Bimanual Grasp Synthesis for Dexterous Robot Hands

Yanming Shao, Chenxi Xiao*

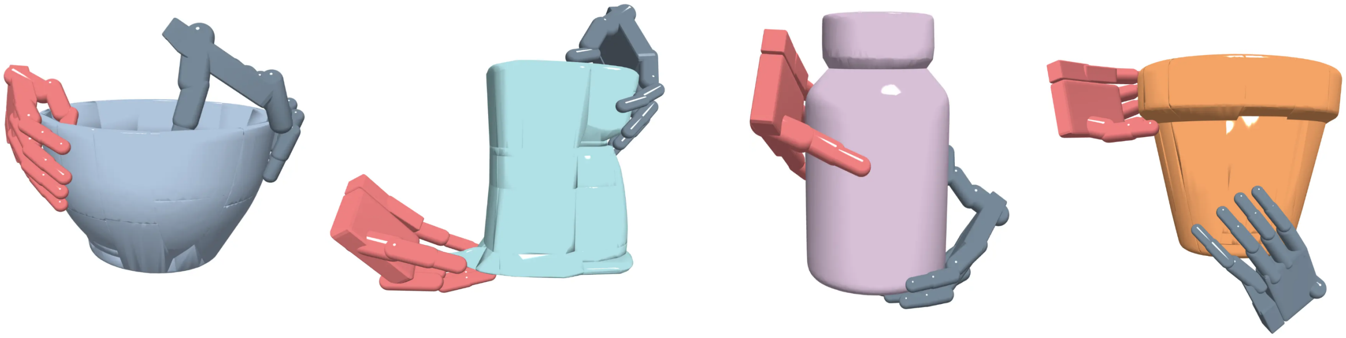


Fig. 1: Bimanual manipulation is necessary for handling large and heavy objects (e.g., basins, kitchen appliances). These objects would otherwise be difficult to handle using single-handed manipulators due to imbalanced contact forces and torques.

***Abstract*—Humans naturally perform bimanual skills to handle large and heavy objects. To enhance robots' object manipulation capabilities, generating effective bimanual grasp poses is essential. Nevertheless, bimanual grasp synthesis for dexterous hand manipulators remains underexplored. To bridge this gap, we propose the BimanGrasp algorithm for synthesizing bimanual grasps on 3D objects. The BimanGrasp algorithm generates grasp poses by optimizing an energy function that considers grasp stability and feasibility. Furthermore, the synthesized grasps are verified using the Isaac Gym physics simulation engine. These verified grasp poses form the BimanGrasp-Dataset, the first large-scale synthesized bimanual dexterous hand grasp pose dataset to our knowledge. The dataset comprises over 150k verified grasps on 900 objects, facilitating the synthesis of bimanual grasps through a data-driven approach. Last, we propose BimanGrasp-DDPM, a diffusion model trained on the BimanGrasp-Dataset. This model achieved a grasp synthesis success rate of 69.87% and significant acceleration in computational speed compared to BimanGrasp algorithm.**

***Index Terms*—Bimanual Manipulation, Grasping, Dexterous Manipulation**

I. Introduction

HUMANS can seamlessly coordinate both hands to perform complex tasks in daily life. This is mainly due to several advantages of bimanual manipulation. For instance, bimanual manipulation enables diverse object interaction skills effortlessly, such as tying ropes, knitting clothes, and performing kitchen chores. Compared to using only a single hand, bimanual manipulation reduces human fatigue and improves body balance by distributing payloads more evenly [1]. This capability is particularly beneficial when handling large and heavy objects, as it enhances both productivity and safety.

In the field of robotics, the growing market for humanoid robots has driven the development and use of multi-fingered dexterous hands for object manipulation. Recent works [2]–[5] have mainly focused on unimanual dexterous grasping based on object geometry. These techniques have simplified the use of dexterous hands with high degrees of freedom (DoF). However, they typically focused on objects of a size and mass suitable for a single hand. This narrow focus overlooks many larger and heavier objects (e.g., heavy bottles, home appliances, and furniture) that exceed the volume of unimanual grasps. Moreover, these works involved only one dexterous hand and did not fully exploit the potential of humanoid robots, which naturally possessed two hands.

On the other hand, there have been prior works focusing on bimanual manipulation. These works focus mainly on learning various object interaction skills. State-of-the-art works in this area have utilized Reinforcement Learning (RL) and imitation learning to acquire skills such as opening bottle lids and closing doors [6]–[9]. Although these studies involve bimanual grasping skills, they target very specific objects with unique functional affordances, requiring an ad-hoc training process for each object in simulation. Therefore, a research gap remains unbridged when designing general grasping skills for arbitrary objects. To our knowledge, there are no existing tools or grasp pose datasets for bimanual dexterous hand grasping. Although some studies have developed frameworks for bimanual object grasping [10], they are limited to parallel-jaw grippers. This highlights a research gap in the use of bimanual dexterous hands.

From a technical perspective, synthesizing bimanual grasp poses for dexterous hands presents greater challenges compared to the conventional grasp search problem. One of the main reasons is the significantly larger action space, which greatly increases the computational cost. For instance, the DoF of a *Shadow Hand* manipulator exceed 20 [11]. When extended to bimanual grasping, the DoF doubles, leading to a substantial increase in computational cost. In practice,

Our project page is available at https://BimanGrasp.github.io/.

Yanming Shao is with the School of Information Science and Technology at ShanghaiTech University, shaoym2023@shanghaitech.edu.cn

Chenxi Xiao (* corresponding author) is with MoE Key Laboratory of Intelligent Perception and Human-Machine Collaboration (KLIP-HuMaCo), and with the School of Information Science and Technology at ShanghaiTech University, xiaochx@shanghaitech.edu.cn

reducing computational cost has been the main focus of most studies on dexterous grasp synthesis [4], [12], where the ability to generate grasps in quasi-real time when encountering new objects is highly desirable.

