Grasp Multiple Objects with One Hand

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

https://multigrasp.github.io

Fig. 1: **The proposed multi-object grasping method *MultiGrasp* drives the Shadow Hand to catch multiple Pokémons simultaneously.**

***Abstract*— The human hand's complex kinematics allow for simultaneous grasping and manipulation of multiple objects, essential for tasks like object transfer and in-hand manipulation. Despite its importance, robotic multi-object grasping remains underexplored and presents challenges in kinematics, dynamics, and object configurations. This paper introduces *MultiGrasp*, a two-stage method for multi-object grasping on a tabletop with a multi-finger dexterous hand. It involves (i) generating pre-grasp proposals and (ii) executing the grasp and lifting the objects. Experimental results primarily focus on dual-object grasping and report a 44.13% success rate, showcasing adaptability to unseen object configurations and imprecise grasps. The framework also demonstrates the capability to grasp more than two objects, albeit at a reduced inference speed.**

I. Introduction

Infants aged 6-9 months transition from "grabbing" objects with their entire hand to "pincer grasp" using only a subset of fingers [1]. This developmental milestone lays the groundwork for advanced object manipulation, including multi-object grasping [2, 3]. In comparison, the field of robotics has seen significant advances in multi-fingered dexterous hands [4–10]. These robotic hands enable complex grasping and in-hand manipulation [11–17], enhancing interaction with the environment for embodied intelligence.

However, most of the existing work on robotic grasping is geared toward single-object grasping [4–9]. Some approaches often employ similar strategies that involve enveloping the hand around the object and squeezing the fingers towards it [6, 18]. Such a paradigm essentially treats dexterous robotic hands as if they were parallel grippers, thereby significantly underestimating and underutilizing the inherent potential offered by their highly articulated structure and kinematic redundancy.

In this work, we extend beyond single-object grasping to explore multi-object grasping, a complex yet under-studied task requiring careful management of dexterous kinematics and dynamics. Given multiple objects on a table, our goal is to control a multi-fingered dexterous hand to grasp and lift them simultaneously. While similar to single-object grasping at first glance, the key difference lies in the need for independent force closure on each object. Unlike single-object grasping where the object is a single entity, multi-object grasping deals with objects lacking a rigid connection among them, introducing unique challenges, including:

Diverse Configurations: Multi-object grasping involves a wide array of object configurations, influenced by their differing geometries, combinations, and placements. The diversity is further compounded by the various hand configurations, requiring the development of adaptable and versatile grasping strategies [3, 10].

Intricate Kinematics: In multi-object grasping, each object takes up a considerable part of the hand's workspace. Simple contacts via the palm or fingertips are inadequate, and the entire length and sides of the fingers must be utilized [3, 10]. This requires carefully configuring the grasping pose for force closure on each object while avoiding collisions.

Complex Dynamics: In multi-object grasping, the traditional *enveloping and squeezing* strategy for single-object grasping is insufficient. Repositioning a finger towards one object could compromise support for another. Therefore, precise control and adjustment of the wrenches at each contact point become crucial.

To tackle these challenges, we present *MultiGrasp*, a computational framework for multi-object grasping shown in Fig. 2. *MultiGrasp* first generates a pre-grasp pose, followed by an execution policy to pick up the object. We construct *Grasp'Em*, a large-scale synthetic dataset of 90k diverse multi-object grasps using the Shadow Hand. A grasp generation model based on [19, 20] produces pre-grasp poses for unseen object configurations. For grasp execution, we propose a two-stage policy that combines motion planning for reaching and a learned policy for lifting. We incorporate

† Corresponding emails: {liutengyu, syhuang}@bigai.ai.
[1] National Key Laboratory of General Artificial Intelligence, Beijing Institute for General Artificial Intelligence (BIGAI). [2] Department of Automation, Tsinghua University (THU). [3] Institute for Artificial Intelligence, Peking University (PKU). [4] School of Electronics Engineering and Computer Science, PKU.

techniques to enhance the lifting policy's generalization and adaptability for unseen object configurations and imprecise pre-grasp poses. Our method achieves a 44.13% success rate in simulation for dual-object grasping with a Shadow Hand and scales to more objects. Real-world experiments confirm its applicability.

In summary, the contributions of this work are fourfold: (i) the introduction of *Grasp'Em*, a large-scale synthetic dataset tailored for multi-object grasping research; (ii) the development of the first Goal-Conditioned Reinforcement Learning (GCRL) policy for concurrent grasping and lifting of multiple objects from a table; (iii) the enhancement of the execution policy for better adaptability to unseen object configurations and imprecise pre-grasp poses, achieved via specialist distillation and curriculum learning; (iv) a comprehensive framework, *MultiGrasp*, that extends existing robotic systems toward robust, accurate multi-object grasping.

