IEEE Robotics and Automation Letters (RAL) paper presented at
2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Grasp Multiple Objects with One Hand

Yuyang Li[1,2,3], Bo Liu[1], Yiran Geng[4], Puhao Li[2,3], Yaodong Yang[1], Yixin Zhu[1], Tengyu Liu[2,†], Siyuan Huang[2,†]

Fig. 1: **Our multi-object grasping method, *MultiGrasp*, drives the Shadow Hand to simultaneously grasp multiple Pokémons.**

***Abstract*— The intricate kinematics of the human hand enable simultaneous grasping and manipulation of multiple objects, essential for tasks such as object transfer and in-hand manipulation. Despite its significance, the domain of robotic multi-object grasping is relatively unexplored and presents notable challenges in kinematics, dynamics, and object configurations. This paper introduces *MultiGrasp*, a novel two-stage approach for multi-object grasping using a dexterous multi-fingered robotic hand on a tabletop. The process consists of (i) generating pre-grasp proposals and (ii) executing the grasp and lifting the objects. Our experimental focus is primarily on dual-object grasping, achieving a success rate of 44.13%, highlighting adaptability to new object configurations and tolerance for imprecise grasps. Additionally, the framework demonstrates the potential for grasping more than two objects at the cost of inference speed.**

I. Introduction

Infants, between 6 to 9 months of age, transition from using a rudimentary "grabbing" technique with their entire hand to a more refined "pincer grasp," involving only a subset of fingers [1]. This developmental progression underpins advanced object manipulation skills, including the ability to grasp multiple objects [2, 3]. Similarly, in robotics, significant advancements have been made in multi-fingered dexterous hands [4–9], enabling intricate grasping and in-hand manipulation tasks [10–15] and enhancing interaction capabilities for embodied intelligence.

However, the majority of existing research in robotic grasping focuses on single-object scenarios [4–8]. Common strategies often mirror the action of enveloping the object with the hand and squeezing the fingers towards it [6, 16], effectively reducing the use of sophisticated dexterous hands to mere parallel grippers. This overlooks the potential of their complex articulated structure and kinematic redundancy.

In this work, we delve into the less-explored domain of multi-object grasping, an intricate task that necessitates meticulous management of the hand's dexterous kinematics and dynamics. Our objective is to manipulate a multi-fingered dexterous hand to simultaneously grasp and lift multiple objects placed on a table. Distinct from single-object grasping, multi-object grasping demands independent force closure on each object. In this scenario, each object is a separate entity with no rigid interconnection, presenting unique challenges:

Diverse Configurations: Multi-object grasping encompasses a broad spectrum of object configurations influenced by varying geometries, combinations, and placements. This diversity is further compounded by the various hand configurations, necessitating the development of adaptable and flexible grasping strategies [3, 9].

Intricate Kinematics: The task of multi-object grasping requires using the full extent of the hand's workspace, as each object occupies a significant portion of it. Simple contacts via the palm or fingertips are insufficient. Instead, the entire length and sides of the fingers must be engaged [3, 9], necessitating a carefully configured grasping pose for effective force closure on each object while avoiding collisions.

Complex Dynamics: The traditional *enveloping and squeezing* approach, typically employed in single-object grasping, is inadequate for multi-object scenarios. Repositioning a finger to better grasp one object might jeopardize the grasp on another. Therefore, precise control and fine-tuning of the wrenches at each contact point become imperative.

To address these challenges, we introduce *MultiGrasp*, a computational framework devised for multi-object grasping. As depicted in Fig. 2, *MultiGrasp* begins by generating a pre-grasp pose for the given target objects. This pose represents a preliminary hand configuration, serving as a goal strategy for execution. Utilizing this strategy, the framework employs an execution policy to control the hand in picking up the objects. To demonstrate this process, we have constructed *Grasp'Em*, a large-scale synthetic dataset comprising 90k diverse multi-object grasps, utilizing the Shadow Hand. Moreover, we devise a new grasp generation model [17, 18], enabling the efficient generation of pre-grasp poses for novel object configurations. For grasp execution, we propose a dual-stage

† Corresponding emails: {liutengyu, syhuang}@bigai.ai.
[1] Institute for AI, Peking University
[2] State Key Lab of General AI, Beijing Institute for General AI
[3] Department of Automation, Tsinghua University
[4] Department of EECS, Peking University
See additional material on https://multigrasp.github.io

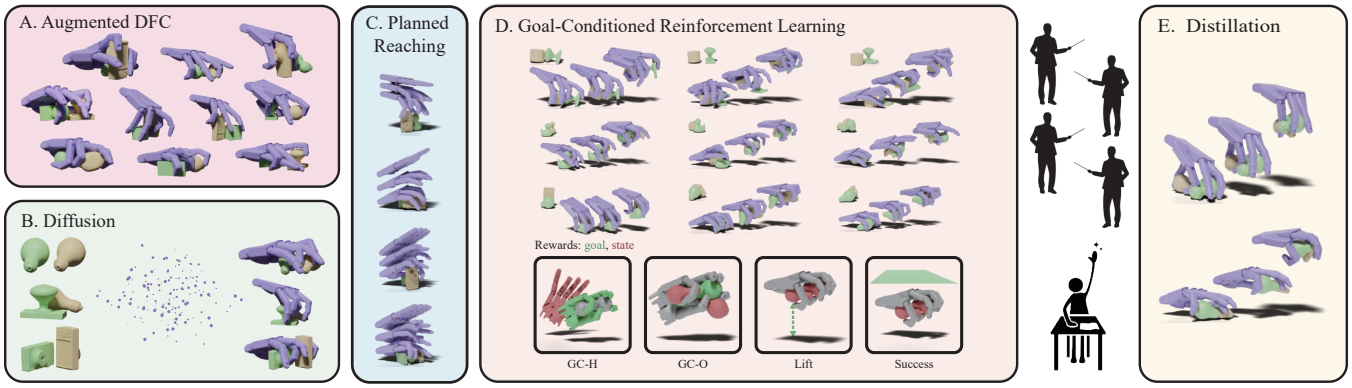


Fig. 2: **Overview of *MultiGrasp*.** The pre-grasp pose proposal module (A, B) generates an optimal hand pose to grasp the target objects. A motion planning module (C) then plans the reaching trajectory from a flat hand to the desired pose. For lifting, a suite of specialist Reinforcement Learning (RL) policies (D) are deployed, tailored for different object configurations. These policies are subsequently distilled (E) to develop a vision-based policy suitable for real-world application.

policy that integrates motion planning for reaching and a learned policy for lifting. In simulations, our method demonstrates a success rate of 44.13% in dual-object grasping with a Shadow Hand and shows scalability to handle more objects. Additionally, real-world experiments validate its practical effectiveness.

