017)
33. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv:1707.06347 (2017)

34. Shadow Robot: Shadow robot hand, https://www.shadowrobot.com/dexterous-hand-series
35. Taheri, O., Choutas, V., Black, M.J., Tzionas, D.: GOAL: Generating 4D whole-body motion for hand-object grasping. In: Computer Vision and Pattern Recognition (CVPR) (2022), https://goal.is.tue.mpg.de
36. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: GRAB: A dataset of whole-body human grasping of objects. In: European Conference on Computer Vision (ECCV). vol. 12349, pp. 581–600 (2020)
37. Tekin, B., Bogo, F., Pollefeys, M.: H+O: Unified egocentric recognition of 3D hand-object poses and interactions. In: Computer Vision and Pattern Recognition (CVPR). pp. 4511–4520 (2019)
38. Toshimitsu, Y., Forrai, B., Cangan, B.G., Steger, U., Knecht, M., Weirich, S., Katzschmann, R.K.: Getting the ball rolling: Learning a dexterous policy for a biomimetic tendon-driven hand with rolling contact joints. arXiv:2308.02453 (2023)
39. Turpin, D., Zhong, T., Zhang, S., Zhu, G., Heiden, E., Macklin, M., Tsogkas, S., Dickinson, S., Garg, A.: Fast-grasp'd: Dexterous multi-finger grasp generation through differentiable simulation. In: International Conference on Robotics and Automation (ICRA) (2023)
40. Wan, W., Geng, H., Liu, Y., Shan, Z., Yang, Y., Yi, L., Wang, H.: UniDexGrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. In: International Conference on Computer Vision (ICCV) (2023)
41. Wonik Robotics: Allegro robot hand, https://www.wonikrobotics.com/robot-hand
42. Xu, Y., Wan, W., Zhang, J., Liu, H., Shan, Z., Shen, H., Wang, R., Geng, H., Weng, Y., Chen, J., et al.: UniDexGrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In: Computer Vision and Pattern Recognition (CVPR) (2023)
43. Yang, L., Zhan, X., Li, K., Xu, W., Li, J., Lu, C.: CPF: Learning a contact potential field to model the hand-object interaction. In: International Conference on Computer Vision (ICCV) (2021)
44. Ye, J., Wang, J., Huang, B., Qin, Y., Wang, X.: Learning continuous grasping function with a dexterous hand from human demonstrations. Robotics and Automation Letters (RA-L) (2023)
45. Ye, Y., Li, X., Gupta, A., Mello, S.D., Birchfield, S., Song, J., Tulsiani, S., Liu, S.: Affordance diffusion: Synthesizing hand-object interactions. In: Computer Vision and Pattern Recognition (CVPR) (2023)
46. Ze, Y., Liu, Y., Shi, R., Qin, J., Yuan, Z., Wang, J., Xu, H.: H-index: Visual reinforcement learning with hand-informed representations for dexterous manipulation. Conference on Neural Information Processing Systems (NeurIPS) (2023)
47. Zhang, H., Ye, Y., Shiratori, T., Komura, T.: ManipNet: Neural manipulation synthesis with a hand-object spatial representation. ACM Trans. Graph. (2021). https://doi.org/10.1145/3450626.3459830, https://doi.org/10.1145/3450626.3459830
48. Zhang, H., Christen, S., Fan, Z., Zheng, L., Hwangbo, J., Song, J., Hilliges, O.: ArtiGrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation. In: International Conference on 3D Vision (3DV) (2024)
49. Zheng, J., Zheng, Q., Fang, L., Liu, Y., Yi, L.: CAMS: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In: Computer Vision and Pattern Recognition (CVPR) (2023)

50. Ziani, A., Fan, Z., Kocabas, M., Christen, S., Hilliges, O.: TempCLR: Reconstructing hands via time-coherent contrastive learning. In: International Conference on 3D Vision (3DV). pp. 627–636 (2022)