A. Related Work

Generating Dexterous Grasping: Generating grasping poses for dexterous robotic hands, specifically conditioned on the target objects, is complex due to the intricate kinematics and physical constraints involved. Research in this area is primarily bifurcated into two paradigms: analytical approaches and data-driven methodologies [21].

Analytical methods have a long-standing history, with early research focusing on algorithmic solutions for hand- and object-specific grasping poses [22–24]. GraspIt! [18] extended this to arbitrary hands and objects but mainly through a reach-and-squeeze strategy, limiting the variety of grasping poses. Recent works have introduced generalized grasp quality metrics like force closure and Q_1 for quicker and more adaptable grasp synthesis [7, 25, 26]. Conversely, data-driven methods utilize models to capture the distribution of grasping poses [20, 27–30] or their proxy representations like contact points [31] and contact maps [4, 6, 32, 33], conditioning on object characteristics.

Recent studies aim for human-level interaction capabilities, including functional grasping [9], generalizable grasping [6], and multi-object grasping [10].

Multi-Object Grasping: Multi-object grasping aims to find optimal configurations for holding multiple objects in one hand. Existing research mainly follows two distinct approaches. The first focuses on grasping a handful of simple objects like balls, bricks, or pencils for grasp efficiency. It usually leverages inter-object contacts for grasp stability [34–39], and does not require much kinematic redundancy. Although efficient in grasping, this limits individual object manipulation. The second approach leverages the hand's kinematic redundancy to grasp each object with different hand regions, allowing nuanced control over individual objects [10]. Our work aligns more with the second approach, aiming to maintain individual object maneuverability while boosting grasp efficiency.

Reinforcement Learning (RL): Robots often operate in complex physical environments, making analytical solutions challenging due to noisy sensory input. RL is commonly used for decision-making and control in these cases [4, 5, 16, 40, 41]. As a specialized form, GCRL [42] focuses on skill acquisition for predefined objectives, but faces challenges in goal generalization and sparse reward handling [42]. To aid RL in robotics, IsaacGym [43] offers a GPU-accelerated simulation environment ideal for parallelized and computationally demanding agent training.

In multi-object grasping, naive approaches fall short due to diverse object configurations. We use GCRL for secure lifting policies, accelerated by IsaacGym, and introduce techniques for improved generalization and dense rewards.

B. Overview of MultiGrasp

To clarify the task, we consider a tabletop environment 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 goal is to find a sequence of hand actions, $\mathcal{A} = \{a^t\}_{t=1}^{T}$, to control the robotic hand to grasp all objects. We focus on cases where objects are close enough for simultaneous grasping.

As shown in Fig. 2, the pipeline is divided into stages. Given $\mathbf{O}$, the first stage (Fig. 2A-B) proposes a pre-grasp pose $H = (p, R, q)$ that encapsulates all objects. p denotes the hand's position, R its orientation, and q the joint angles. We have two solutions to sample H from conditional distribution $H \sim p(H|\mathbf{O})$: a synthetic algorithm for high-quality but slower synthesis (Sec. II-A), and a faster generative model that compromises quality (Sec. II-B). Then, the pose is executed (Fig. 2C-E) with an optimization-based motion plan to guide the hand to the pre-grasp pose, followed by an RL policy for lifting the objects (Sec. III).