To summarize, the primary contributions of this work are: (i) the creation of *Grasp'Em*, a large-scale synthetic dataset explicitly curated for multi-object grasping research; (ii) the formulation of the first Goal-Conditioned Reinforcement Learning (GCRL) policy dedicated to the simultaneous grasping and lifting of multiple objects; (iii) the augmentation of the execution policy for enhanced adaptability to novel object configurations and imprecise pre-grasp poses, achieved through specialist distillation and curriculum learning; (iv) the comprehensive framework *MultiGrasp*, which advances current robotic systems to achieve multi-object grasping.

A. Related Work

Generating dexterous grasping: Conditioned on the target objects, the generation of grasping poses for dexterous hands presents a complex challenge due to intricate kinematics and physical constraints. Research in this domain primarily bifurcates into analytical and data-driven approaches [19].

Analytical methods have a long history, with early research focusing on algorithmic solutions for hand- and object-specific grasping poses [20–23]. GraspIt! [16] extended this to arbitrary hands and objects, albeit predominantly through a reach-and-squeeze strategy, limiting the diversity of grasping poses. Recent works have introduced generalized grasp quality metrics, such as force closure and Q_1, enabling faster and more adaptable grasp synthesis [7, 24, 25].

Conversely, data-driven approaches learn the distribution of feasible grasping poses [18, 26, 27], or utilize proxy representations such as contact points [28] and contact maps [4, 6, 29, 30], conditioned on the characteristics of the objects.

Contemporary research endeavors to reach human-level interaction capabilities, including functional grasping [31], generalizable grasping [6], and multi-object grasping [9].

Multi-object grasping: The goal of multi-object grasping is to identify optimal hand configurations for holding multiple objects simultaneously. Research in this area predominantly follows two distinct methodologies. The first approach emphasizes grasping a collection of simple objects, such as balls, bricks, or pencils, with a focus on efficiency in the grasping process. This method often relies on the contacts between objects for grasp stability [32–36]. It typically does not necessitate extensive kinematic redundancy, offering efficient grasping capabilities but at the cost of limited individual object manipulation. The second approach, however, utilizes the hand's kinematic redundancy by engaging different hand regions to grasp each object. This method allows for more detailed control over individual objects [9]. Our work is more in line with this second approach, where the focus is on maintaining the ability to maneuver each object independently while also enhancing overall grasp efficiency.

Reinforcement learning: Robotic operation in complex physical environments often presents significant challenges for analytical solutions, particularly due to noisy sensory inputs. In such scenarios, RL has become a favored choice for decision-making and control [37–40]. The emphasis of RL-based manipulation has largely been on single-object interactions [4, 5, 15, 41, 42]. These approaches tend to be inadequate for learning the intricate grasping techniques necessary for dexterous hands. For the complex task of multi-object grasping using dexterous hands where diverse object configurations pose additional challenges, conventional RL strategies are often insufficient. In our work, we employ GCRL [43] to develop robust lifting policies, accelerated by IsaacGym [44].

B. Overview of MultiGrasp

Formally, we consider a tabletop scenario populated with multiple objects, denoted as $\mathbf{O} = \{O_j\}_{j=1}^{N_o}$. Each object O_j is represented as a point cloud in $\mathbb{R}^{N\times 3}$, sampled from its surface $S(O_j)$. The objective is to identify a sequence of hand actions, $\mathcal{A} = \{a^t\}_{t=1}^{T}$, enabling the robotic hand to simultaneously grasp all objects. We primarily concentrate on scenarios where the objects are sufficiently proximate for

simultaneous grasping.

Fig. 2 illustrates the framework. For the given objects $\mathbf{O}$, a pre-grasp pose $H = (p, R, q)$ is proposed, encapsulating all targets; p denotes the hand's position, R its orientation, and q the joint angles. Two methods sample H: a detailed synthetic algorithm (Sec. II-A) and a fast generative model (Sec. II-B). The grasp execution involves following a trajectory to H and then lifting the objects with a learned policy (Sec. III).

II. Pre-Grasp Pose Generation

A. Preliminaries

In multi-object grasping, a well-crafted pre-grasp pose is essential to meet the static force-closure conditions, serving as a goal for dynamic grasp execution. Leveraging the Differentiable Force Closure (DFC) algorithm [7], we generate diverse and stable pre-grasp poses for multiple objects. The hand configuration H is derived from the Gibbs distribution $p(H|\mathbf{O}) = \frac{p(H,\mathbf{O})}{p(\mathbf{O})} \propto p(H,\mathbf{O}) \sim \frac{1}{Z} e^{-E(H,\mathbf{O})}$, where the energy function $E(H,\mathbf{O})$ integrates various terms:

$$\begin{aligned} E(H,\mathbf{O}) = &\sum_{j=1}^{N_o} \min_{x_j \subset S(H)} E_{\mathrm{FC}}(x_j, O_j) \\ &+ \lambda_{\mathrm{p}} E_{\mathrm{p}}(H,\mathbf{O}) + \lambda_{\mathrm{sp}} E_{\mathrm{sp}}(H) + \lambda_{\mathrm{q}} E_{\mathrm{q}}(H), \end{aligned} \tag{1}$$

where $E_{\mathrm{FC}}(x_j, O_j)$ computes the force-closure error for object O_j, with x_j representing the contact points on the hand that minimize this error. $E_{\mathrm{p}}(H,\mathbf{O})$ penalizes any penetration between the hand and each object. Regarding the hand's configuration, E_{sp} minimizes the self-collision of the hand, and $E_{\mathrm{q}}(H)$ penalizes deviations of joint angles from their limits. The weights $\lambda_{(\cdot)}$ are employed to balance these diverse energy components. A gradient-based approach supplemented by the Metropolis-Adjusted Langevin Algorithm (MALA) optimizes Eq. (1) to avoid suboptimal local minima. The optimization process is parallelized across multiple initial states for efficiency. A filtering step is further applied to eliminate cases exceeding a predefined threshold. Fig. 3 showcases the results of this synthesis for various object counts. For algorithmic details and force-closure estimation, we direct readers to [7].

