Grasping a Handful: Sequential Multi-Object Dexterous Grasp Generation

Haofei Lu, Yifei Dong, Zehang Weng, Jens Lundell, and Danica Kragic

***Abstract*— We introduce the sequential multi-object robotic grasp sampling algorithm SeqGrasp that can robustly synthesize stable grasps on diverse objects using the robotic hand's partial Degrees of Freedom (DoF). We use SeqGrasp to construct the large-scale Allegro Hand sequential grasping dataset SeqDataset and use it for training the diffusion-based sequential grasp generator SeqDiffuser. We experimentally evaluate SeqGrasp and SeqDiffuser against the state-of-the-art non-sequential multi-object grasp generation method MultiGrasp in simulation and on a real robot. The experimental results demonstrate that SeqGrasp and SeqDiffuser reach an 8.71%-43.33% higher grasp success rate than MultiGrasp. Furthermore, SeqDiffuser is approximately 1000 times faster at generating grasps than SeqGrasp and MultiGrasp.**

I. Introduction

Generation of dexterous grasps has been studied for a long time, both from a technical perspective on generating grasps on robots [1]–[11] and understanding human grasping [12]–[15]. Most of these methods rely on bringing the robotic hand close to the object and then simultaneously enveloping it with all fingers. While this strategy often results in efficient and successful grasp generation, it simplifies dexterous grasping to resemble parallel-jaw grasping, thereby underutilizing the many DoF of multi-fingered robotic hands [10]. In contrast, grasping multiple objects with a robotic hand, particularly in a sequential manner that mirrors human-like dexterity, as shown in Fig. 1, is still an unsolved problem.

In this work, we introduce SeqGrasp, a novel hand-agnostic algorithm for generating sequential multi-object grasps. Our approach utilizes an optimization-based method to sequentially determine single-object grasp poses using a subset of the hand's DoF. As the grasp sequence progresses, the DoF engaged in previous grasps are frozen, leaving only the remaining DoF available for subsequent object grasps. To only engage a subset of the hand's DoF for each grasp, we propose an Opposition Space (OS) selection strategy that enables stable grasping using only a pair of links. Using SeqGrasp, we construct the large-scale dataset SeqDataset containing 870K penetration-free Allegro hand grasps across 509 objects, with up to four sequentially grasped objects. Finally, we train the conditional sequential grasping diffusion model SeqDiffuser on SeqDataset to enable sequential grasping on novel objects.

We experimentally evaluate SeqGrasp, SeqDiffuser, and the state-of-the-art simultaneous multi-object grasping method MultiGrasp [16] in simulation and on physical hardware. The simulation results revealed that SeqGrasp and SeqDiffuser perform on par with MultiGrasp for picking one or two objects while outperforming it when picking three to four objects. Moreover, SeqDiffuser demonstrates superior efficiency, generating 256 grasps within one second compared to approximately 1000 seconds for SeqGrasp and MultiGrasp. For the real-world experiments, we replicated the grasp sequences proposed by the methods on a real Allegro Hand attached to a Franka Panda Robot. These results align with the simulation findings, demonstrating that SeqGrasp reaches a 43.33% higher grasp success rate than MultiGrasp.

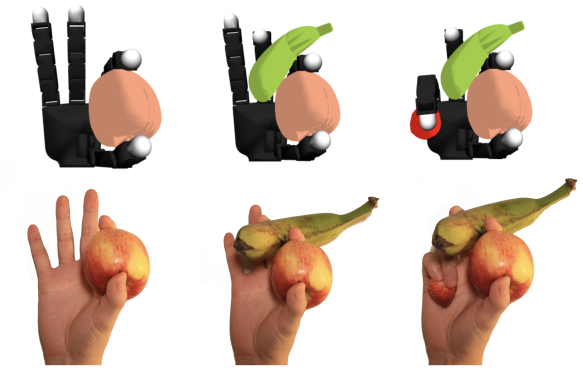


Fig. 1: **Sequential multi-object grasping.**

Our contributions can be summarized as follows:

- SeqGrasp, a novel hand-agnostic algorithm for sequential multi-object grasp generation.
- SeqDataset, a large-scale dataset for sequential multi-object dexterous grasping.
- SeqDiffuser, an efficient conditional sequential grasp diffusion model.
- Extensive simulation and real-world experiments demonstrating the feasibility and effectiveness of SeqGrasp.

II. Related Work

The work presented here spans dexterous grasping, multi-object grasping, and datasets for dexterous grasping and we review these areas below.

A. Dexterous Grasping

Analytical methods. Early research in dexterous grasping generated stable grasps by optimizing a grasp quality metric such as the force-closure metric [18], [19]. Although these methods are theoretically sound, they are computationally demanding because (i) the many DoF dexterous hands cause high-dimensional search spaces [20], and (ii) the quality metrics are expensive to compute [21]. Consequently, some methods have focused on reducing the search space by imposing constraints on the hand [22], [23] or restricting it

Dataset	Hand	Data Collection	Objects	Obj. Set/Seq.	Grasps	Strategy	Max. Obj. Grasped	Open-source
MOG [17]	Human&Barrett	Sim./Real	11	\	28K	Simu./Seq.	\	✗
Grasp'Em [16]	Shadow	Sim.	8	36	90K	Simultaneous	2	✓
SeqDataset(Ours)	Allegro	Sim.	509	2400	870K	Sequential	4	✓*

TABLE I: **Multi-object grasping datasets. *Our dataset will be made publicly available upon acceptance.**

to joint configuration subspaces [20]. Another line of work has proposed a computationally cheap differentiable force-closure estimator [10], [21], which has the advantage of being hand-agnostic. This work extends the differentiable force-closure measure from [10] to sequential multi-object grasping.

Data-driven methods. Recent advancements in machine learning have significantly improved dexterous grasp generation [5], [6], [8], [24]. Nowadays, deep generative models can generate thousands of dexterous grasps on previously unseen and partially observed objects within seconds [5], [6], [8], [24], something analytical methods cannot. Still, only a few data-driven dexterous grasping methods have been developed for multi-object grasping [16], [25]. In this work, we train a new diffusion-based sequential multi-object grasp sampler SeqDiffuser inspired from [5] on our own optimization-generated sequential multi-object dataset SeqDataset.

B. Multi-Object Grasping

Multi-object grasping presents unique challenges due to the complex multi-object interactions and the high-dimensional configuration space spanned by the hand and the objects. Some prior parallel-jaw multi-object grasping methods [26], [27] explored multi-object push grasps where scattered objects are first pushed together to facilitate multi-object grasping. However, these methods are limited by their reliance on simple shape primitives and parallel-jaw grippers. In comparison, our work can handle objects of diverse shapes and sizes.