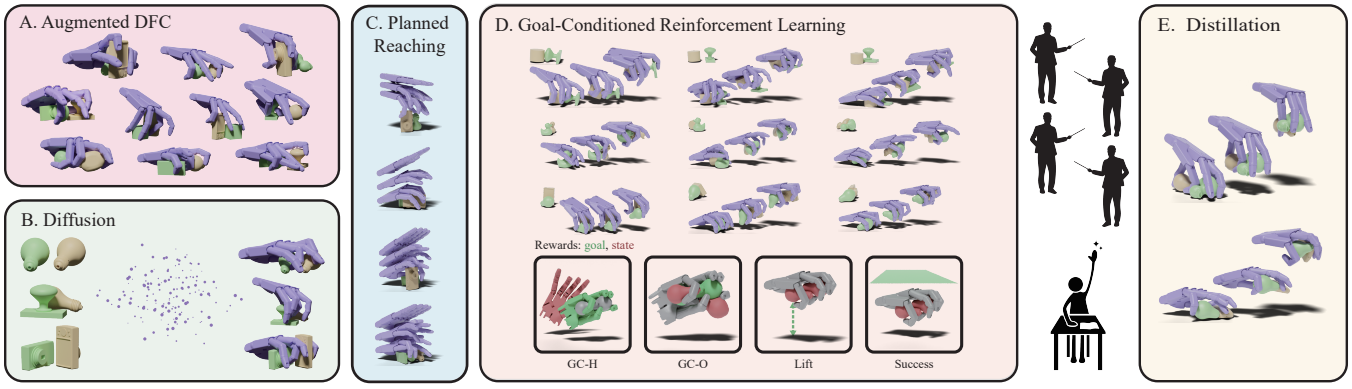


Fig. 2: **The pipeline of *MultiGrasp*.**

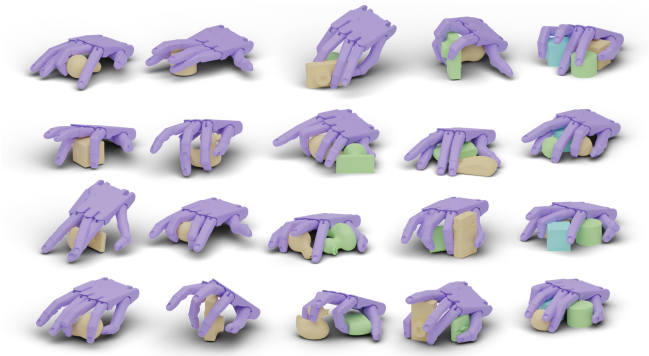


Fig. 3: **Synthetic grasps of the augtmented DFC.** Columns 1-2: single object. Columns 3-4: duel objects. Column 5: triple objects.

II. Multi-object Grasp Pose Generation

A. Preliminaries

Multi-object grasping necessitates a carefully crafted pre-grasp pose. By augmenting the Differentiable Force Closure (DFC) algorithm [7], we generate diverse and stable pre-grasp poses for multiple objects. The hand configuration H is sampled 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 $E(H,\mathbf{O})$ aggregates various energy terms for grasping:

$$E(H,\mathbf{O}) = \sum_{j=1}^{N_o} \min_{x_j \subset S(H)} E_{\text{FC}}(x_j, O_j) + \lambda_{\text{pen}} E_{\text{pen}}(H,\mathbf{O}) + \lambda_{\text{q}} E_{\text{q}}(H), \quad (1)$$

where $E_{\text{FC}}(x_j, O_j)$ calculates the force-closure error for object O_j, with x_j as the hand's contact points that minimize this error. $E_{\text{pen}}(H,\mathbf{O})$ penalizes object-hand penetration, and $E_{\text{q}}(H)$ penalizes joint angles exceeding their limits. We employ gradient-based optimization on Eq. (1), adjusted by Metropolis-Adjusted Langevin Algorithm (MALA) to evade poor local minima. For efficiency, we parallelize optimization from multiple initial states, discarding those energies exceeding acceptance thresholds. For algorithmic details and force-closure estimation, we refer to [7].

B. Multi-object Grasp Generation

Using DFC for multi-object pre-grasp pose generation on the fly is computationally intensive, particularly in diverse real-world scenarios. To expedite this, we employ a conditional diffusion model, sidestepping the cumbersome MALA synthesis. In line with Denoising Diffusion Probabilistic Model (DDPM) [19], given the object point clouds $\mathbf{O}$, the grasp pose $H = H^{(0)}$ is sampled through a denoising process:

$$\begin{aligned} &p(H^{(0)}|\mathbf{O}) = \prod_{t=1}^{T} p(H^{(t-1)}|H^{(t)},\mathbf{O}), \\ &p(H^{(t-1)}|H^{(t)},\mathbf{O}) = \mathcal{N}\left(H^{(t-1)}; \mu^{(t-1)}, \Sigma^{(t-1)}\right), \\ &\mu^{(t-1)} = \mu_\theta(H^{(t)}, t, f_\theta(\mathbf{O})), \quad \Sigma^{(t-1)} = \Sigma(t). \end{aligned} \quad (2)$$

We utilize SceneDiffuser [20] to learn Eq. (2). The model takes as input object point clouds and proposes a pre-grasp pose. To capture local object geometry crucial for grasping [44], PointNet++ [45] is employed to extract N_{feat} feature vectors from each object's point cloud. These $N_o \times N_{\text{feat}}$ features are concatenated to form object conditions $f_\theta(\mathbf{O})$. During each sampling step, cross-attention is computed with $H^{(t)}$ as queries and the object conditions as keys and values.

To differentiate features from various objects, we append a learnable embedding to each feature vector. Identical embeddings are used for the same object instance, while distinct ones are used for different objects. To facilitate part-level interaction reasoning between finger links and objects in Cartesian space, we represent the hand configuration using 31 keypoints rather than joint angles. An optimization-based Inverse Kinematics (IK) solver is employed to derive joint angles from these keypoints.

To train the model, we use our synthesis algorithm to create *Grasp'Em*, a large-scale synthetic dataset for multi-object grasping. The dataset comprises $\approx 90k$ synthetic pre-grasp poses, including $16.4k$ single- and $73.7k$ dual-object grasps, featuring 8 diverse objects from YCB [46] and ContactDB [47]. Objects are rescaled and randomly placed on a table. The augmented DFC algorithm runs for 7,500 optimization steps, with the first 6,000 adjusted by MALA. Large-scale synthesis for two objects from 8192 initial proposals takes about 1.25 hours on a single NVIDIA A100. Fig. 3 displays synthesis results for varying object counts.