B. Multi-object Grasp Generation

Using DFC for multi-object pre-grasp pose generation is computationally demanding. To bypass this intensive optimization, we employ diffusion models, a family of generative models recently shown to be effective for complex data [18, 45]. Following Denoising Diffusion Probabilistic Model (DDPM) [17], given object point clouds $\mathbf{O}$, the grasp pose $H = H^{(0)}$ is sampled through a denoising process:

$$p(H^{(0)}|\mathbf{O}) = \prod_{t=1}^{T} p(H^{(t-1)}|H^{(t)}, \mathbf{O}), \tag{2}$$

where

$$\begin{aligned} &p(H^{(t-1)}|H^{(t)}, \mathbf{O}) = \mathcal{N}\left(H^{(t-1)}; \mu^{(t-1)}, \Sigma^{(t-1)}\right), \\ &\mu^{(t-1)} = \mu_\theta(H^{(t)}, t, f_\theta(\mathbf{O})), \quad \Sigma^{(t-1)} = \Sigma(t). \end{aligned} \tag{3}$$

We adopt SceneDiffuser [18] to learn Eq. (3). This model takes object point clouds as input and proposes a pre-grasp pose. PointNet++ [46] is employed to extract feature vectors from each object's point cloud. These features, aggregated as $N_o \times N_{\mathrm{feat}}$, act as the object conditions $f_\theta(\mathbf{O})$. Cross-attention, computed with $H^{(t)}$ as queries and object conditions as keys and values, is utilized in each sampling step.

Differentiating features from distinct objects, we append a learnable embedding to each feature vector, with identical embeddings for the same object and different ones for varied objects. To support part-level interaction reasoning between finger links and objects, drawing inspiration from several studies [47–49], the hand configuration is represented by 31 keypoints on its links in Cartesian space, rather than joint angles in joint space. An optimization-based Inverse Kinematics (IK) solver derives joint angles from these keypoints.

For model training, we generated *Grasp'Em* using our synthesis algorithm. The dataset for multi-object grasping includes $\approx 90k$ synthetic pre-grasp poses, with a mix of $16.4k$ single- and $73.7k$ dual-object grasps featuring 8 objects (36 combinations) from YCB [50] and ContactDB [51]. Objects are rescaled so that multiple ones can be fitted in one hand. Object pairs are randomly placed on a table with stable positions and orientations.

While training, the data is preprocessed to align the palm direction (projected on the tabletop) by rotating the hand and objects around the z-axis, which simplifies learning by reducing one Degrees of Freedom (DoF). The model learns to generate palm-aligned hand poses, allowing control over the palm directions. This is done by rotating the target objects around the z-axis to align the desired palm direction with the alignment direction, and inverse the rotation on the generated hand pose to obtain the final hand poses (Fig. 4).

C. Multi-object Grasp Refinement

While the diffusion model shows potential in data generation, it occasionally results in imperfect grasps with penetration objects or insufficient contact. To rectify these shortcomings, we refine the hand configuration with optimization:

$$\begin{aligned} \min_H E_g(H,\mathbf{O}) = &E_{\mathrm{p}}(H,\mathbf{O}) \\ &- \frac{\lambda_c}{N_o|S(H)|} \sum_{j=1}^{N_o} \sum_{\substack{x \in S(H) \\ d(x,O_j) \leqslant \tau}} d(x, O_j), \end{aligned} \tag{4}$$

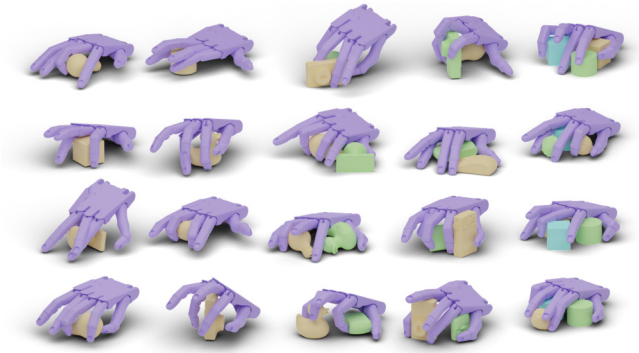


Fig. 3: **Synthetic grasps using the augtmented DFC.** From left to right: one (cols. 1-2), two (cols. 3-4), and three objects (col. 5).

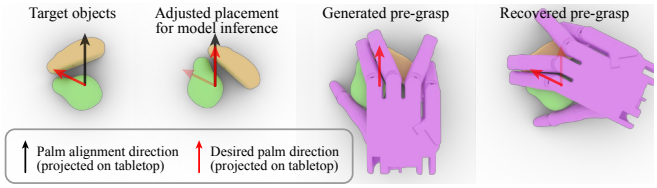


Fig. 4: **Generate grasps with desired palm direction.** Given the desired palm direction (red arrow), the objects are rotated around the z-axis to align with the desired palm direction (black arrow). The pre-grasp pose is generated on the adjusted objects and is rotated to recover the pre-grasp under the initial object placement.

where the first term is as defined in Eq. (1), and the second term guides floating fingers closer to the surfaces of nearby objects for contact. Here, $d(\cdot)$ is the distance from a potential contact point x on the hand to the surface of the object O_j. To achieve a coarse-to-fine refinement, we linearly decrease the threshold τ from 2.0mm to 1.0mm during optimization.