A few works have addressed dexterous multi-object grasping [16], [25], [28] where [16] targets simultaneous multi-object grasping while [25], [28] targets sequential multi-object grasping. Li et al. [16] proposed MultiGrasp a two-stage simultaneous multi-object dexterous grasping framework where a generative grasp sampler proposed poses to simultaneously pick many objects, followed by a learned policy for executing the pick. The main limitation of [16] is that objects must be spatially close and of similar size and shape. In comparison, our method can handle scattered objects of different shapes and sizes by sequentially picking one at a time. The other works that do sequential multi-object grasping [25], [28] restrict the grasping to a maximum of two objects [25] or to primitive object shapes such as cylinders or spheres [28]. In comparison, our method can handle up to four objects of complex shapes and sizes. We also use our method to collect the largest sequential multi-object grasping dataset to date.

C. Datasets for Dexterous Grasping

Large-scale dexterous grasp datasets [7], [10], [16], [17], [24], [29]–[31] have significantly advanced the training of data-driven dexterous grasping methods. However, most of these datasets target single-object grasping [7], [10], [24], [29]–[31], with only a few for multi-object grasping [16], [17]. As shown in Table I, these existing multi-object datasets are small and predominantly focus on simultaneous rather than sequential grasping. Therefore, we collect the new large-scale sequential multi-object grasping dataset SeqDataset using our method SeqGrasp. SeqDataset is, to date, by far the largest multi-object grasping dataset.

III. Problem Formulation

The problem addressed in this work is sequential multi-object grasping, which we define as follows:

Definition 1 (Sequential multi-object grasping)**.** A sequential multi-object grasp is a grasp where one object is grasped at a time using a subset of the dexterous hand's DoF, while previously grasped objects, if any, remain fixed to the hand.

To contrast, *simultaneous multi-object grasping* addresses how to grasp multiple objects *simultaneously*, typically utilizing all the DoF of the hand [16].

We formulate the sequential multi-object grasping problem as generating a sequence of N grasps $\mathcal{G} = \{\mathbf{g}_i\}_{i=1}^{N}$ for picking a sequence of N objects $\mathcal{O} = \{\mathbf{O}_i\}_{i=1}^{N}$, where each $\mathbf{g}_n \in \mathcal{G}$ is restricted to a specific subset $\mathcal{OS}_n$ of the hand's total DoF and $N \geq 2$. Mathematically, this can be described as

$$\mathbf{g}_n = \underset{\mathbf{g}_n}{\operatorname{argmin}} E(\mathbf{g}_n, \mathbf{O}_n, \mathcal{G}_{n-1}, \mathcal{O}_{n-1}, \mathcal{OS}_n), \quad \forall n = 1, \ldots, N, \qquad (1)$$

where $\mathcal{G}_{n-1} = \{\mathbf{g}_i\}_{i=1}^{n-1}$, $\mathcal{O}_{n-1} = \{\mathbf{O}_i\}_{i=1}^{n-1}$, $\mathcal{G}_0 = \varnothing$, and $\mathcal{O}_0 = \varnothing$. E in Eq. 1 is a differentiable function that quantifies how well grasp $\mathbf{g}_n$ can pick object $\mathbf{O}_n$ with the DoF $\mathcal{OS}_n$ given all previously generated grasps $\mathcal{G}_{n-1}$ and objects $\mathcal{O}_{n-1}$.

In this work, we represent $\mathcal{OS}_n$ as an opposition space (Section IV-A), each object $\mathbf{O} \in \mathcal{O}$ as a triangular mesh, and each grasp $\mathbf{g} \in \mathcal{G}$ as a vector $\mathbf{g} = [\mathbf{p}, \mathbf{r}, \boldsymbol{\theta}] \in \mathbb{R}^{9+K}$, where $\mathbf{p} \in \mathbb{R}^3$ is the hand's base position, $\mathbf{r} \in \mathbb{R}^6$ is the hand's base orientation in a 6D continuous reprsentation [32], and $\boldsymbol{\theta} \in \mathbb{R}^K$ is the K-dimensional hand joint angles which are 16 for the Allegro Hand. We assume the shape of all objects in $\mathcal{O}$ to be fully known. Next, we will introduce SeqGrasp our algorithm for solving Eq. 1.

IV. Sequential Grasp Generation

Here, we present Algorithm 1 for sequential grasp generation. It includes (i) an opposition space selection strategy (Section IV-A), (ii) an optimization-based grasp synthesis

Algorithm 1: SeqGrasp

Input : Object sequence $\mathcal{O}$, OSes $\mathcal{OS}$, N_{step}, and p_{accept}.
Output: The optimized grasp sequence ${\mathcal{G}^*}_n$.

1 $n = 1$;
2 **while** $\mathcal{OS} \neq \varnothing$ *and* $n \leq N$ **do**
3 $\quad \mathcal{OS}_n \sim \mathcal{U}(\mathcal{OS})$;
4 $\quad \{\mathbf{x}_j\}_{j=1}^2 \sim \mathcal{U}(\mathbf{S}_n)$;
5 $\quad$ **for** $s = 1$ **to** N_{step} **do**
6 $\quad\quad \Delta = \partial E(\mathbf{g}_n, \mathbf{O}_n, \mathcal{G}_{n-1}, \mathcal{O}_{n-1}, \{\mathbf{x}_j\}_{j=1}^2)/\partial \mathbf{g}_n$;
7 $\quad\quad \mathbf{g}_n \leftarrow \text{MALA}(\mathbf{g}_n, \mathbf{J}_n, \Delta)$;
8 $\quad\quad \{\mathbf{x}_j\}_{j=1}^2 \sim f(\mathbf{S}_n, p_{\text{accept}})$;
9 $\quad$ **end**
10 $\quad \mathcal{OS} \leftarrow \mathcal{OS} \setminus \mathcal{OS}_n$;
11 $\quad$ **for** $\mathcal{OS}_j \in \mathcal{OS}$ **do**
12 $\quad\quad \mathbf{J}_j \leftarrow \mathbf{J}_j \odot (\mathbf{1} - \mathbf{J}_n)$;
13 $\quad\quad$ **if** $\mathbf{J}_j = \mathbf{0}$ **then**
14 $\quad\quad\quad \mathcal{OS} \leftarrow \mathcal{OS} \setminus \mathcal{OS}_j$;
15 $\quad\quad$ **end**
16 $\quad$ **end**
17 $\quad n \mathrel{+}= 1$;
18 **end**

method (Section IV-B), and (iii) an energy-based cost function (Section IV-C). Fig. 2 shown an example of running Algorithm 1 to grasp three different objects with three different dexterous hands.