C. Multi-object Grasp Refinement

Although the diffusion model offers promise, it sometimes produces imprecise grasps, such as hand penetration into objects or lack of finger contact. To address these issues, we optimize the hand configuration to minimize penetration and promote close contact. The optimization is formulated:

$$\min_H E_g(H,\mathbf{O}) = E_{\text{pen}}(H,\mathbf{O}) - \frac{\lambda_c}{N_o|S(H)|} \sum_{j=1}^{N_o} \sum_{\substack{x \in S(H) \\ d(x,O_j) \leqslant \tau}} d(x, O_j), \quad (3)$$

where the penetration energy is as defined in Eq. (1). The second term aims to guide floating fingers toward nearby object surfaces, where $d(\cdot)$ is the distance from the contact point x to the object surface O_j. The threshold τ linearly decreases from 2.0mm to 1.0mm for coarse-to-fine refinement.

III. Multi-object Grasp Execution

The execution of the proposed grasping is divided into a reaching and a lifting phase, as depicted in Fig. 2C-E. The reaching phase is solved using a motion planner, while the lifting phase employs a learned RL policy.

Reaching: To plan a collision-free trajectory from the initial to the pre-grasp pose, we start with linear interpolation as an initial guess and then optimize to eliminate penetration while ensuring temporal smoothness.

Lifting: Conventional execution strategies, such as squeezing fingers and lifting from pre-grasp poses [8, 18], or squeezing fingers toward the nearest object surface [6], often fail due to task complexity. To achieve precise, adaptive control, we employ GCRL to address the intricate dynamics of object interactions.

A. Learning a Multi-Object Lifting Policy

To manage the complex dynamics of lifting objects, we employ Proximal Policy Optimization (PPO) [48] to learn a lifting policy in simulation. Starting from the pre-grasp pose of both hands and objects, the policy aims to control hand pose and joint angles to lift *all* objects, with reward defined:

$$r = \omega_{\text{lift}} r_{\text{lift}} + \omega_{\text{succ}} \mathbf{1}_{\text{succ}} + \omega_{\text{r}} r_{\text{r}} + \omega_{\text{q}} r_{\text{q}} + \omega_{\text{obj}} r_{\text{obj}}. \quad (4)$$

To promote successful lifting, we offer a dense reward based on the object elevation $r_{\text{lift}} = \min_j h_j$, where h_j is the height of the j-th object. A bonus r_{succ} is given for lifting all objects above 15cm. Empirically, rewards for maintaining goal states for hand and object positions enhance learning. Specifically, we linearly penalize deviations from goal states in hand orientation (r_{r}), joint angles (r_{q}), and object poses in hand coordinates (r_{obj}). Fig. 2D visualizes these rewards.

Following Chen *et al*. [16], we enrich the policy's observations with current and goal states of the hand and objects, as well as their residues. Geometrical cues are provided by the hand and object features extracted from point clouds using a pretrained PointNet [49]. These features and states are computed in hand coordinates for consistent observations during lifting. To enhance sample efficiency, we train the policy across 512 parallel IsaacGym environments [43], each initialized with a unique pre-grasp pose, and periodically update these poses for broader training coverage.

Importantly, in real-world applications, only the hand's state information and depth camera-captured point cloud are available. Therefore, we distill the policy into a vision-based version using DAgger [50], replacing object features and states in the observation with scene point cloud features.

B. Learning a Generalist from Specialists

We find that the lifting policy varies depending on object configurations and pre-grasp poses, such as spheres versus cylinders or parallel versus perpendicular placements to the forearm. To create a generalist policy adaptable to diverse scenarios, we maintain the settings from Sec. III but empirically cluster the grasp data into bins based on object combinations and placements. We then train a specialist policy for each bin. During the distillation of the vision-based policy, expert demonstrations for each grasp are sourced from the corresponding specialist policy.

C. Adapting to Imprecise Pre-Grasp Poses

Even after refinement, generated grasps may still have penetration and lack of contact. Additionally, executing the planned trajectories can lead to collisions and objects being moved. Policies trained solely on high-quality synthetic data are ill-suited for these imprecise scenarios, and direct training on such poses is ineffective due to the misleading goal.

To tackle this, we employ a structured learning curriculum. We generate imprecise poses using the trained DDPM on random object placements. Importantly, we record the hand and object states at the final frame of the reaching trajectory for both synthetic and generated grasps, capturing instances where objects are touched and moved by fingers. The first training period focuses solely on synthetic pre-grasp poses. In the second period, we introduce these imprecise poses from both synthetic and generated data with objects moved. The final period primarily includes generated data with objects moved.