In this paper, we aim to develop a bimanual grasp synthesis pipeline optimized for both grasp quality and generation speed. To achieve this, we first propose BimanGrasp algorithm, a grasp synthesis method that searches for bimanual grasp poses in a high-dimensional configuration space using stochastic optimization. Secondly, by implementing BimanGrasp algorithm with GPU-based optimization, we have synthesized the BimanGrasp-Dataset, which comprises over 150k grasps. Each grasp has been verified through simulations in the Isaac Gym environment. [13]. The validation results demonstrate that the bimanual grasp strategy can handle large and heavy objects, which were previously unattainable with unimanual grasping techniques. Lastly, by utilizing the proposed dataset, we have significantly accelerated bimanual grasp synthesis by transforming it into a data-driven paradigm. We introduce BimanGrasp-DDPM, the first diffusion model capable of efficiently generating diverse bimanual grasp poses.

To summarize, the contributions of this paper are as follows:

- BimanGrasp algorithm: an offline bimanual grasp synthesizer based on stochastic optimization.
- BimanGrasp-Dataset: A dataset of bimanual robot grasping poses, validated through physics simulation.
- BimanGrasp-DDPM: a quasi-real time bimanual grasp generator based on DDPM.
- Quantitative studies comparing the performance of bimanual grasping with unimanual grasping.

II. Related Works

A. Robot Manipulation

Robots are cyber-physical systems that possess the ability to interact with various objects in the physical environments. Hence, object manipulation has long been a key research area. Alone this line of research, previous works have focused on developing new skills, improving efficiency and safety. Due to the complex nature of environments, object manipulation encompasses diverse forms. One category is prehensile manipulation, where robots aim to grasp objects. The other category is non-prehensile manipulation, which primarily includes non-grasping skills [14]. Typical applications include planar pushing [15], throwing and catching [16], solving a Rubik's Cube [17], and playing instruments [18].

In this paper, we focus on developing bimanual object grasping skills, which belong to the prehensile manipulation category. To enable bimanual grasping, it requires obtaining cooperative grasp poses that enclose objects inside. This process is known as bimanual grasp synthesis. Previous research has extensively studied the synthesis of grasp poses, but mainly for grippers [19]. These studies have reportedly achieved both high grasp success rates and real-time computational efficiency [20], [21]. However, due to the limited capability of low-DoF grippers, there is a growing need for synthesizing grasps for high-DoF manipulators, such as dexterous hands.

Synthesizing grasps for dexterous hands is more challenging than for grippers. Early researches used a grasp synthesis pipeline similar to the conventional approach for grippers. This method involves sampling grasp poses and analyzing the likelihood that each grasp pose could satisfy the force-closure condition [12], [19]. The advantage of these approaches is the ability to synthesize grasp poses for almost arbitrary object shapes. However, they are computationally expensive for manipulators with high DoF due to a significantly larger action space.

To accelerate grasp generation, recent works have adopted data-driven approaches to generate grasp poses directly [22]. Common data-driven generative models include variational autoencoders [23], [23]–[25], normalizing flows [22], and diffusion models [26]–[29]. This line of research has demonstrated improved speed and quality in the synthesis of unimanual grasps [22], [27]. However, it has not yet been applied to bimanual manipulation.

B. Bimanual Manipulation

Bimanual manipulation refers to the coordinated use of both hands to manipulate objects. This type of manipulation is essential for a wide range of activities, from daily tasks to complex professional operations. In everyday life, actions such as tying shoelaces, opening jars, and typing on a keyboard depend on the synchronized use of both hands. In professional fields, bimanual manipulation is crucial in areas like surgery and component assembly, where precision and coordination between both hands are paramount [30].

Bimanual manipulation is also a critical skill for robots. Historically, humanoid robots capable of bimanual object manipulation emerged in the 2000s [31], [32]. Since then, bimanual skills such as moving kitchen cookware [33], dishwashing [34], and object pick-and-place [35] have been developed. However, these early strategies were based on two parallel grippers rather than dexterous hands.

Recent advancements in robotics have increasingly focused on learning dexterous bimanual manipulation with multi-fingered hands, often using human demonstrations. This progress has enabled robots to perform coordinated actions with greater precision than unimanual approaches [7], [9], [36]. Despite these advancements, current researches on dexterous bimanual manipulation target very specific objects and lack large-scale datasets or diverse hand pose types [37], [38]. To bridge the gap, we propose techniques and a dataset that aim to help humanoid robots develop more general bimanual grasping skills.

III. Method

A. Problem Definition and Overview

The problem of bimanual grasp synthesis is formulated as Eq. (1). Given an object mesh denoted as O, our goal is to obtain grasp poses by maximizing the grasp quality score S that is empirically defined. The metric $\mathcal{G}$ used for calculating the score also considers rigid body poses $\boldsymbol{T}_l \in SE(3)$ and joint configurations $\boldsymbol{\theta}_l \in \mathbb{R}^{22}$, where $l \in \{1, 2\}$ denotes the left and right manipulators, respectively. This optimization paradigm