A. Opposition Space Selection Strategy

The primary objective in sequential multi-object grasping is to maximize the hand's remaining DoF after each grasp. For this purpose, we propose a grasp planning strategy guided by OSes [14], [28], [33], [34]. An OS is a functional subspace within the hand's kinematic structure formed by pairs of opposing surfaces (such as fingertips, lateral surfaces of fingers, or palm surfaces) along with the joints that control these surfaces [28]. It represents regions where opposing forces can be applied to create stable grasps. The number of OSes is hand-dependent and varies based on the kinematic structure. Fig. 3a shows the seven different OSes for the Allegro Hand.

Mathematically, each opposition space can be represented as a pair $\mathcal{OS}_i = \{\mathbf{J}_i, \mathbf{S}_i\}$, where $\mathbf{J}_i \in \{0,1\}^K$ is a binary vector indicating which joints are involved in controlling the opposition space, and $\mathbf{S}_i \in \mathbb{R}^{3\times M_i}$ represents the 3D points on the hand where opposing forces can be applied. Fig. 3b shown an example of two different $\mathbf{S}_i$ for the Allegro Hand,

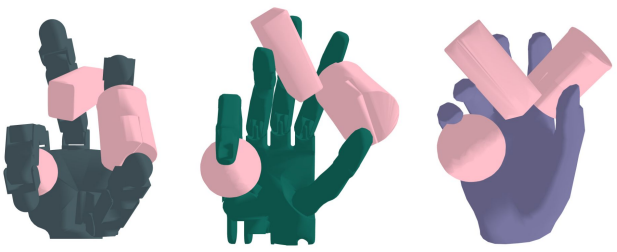


Fig. 2: **Multi-object grasping visualizations** for different hands. From left to right: Allegro Hand, Shadowhand, MANO.

where palm and pad oppositions have contact points located on the inner surfaces of fingers and palm and side oppositions have contact points on the fingers' lateral surfaces.

Let $\mathcal{OS} = \{\mathcal{OS}_i\}_{i=1}^L$ be the set of all OSes. Given this set, Algorithm 1 samples a random OS from it (Line 3) and uses it for subsequent grasp generation (Section IV-B). Once grasp generation is complete, the sampled OS can no longer be used and is thus removed from the available OSes (Line 10). $\mathbf{J}_i$ of all the remaining OSes are also updated by zeroing out the joints used in $\mathcal{OS}_n$ (Line 12). Subsequently, all OSes with $\mathbf{J} = \mathbf{0}$, meaning that no more controllable joints exist, are removed (Line 14). For instance, in the case of the Allegro Hand, if the thumb-index OS is selected, then both the thumb-palm and index-palm OSes become unavailable due to shared joint constraints.

B. Optimization-based Grasp Generation

The next step in the algorithm (Lines 5-9) is to generate a stable, physically plausible, and collision-free grasp that respects the sampled $\mathcal{OS}_n$. To achieve this, we formulate E in Eq. 1 as an energy function (Section IV-C) and numerically optimize it using the Metropolis-Adjusted Langevin Algorithm (MALA) [35] (Line 7).

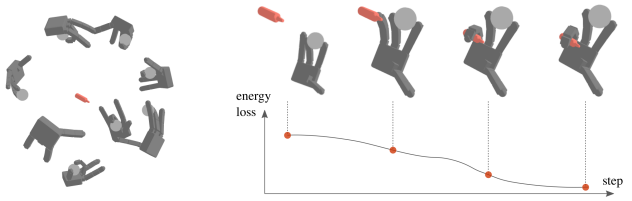
In robotic grasping, MALA has been used to optimize single object grasps $\mathbf{g}_n$ by iteratively refining $\mathbf{p}_n$, $\mathbf{r}_n$ and $\boldsymbol{\theta}_n$ according to Langevin dynamics [10], [21]. However, we must adapt MALA to sequential multi-object grasping. To this end, we propose a new grasp as $\hat{\mathbf{g}}_n \leftarrow \mathbf{g}_n - \gamma\,[\mathbf{1}, \mathbf{J}_n] \odot \Delta$, where γ is the step size, $\mathbf{1} \in \mathbb{R}^9$ is a padding vector to align the length of $\mathbf{J}$ with $\mathbf{g}$, $\Delta = \partial E/\partial \mathbf{g}_n$ is the energy gradient, and $\odot$ is the element-wise (Hadamard) product. $\hat{\mathbf{g}}_n$ is accepted if $\alpha \geq u$, where $u \sim \mathcal{U}([0,1])$ and

$$\alpha = \frac{E(\hat{\mathbf{g}}_n, \mathbf{O}_n, \mathcal{G}_{n-1}, \mathcal{O}_{n-1}, \{\mathbf{x}_j\}_{j=1}^2)}{E(\mathbf{g}_n, \mathbf{O}_n, \mathcal{G}_{n-1}, \mathcal{O}_{n-1}, \{\mathbf{x}_j\}_{j=1}^2)}. \tag{2}$$

The above procedure is repeated for a fixed number of steps where, at each step, $\{\mathbf{x}_j\}_{j=1}^2$ is re-sampled with probability p_{accept} (Line 8). This resampling process helps accelerate convergence and escape from local minimas [10], [21].

(a) (b)

Fig. 3: (a) **Grasps using all seven OSes**. From left to right, first row: middle-ring, index-middle, and thumb-index, second row: ring-palm, middle-palm, index-palm, and thumb-palm. (b) **Visualization of contact point candidates** on Allegro Hand surface. Cyan and pink points denote palm opposition and side opposition contacts, respectively.



(a) (b)

Fig. 4: (a) **Initialization.** The initial grasp configurations are randomly sampled on the expanded convex hull of the object (bottle) while the previously grasped object (ball) remains in the hand. (b) **Optimization.** During optimization, the grasp is incrementally refined, ultimately securing the target object using the ring-palm OS.

We initialize $\mathbf{g}_n$ at a randomly sampled position on the expanded convex hull of the target object $\mathbf{O}_n$ as exemplified in Fig. 4a,. If $n = 1$, then $\boldsymbol{\theta}_1$ is set to a natural open-hand and collision-free posture, while for $n \geq 2$, $\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1}$. A visual example of the optimization process when grasping a second object is shown in Fig. 4b.

C. Energy Function

Numerically optimizing the energy function in Eq. 1 should result in stable, collision-free, joint-respecting, and OS-respecting grasps. We design the following energy function to capture all of these behaviors

$$E = \mathbf{w}^T \begin{bmatrix} E_{\text{fc}} & E_{\text{dis}} & E_{\text{hop}} & E_{\text{hsp}} & E_{\text{joint}} & E_{\text{oop}} \end{bmatrix}^T, \quad (3)$$

where $\mathbf{w} \in \mathbb{R}^6$ is a weight vector controlling the relative importance of the force-closure E_{fc}, contact distance E_{dis}, hand-object penetration E_{hop}, hand self-penetration E_{hsp}, joint limits E_{joint}, and object-object penetration E_{oop} energy terms.