III. Multi-object Grasp Execution

We execute the grasp with two phases: reaching and lifting (see Fig. 2C-E). The reaching phase involves motion planning, whereas the lifting phase utilizes a learned RL policy.

Reaching: To plan a collision-free trajectory from the initial hand pose to the pre-grasp pose, we first compute their linear interpolation as a preliminary trajectory. This trajectory is further optimized to remove any hand-object penetrations at each timestep while ensuring a smooth temporal transition. Of note, while the grasp refinement (Sec. II-C) significantly reduces penetration, minor occurrences may still arise. However, the reaching phase could effectively compensate for such slight penetrations by causing minor object displacements. These disturbances will be accommodated by adapting the lifting policy (detailed in Sec. III-B).

Lifting: In real-world applications, conventional execution strategies like directly lifting from pre-grasp poses [8, 16], or squeezing fingers toward the closest object surface before lifting [6], often prove inadequate for multi-object grasping due to the complex nature of hand-object contacts. To address this, we employ GCRL for precise and adaptable control over the intricate dynamics involved in hand-object interactions.

A. Learning a Multi-Object Lifting Policy

For our lifting policy, we utilize Proximal Policy Optimization (PPO) [52] for learning within a simulated environment. Beginning from the pre-grasp pose of both the hands and objects, this policy controls the hand's pose and joint angles to lift all objects, guided by a reward:

$$r = \omega_{\text{lift}} r_{\text{lift}} + \omega_{\text{succ}} \mathbf{1}_{\text{succ}} + \omega_{\text{r}} r_{\text{r}} + \omega_{\text{q}} r_{\text{q}} + \omega_{\text{obj}} r_{\text{obj}}, \quad (5)$$

$r_{\text{lift}} = \min_j h_j$ provides a dense reward based on object elevation, where h_j indicates the height of the j-th object). A success bonus (r_{succ}) is awarded for elevating all objects above 15cm. We also find it beneficial to include rewards for maintaining goal hand rotation (r_{r}), joint angles (r_{q}), and object positions relative to the hand (r_{obj}). These rewards are visualized in Fig. 2D.

The observation space for our policy is outlined in Tab. I(a). It incorporates the current and goal states of both the hand and objects, along with their respective residues. Geometric information is sourced from hand and object features, extracted from point clouds using a pre-trained PointNet [53], all within the palm's coordinate frame. To optimize sample efficiency, the policy is trained across 512 parallel environments in IsaacGym [44], each with a unique pre-grasp pose from *Grasp'Em*'s training set. We periodically refresh these poses to ensure comprehensive dataset coverage. The entire training process spans 8000 iterations, approximately taking 4 hours on a single NVIDIA A100.

In real-world scenarios, only the hand's state and the point cloud captured by depth cameras are typically available. To adapt the policy for such conditions, we distill it into a vision-based version using DAgger [54]. This involves replacing object features and states in the observation space (marked with $*$ in Tab. I) with PointNet features of the hand-object scene, captured by three RGB-D cameras. The point cloud is also re-sampled to ensure even point distribution.

B. Learning a Generalist from Specialists

We find that the lifting policy's efficacy varies with object configurations and pre-grasp poses; *e.g.*, when handling spheres versus cylinders or objects placed in various poses. To develop a generalist policy capable of adapting to various scenarios, we draw inspiration from previous works [5, 38, 55]. While retaining the settings from Sec. III, we categorize the grasp data into distinct clusters based on object combinations and their placements relative to the palm. A dedicated specialist policy is trained for each cluster to master the grasps within, which provides demonstrations in distilling the vision-based generalist policy.

TABLE I: **Observation and action definitions of our lifting policy.** The policy's 12-dimensional *state* encompasses the position, orientation (represented by XYZ Euler angles), linear velocity, and angular velocity. The *residue* is calculated as the difference between the current and goal values. Markers * and ** indicate elements specific to the state-based and vision-based policies, respectively.

(a) Observation space.

Observation	Dimensions
Hand joint angles, velocities, forces	22×3
Hand fingertip wrenches	5×6
Hand base state	12
Last actions	24
Hand joint angle goals and residues	22 + 22
Hand orientation goal and residue	3 + 3
Hand point cloud feature	256
Object O_i state *	12
Object O_i pose goal and residue *	6 + 6
Object O_i point cloud feature *	256
Scene point cloud feature **	256

(b) Action space.

Action	Dimensions
Hand joint angle targets (actuated)	18
Hand base wrench	6

C. Adapting to Imprecise Pre-Grasp Poses

Even with refinement, the generated grasps may be imperfect, and the execution of the planned reaching trajectories might inadvertently cause collisions that displace objects. Policies trained exclusively on high-quality synthetic data are ill-suited for such scenarios, and training directly on these poses is ineffective due to the distorted goal representation practically. To address this challenge, we implement a structured learning curriculum. First, we generate imprecise poses by using the trained DDPM on randomly positioned objects. Critically, we capture the states of the hand and objects at the end of the reaching phase, including moments where fingers may have shifted the objects. Next, the training regimen is divided into three phases: the first phase focuses on synthetic pre-grasp poses; the second phase introduces these imprecise poses, encompassing both synthetic and generated instances where objects have been displaced; the final phase predominantly incorporates generated data with moved objects.

Of note, while our approach builds upon existing methodologies [4, 5, 18, 26, 27, 41, 42], our concentration on dexterous robotic hands differentiates it from the referenced studies. The complexities involved in designing and controlling high-DoF dexterous robotic hands are distinct and more intricate. This necessitates substantial modifications and enhancements to the established methods, aligning them with the unique demands of dexterous manipulation.

IV. Simulations and Experiments

In accordance with the evaluation protocol detailed in Sec. IV-A, we conduct a comprehensive validation of the proposed method. Quantitative results pertaining to the pre-grasp poses are elaborated in Sec. IV-B. The execution phase of the method is assessed in Sec. IV-C, and ablation studies that examine the impact of various components of our approach are presented in Sec. IV-D. To demonstrate the method's ability to generalize, we provide results from simulations involving multiple objects and from real-world tests with dual-object setups (Sec. IV-E). We finally discuss observed failure cases and their implications.