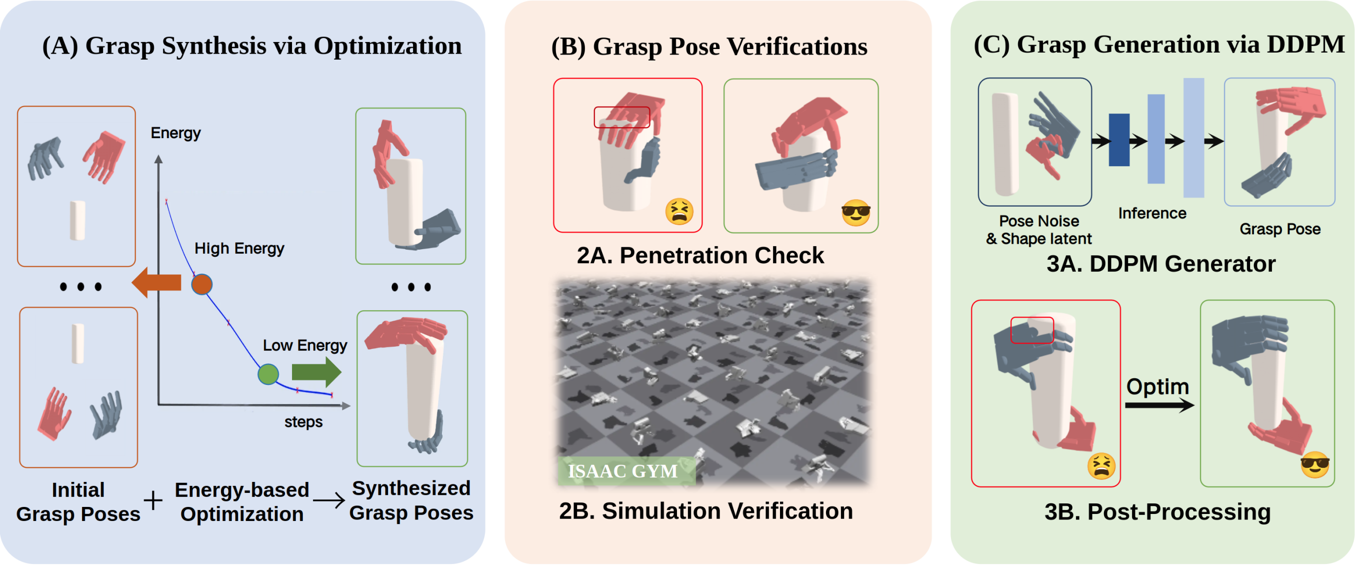

(A) Grasp Synthesis via Optimization
Energy
High Energy
Low Energy
steps
Initial Grasp Poses
+
Energy-based Optimization
→
Synthesized Grasp Poses
(B) Grasp Pose Verifications
2A. Penetration Check
ISAAC GYM
2B. Simulation Verification
(C) Grasp Generation via DDPM
Pose Noise & Shape latent
Inference
Grasp Pose
3A. DDPM Generator
Optim
3B. Post-Processing

Fig. 2: Our pipeline for synthesizing stable bimanual grasps, which includes: (A) generating grasp poses by initializing the bimanual grasp poses around the objects and then improving their quality through optimization; (B) verifying the grasp poses based on shape penetration and physics simulation using Isaac Gym; (C) utilizing the verified grasps (BimanGrasp-Dataset) to train a generative model (BimanGrasp-DDPM), with post-processing techniques to remove penetrations.

is employed for synthesizing bimanual grasps, corresponding to the BimanGrasp algorithm proposed in Sec. III-B.

$$\max_{\boldsymbol{T}_1,\boldsymbol{\theta}_1,\boldsymbol{T}_2,\boldsymbol{\theta}_2} S=\mathcal{G}(O,\boldsymbol{T}_1,\boldsymbol{\theta}_1,\boldsymbol{T}_2,\boldsymbol{\theta}_2). \tag{1}$$

Although this approach has been commonly adopted for grasp synthesis problems involving manipulators with lower DoF (e.g., [12], [39], to mention a few), one issue lies in computational efficiency. Given the high DoF of two dexterous hands, obtaining feasible bimanual grasps in quasi-real time is difficult. This limitation hinders the applications of such methods, especially for scenarios where timely responses are crucial.

Compared to the optimization-based paradigm discussed above, synthesizing bimanual grasps through deep learning models could be more efficient. To verify this idea, we propose BimanGrasp-DDPM model (Sec. III-D). Nevertheless, training such a generative model requires a dataset of bimanual grasp poses, which does not currently exist. To bridge this gap, we introduce the BimanGrasp-Dataset, which is synthesized offline using the optimization-based BimanGrasp algorithm. The overall system architecture for achieving this is shown in Fig. 2.

B. Synthesis Bimanual Grasp via Stocastic Optimization

This section describes the BimanGrasp algorithm, which is capable of synthesizing grasp poses conditioned on object meshes. The algorithm consists of two steps: 1) initialize grasp poses around the target object, and 2) iteratively optimize an energy function, during which the poses are adjusted based on grasp stability and penetration. The detailed procedures for achieving this are as follows:

Step 1: Initialize Bimanual Hand Poses. Humans naturally grasp objects by facing them and approaching from two opposite sides. In our algorithm, we empirically initialize each hand's pose symmetrically around the object's center to increase the similarity to human grasps and reduce the chance of penetration. Specifically, we adopt the initialization procedures from [4]. First, two hands are placed on an inflated convex hull enveloping the object, with the palms facing the object. Then, we progressively decrease the hull's size, reducing the gap between the hands and the object until they make contact. Note that it's important to introduce variations in the initial hand poses and joint angles. This randomization technique enlarges the search space, allowing for more diverse grasps and helping avoid sub-optimal local minima.

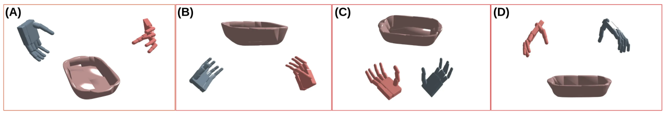


Fig. 3: Initialization of grasp poses (with randomization applied to joint angles and poses). We demonstrate four examples, which approach the object from different directions.

Step 2: Improve Grasp Quality. The grasp quality is improved by optimizing an energy function, defined as the weighted sum of all terms in Table. I. In the table, we denote $d(p,q)$ for $p,q\in\mathbb{R}^3$ as the Euclidean distance between two points, and $d(p,O)=\min_{q\in O}(p,q)$ as the distance between p and the object mesh. We denote H_l as the mesh of each hand with $l\in\{1,2\}$, and $P(H_l)$ as the anchor points selected from the hand mesh to compute penetration.

TABLE I: Energy function for grasp search problem. The minimization objective of the algorithm is the weighted sum of all terms.

Term	Formulation
E_{dis}: Hand-object distance	$\sum_{a=1}^{n} d(x_a, O)$
E_{fc}: Force Closure	$\|\|\boldsymbol{Gc}\|\|_2$
E_{vew}: Wrench Ellipse Volume	$\left(\det\left(\mathbf{GG}^T\right)\right)^{-\frac{1}{2}}$
E_{objpen}: Hand-Object Penetration	$\sum_{l \in \{1,2\}} \sum_{p_l \in P(H_l)} \max(\delta - d(p_l, O), 0)$
E_{selfpen}: Hand Self-Penetration	$\sum_{l \in \{1,2\}} \sum_{p,q \in P(H_l)} \max(\delta - d(p, q), 0)$
E_{bimpen}: Inter-Hands Penetration	$\sum_{p \in P(H_1), q \in P(H_2)} \max(\delta - d(p, q), 0)$
E_{joint}: Violation of Joint Limits	$\sum_{i=1}^{44} (\max(\theta_i - \theta_i^{max}, 0) + \max(\theta^{min} - \theta_i, 0))$

The empirical quality metric considers both the grasp stability and feasibility. One main goal is to keep hands close to object's surface. To achieve this, we construct the term E_{dis}, which quantifies the distance between the two hands and the object. By minimizing this term, the fingers and palms land close to the object surface. Here x_a with $a \in \{1, 2, \ldots, n\}$ denotes a point cloud with $n = 4000$ points sampled from both hands' surfaces.