The force-closure term (E_{fc}) encourages the grasp to be in force-closure equilibrium [36]. Following [10] and assuming zero friction and uniform contact force magnitudes, we define it as

$$E_{\text{fc}}(\{\mathbf{x}_j\}_{j=1}^2) = \|\mathbf{G}\mathbf{c}\|^2, \quad (4)$$

where $\mathbf{c} = [\mathbf{c}_1^T, \mathbf{c}_2^T]^T \in \mathbb{R}^{6\times 1}$ represents the concatenated contact normals at each contact point $\{\mathbf{x}_j\}_{j=1}^2$. $\mathbf{G}$ is defined as:

$$\mathbf{G} = \begin{bmatrix} \mathbf{I} & \mathbf{I} \\ [\mathbf{x}_1]_\times & [\mathbf{x}_2]_\times \end{bmatrix}, \quad (5)$$

where $\mathbf{I}$ represents the identity matrix, and $[\mathbf{x}_j]_\times$ $(1 \leq j \leq 2)$ denotes the skew-symmetric matrix formed from the contact point $\mathbf{x}_j$.

The contact distance and penetration terms (E_{dis} & E_{hop}) encourage the hand-object contacts to occur close to the object surface without penetrating it. The contact distance is mathematically defined as

$$E_{\text{dis}}(\{\mathbf{x}_j\}_{j=1}^2, \mathbf{O}_n) = \sum_{j=1}^{2} d(\mathbf{x}_j, \mathbf{O}_n), \quad (6)$$

where $d(\mathbf{x}_j, \mathbf{O}_n) = \min_{\mathbf{v}\in\mathbf{O}_n} \|\mathbf{x}_j - \mathbf{v}\|_2$ is the shortest point-mesh distance. Similarly, the hand-object penetration term is defined as:

$$E_{\text{hop}}(\mathbf{g}_n, \mathbf{O}_n) = \sum_{\mathbf{v}\in\mathcal{V}_{\text{hop}}(\mathbf{H_g}, \mathbf{O}_n)} d(\mathbf{v}, \mathbf{O}_n), \quad (7)$$

where $d(\mathbf{v}, \mathbf{O}_n) = \min_{\mathbf{v}_1\in\mathbf{O}_n} \|\mathbf{v} - \mathbf{v}_1\|_2$ and $\mathcal{V}_{\text{hop}}(\mathbf{H_g}, \mathbf{O}_n)$ is the set of points on the hand surface pointcloud $\mathbf{H_g} \in \mathbb{R}^{3\times M_h}$ that penetrate the object $\mathbf{O}_n$.

The self-collision and joint limit terms (E_{hsp} & E_{joint}) encourage physical feasibility. We define these as

$$E_{\text{hsp}}(\mathbf{g}_n) = \sum_{\mathbf{v}_1,\mathbf{v}_2\in\mathcal{V}_{\text{hsp}}(\mathbf{H_g}),\mathbf{v}_1\neq\mathbf{v}_2} \max(\|\mathbf{v}_1 - \mathbf{v}_2\|_2, 0), \quad (8)$$

$$E_{\text{joint}}(\mathbf{g}_n) = \|(\boldsymbol{\theta} - \boldsymbol{\theta}^{\text{upper}})^+\|_1 + \|(\boldsymbol{\theta}^{\text{lower}} - \boldsymbol{\theta})^+\|_1, \quad (9)$$

where $\mathcal{V}_{\text{hsp}}(\mathbf{H_g})$ denotes all surface points of the hand that are self-penetrating, $(\cdot)^+$ denotes the element-wise operation $\max(\cdot, 0)$, and $\boldsymbol{\theta}^{\text{upper}}$ and $\boldsymbol{\theta}^{\text{lower}}$ denote the upper and lower limits of all joints.

Finally, the term (E_{oop}) minimizes object-object penetration. It is defined as

$$E_{\text{oop}}(\{\mathbf{O}_i\}_{i=1}^n) = \sum_{i=1}^{n-1} \sum_{\mathbf{v}\in\mathcal{V}_{\text{oop}}(\mathbf{O}_i, \mathbf{O}_n)} d(\mathbf{v}, \mathbf{O}_n), \quad (10)$$

where $d(\mathbf{v}, \mathbf{O}_n) = \min_{\mathbf{v}_1\in\mathbf{O}_n} \|\mathbf{v} - \mathbf{v}_1\|_2$, and $\mathcal{V}_{\text{oop}}(\mathbf{O}_i, \mathbf{O}_n)$ are the inter-penetrating surface points between the previously grasped object $\mathbf{O}_i$ and the current object $\mathbf{O}_n$.

V. Dataset Generation

We use SeqGrasp to generate our large-scale dataset SeqDataset containing 4.9 million sequential Allegro Hand grasps on over 509 objects from DexGraspNet [10]. Each object is resized to fit within the Allegro Hand by scaling its axis-aligned bounding box to be between 0.06 and 0.10 meters. Then, from the 509 resized objects, we generate 600 unique object sets containing four objects each. The objects in these sets are randomly permuted four times, resulting in 2,400 unique object sequences.

We run Algorithm 1 on the 2,400 unique object sequences with $\mathbf{w} = [50, 50, w_{\text{hop}}, 5, 1, 5]^T$, where w_{hop}, that penalizes hand-object penetration, grows linearly from 5 to $5e^2$. The optimization runs for 6,000 iterations per grasp. We validate the generated grasp sequences in the physics simulator Isaac Gym following the setup in [10]. This setup initializes all object densities to 500 kg/m^3, the friction coefficient to 2.0, the hand to the generated grasp configuration, and objects as free-floating. Then, the hand closes until contact is established with the object, which happens when the distance between the hand and the object is less than 2 mm. Finally, a grasp stability test is evaluated by linearly applying an acceleration of 9.8 m/s^2 in all six orthogonal directions for 100 consecutive simulation steps. A grasp sequence is successful if all grasped objects remain in contact with the hand after the grasp stability test and the maximum penetration depth is less than 1 cm. Otherwise, the grasp sequence is a failure. This procedure ultimately produced 4.9

OSes	Success	Total	Success Rate (%)
Middle-Ring	100.83	704.97	14.30
Index-Middle	97.43	700.90	13.90
Thumb-Index	97.29	1691.41	5.75
Ring-Palm	147.33	323.31	45.57
Middle-Palm	152.77	404.84	37.74
Index-Palm	194.13	546.60	35.52
Thumb-Palm	85.86	485.80	17.67
Num. Obj. Grasped	**Consumed**	**Total**	**Consumed Rate (%)**
One	0.00	444.72	0.00
Two	9.75	251.52	3.88
Three	69.20	139.31	49.68
Four	40.09	40.09	100.00

TABLE II: **Statistics of SeqDataset.** All numbers of grasps are shown in thousand.

million grasps, of which 870K (17.82%) were successful. Only successful grasps were retained in SeqDataset. The entire data generation process requires approximately 2,500 GPU hours on an NVIDIA A100[1].