IV. Simulations and Experiments

Following the evaluation protocol outlined in Sec. IV-A, we rigorously validate the proposed method. We present quantitative results for pre-grasp poses in Sec. IV-B, execution in Sec. IV-C, and ablation studies in Sec. IV-D. The method's generalization is showcased through simulations for multiple objects (Sec. IV-E) and real-world dual-object tests (Sec. IV-F). Failure cases are discussed in Sec. IV-G.

A. Evaluation Protocols

We generate random table-top object placements for each multi-object combination, following the same procedure as in our dataset (Sec. II-B). Pre-grasp poses are generated for each placement, and those with severe penetration or insufficient contact ratio are discarded. A grasp is deemed successful if all objects are lifted above 10cm. We test the policy on 512 poses, each with five attempts, and calculate the average success rate.

For dual-object grasping, our dataset's eight objects yield $C_8^1 + C_8^2 = 36$ unique combinations. We designate eight pairs as unseen combinations to assess generalization, ensuring all eight objects are represented. Additionally, we introduce six combinations with three out-of-domain objects to further test generalization. The direction of the line connecting object centers in the hand's local coordinate frame is used to cluster grasps into bins, with a specialist policy trained for each bin.

Baselines: While no methods directly align with our task, we adopt two recent dexterous grasping approaches for comparison. *GenDexGrasp* [6] uses contact maps for single-object grasping, which we extend to multi-object scenarios using *Grasp'Em* and test with our execution pipeline. *Intersection Bisector Surface (IBS)-Grasping* [8] learns single-object grasps with observation on the IBS using RL. We adapt it by using *Grasp'Em* as initial off-policy demonstrations and then training with a modified reward for grasping multiple objects.

Protocols: For baselines, we evaluate their performance using pre-grasp poses generated on the same set of unseen object placements. We assess the effectiveness of these pre-grasp poses by training a single state-based policy for each method. Specifically, for our approach and GenDexGrasp, we train one state-based policy as described in Sec. III. In the case of IBS-Grasping, we retain its original execution policy, which involves lifting directly from the pre-grasp pose once the policy predicts a stop action.

B. Pre-Grasp Proposals

We evaluate the quality of synthesized or generated static grasping poses using four metrics: (1) $\mathbf{Q_1}$ **Metric:** Measures the largest origin-centered 6D sphere radius in the resistive wrench space [51]. For multi-object grasps, we report the

minimum generalized Q_1 [52] among all objects. (2) **Penetration (PN):** The maximum intersection depth (in mm) between the hand and all objects. A physically plausible grasp should minimize this metric. (3) **Grasp Diversity (Div):** The average variance of joint angles (in deg) across all grasp samples, indicating the diversity of grasping strategies. (4) **Inference Time:** Measured in seconds on a single RTX 3090Ti GPU with a batch size of 256, evaluating efficiency.

Quantitative results in Tab. I demonstrate the efficacy of our methods in generating viable multi-object grasps within reasonable time constraints. While synthetic grasps (Syn-Pl) offer the highest quality but are time-intensive, generated grasps (Gen) offer a trade-off between quality and inference time, yet still achieve acceptable success rates. Notably, the generative model effectively generalizes to unseen object placements (Gen-Pl), combinations (Gen-Com), and geometries (Gen-Geo), despite being trained on only 8 objects. This robust performance is attributed to the diversity in object configurations used during training.

In contrast, Tab. Ib shows GenDexGrasp [6] (GDG) offers similar grasp quality but lacks diversity and often penetrates the table, as it was not designed for tabletop grasping. IBS-Grasping [8] (IBS) has more severe penetration issues, likely due to its unstable stochastic policy.

C. Execution Policy

Tabs. Ia and Ib reports execution results in the last columns. Synthetic grasps on unseen object placements (Syn-Pl) achieve the highest success rate. Generated grasps (Gen) of lower quality still maintain a reasonable success rate. The success rates decrease for out-of-domain object combinations (Gen-Com) and geometries (Gen-Geo). In distillation, student policies exhibit slightly lower success rates compared to their teachers but are viable for real-world applications.

In comparison, baseline methods perform worse in execution. The main reason is that the representation for grasping they use contains insufficient and coarse information for careful kinematics management in multi-object grasping. Besides, deep penetration causes more collisions in reaching that severely perturb the objects.

TABLE I: **Quantitative results.** Specialist and generalist success rates are separated by "/". *Same execution method as "w/o Spe" in Tab. IIb. **Evaluated in PyBullet; otherwise in Isaac Gym.

(a) Quantitative evaluations on our method.