The term E_{fc} represents the force closure, which serves as the main heuristic for grasp stability. In E_{fc}, c is the contact normal vector at the contact points $\boldsymbol{x}_j = (x_j, y_j, z_j)$, where $j \in \{1, 2, \ldots, 8\}$ is the index of contact points. Note that unlike unimanual grasping, our approach leverages 8 contact points for grasp collaboration (4 from each hand). This formulates the grasp matrix $\boldsymbol{G}$:

$$\boldsymbol{G} = \left[\begin{array}{ccc|ccc} \boldsymbol{I} & \ldots & \boldsymbol{I} & \boldsymbol{I} & \ldots & \boldsymbol{I} \\ \boldsymbol{R}_1 & \ldots & \boldsymbol{R}_4 & \boldsymbol{R}_5 & \ldots & \boldsymbol{R}_8 \end{array}\right] \tag{2}$$

where

$$\boldsymbol{R}_j = \begin{bmatrix} 0 & -z_j & y_j \\ z_j & 0 & -x_j \\ -y_j & x_j & 0 \end{bmatrix}, \mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{3}$$

In addition, E_{vew} is defined to prevent the Gram matrix $\boldsymbol{GG}^T$ from being ill-conditioned. This ensures that the grasp can effectively resist small external wrench disturbances from any direction. Notably, optimizing E_{fc} and E_{vew} establishes the differential force closure condition proposed by [12].

We also prevent penetration failures by accounting for the following penetration patterns: (1) between two hands and the object; (2) between the left and right hands; and (3) within each hand and itself. The energy terms that prevent these three types of penetrations are E_{objpen}, E_{bimpen}, and E_{selfpen}, respectively. Each of these energy terms is calculated with the distance between some anchor points from $P(H_1)$ and $P(H_2)$, unless it is lower than a fixed small threshold ϵ.

In addition, an energy term E_{joint} is introduced to handle joint limit violations. For each joint angle θ_i with $i \in \{1, 2, \ldots, 44\}$, we denote θ_i^{min} as its lower limit and θ_i^{max} as its upper limit. If it is outside $[\theta_i^{\text{min}}, \theta_i^{\text{max}}]$, the violation is penalized using the out-of-range value, in form of $\max(\theta_i - \theta_i^{\text{max}}, 0) + \max(\theta^{\text{min}} - \theta_i, 0)$.

We jointly optimize the weighted sum of all the aforementioned energy terms. Given the non-convexity of the energy function, we employ the Metropolis-adjusted Langevin algorithm (MALA) optimizer, which introduces stochasticity to circumvent local optima [12]. The hand configurations during optimization process are showcased in Fig. 4.

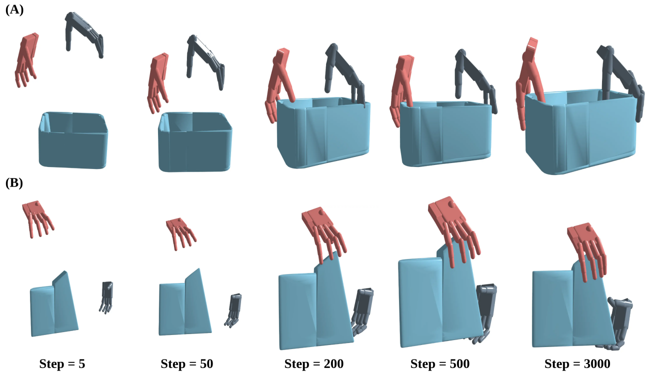


Fig. 4: Visualization of the BimanGrasp algorithm's optimization process on two objects: (A) a square container, and (B) a home appliance. Both hands started by approaching the object from initialized poses. As optimization proceeds, the hands gradually landed on object surfaces.

C. Dataset Generation

To prepare data for training generative models, we synthesized grasp poses for a dataset of objects from Google's Scanned Objects (GSO) Dataset [40]. The manipulator used is a pair of *Shadow Hand*s. Each *Shadow Hand* has 22 actuated joints (denoted as $\boldsymbol{\theta}_i$ each), and a 6 dimensional rigid body pose. For both manipulators, our action space has $(22 + 6) \times 2 = 56$ dimensions in total.

We followed the grasp synthesis procedures outlined in Sec. III-B to generate an initial set of grasp poses. Then, we used the Isaac Gym environment [13] to label whether a grasp can successfully lift and hold objects, as shown in Fig. 5. The friction coefficient for both the objects and hands are fixed at 3 (same as [4], [26]). To control the motor output, we employed a PD controller in Isaac Gym, with stiffness $K_p = 1000.0$ and damping $K_d = 10.0$.

A grasp configuration is labeled as successful if the object remained in the hand for 2.0 seconds of simulation time (*i.e.*, 120 steps at 60 Hz) over 6 evaluation trials under a gravity of 9.8 m$\cdot$ s^{-2}. During each evaluation, the object and hands were randomly rotated together to verify the grasp under varying gravity force directions.

We also checked for the three types of penetration described in Sec. III-B. If the total penetrations exceeded 1.5 mm, the grasp configuration fails the evaluation. All successful grasps, along with the objects they were conditioned on, were saved into our BimanGrasp-Dataset.

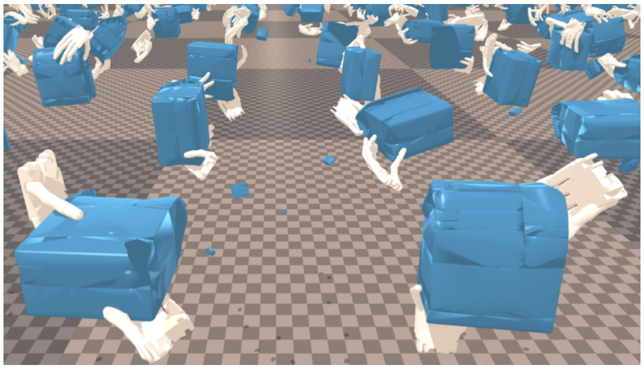


Fig. 5: The visualization of our physical verification of the bimanual grasps using Isaac Gym [13]. Objects are held firmly by stable grasps, while they slip away from unstable grasps.