The statistics of SeqDataset are presented in Table II. The results demonstrate that grasps using palm opposition achieved significantly higher success rates, suggesting that these OSes are important in sequential multi-object grasping. Notably, grasps using thumb-index and thumb-palm OSes display lower success rates than other OSes, which deviates from previous research findings that underscore the thumb's dominant role in human hand manipulation tasks [37]. We hypothesize that this discrepancy may stem from the biomechanical differences between the Allegro and the human hand. Finally, SeqDataset's grasp consumption, which indicates no more available OSes ($\mathcal{OS} = \emptyset$), aligns with its objective of supporting multi-object grasping, with a significant portion of cases (96.12%) showing potential for sequential grasping of three to four objects, indicating efficient utilization of the hand's DoF.

VI. Conditional Sequential Grasp Diffuser

Finally, we introduce SeqDiffuser, a diffusion-based sequential grasp generation method trained on SeqDataset. The architecture of SeqDiffuser is similar to the sampler from [5], with the distinction that SeqDiffuser is also conditioned on the OS to enable sequential multi-object grasping.

For generating the grasp $\mathbf{g}_n$ on the n-th object $\mathbf{O}_n$, we first sample $\mathcal{OS}_n \sim \mathcal{OS}$. The selected OS is encoded as a one-hot feature vector $\mathbf{f}_n \in \{\mathbf{f}^i\}_{i=1}^L$. This feature vector is concatenated with the grasp $\mathbf{g}_n$ forming the augmented grasp representation $\tilde{\mathbf{g}}_n = [\mathbf{f}_n, \mathbf{p}_n, \mathbf{r}_n, \boldsymbol{\theta}_n]$. Of all the elements in $\tilde{\mathbf{g}}_n$,

we only want to diffuse $\mathbf{p}_n$, $\mathbf{r}_n$, and the subset of $\boldsymbol{\theta}$ that corresponds to $\mathcal{OS}_n$, *i.e.*, $\mathbf{J}_n$. Therefore, we create the binary diffusion selection vector $\mathbf{k}_n = [\mathbf{0}, \mathbf{1}, \mathbf{J}_n]$, where $\mathbf{0} \in \mathbb{R}^L$ and $\mathbf{1} \in \mathbb{R}^9$. Note that in the following, the subscript t of $\mathbf{g}_t$ denotes the timestep in the diffusion process, while the sequential grasp step n is omitted from in $\tilde{\mathbf{g}}_n$ and $\mathbf{k}_n$ for clarity.

[1]Due to this computational overhead, we do not generate a dataset for MANO nor Shadow hand. However, we will make the code publicly available for others to create such datasets.

The forward process for adding noise to a successful grasp $\tilde{\mathbf{g}}_0$ over T timesteps is

$$q(\tilde{\mathbf{g}}_{1:T}|\tilde{\mathbf{g}}_0) = \prod_{t=1}^{T} q(\tilde{\mathbf{g}}_t|\tilde{\mathbf{g}}_{t-1}), \tag{11}$$

$$q(\tilde{\mathbf{g}}_t|\tilde{\mathbf{g}}_{t-1}) = \mathcal{N}(\tilde{\mathbf{g}}_t; \mathbf{k} \odot \sqrt{1-\beta_t}\tilde{\mathbf{g}}_{t-1}, \beta_t \mathbf{I}), \tag{12}$$

where β_t is the scheduled noise variance at time step t. To reconstruct the original $\tilde{\mathbf{g}}_0$ from $\tilde{\mathbf{g}}_T$, SeqDiffuser learns the reverse diffusion process by estimating the Gaussian noise at each step t and progressively removing it:

$$p_{\boldsymbol{\psi}}(\tilde{\mathbf{g}}_{0:T}|\mathbf{h}_\mathbf{O}) = p(\tilde{\mathbf{g}}_T) \prod_{t=1}^{T} p_{\boldsymbol{\psi}}(\tilde{\mathbf{g}}_{t-1}|\tilde{\mathbf{g}}_t, \mathbf{h}_\mathbf{O}), \tag{13}$$

$$p_{\boldsymbol{\psi}}(\hat{\mathbf{g}}_{t-1}|\tilde{\mathbf{g}}_t, \mathbf{h}_\mathbf{O}) = \mathcal{N}(\tilde{\mathbf{g}}_{\mathbf{t-1}}; \mathbf{k} \odot \hat{\mu}_{\boldsymbol{\psi}}(\tilde{\mathbf{g}}_t, \mathbf{h}_\mathbf{O}, t), \hat{\Sigma}_{\boldsymbol{\psi}}(\tilde{\mathbf{g}}_t, \mathbf{h}_\mathbf{O}, t)), \tag{14}$$

where $\hat{\mu}_{\boldsymbol{\psi}}(\tilde{\mathbf{g}}_t, \mathbf{h}_\mathbf{O}, t)$ and $\hat{\Sigma}_{\boldsymbol{\psi}}(\tilde{\mathbf{g}}_t, \mathbf{h}_\mathbf{O}, t)$ are the learnable mean and variance of a Gaussian distribution and $\mathbf{h}_\mathbf{O}$ is the Basis Point Set (BPS) [38] encoded feature of the target object $\mathbf{O}$. In the forward and reverse processes, noise is only added or removed at positions where $\mathbf{k}$ is nonzero. The loss function is thereby formulated as:

$$\mathcal{L}_\epsilon = ||\mathbf{k} \odot \hat{\epsilon}_t - \mathbf{k} \odot \epsilon_t||^2, \tag{15}$$

where $\hat{\epsilon}_t = \epsilon_\psi(\tilde{\mathbf{g}}_t, \mathbf{h}_\mathbf{O}, t)$ and ϵ_t are the estimated noise and ground-truth noise, respectively.

VII. Experiments

We experimentally evaluate SeqGrasp and SeqDiffuser in both simulation and the real world. The specific questions we wanted to address with the experiments were:

1) How well can SeqGrasp and SeqDiffuser generate successful and diverse grasps?
2) What is the difference between simultaneously and sequentially grasping multiple objects?
3) Are the generated grasps executable on real hardware?