A. Evaluation Protocols

We generate random table-top object arrangements for each multi-object combination, following the same procedure as in our dataset (Sec. II-B). For each arrangement, we generate pre-grasp poses, discarding those with significant penetration or insufficient contact ratio. In the execution phase, a grasp is deemed successful if all objects are elevated above 10cm. The policy is evaluated on 512 unique poses, with five trials per pose, to compute the average success rate.

For dual-object grasping, the 8 objects in our dataset result in $C_8^1 + C_8^2 = 36$ unique combinations. We designate 8 pairs as unseen combinations, covering all objects, to evaluate our framework's generalization capabilities. Further, 6 combinations with 3 out-of-domain objects are introduced to challenge the framework with unfamiliar geometries. To cluster the grasps into bins (Sec. III-B), we first group them by object combination and then further subdivide them based on the direction line connecting object centers in the palm's frame, resulting in 6 bins. We observe that grasping becomes particularly challenging when objects are aligned parallel to the forearm, primarily due to difficulties in achieving secure force closure. To mitigate this, we utilize the palm-alignment strategy from Sec. II-B to limit palm directions in generated pre-grasp poses, thereby avoiding challenging configurations.

For comparison, we select three dexterous grasping methods, despite none being directly comparable to our task. We first contrast with a vision-only variant of our framework. GenDexGrasp [6], initially designed for single-object grasping using contact maps, is extended to multi-object scenarios with *Grasp'Em* and evaluated using our execution framework. IBS-Grasping [8] learns single-object grasps considering the Intersection Bisector Surface (IBS) with RL. We adapt it by integrating *Grasp'Em* as initial off-policy demonstrations and modifying the reward structure for multi-object grasping.

The baselines are tested with pre-grasp poses generated for the same set of unseen object placements. We assess the effectiveness of these poses by training a single state-based policy for each method. For our approach and GenDexGrasp, we implement a state-based policy as outlined in Sec. III. For IBS-Grasping, we adhere to its original approach, directly lifting objects from the generated pre-grasp poses.

B. Pre-Grasp Proposals

We assess the quality of synthesized or generated static poses using four metrics:

1) **$\mathbf{Q_1}$ Metric**: This metric evaluates the largest radius of an origin-centered 6D sphere within the resistive wrench space [56], reflecting grasp stability. We compute Q_1 for each object following Liu *et al.* [57]. For multi-object grasps, we report the minimum Q_1 across all objects.
2) **Penetration (PN)**: This is the maximal intersection depth (in mm) between the hand, objects, and the table. Although a physically plausible grasp should be collision-free, slight penetrations are sometimes unavoidable due to numerical precision in computations.
3) **Grasp Diversity (Div)**: This measures the average variance of joint angles (in deg) across all grasp samples, indicating the range of grasping strategies. A versatile grasp generator should offer a variety of strategies.
4) **Inference Time (Time)**: Measured in seconds on a single RTX 3090Ti GPU, this metric assesses efficiency. We test our method with a batch size of 256, GenDexGrasp with 32 to fit GPU memory capacity, and IBS-Grasping with 1 following its original implementation.

Quantitative results in Tab. II highlight our method's proficiency in producing viable multi-object grasps efficiently. Synthetic grasps (`Syn`) provide high quality but are time-consuming, while generated grasps (`Gen`) balance quality and speed, maintaining satisfactory success rates. Remarkably, the generative model shows generalization to new object placements (`Gen-Pl`), combinations (`Gen-Com`), and geometries (`Gen-Geo`), although trained on only 8 objects. This is attributed to the diverse object configurations in the training data. In comparison, GenDexGrasp [6] (`GDG`) offers comparable quality but less diversity and tends to penetrate

TABLE II: **Quantitative evaluations on our method and adapted baselines [6, 8].** Abbreviations are defined in Sec. IV-B. Success rates for specialist and generalist approaches are denoted separately, divided by "/". Notably, the performance of the baseline marked with * was evaluated using its original implementation [8] within the PyBullet environment.

Setting	Pre-Grasp Pose				Execution
	$Q_1 \uparrow$	PN $\downarrow$	Div $\rightarrow$	Time $\downarrow$	Succ (%)
Syn-Pl	**0.30**	**1.64**	8.54	$\approx$ 1000	**68.34 / 44.13**
Syn-Com	**0.31**	1.78	8.58	$\approx$ 1000	26.73
Syn-Geo	**0.31**	1.81	8.66	$\approx$ 1000	22.55
Gen-Pl	0.29	**1.67**	9.24	**12.28**	40.20 / **30.24**
Gen-Com	0.25	2.65	8.45	12.83	23.32
Gen-Geo	0.29	**1.54**	9.14	**12.08**	15.65
Syn-Pl-V	**0.30**	**1.64**	8.54	$\approx$ 1000	0.59
Gen-Pl-V	0.29	**1.67**	9.24	**12.28**	0.04
GDG-Pl [6]	0.27	27.75	4.23	25.67 (32 samples)	25.55
GDG-Com [6]	**0.33**	25.05	3.91		14.57
GDG-Geo [6]	0.27	19.12	3.98		6.91
IBS-Pl [8]	0.23	36.29	7.09	4.45 (1 sample)	12.20*
IBS-Com [8]	0.22	35.92	7.40		13.09*
IBS-Geo [8]	0.23	36.32	7.36		14.18*

the table, indicating its limitation in tabletop scenarios. IBS-Grasping [8] (IBS) experiences more severe penetration, possibly due to its unstable stochastic policy.

C. Execution Policy

The execution phase is assessed in a simulator, with success defined as lifting all objects over 10 cm. Each grasp undergoes five trials, and the average success rates are reported in Tab. II. Synthetic grasps (Syn) yield the highest success rate. For unseen placements, state-based specialists achieve an average success rate of **68.34%**, while the vision-based generalist policy attains **44.13%**. Despite their lower quality, generated grasps (Gen) still maintain reasonable success. However, success rates decrease for out-of-domain combinations (-Com) and geometries (-Geo).