D. Grasp Generation Through Data Driven approach

We then developed a generative model based on the dataset of successfully grasped poses. The proposed generative model employed the Denoising Diffusion Probabilistic Models (DDPM) architecture [41], which is widely used for various generative tasks. Conditioned on a feature vector $\mathcal{O} \in \mathbb{R}^{1024}$ (extracted by PointNet [42] from a point cloud sampled from O), the DDPM model aims to transform standard Gaussian noise $\boldsymbol{h}^T \sim \mathcal{N}(0; I)$ into a grasp pose $\boldsymbol{h}^0$. This was achieved through an iterative denoising process, as described in Eq. (4):

$$p_\theta(\boldsymbol{h}^0|\mathcal{O}) = p(\boldsymbol{h}^T)\prod_{t=1}^{T} p(\boldsymbol{h}^{t-1}|\boldsymbol{h}^t, \mathcal{O}) \tag{4}$$

For each stage of the denoising process, the implementation was based on the reparameterization trick [43], as given in Eq. (5). We leveraged the U-Net model to predict the mean μ_θ and the standard deviation $\boldsymbol{\Sigma}_\theta$ of the noise added at each diffusion step.

$$p(\boldsymbol{h}^{t-1}|\boldsymbol{h}^t, \mathcal{O}) = \mathcal{N}(\boldsymbol{h}^t; \boldsymbol{\mu}_\theta(\boldsymbol{h}^t, t, \mathcal{O}), \boldsymbol{\Sigma}_\theta(\boldsymbol{h}^t, t, \mathcal{O})) \tag{5}$$

Then, the learning objective is given in Eq. (6):

$$L_\theta(\boldsymbol{h}^0|\mathcal{O}) = E_{t,\epsilon,\boldsymbol{h}^0}\left[||\epsilon - \epsilon_\theta(\sqrt{\overline{a_t}}\boldsymbol{h}^0 + \sqrt{1-\overline{a_t}}\epsilon, t, \mathcal{O})||\right]. \tag{6}$$

where ϵ is the noise to estimate; ϵ_θ is the noise predicted by the model; t represents the denoising steps; $\overline{a_t}$, defined as $\prod_{k=1}^{t} a_k$, represents the noise intensity.

Nevertheless, grasp poses directly obtained from the DDPM model could be infeasible. This is mainly because penetrations are not explicitly considered during the generative process. While preventing DDPM from generating penetrated grasps requires nontrivial customization, we leverage a post-processing technique. That is, after obtaining the grasp poses from DDPM, we improve the poses by optimizing on energy terms defined in Table I. We use much fewer (100) steps, compared to 10000 steps in BimanGrasp Algorithm, to keep the computational efficiency.

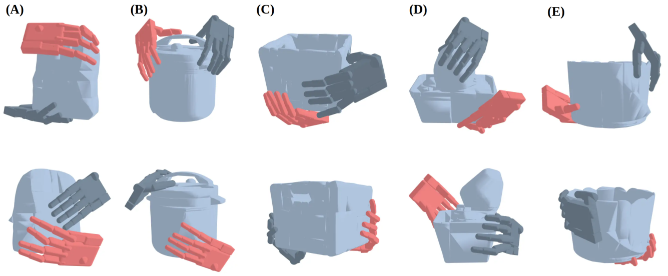


Fig. 6: Visualization of the grasp poses synthesized on daily-life objects using BimanGrasp algorithm. Objects include: (A) backpack, (B) pot, (C) box, (D) bucket, and (E) container. All object models are from Google Scanned Dataset [40].

IV. Experiments

To validate our proposed approaches, we conducted experiments to: 1) evaluate the performance of the BimanGrasp algorithm quanlitatively and quantitatively (Sec. IV-A); 2) evaluate the performance of the DDPM trained on the BimanGrasp-Dataset (Sec. IV-B). Last but not least, 3) we provide discussions on experiments, ablation studies, and broader insights (Sec. IV-C).

A. Evaluation on Bimanual Grasp Synthesis

Visualization of Grasps. First, we evaluate our proposed BimanGrasp algorithm on everyday objects from the GSO [40] dataset. These objects include various household items, such as containers and kitchen utensils, as shown in Fig. 6. The generated grasps for these daily-life objects are diverse. For instance, some grasps secure the body of a cylindrical container, while others pinch the edges of a box. These results successfully demonstrate the reliable generation of grasp poses for objects with diverse shapes.

One question is whether the generated grasps are human-like [4], [25]. To validate this, we adopted the evaluation approach used in [25]. We employed *GPT-4 Vision* to score each bimanual grasp on a scale of 1 to 3 points (evaluating 1,000 grasps, with 3 views per grasp). The average score obtained was 2.67.

Grasp Success Rate. Next, we quantitatively evaluate the quality of the grasp poses synthesized by BimanGrasp. The evaluation was conducted on the synthesized 450k bimanual grasps (900 objects, with 500 poses per object). A grasp pose is considered successful if it meets two criteria: 1) no penetration occurs, and 2) object does not slip away during the physics verification process, as outlined in Sec. III-C. During physics verification, all 900 objects were used in their original sizes provided in [40], with a density of $\rho = 2500~\text{kg}\cdot\text{m}^{-3}$ and object friction coefficient of 3 (following setting [4], [26]).