Setting	Pre-Grasp Pose				Execution
	$Q_1 \uparrow$	PN $\downarrow$	Div $\rightarrow$	Time $\downarrow$	Succ (%)
Syn-Pl	**0.30**	1.64	8.54	> 1200	68.34 / **44.13**
Gen-Pl	0.29	1.67	9.24	**12.29**	*40.20* / **30.24**
Gen-Com	0.25	2.65	8.45	**12.29**	- / 23.32
Gen-Geo	0.29	1.54	9.14	**12.29**	- / 15.65

(b) Comparisons with baseline models [6, 8]

Method	$Q_1 \uparrow$	PN $\downarrow$	Div $\uparrow$	Succ (%)
Ours-1*	**0.29**	**1.67**	**9.24**	**37.34**
GDG [6]	0.27	27.75	4.23	25.55
IBS [8]	0.23	36.29	7.09	12.20**

D. Ablations

We conduct ablation studies on our pipeline, focusing on grasp generation (Tab. IIa) and execution policy (Tab. IIb). For simplicity, we evaluate only generated grasps on unseen object placements and report specialist performances, which generally correlate positively with student policy outcomes. We limit our tests to four randomly selected specialists, covering all four object placement bins, to ablate other execution techniques. These results offer key insights:

Reasoning Interactions in Cartesian Space: Inspired by Zhang *et al*. [53], our model generates 31 keypoints on the hand and uses IK to obtain the grasp. Directly generating joint angles led to a performance drop. This is likely because generating joint angles requires the model to reason in the robot's joint space, whereas our keypoint approach reasons in Cartesian space, capturing part-level interactions more effectively. Additionally, the generated keypoints adhere closely to kinematic constraints, with average IK errors below $1\mathrm{mm}$ per keypoint.

From Specialists to Generalists: Tab. IIb reveals that using specialists for similar object placements and combinations marginally improves expert demonstration quality. However, using specialists individually leads to performance declines. While Xu *et al*. [4] showed the benefits of multiple specialists for diverse object geometries, our dataset's limited diversity—36 combinations from 8 mostly convex objects—may be sufficient for a single teacher policy. The effectiveness of specialists might become apparent with more diverse object configurations and better clustering.

Training Adaptive Policy: Tab. IIc underscores the importance of our design choices in the execution policy. First, the near-complete failure of lifting without an RL policy (w/o RL) highlights its necessity. Second, omitting observations and rewards for maintaining the pre-grasp pose (w/o Goal) complicates learning by enlarging the search space. Finally, training solely on synthetic data without adaptation to imprecise poses (w/o Adpt.), or directly training on these poses without a structured curriculum (w/o Curr.), leads to suboptimal performance.

TABLE II: **Ablation Studies.** *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 Pos.	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) Other RL designs.

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

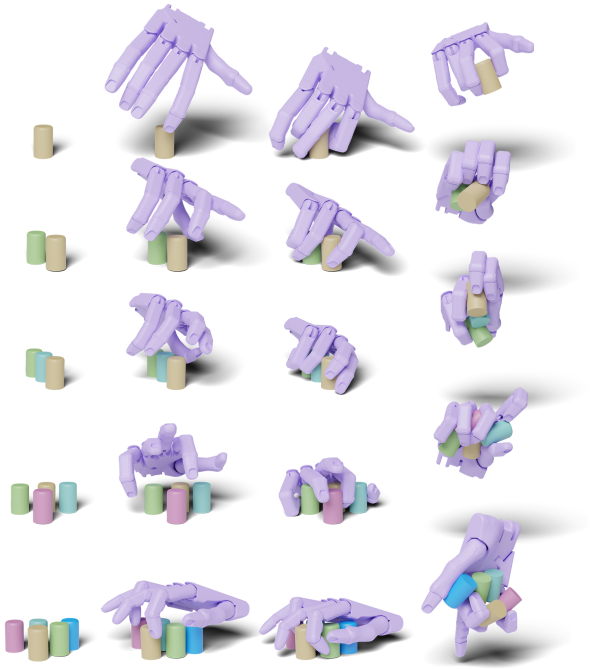


Fig. 4: **Our pipeline supports grasping different amounts (1-5) of objects.** Each row depicts object placement and execution for varying object counts.

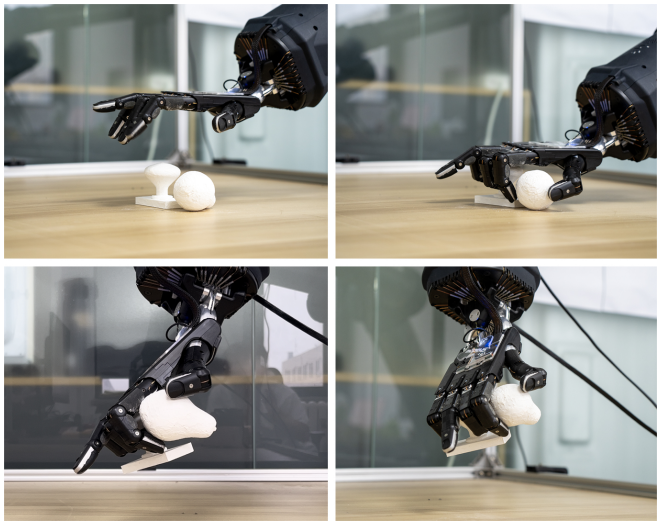


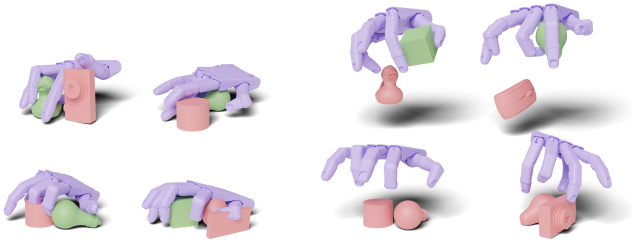
Fig. 5: **Executing a grasp with a Shadow Hand in the real world.** Pictures above show reaching, grasping (top), and lifting (bottom).