We compare our method to the optimization-based sampler MultiGrasp from [16], which generates simultaneous grasps on clustered objects. As such, for MultiGrasp, we must first sample clustered object configurations and then generate multi-object grasps directly on the object cluster. In contrast, SeqGrasp and SeqDiffuser do not require objects to be spatially close as they generate grasps sequentially based on previously successful ones. While this comparison is not entirely fair, we still believe comparing these two strategies offers valuable insights.

For training SeqDiffuser, we split SeqDataset into an 80% training set and a 20% test set, ensuring that no training objects were used in the experimental evaluation. The object point clouds are obtained by sampling 2048 points on the object mesh surfaces using Farthest Point Sampling (FPS).

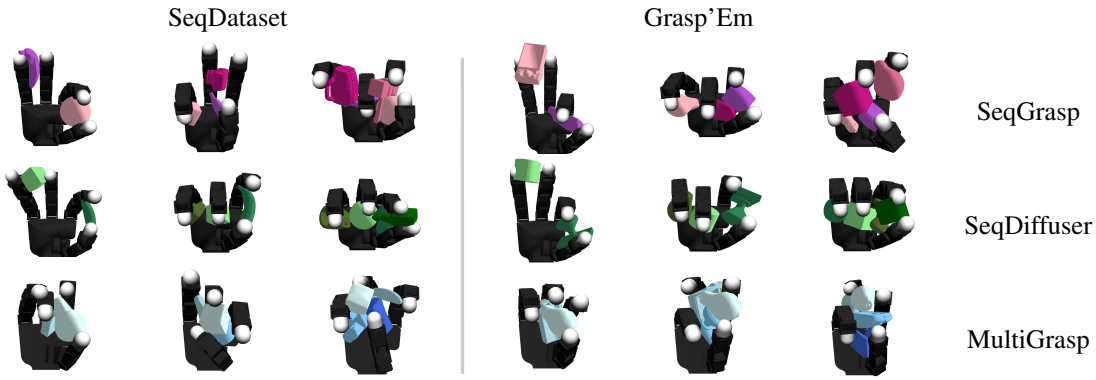


Fig. 5: **Qualitative results.** For SeqGrasp and SeqDiffuser, we only show consumed grasps, that is, when $\mathcal{OS} = \emptyset$. For SeqGrasp and SeqDiffuser, the grasp sequences are visually indicated by a color gradient, transitioning from lighter to darker shades. In contrast, for MultiGrasp, the color gradient is only used to differentiate the objects.

A. Simulation Experiments

We evaluated all generated grasps in the simulation experiments in Isaac Gym [39]. We used two object sets: (1) all eight objects from Grasp'Em [16] and (2) a random selection of eight validation objects from SeqDataset. We randomly generated ten four-object sequences for each object set, and, per object, we generated 256 grasps, resulting in 10,240 grasps per method.

We used the following metrics to assess the quality of the generated grasps:

1) **Success rate (SR) in percent**: The same success criteria as described in Section V.
2) **Maximum penetration depth (Pene.) in mm**: The maximum interpenetration distance between the hand and all grasped objects.
3) **Diversity (Div.) in radian**: Grasp diversity is determined by calculating the standard deviation of $\mathbf{g}$ across all successful grasps.
4) **Efficiency (Eff.) in second**: The computational time required to generate a batch of 256 grasps on an NVIDIA A100.

The quantitative results are presented in Table III, while Fig. 5 qualitatively illustrates a few grasps. The results demonstrate that SeqGrasp achieves the highest success rate and the lowest penetration depth when grasping two or more objects. MultiGrasp performs well for one- and two-object grasps, as it utilizes all of the hand's available DoF to grasp the objects. However, because MultiGrasp requires all objects to be initialized nearby, the success rate of the generated grasps is susceptible to the initial object placements. In contrast, SeqGrasp and SeqDiffuser do not suffer from this limitation.

We observe a significant performance drop when transitioning from three-object to four-object grasps across all methods. We hypothesize that this decline occurs because the three previously grasped objects occupy substantial space within the Allegro Hand, pushing the fourth object grasp to the limits of the hand's kinematic redundancy. Additionally, as the number of grasped objects increases, object-object interactions grow exponentially, making the task considerably more challenging, a finding also reported in [16]. Nevertheless, SeqGrasp demonstrates superior performance in scenarios involving three or more objects.

	SR ↑		Pene. ↓		Eff. ↓	Div. ↑
Method	**SData**	**G'Em**	**SData**	**G'Em**	**Avg.**	**Avg.**
MulG-1	**66.84**	**65.39**	**1.14**	**1.27**	600	0.284
SeqD-1	46.95	45.23	5.55	5.59	**0.8**	0.321
SeqG-1	50.04	40.78	1.73	2.16	900	**0.332**
MulG-2	**22.46**	16.48	2.30	2.83	750	0.347
SeqD-2	18.83	23.00	5.84	6.28	**0.8**	0.359
SeqG-2	21.21	**32.03**	**2.14**	**1.78**	900	**0.367**
MulG-3	10.78	3.55	3.39	4.04	900	0.340
SeqD-3	11.05	9.22	5.93	6.47	**0.8**	0.334
SeqG-3	**19.49**	**21.05**	**2.23**	**2.23**	900	**0.349**
MulG-4	0.90	0.47	5.17	6.27	1000	**0.329**
SeqD-4	**3.01**	1.68	6.18	6.69	**0.8**	0.293
SeqG-4	2.93	**5.04**	**2.70**	**2.62**	900	0.312

TABLE III: **Simulation results**. MulG, SeqG, SeqD, G'Em, and SData are short the MultiGrasp, SeqGrasp, SeqDiffuser, Grasp'Em, and SeqDataset, respectively. The $-i$ following the method name denotes the number of objects used for grasp generation. ↑(↓), the higher (lower), the better.

Another notable observation is that SeqDiffuser generates grasps with high penetration, which aligns with previous work on single object diffusion-based grasp sampling [5]. Still, SeqDiffuser is valuable because it generates grasps 750-1250 times faster than SeqGrasp and MultiGrasp.

B. Real-World Experiments

To address the final question, we evaluated the stability of the generated grasps on a real Allegro Hand mounted to a Franka Emika Panda robotic arm. For the evaluation, we 3D printed the same eight SeqDataset test objects from the simulation experiments. The printed objects are shown in Fig. 6. We applied a silicone gel coating to the object surfaces and attached an anti-slip grip to the palm to increase friction.