Our experimental results visualize the relationship between the object's diameter (d) versus grasp success rate, as illustrated in Fig. 7. We compared our bimanual grasping strategy with two unimanual grasp baselines from a current state-of-the-art approach [4]. After benchmarking all grasp poses, we visualized the success rate distribution in Fig. 7. Objects were grouped into seven categories based on their diameters. The

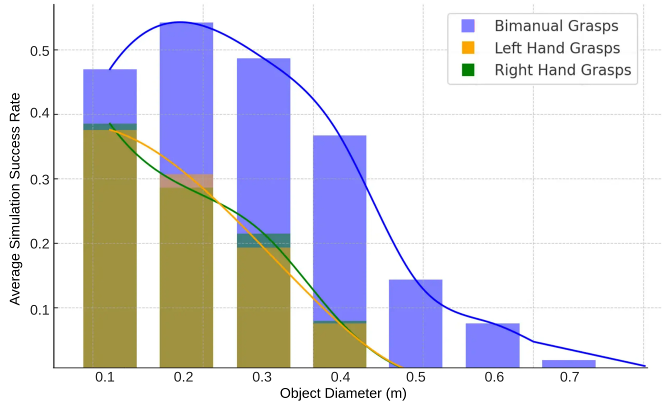


Fig. 7: The average grasp success rate across various object diameters. As object diameters increase, the success rates decrease due to the increased grasp difficulty. In all experiments, bimanual grasps outperform unimanual baselines.

results show that our bimanual grasping strategy consistently achieves a higher success rate than unimanual strategies across all object sizes. Furthermore, the performance advantage of the bimanual strategy increases with object diameter. Unimanual grasps almost entirely fail for objects larger than $d = 0.5$ m, while bimanual grasping remains effective for objects up to $d = 0.7$ m. This highlights the superior effectiveness of bimanual grasps, especially for larger objects.

TABLE II: Comparison of the average grasp success rate (%) under different object densities ($kg \cdot m^{-3}$).

Density	$\rho = 5000$	$\rho = 2500$	$\rho = 500$
Both hands	**41.02**	**54.03**	**71.42**
Uni2Bim (opt)	32.87	45.26	56.69
Left Hand Only	23.38	41.48	68.42
Right Hand Only	21.85	41.95	68.48

Next, we evaluated the grasp success rates for objects of different masses. To minimize the influence of object size, all objects were normalized to diameter $d = 0.2$ m. Then, we created three comparative groups by varying the object density: $\rho = 5000, 2500,$ and 500 kg$\cdot$m^{-3}. Within each group, we evaluated the average grasp success rate across all 900 objects in the Isaac Gym simulation.

From results in Table. II, we conclude that bimanual grasping offers significant advantages in object manipulation, particularly when handling heavy objects. Across all groups, bimanual grasp strategies consistently achieved higher success rates than unimanual strategies. This advantage is especially significant when grasping heavy objects. For instance, when $\rho = 5000$ kg $\cdot$ m^{-3}, the bimanual grasp strategy achieved a success rate of 41.02%, while the unimanual grasp strategies only achieved 23.38% and 21.85%. Note that this advantage is not solely due to the larger forces generated by more motors from two hands, but also because of the cooperation between them. To demonstrate this, we implemented a baseline named Uni2Bim (opt) for comparison with the BimanGrasp algorithm. Uni2Bim (opt) optimizes each hand separately, following [4], without any joint optimization between the two hands. As shown in results, Uni2Bim (opt) achieved lower success rates across all three object densities, highlighting the importance of joint optimization architecture.

In addition, we evaluate the grasp robustness by varying the friction coefficient. To do this, we fix the object density at 2500 $kg \cdot m^{-3}$ and allow the object friction coefficient to range from 0.5 to 3.0 (following [44]). The overall simulation success rate is shown in Table III. When the friction coefficient was significantly reduced to 0.5, the grasp success rate decreased to 45.40% from 54.03%, indicating that around 84% of all grasps are still valid. This demonstrates the robustness of our generated grasps.

TABLE III: Success rate (%) of BimanGrasp in IsaacGym under different friction coefficient settings. Object density is at $\rho = 2500$ kg $\cdot$ m^{-3}.

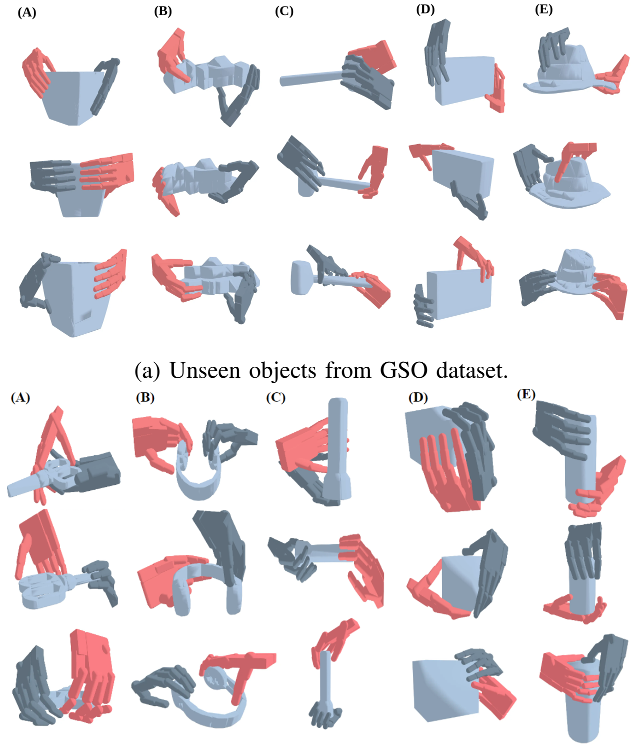
Friction Coeff.	0.5	1.0	1.5	2.0	2.5	3.0
Succ. Rate	45.40	47.04	49.32	51.14	52.44	54.03

B. Evaluation on the Bimanual DDPM

Using the synthesized dataset, we trained a DDPM model following procedures described in Sec. III-D. The training was performed on the successful grasps from BimanGrasp-Dataset, (randomly selected $900 \times 75\% = 675$ objects). We then evaluated the grasping success rate on the remaining 25% of objects (225 unseen objects) using Isaac Gym with 500 grasps per object. The performance of the synthesized grasps for unseen objects was evaluated under a uniform diameter of $d = 0.2$ m and varying object densities $\rho = 5000, 2500,$ and 500 kg $\cdot$ m^{-3}. The average success rate was 42.39% for $\rho = 5000$ kg$\cdot$m^{-3}, 54.06% for $\rho = 2500$ kg$\cdot$m^{-3}, and 69.87% for $\rho = 2500$ kg $\cdot$ m^{-3}. These results are comparable to the success rates of the analytically synthesized bimanual grasp poses from BimanGrasp algorithm, as shown in Table. II.