During distillation, student policies show an approximate 30% decrease in success rates compared to their teachers, mainly due to two factors: (i) The vision-based generalist policy lacks direct access to object states, relying instead on scene observation through cameras, which leads to less precise observations. (ii) While each specialist focuses on a narrow set of similar grasping strategies, the generalist policy must adapt to a much wider range of pre-grasp poses and strategies. Although the performance drop, this policy is practical for real-world scenarios.

Comparatively, the baseline methods (GDG and IBS) demonstrate poorer performance in both grasp generation and execution across all generalization levels. Their limited grasp representations lack detailed kinematic information necessary for multi-object grasping. Furthermore, deep penetration during grasping leads to more collisions in reaching, resulting in significant object displacement. We also evaluated two visual baselines (-Vis), where a vision-based policy is learned directly from scratch using RL. The coarse observation in the beginning of learning causes the failure in learning hand control, which highlights the importance of the distillation process.

D. Ablations

We conduct ablation studies to dissect the components of our framework, focusing on grasp generation (Tab. IIIa) and execution policy (Tab. IIIb). For brevity, we evaluate generated grasps on unseen object placements (-Pl), reporting only specialist performances which typically mirror student policy outcomes. We select four specialists covering all object placement bins to evaluate different execution techniques. These studies yield significant insights:

Pre-grasp pose reasoning: As discussed in Sec. II-B, our model generates 31 keypoints to represent the hand. In contrast, directly generating joint angles (Joint Angles), a common approach in literature [4, 5, 26], resulted in decreased performance. This suggests that reasoning in Cartesian space, as done in our keypoint-based method, captures part-level hand-object interactions more effectively than joint space reasoning. Additional ablations validate the impact of object embedding attachment (w/o Obj Embd) and the grasp refinement process (w/o Refinement).

From specialists to generalists: Our approach includes clustering the training data by object combinations and placements, training a specialist policy for each cluster. We explore variations including clustering solely by object combination (w/o Spe-Pl) or placement (w/o Spe-Com), and training a single policy for all grasps (w/o Spe). As per Tab. IIIb, training specialists for specific placements and combinations marginally enhances expert demonstration quality. However, using specialists in isolation leads to reduced performance. While Xu *et al.* [4] highlighted the advantages of multiple specialists for varied object geometries, our dataset's limited diversity from 8 mostly convex objects might not necessitate distinct specialist policies. The true potential of specialists may emerge with more diverse object configurations.

Training adaptive policy: Tab. IIIc emphasize the significance of our execution policy's design. The near-total failure of lifts without an RL policy (w/o RL) underlines its critical role. Omitting observations and rewards for maintaining pre-grasp pose (w/o Goal) lowers the learning efficiency. Training exclusively on synthetic data without adapting to imprecise poses (w/o Adpt) or adapting without a structured curriculum (w/o Curr), results in

TABLE III: **Ablation studies.** Abbreviations are explained in Secs. IV-B and IV-D. *Evaluated on a subset of objects for efficiency.

(a) Generative model

Setting	$Q_1 \uparrow$	PN $\downarrow$	Succ (%)
Ours	**0.29**	**1.67**	**40.20**
Joint Angles	0.18	5.50	19.31
w/o Obj Embd	0.27	**1.38**	37.21
w/o Refinement	0.29	7.68	16.24

(b) Specialist settings in RL

Setting	Succ (%)
Ours	**40.20**
w/o Spe-Pl	29.09
w/o Spe-Com	26.23
w/o Spe	37.34

(c) Training settings in RL

Setting	Succ (%)
Ours*	**45.25**
w/o RL	1.37
w/o Goal	16.79
w/o Adpt	25.05
w/o Curr	24.88

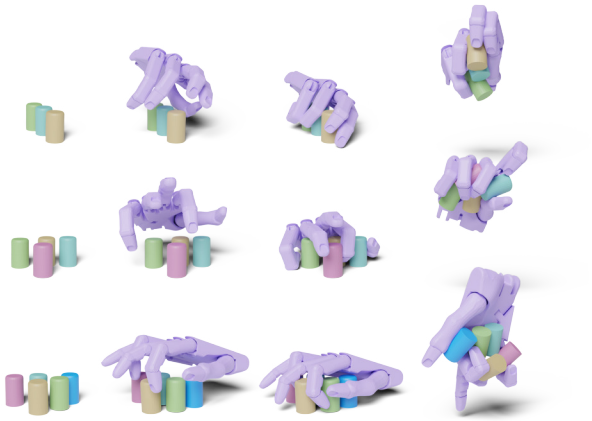


Fig. 5: **Our framework supports grasping varying amounts (3-5) of objects.** Each row demonstrates the object placement and the execution process for different numbers of objects.

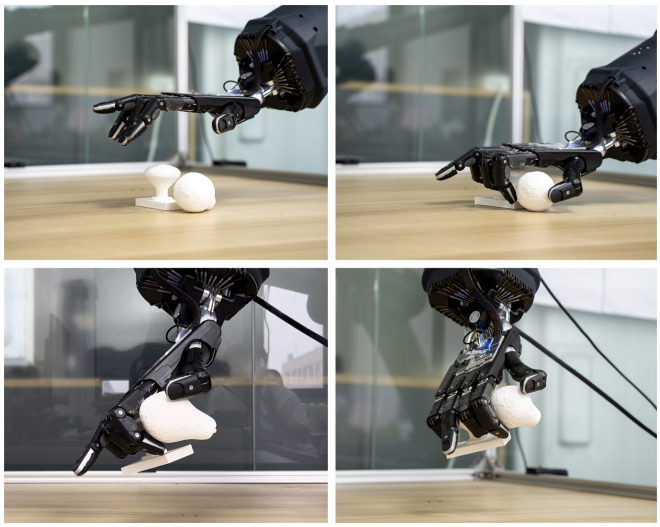


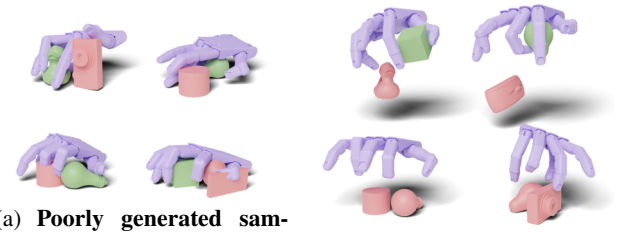
Fig. 6: **Experiment with a Shadow Hand.** The sequence shows the phases of reaching, grasping, and lifting during the execution process.

suboptimal outcomes.