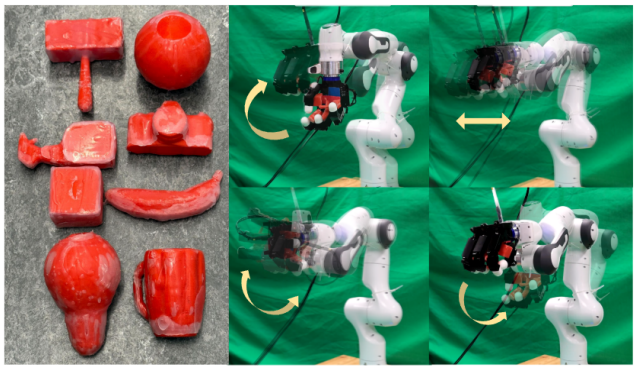


Fig. 6: Left: **Printed object set** with silicone gel coating over the surface. Right: **The stability test.**

Real exp.	SeqG-3	SeqG-4	SeqD-3	SeqD-4	MulG-3	MulG-4
Succ. trials	6/10	6/10	2/10	1/10	2/10	0/7
Succ. objects	24/30	33/40	20/30	21/40	11/30	12/28

TABLE IV: **Real-world experimental results.** Succ. trials mean successful trials and succ. objects mean how many objects were still grasped even if the trial failed.

We followed the procedure outlined in [28] to replicate the grasp on the real hardware. This procedure involves positioning each object as closely as possible with its generated pose and then closing the relevant joints of the hand to retain the objects. For grasps with severe penetration issues, we placed the object in the closest feasible real-world position. We only evaluated three- and four-object grasps. For MultiGrasp and SeqGrasp, we replicated the successful grasps with the lowest energy, while for SeqDiffuser, we randomly chose one of the generated grasps.

We evaluated the grasps with the grasp stability test shown in Fig. 6. In this test, the arm was first moved to a holding position with the palm facing downward. Next, the arm moved left-to-right, and then the last joint rotated $\pm 90°$. A trial is considered successful if all objects remain grasped.

The experimental results are presented in Table IV. Similar to the simulation results, SeqGrasp outperformed SeqDiffuser and MultiGrasp in both three-object and four-object grasping tasks, achieving an average success rate of 60%. The primary factor contributing to SeqDiffuser's low success rate (15%) was significant object interpenetration in the generated grasps, which meant that the closest physically feasible grasp replicable on the real hand differed substantially from the intended configuration. MultiGrasp, with an average success rate of (12%), mainly failed because the stability of the generated grasps relied on many object-object contacts, which are sensitive to minor object displacements. This resulted in frequent failures, particularly in four-object grasping trials where MultiGrasp achieved no successful trials and was even unsuccessful in finding solutions in three trials, which is why the last column in Table IV is seven and not ten. If we also count these as failed trials, the average success rate for MultiGrasp decreases to 10%.

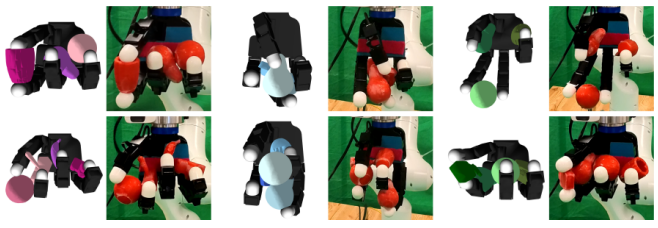


Fig. 7: **Grasp replication on real hardware.** First row: three-object grasp, second row: four-object grasp.

Our real-world experiments highlight the robustness of sequential grasping as a strategy for dexterous multi-object manipulation. By generating grasps iteratively and independently of object proximity, SeqGrasp and SeqDiffuser mitigate the challenges posed by object interactions, making it a more reliable and practical approach for real-world applications.

VIII. Limitations

Although both SeqGrasp and SeqDiffuser can produce high-quality and diverse grasps, the generated grasps represent only a subset of possible multi-object grasp solutions. Specifically, our solution relies on only one or two fingers for a grasp. In contrast, humans can flexibly adjust their fingers and distribute space within the hand to grasp objects of varying sizes, shapes, and weights. Moreover, our approach does not account for object sequencing, whereas humans may prefer a specific order to facilitate multi-object grasping. Addressing these limitations is an exciting future research direction.

IX. Conclusion

We proposed SeqGrasp, an algorithm for sequentially grasping multiple objects with a dexterous hand. SeqGrasp combines OSes and differentiable-force closure to generate stable grasps that maximize the hand's remaining DoF after each grasp. Using SeqGrasp, we constructed SeqDataset, currently the largest sequential grasping dataset, comprising 870K validated grasps across 509 diverse objects. This dataset enabled the training of SeqDiffuser, our diffusion-based sequential multi-object grasp sampler. The experimental evaluations demonstrated that SeqGrasp and SeqDiffuser outperformed the simultaneous multi-object grasping baseline MultiGrasp, achieving an 8.71%-43.33% higher average success rate. Moreover, SeqDiffuser proved to be 750-1250 times faster at generating grasps than SeqGrasp and MultiGrasp. In conclusion, this work demonstrated a fast and stable sequential multi-object grasp generation solution, which we hope can pave the way for more research in multi-object grasping.

References

[1] J. Bohg, A. Morales, T. Asfour, and D. Kragic, "Data-driven grasp synthesis—a survey," *IEEE Transactions on robotics*, vol. 30, no. 2, pp. 289–309, 2013.

[2] M. Li, K. Hang, D. Kragic, and A. Billard, "Dexterous grasping under shape uncertainty," *Robotics and Autonomous Systems*, vol. 75, pp. 352–364, 2016.

[3] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6446, p. eaat8414, 2019.

[4] R. Newbury, M. Gu, L. Chumbley, A. Mousavian, C. Eppner, J. Leitner, J. Bohg, A. Morales, T. Asfour, D. Kragic *et al.*, "Deep learning approaches to grasp synthesis: A review," *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3994–4015, 2023.

[5] Z. Weng, H. Lu, D. Kragic, and J. Lundell, "Dexdiffuser: Generating dexterous grasps with diffusion models," *IEEE Robotics and Automation Letters*, vol. 9, no. 12, pp. 11 834–11 840, 2024.

[6] J. Lu, H. Kang, H. Li, B. Liu, Y. Yang, Q. Huang, and G. Hua, "Ugg: Unified generative grasping," in *European Conference on Computer Vision*. Springer, 2024, pp. 414–433.

[7] J. Lundell, F. Verdoja, and V. Kyrki, "Ddgc: Generative deep dexterous grasping in clutter," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6899–6906, 2021.