Next, we aimed to assess whether the performance of Bimanual-DDPM is comparable to other methods. Since no existing methods currently address bimanual grasp synthesis, we manually crafted two baselines. (1) We customized a Conditional Variational Autoencoder (CVAE) [23] and retrained the network using a bimanual grasping protocol based on the BimanGrasp dataset. (2) We used the approach from [26] to generate grasps for each hand separately (referred to as Uni2Bim (dm)), without utilizing our dataset priors that account for hand collaboration. When evaluated on objects with a density of 2500 kg $\cdot$ m^{-3}, CVAE achieved a success rate of 11.85%, while Uni2Bim (dm) reached 36.52%. Both baselines performed significantly worse than BimanGrasp-DDPM, although CVAE has advantages in its running time (only 18 ms per grasp).

Last, we showcase the generalizability of our model to objects from other datasets. For this, we selected 60 objects from the DDG, YCB, and ContactDB datasets (which differ from the GSO Dataset used for training). We randomly scaled object diameters within the range $[0.2$ m$, 0.4$ m$]$. The object density was fixed at 1000 kg $\cdot$ m^{-3}. We achieved an average



(a) Unseen objects from GSO dataset.

(b) Unseen objects from DDG, YCB, and contactDB datasets.

Fig. 8: The diverse bimanual stable grasps synthesized with BimanGrasp-DDPM. All objects are unseen during training.

success rate of 63.23%. Together with grasp poses generated for GSO objects, the visualizations are shown in Fig. 8.

C. Discussions

Limitations. While we have demonstrated the effectiveness of our BimanGrasp-DDPM model, the DDPM's grasp synthesis process does not explicitly consider penetration. Consequently, the generative model can still produce infeasible grasp poses. This issue is currently mitigated through a post-processing step. As previously described in Sec. III-D. We anticipate that this limitation can be addressed by leveraging more recent diffusion models that account for physical constraints, such as the approaches described in [45].

In addition, the algorithms may generate grasp poses that are not human-like. Given that our dataset provides a large number of diverse grasps, we believe a possible solution is to automatically select a subset of the dataset using a visual language model scorer and then retrain the DDPM model under a human-like grasp data distribution.

Diversity. The diversity of grasps was evaluated using an entropy metric H_{mean} adapted from [4], [26]. The mean entropy H_{mean} of grasps generated by BimanGrasp algorithm and BimanGrasp-DDPM is 4.39 and 3.72, with standard deviations H_{std} of 0.47 and 0.49, respectively. Together with the visualization results in Sec. IV-B, this proves that DDPM can generate multi-modal and diverse bimanual grasps.

Failure Cases. As part of an ablation study, we highlight failure cases from the BimanGrasp algorithm synthesized grasps in Fig. 9. All these failure patterns could happen both in BimanGrasp algorithm and BimanGrasp-DDPM. Our observations indicate that penetration remains the primary cause of grasp failure. The optimization procedure occasionally fails to retract the hand from objects due to local optima. Additionally, in some instances, the hands may become detached from the objects' surface, causing the object to slip away before finger attached to the object surface.

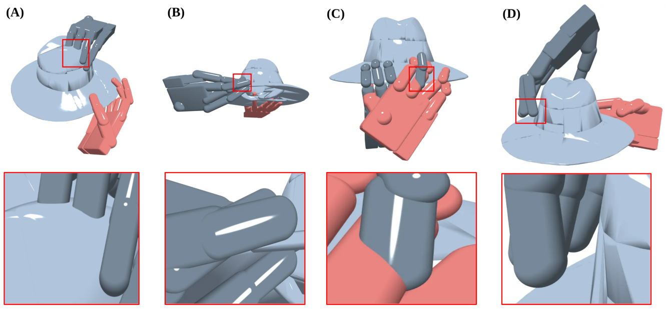


Fig. 9: Visualization of four most common failure patterns: (A) hand-object penetration, (B) hand's self-penetration, (C) inter-hand penetration, and (D) failure to establish contact.

Computational Cost. The synthesis of the BimanGrasp Dataset was achieved on a server with four Nvidia A40 GPUs. It requires 170 GB of GPU memory and takes 117 minutes to generate 4,500 grasps per batch. For DDPM, the inference stage can be accomplished only on a commercial GPU. Using an RTX 4090 GPU, we can parallelize the inference of 64 grasps, with the inference time being 8.19 seconds.

V. Conclusion

Humans naturally utilize bimanual actions for various object manipulation tasks. However, algorithms for synthesizing bimanual dexterous hands grasping poses were barely studied so far. To bridge this gap, we proposed the BimanGrasp algorithm, which successfully synthesizes bimanual grasps for diverse objects by leveraging stochastic optimization algorithms guided by energy-based heuristics. These grasps are then refined through physical validation in the Isaac Gym environment. Through this process, we obtained the BimanGrasp-Dataset, which contains over 150k verified pairs of grasps for 900 objects. This dataset enables the training of data-driven models capable of accelerating the grasp synthesis process. To prove this, we developed a BimanGrasp-DDPM model that excels in offering efficient grasp poses and with grasp success rate comparable to that of the BimanGrasp algorithm.

In the future, we plan to further improve the grasp success rate. We also plan to conduct experimental validations of the proposed algorithms using a real-world bimanual humanoid robot. We believe that the proposed technique can provide impacts to humanoid robots by enhancing their ability to handle various daily life objects in homes and other unstructured environments.

References

[1] N. Vahrenkamp, M. Przybylski, T. Asfour, and R. Dillmann, "Bimanual grasp planning," in *2011 11th IEEE-RAS International Conference on Humanoid Robots*. IEEE, 2011, pp. 493–499.

[2] 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.

[3] 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)*. IEEE, 2023, pp. 8068–8074.

[4] 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)*. IEEE, 2023, pp. 11 359–11 366.

[5] 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*, 2024.

[6] 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," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5150–5163, 2022.

[7] H. Zhang, S. Christen, Z. Fan, L. Zheng, J. Hwangbo, J. Song, and O. Hilliges, "Artigrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation," *arXiv preprint arXiv:2309.03891*, 2023.