E. Grasping More Objects

We evaluate our pipeline's ability to grasp varying numbers of objects, specifically 1-5 cylinders. Grasps and corresponding state-based execution policies are synthesized for each case, as shown in Fig. 4. As the object count increases, both synthesis and execution become more challenging. For four objects, object-object contact becomes crucial for stability. With five objects, the hand must invert to scoop them up due to kinematic constraints. These results highlight our method's scalability and performance boundary.

F. Real-World Experiment

We validate our method with a Shadow Hand mounted on a UR10e arm in the real world. Due to the task's complexity, we precompute execution trajectories in simulation and implement them on the physical robot. After hand-eye calibration, the hand and arm are jointly controlled along the trajectory. As shown in Fig. 5, our method successfully enables the robot to grasp two objects from a table, showcasing its applicability to real-world robotic systems.



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

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

Fig. 6: **Common grasp failures in generation and execution.**

G. Failure Cases

We show examples of typical failure cases in our pipeline in Fig. 6. Main failures come from (1) bad grasp generation samples and (2) failure to lift the objects or drop the objects in the air in execution.

V. Discussions

This paper introduces *MultiGrasp*, a comprehensive framework for simultaneous multi-object grasping using multi-finger hands. Our approach demonstrates scalability to varying object quantities and feasibility for real-world deployment. This work paves the way for further advancements in multi-object grasp planning and execution, aiming to empower robots with versatile and efficient grasping capabilities in the real world. Although we focus mainly on simultaneously grasping multiple objects, we also envision the potential of another practical and anthropomorphic approach that involves grasping objects sequentially. We plan to extend our framework to support efficient sequential object grasping, enhancing its versatility.

Additionally, future exploration includes bimanual multi-object manipulation, where one hand holds objects while the other inserts them. This approach relaxes geometrical constraints and opens new avenues for versatile object interactions. Furthermore, we aim to equip robots with in-hand manipulation and tool usage abilities, broadening their utility in real-world scenarios. These advances will enable robots to perform more sophisticated tasks and interactions, bridging the gap between robots and humans.

Acknowledgement: The authors thank Qianxu Wang (PKU), Zihang Zhao (PKU), Junfeng Ni (THU), Nan Jiang (PKU), and Wanlin Li (BIGAI) for their helpful discussions and assistance in models, experiments, and training. This work is supported in part by the National Key R&D Program of China (2022ZD0114900), the Beijing Municipal Science & Technology Commission (Z221100003422004), and the 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. 1

[2] H. Moll and M. Tomasello, "Infant cognition," *Current Biology*, vol. 20, no. 20, pp. R872–R875, 2010. 1

[3] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6446, p. eaat8414, 2019. 1

[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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 4737–4746. 1, 2, 5

[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 *International Conference on Computer Vision (ICCV)*, 2023. 1, 2

[6] P. Li, T. Liu, Y. Li, Y. Zhu, Y. Yang, and S. Huang, "Gendexgrasp: Generalizable dexterous grasping," in *International Conference on Robotics and Automation (ICRA)*, 2023. 1, 2, 3, 4, 5

[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," *IEEE Robotics and Automation Letters (RA-L)*, vol. 7, no. 1, pp. 470–477, 2021. 1, 2, 3

[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," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, jul 2022. [Online]. Available: https://doi.org/10.1145/3528223.3530091 1, 3, 4, 5

[9] A. Agarwal, S. Uppal, K. Shaw, and D. Pathak, "Dexterous functional grasping," in *Conference on Robot Learning (CoRL)*, 2023. 1, 2

[10] K. Yao and A. Billard, "Exploiting kinematic redundancy for robotic grasping of multiple objects," *Transactions on Robotics (T-RO)*, 2023. 1, 2

[11] M. T. Mason and J. K. Salisbury Jr, *Robot hands and the mechanics of manipulation*. The MIT Press, Cambridge, MA, 1985. 1

[12] 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 *International Conference on Robotics and Automation (ICRA)*, 2014. 1

[13] D. Rus, "In-hand dexterous manipulation of piecewise-smooth 3-d objects," *International Journal of Robotics Research (IJRR)*, vol. 18, no. 4, pp. 355–381, 1999. 1