[8] Z. Wei, Z. Xu, J. Guo, Y. Hou, C. Gao, Z. Cai, J. Luo, and L. Shao, "D(r,o) grasp: A unified representation of robot and object interaction for cross-embodiment dexterous grasping," *arXiv preprint arXiv:2410.01702*, 2024.

[9] Y. Xu, W. Wan, J. Zhang, H. Liu, Z. Shan, H. Shen, R. Wang, H. Geng, Y. Weng, J. Chen, T. Liu, L. Yi, and H. Wang, "Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 4737–4746.

[10] R. Wang, J. Zhang, J. Chen, Y. Xu, P. Li, T. Liu, and H. Wang, "Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 11 359–11 366.

[11] P. Li, T. Liu, Y. Li, Y. Geng, Y. Zhu, Y. Yang, and S. Huang, "Gendexgrasp: Generalizable dexterous grasping," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 8068–8074.

[12] J. Romero, H. Kjellström, and D. Kragic, "Hands in action: real-time 3d reconstruction of hands in interaction with objects," in *2010 IEEE International Conference on Robotics and Automation*. IEEE, 2010, pp. 458–463.

[13] D. Song, N. Kyriazis, I. Oikonomidis, C. Papazov, A. Argyros, D. Burschka, and D. Kragic, "Predicting human intention in visual observations of hand/object interactions," in *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2013, pp. 1608–1615.

[14] T. Feix, J. Romero, H.-B. Schmiedmayer, A. M. Dollar, and D. Kragic, "The grasp taxonomy of human grasp types," *IEEE Transactions on human-machine systems*, vol. 46, no. 1, pp. 66–77, 2015.

[15] M. Kokic, D. Kragic, and J. Bohg, "Learning task-oriented grasping from human activity datasets," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3352–3359, 2020.

[16] Y. Li, B. Liu, Y. Geng, P. Li, Y. Yang, Y. Zhu, T. Liu, and S. Huang, "Grasp multiple objects with one hand," *IEEE Robotics and Automation Letters*, vol. 9, no. 5, pp. 4027–4034, 2024.

[17] Y. Sun, E. Amatova, and T. Chen, "Multi-object grasping-types and taxonomy," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 777–783.

[18] A. Miller and P. Allen, "Graspit! a versatile simulator for robotic grasping," *IEEE Robotics & Automation Magazine*, vol. 11, no. 4, pp. 110–122, 2004.

[19] C. Rosales, R. Suárez, M. Gabiccini, and A. Bicchi, "On the synthesis of feasible and prehensile robotic grasps," in *IEEE International Conference on Robotics and Automation*. IEEE, 2012, pp. 550–556.

[20] M. T. Ciocarlie and P. K. Allen, "Hand posture subspaces for dexterous robotic grasping," *The International Journal of Robotics Research*, vol. 28, no. 7, pp. 851–867, 2009.

[21] T. Liu, Z. Liu, Z. Jiao, Y. Zhu, and S.-C. Zhu, "Synthesizing diverse and physically stable grasps with arbitrary hand structures using differentiable force closure estimator," *IEEE Robotics and Automation Letters*, vol. 7, no. 1, pp. 470–477, 2022.

[22] J. Ponce, S. Sullivan, J.-D. Boissonnat, and J.-P. Merlet, "On characterizing and computing three- and four-finger force-closure grasps of polyhedral objects," in *Proceedings IEEE International Conference on Robotics and Automation*, 1993, pp. 821–827 vol.2.

[23] J.-W. Li, H. Liu, and H.-G. Cai, "On computing three-finger force-closure grasps of 2-d and 3-d objects," *IEEE Transactions on Robotics and Automation*, vol. 19, no. 1, pp. 155–161, 2003.

[24] V. Mayer, Q. Feng, J. Deng, Y. Shi, Z. Chen, and A. Knoll, "Ffhnet: Generating multi-fingered robotic grasps for unknown objects in real-time," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 762–769.

[25] S. He, Z. Shangguan, K. Wang, Y. Gu, Y. Fu, Y. Fu, and D. Seita, "Sequential multi-object grasping with one dexterous hand," *arXiv preprint arXiv:2503.09078*, 2025.

[26] W. C. Agboh, J. Ichnowski, K. Goldberg, and M. R. Dogar, "Multi-object grasping in the plane," in *The International Symposium of Robotics Research*. Springer, 2022, pp. 222–238.

[27] T. Yonemaru, W. Wan, T. Nishimura, and K. Harada, "Learning to group and grasp multiple objects," *arXiv preprint arXiv:2502.08452*, 2025.

[28] K. Yao and A. Billard, "Exploiting kinematic redundancy for robotic grasping of multiple objects," *IEEE Transactions on Robotics*, vol. 39, no. 3, pp. 1982–2002, 2023.

[29] J. Chen, Y. Ke, and H. Wang, "Bodex: Scalable and efficient robotic dexterous grasp synthesis using bilevel optimization," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[30] H. Zhang, S. Christen, Z. Fan, O. Hilliges, and J. Song, "GraspXL: Generating grasping motions for diverse objects at scale," in *European Conference on Computer Vision (ECCV)*, 2024.

[31] Y. Liu, Y. Yang, Y. Wang, X. Wu, J. Wang, Y. Yao, S. Schwertfeger, S. Yang, W. Wang, J. Yu, X. He, and Y. Ma, "Realdex: towards human-like grasping for robotic dexterous hand," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, ser. IJCAI '24, 2024.

[32] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5745–5753.

[33] T. Iberall, "Opposition space as a structuring concept for the analysis of akilled hand movements," *Experimental brain research series*, vol. 15, pp. 158–173, 1986.

[34] J. B. Smeets, K. van der Kooij, and E. Brenner, "A review of grasping as the movements of digits in space," *Journal of neurophysiology*, vol. 122, no. 4, pp. 1578–1597, 2019.

[35] G. O. Roberts and R. L. Tweedie, "Exponential convergence of langevin distributions and their discrete approximations," *Bernoulli*, vol. 2, pp. 341–363, 1996.

[36] E. Rimon and J. Burdick, "On force and form closure for multiple finger grasps," in *Proceedings of IEEE International Conference on Robotics and Automation*, vol. 2. IEEE, 1996, pp. 1795–1800.

[37] G. Cotugno, K. Althoefer, and T. Nanayakkara, "The role of the thumb: Study of finger motion in grasping and reachability space in human and robotic hands," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 7, pp. 1061–1070, 2017.

[38] S. Prokudin, C. Lassner, and J. Romero, "Efficient learning on point clouds with basis point sets," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4332–4341.

[39] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa *et al.*, "Isaac gym: High performance gpu-based physics simulation for robot learning," *arXiv preprint arXiv:2108.10470*, 2021.