[8] Y. Chen, Y. Geng, F. Zhong, J. Ji, J. Jiang, Z. Lu, H. Dong, and Y. Yang, "Bi-dexhands: Towards human-level bimanual dexterous manipulation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[9] Z. Fan, O. Taheri, D. Tzionas, M. Kocabas, M. Kaufmann, M. J. Black, and O. Hilliges, "Arctic: A dataset for dexterous bimanual hand-object manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 943–12 954.

[10] G. Zhai, Y. Zheng, Z. Xu, X. Kong, Y. Liu, B. Busam, Y. Ren, N. Navab, and Z. Zhang, "Da^2 dataset: Toward dexterity-aware dual-arm grasping," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 8941–8948, 2022.

[11] 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," *The International Journal of Robotics Research*, vol. 39, no. 1, pp. 3–20, 2020.

[12] 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, 2021.

[13] 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.

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

[15] J. K. Li, W. S. Lee, and D. Hsu, "Push-net: Deep planar pushing for objects with unknown physical properties." in *Robotics: Science and Systems*, vol. 14, 2018, pp. 1–9.

[16] F. Lan, S. Wang, Y. Zhang, H. Xu, O. Oseni, Y. Gao, and T. Zhang, "Dexcatch: Learning to catch arbitrary objects with dexterous hands," *arXiv preprint arXiv:2310.08809*, 2023.

[17] I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas *et al.*, "Solving rubik's cube with a robot hand," *arXiv preprint arXiv:1910.07113*, 2019.

[18] K. Zakka, L. Smith, N. Gileadi, T. Howell, X. B. Peng, S. Singh, Y. Tassa, P. Florence, A. Zeng, and P. Abbeel, "Robopianist: A benchmark for high-dimensional robot control," *arXiv preprint arXiv:2304.04150*, 2023.

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

[20] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 1316–1322.

[21] M. Gualtieri, A. Ten Pas, K. Saenko, and R. Platt, "High precision grasp pose detection in dense clutter," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 598–605.

[22] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4737–4746.

[23] H. Jiang, S. Liu, J. Wang, and X. Wang, "Hand-object contact consistency reasoning for human grasps generation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 11 107–11 116.

[24] W. Wei, D. Li, P. Wang, Y. Li, W. Li, Y. Luo, and J. Zhong, "Dvgg: Deep variational grasp generation for dextrous manipulation," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1659–1666, 2022.

[25] Y. Liu, Y. Yang, Y. Wang, X. Wu, J. Wang, Y. Yao, S. Schwertfeger, S. Yang, W. Wang, J. Yu *et al.*, "Realdex: Towards human-like grasping for robotic dexterous hand," *arXiv preprint arXiv:2402.13853*, 2024.

[26] J. Lu, H. Kang, H. Li, B. Liu, Y. Yang, Q. Huang, and G. Hua, "Ugg: Unified generative grasping," *arXiv preprint arXiv:2311.16917*, 2023.

[27] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 750–16 761.

[28] Z. Weng, H. Lu, D. Kragic, and J. Lundell, "Dexdiffuser: Generating dexterous grasps with diffusion models," *arXiv preprint arXiv:2402.02989*, 2024.

[29] J. Cao, J. Liu, K. Kitani, and Y. Zhou, "Multi-modal diffusion for hand-object grasp generation," *arXiv preprint arXiv:2409.04560*, 2024.

[30] F. Krebs and T. Asfour, "A bimanual manipulation taxonomy," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 031–11 038, 2022.

[31] R. Platt, R. A. Grupen, and A. H. Fagg, "Learning grasp context distinctions that generalize," in *2006 6th IEEE-RAS International Conference on Humanoid Robots*. IEEE, 2006, pp. 504–511.

[32] C. Smith, Y. Karayiannidis, L. Nalpantidis, X. Gratal, P. Qi, D. V. Dimarogonas, and D. Kragic, "Dual arm manipulation—a survey," *Robotics and Autonomous systems*, vol. 60, no. 10, pp. 1340–1353, 2012.

[33] N. Vahrenkamp, M. Do, T. Asfour, and R. Dillmann, "Integrated grasp and motion planning," in *2010 IEEE International Conference on Robotics and Automation*. IEEE, 2010, pp. 2883–2888.

[34] J. Bohg, K. Welke, B. León, M. Do, D. Song, W. Wohlkinger, M. Madry, A. Aldóma, M. Przybylski, T. Asfour *et al.*, "Task-based grasp adaptation on a humanoid robot," *IFAC Proceedings Volumes*, vol. 45, no. 22, pp. 779–786, 2012.

[35] J.-P. Saut, M. Gharbi, J. Cortés, D. Sidobre, and T. Siméon, "Planning pick-and-place tasks with two-hand regrasping," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2010, pp. 4528–4533.

[36] T. Lin, Y. Zhang, Q. Li, H. Qi, B. Yi, S. Levine, and J. Malik, "Learning visuotactile skills with two multifingered hands," *arXiv preprint arXiv:2404.16823*, 2024.

[37] 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.

[38] K. Harada, T. Foissotte, T. Tsuji, K. Nagata, N. Yamanobe, A. Nakamura, and Y. Kawai, "Pick and place planning for dual-arm manipulators," in *2012 IEEE International Conference on Robotics and Automation*. IEEE, 2012, pp. 2281–2286.

[39] H. Dai, A. Majumdar, and R. Tedrake, "Synthesis and optimization of force closure grasps via sequential semidefinite programming," *Robotics Research: Volume 1*, pp. 285–305, 2018.

[40] L. Downs, A. Francis, N. Koenig, B. Kinman, R. Hickman, K. Reymann, T. B. McHugh, and V. Vanhoucke, "Google scanned objects: A high-quality dataset of 3d scanned household items," 2022.

[41] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[42] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[43] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[44] M. Liu, Z. Chen, X. Cheng, Y. Ji, R. Yang, and X. Wang, "Visual whole-body control for legged loco-manipulation," *arXiv preprint arXiv:2403.16967*, 2024.

[45] Y. Yuan, J. Song, U. Iqbal, A. Vahdat, and J. Kautz, "Physdiff: Physics-guided human motion diffusion model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 010–16 021.