[14] B. Calli, K. Srinivasan, A. Morgan, and A. M. Dollar, "Learning modes of within-hand manipulation," in *International Conference on Robotics and Automation (ICRA)*, 2018. 1

[15] 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," *International Journal of Robotics Research (IJRR)*, vol. 39, no. 1, pp. 3–20, 2020. 1

[16] T. Chen, J. Xu, and P. Agrawal, "A system for general in-hand object re-orientation," in *Conference on Robot Learning (CoRL)*, 2021. 1, 2, 4

[17] Z.-H. Yin, B. Huang, Y. Qin, Q. Chen, and X. Wang, "Rotating without seeing: Towards in-hand dexterity through touch," *arXiv preprint arXiv:2303.10880*, 2023. 1

[18] A. T. Miller and P. K. Allen, "Graspit! a versatile simulator for robotic grasping," *IEEE Robotics and Automation Magazine (RA-M)*, vol. 11, no. 4, pp. 110–122, 2004. 1, 2, 3

[19] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1, 3

[20] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 3

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

[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 *International Conference on Robotics and Automation (ICRA)*, 1993. 2

[23] 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," *International Journal of Robotics Research (IJRR)*, vol. 16, no. 1, pp. 11–35, 1997. 2

[24] 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. 2

[25] 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 *International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 359–11 366. 2

[26] 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 *European Conference on Computer Vision (ECCV)*, 2022. 2

[27] H. Jiang, S. Liu, J. Wang, and X. Wang, "Hand-object contact consistency reasoning for human grasps generation," in *International Conference on Computer Vision (ICCV)*, 2021. 2

[28] 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 *International Conference on Robotics and Automation (ICRA)*, 2021. 2

[29] E. Corona, A. Pumarola, G. Alenya, F. Moreno-Noguer, and G. Rogez, "Ganhand: Predicting human grasp affordances in multi-object scenes," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[30] J. Lundell, F. Verdoja, and V. Kyrki, "Ddgc: Generative deep dexterous grasping in clutter," *IEEE Robotics and Automation Letters (RA-L)*, vol. 6, no. 4, pp. 6899–6906, 2021. 2

[31] 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," *IEEE Robotics and Automation Letters (RA-L)*, 2020. 2

[32] S. Brahmbhatt, A. Handa, J. Hays, and D. Fox, "Contactgrasp: Functional multi-finger grasp synthesis from contact," in *International Conference on Intelligent Robots and Systems (IROS)*, 2019. 2

[33] P. Grady, C. Tang, C. D. Twigg, M. Vo, S. Brahmbhatt, and C. C. Kemp, "Contactopt: Optimizing contact to improve grasps," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[34] T. Yamada, T. Taki, M. Yamada, and H. Yamamoto, "Grasp stability analysis of two objects by considering contact surface geometry in 3d," in *International Conference on Robotics and Biomimetics*, 2011. 2

[35] 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. 2

[36] B. Donald, L. Gariepy, and D. Rus, "Distributed manipulation of multiple objects using ropes," in *International Conference on Robotics and Automation (ICRA)*, 2000. 2

[37] K. Harada and M. Kaneko, "Kinematics and internal force in grasping multiple objects," in *International Conference on Intelligent Robots and Systems (IROS)*, 1998. 2

[38] 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. 2

[39] Y. Sun, E. Amatova, and T. Chen, "Multi-object grasping-types and taxonomy," in *International Conference on Robotics and Automation (ICRA)*, 2022. 2

[40] Y. Geng, B. An, H. Geng, Y. Chen, Y. Yang, and H. Dong, "Rlafford: End-to-end affordance learning for robotic manipulation," in *International Conference on Robotics and Automation (ICRA)*, 2023. 2

[41] 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 *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[42] M. Liu, M. Zhu, and W. Zhang, "Goal-conditioned reinforcement learning: Problems and solutions," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2022. 2

[43] 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. 2, 4

[44] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *Transactions on Robotics (T-RO)*, 2023. 3

[45] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3

[46] 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," *International Journal of Robotics Research (IJRR)*, vol. 36, no. 3, pp. 261–268, 2017. 3

[47] S. Brahmbhatt, C. Ham, C. C. Kemp, and J. Hays, "Contactdb: Analyzing and predicting grasp contact via thermal imaging," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[48] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017. 4

[49] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4

[50] 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. 4

[51] C. Ferrari, J. F. Canny *et al.*, "Planning optimal grasps." in *International Conference on Robotics and Automation (ICRA)*, 1992. 4

[52] M. Liu, Z. Pan, K. Xu, K. Ganguly, and D. Manocha, "Deep differentiable grasp planner for high-dof grippers," in *Robotics: Science and Systems (RSS)*, 2020. 5

[53] Y. Zhang, M. J. Black, and S. Tang, "We are more than our joints: Predicting how 3d bodies move," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3372–3382. 5