E. Additional Results

Grasping more objects: We test our framework's capability to grasp more than two objects. By synthesizing grasps for small cylinders and learning a state-based execution policy for each case, we assess the system's performance, as depicted in Fig. 5. The complexity of both pre-grasp proposal and execution escalates with the increase in object count. In scenarios with four objects, the hand's kinematic redundancy is increasingly utilized, and inter-object contact becomes essential for maintaining stability. When handling five objects, the hand needs to invert its pose to scoop the objects. These results underscore the scalability and probe the boundary of our approach.

Real-world experiment: Our method is further tested with a physical robot, using a Shadow Hand attached to a UR10e arm. Given the complexity of the task, we precompute execution trajectories in a simulation and then replicate them on the robot. As illustrated in Fig. 6, our approach successfully enables the robot to pick up two objects from a table, demonstrating its potential for real-world applications.



(a) **Poorly generated samples:** Missing force-closure (top) and penetration (bottom).

(b) **Execution failures:** Dropping objects (top) and lift failures (bottom).

Fig. 7: **Common grasp failures in generation and execution.** Objects marked in red indicate unsuccessful grasping attempts.

Failure cases: The primary causes are inadequate pre-grasp proposal quality and imprecise control during execution. Fig. 7 presents typical examples, including (i) poorly generated grasps, and (ii) objects dropped or not lifted.

V. Discussions

We introduced *MultiGrasp*, a novel framework designed for the simultaneous grasping of multiple objects using multi-fingered robotic hands. Demonstrating the capability to manage various object counts, our approach underscores its potential for real-world implementations. This initiative lays the groundwork for future advancements in multi-object grasp planning and execution, aiming to boost the efficiency and adaptability of robotic grasping in real-world settings. Although our focus primarily lies on concurrent multi-object grasping, the feasibility of sequential object grasping, adopting an anthropomorphic strategy, is also acknowledged. Accordingly, we have refined our grasp synthesis algorithm to support efficient sequential grasping across different environmental conditions; please refer to the code repository.

The current study opens avenues for further exploration, particularly in narrowing the sim2real gap regarding perception and object dynamics, which could improve overall performance and enable more thorough real-world testing. Vision and proprioception, limited by significant occlusion and complex contact scenarios, hint at the potential enhancements tactile sensing could bring. Future research directions include exploring bimanual multi-object manipulation, where one hand stabilizes objects while the other performs precise tasks, such as insertion. Furthermore, our goal extends to equipping robots with capabilities for in-hand manipulation and tool use. Advancements in these domains aim to close the gap between robotic and human capabilities, empowering robots to undertake more sophisticated tasks and interactions.

Acknowledgments

The authors would like to thank Qianxu Wang (PKU), Zihang Zhao (PKU), Junfeng Ni (THU), Nan Jiang (PKU), and Wanlin Li (BIGAI) for their helpful assistance. This work was supported in part by the National Science and Technology Major Project under Grant 2022ZD0114900 and in part by Beijing Nova Program.

References

[1] C. Von Hofsten, "An action perspective on motor development," *Trends in Cognitive Sciences*, vol. 8, no. 6, pp. 266–272, 2004.

[2] H. Moll and M. Tomasello, "Infant cognition," *Current Biology*, vol. 20, no. 20, pp. R872–R875, 2010.

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

[4] Y. Xu, W. Wan, J. Zhang, H. Liu, Z. Shan, H. Shen, R. Wang, H. Geng, Y. Weng, J. Chen *et al.*, "Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy," in *CVPR*, 2023, pp. 4737–4746.

[5] W. Wan, H. Geng, Y. Liu, Z. Shan, Y. Yang, L. Yi, and H. Wang, "Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning," in *ICCV*, 2023.

[6] P. Li, T. Liu, Y. Li, Y. Zhu, Y. Yang, and S. Huang, "Gendexgrasp: Generalizable dexterous grasping," in *ICRA*, 2023.

[7] 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," *RA-L*, vol. 7, no. 1, pp. 470–477, 2021.

[8] Q. She, R. Hu, J. Xu, M. Liu, K. Xu, and H. Huang, "Learning high-dof reaching-and-grasping via dynamic representation of gripper-object interaction," *TOG*, vol. 41, no. 4, pp. 1–14, 2022.

[9] K. Yao and A. Billard, "Exploiting kinematic redundancy for robotic grasping of multiple objects," *T-RO*, 2023.

[10] M. T. Mason and J. K. Salisbury Jr, *Robot hands and the mechanics of manipulation*. The MIT Press, Cambridge, MA, 1985.

[11] N. C. Dafle, A. Rodriguez, R. Paolini, B. Tang, S. S. Srinivasa, M. Erdmann, M. T. Mason, I. Lundberg, H. Staab, and T. Fuhlbrigge, "Extrinsic dexterity: In-hand manipulation with external forces," in *ICRA*, 2014.

[12] D. Rus, "In-hand dexterous manipulation of piecewise-smooth 3-d objects," *IJRR*, vol. 18, no. 4, pp. 355–381, 1999.

[13] B. Calli, K. Srinivasan, A. Morgan, and A. M. Dollar, "Learning modes of within-hand manipulation," in *ICRA*, 2018.

[14] O. M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray *et al.*, "Learning dexterous in-hand manipulation," *IJRR*, vol. 39, no. 1, pp. 3–20, 2020.

[15] T. Chen, J. Xu, and P. Agrawal, "A system for general in-hand object re-orientation," in *CoRL*, 2021.

[16] A. T. Miller and P. K. Allen, "Graspit! a versatile simulator for robotic grasping," *RA-M*, vol. 11, no. 4, pp. 110–122, 2004.

[17] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, 2020.

[18] S. Huang, Z. Wang, P. Li, B. Jia, T. Liu, Y. Zhu, W. Liang, and S.-C. Zhu, "Diffusion-based generation, optimization, and planning in 3d scenes," in *CVPR*, 2023.

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

[20] 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 *ICRA*, 1993.

[21] J. Ponce, S. Sullivan, A. Sudsang, J.-D. Boissonnat, and J.-P. Merlet, "On computing four-finger equilibrium and force-closure grasps of polyhedral objects," *IJRR*, vol. 16, no. 1, pp. 11–35, 1997.

[22] 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.

[23] C. Eppner, A. Mousavian, and D. Fox, "Acronym: A large-scale grasp dataset based on simulation," in *ICRA*, 2021.

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

[25] D. Turpin, L. Wang, E. Heiden, Y.-C. Chen, M. Macklin, S. Tsogkas, S. Dickinson, and A. Garg, "Grasp'd: Differentiable contact-rich grasp synthesis for multi-fingered hands," in *ECCV*, 2022.

[26] H. Jiang, S. Liu, J. Wang, and X. Wang, "Hand-object contact consistency reasoning for human grasps generation," in *ICCV*, 2021.

[27] J. Lundell, E. Corona, T. N. Le, F. Verdoja, P. Weinzaepfel, G. Rogez, F. Moreno-Noguer, and V. Kyrki, "Multi-fingan: Generative coarse-to-fine sampling of multi-finger grasps," in *ICRA*, 2021.

[28] L. Shao, F. Ferreira, M. Jorda, V. Nambiar, J. Luo, E. Solowjow, J. A. Ojea, O. Khatib, and J. Bohg, "Unigrasp: Learning a unified model to grasp with multifingered robotic hands," *RA-L*, 2020.

[29] S. Brahmbhatt, A. Handa, J. Hays, and D. Fox, "Contactgrasp: Functional multi-finger grasp synthesis from contact," in *IROS*, 2019.

[30] P. Grady, C. Tang, C. D. Twigg, M. Vo, S. Brahmbhatt, and C. C. Kemp, "Contactopt: Optimizing contact to improve grasps," in *CVPR*, 2021.

[31] A. Agarwal, S. Uppal, K. Shaw, and D. Pathak, "Dexterous functional grasping," in *CoRL*, 2023.

[32] T. Yamada and H. Yamamoto, "Static grasp stability analysis of multiple spatial objects," *Journal of Control Science and Engineering*, vol. 3, pp. 118–139, 2015.

[33] B. Donald, L. Gariepy, and D. Rus, "Distributed manipulation of multiple objects using ropes," in *ICRA*, 2000.

[34] K. Harada and M. Kaneko, "Kinematics and internal force in grasping multiple objects," in *IROS*, 1998.

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

[36] Y. Sun, E. Amatova, and T. Chen, "Multi-object grasping-types and taxonomy," in *ICRA*, 2022.

[37] J. Hwangbo, J. Lee, A. Dosovitskiy, D. Bellicoso, V. Tsounis, V. Koltun, and M. Hutter, "Learning agile and dynamic motor skills for legged robots," *Science Robotics*, vol. 4, no. 26, p. eaau5872, 2019.

[38] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," *Science Robotics*, vol. 5, no. 47, p. eabc5986, 2020.

[39] T. Miki, J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning robust perceptive locomotion for quadrupedal robots in the wild," *Science Robotics*, vol. 7, no. 62, p. eabk2822, 2022.

[40] F. Jenelten, J. He, F. Farshidian, and M. Hutter, "Dtc: Deep tracking control," *Science Robotics*, vol. 9, no. 86, p. eadh5401, 2024.

[41] Y. Geng, B. An, H. Geng, Y. Chen, Y. Yang, and H. Dong, "Rlafford: End-to-end affordance learning for robotic manipulation," in *ICRA*, 2023.

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

[43] M. Liu, M. Zhu, and W. Zhang, "Goal-conditioned reinforcement learning: Problems and solutions," in *IJCAI*, 2022.

[44] 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.

[45] J. Urain, N. Funk, J. Peters, and G. Chalvatzaki, "Se (3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion," in *ICRA*, 2023.

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

[47] Y. Zhang, M. J. Black, and S. Tang, "We are more than our joints: Predicting how 3d bodies move," in *CVPR*, 2021.

[48] A. Simeonov, Y. Du, A. Tagliasacchi, J. B. Tenenbaum, A. Rodriguez, P. Agrawal, and V. Sitzmann, "Neural descriptor fields: Se (3)-equivariant object representations for manipulation," in *ICRA*, 2022.

[49] D. Seita, Y. Wang, S. J. Shetty, E. Y. Li, Z. Erickson, and D. Held, "Toolflownet: Robotic manipulation with tools via predicting tool flow from point clouds," in *CoRL*, 2023.

[50] B. Calli, A. Singh, J. Bruce, A. Walsman, K. Konolige, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Yale-cmu-berkeley dataset for robotic manipulation research," *IJRR*, vol. 36, no. 3, pp. 261–268, 2017.

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

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

[53] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *CVPR*, 2017.

[54] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *International Conference on Artificial Intelligence and Statistics*, 2011.

[55] D. Chen, B. Zhou, V. Koltun, and P. Krähenbühl, "Learning by cheating," in *CoRL*, 2020.

[56] C. Ferrari, J. F. Canny *et al.*, "Planning optimal grasps." in *ICRA*, 1992.

[57] M. Liu, Z. Pan, K. Xu, K. Ganguly, and D. Manocha, "Deep differentiable grasp planner for high-dof grippers," in *RSS*, 2020.