Benchmarking Multi-Object Grasping

Tianze Chen, Ricardo Frumento, *Graduate Student Member, IEEE*, Giulia Pagnanelli, *Graduate Student Member, IEEE*, Gianmarco Cei, *Graduate Student Member, IEEE*, Villa Keth, Shahaddin Gafarov, Jian Gong, Zihe Ye, *Graduate Student Member, IEEE*, Marco Baracca, *Student Member, IEEE*, Salvatore D'Avella, *Member, IEEE*, Matteo Bianchi, *Member, IEEE*, and Yu Sun, *Senior Member, IEEE*

***Abstract*—In this work, we describe a multi-object grasping benchmark to evaluate the grasping and manipulation capabilities of robotic systems in both pile and surface scenarios. The benchmark introduces threerobot multi-object grasping benchmarking protocols designed to challenge different aspects of robotic manipulation. These protocols are: 1) the Only-Pick-Once protocol, which assesses the robot's ability to efficiently pick multiple objects in a single attempt; 2) the Accurate pick-transferring protocol, which evaluates the robot's capacity to selectively grasp and transport a specific number of objects from a cluttered environment; and 3) the Pick-transferring-all protocol, which challenges the robot to clear an entire scene by sequentially grasping and transferring all available objects. These protocols are intended to be adopted by the broader robotics research community, providing a standardized method to assess and compare robotic systems' performance in multi-object grasping tasks. We establish baselines for these protocols using standard planning and perception algorithms on a Barrett hand, Robotiq parallel jar gripper, and the Pisa/IIT Softhand-2, which is a soft underactuated robotic hand. We discuss the results in relation to human performance in similar tasks as well.**

***Index Terms*—Benchmarking, dexterous manipulation, grasping, multifingered hands.**

I. Introduction

ROBOTIC grasping is a key area of research in robotics and automation. Traditional robotic grasping methods typically handle one object at a time. If a task requires handling multiple items, the robot must repeat the process for each one. In contrast, humans can naturally grasp several items simultaneously. For example, a person might grab a handful of apples from a basket or multiple screws from a tray, depending on the need and what their hand can hold. In logistics and manufacturing, workers often perform batch picking, picking up several small items or containers and putting them into a bin to speed up the processes. However, enabling robots to grasp and hold multiple objects poses significant challenges. This requires more sophisticated techniques in perception and planning.

Grasping tasks could have multiple scenarios. Among them, surface picking and pile picking are the most common. Surface picking involves grasping objects from a flat surface, while pile picking deals with stacked objects. The flat surface could be a tabletop, a conveyor belt, or the bottom of a bin. The object pile could be in a bin, a bag, or a jar. The objects could be identical products of the same size, such as a box of the same candies or cosmetic jars with the same Stock Keeping Unit (SKU). They could also be the same organic object with the same shape but slightly different sizes, such as a bowl of eggs or a bag of apples. They could be a collection of different objects, such as a tabletop of random objects for tabletop arrangement tasks, a shopping basket of items, and a tote of items with different SKUs before re-binning in a warehouse.

Several benchmarks have been proposed to measure the performance of robotic systems for picking one item at a time in many of those scenarios. In [1], a benchmark was designed for measuring the performance of picking and transferring a one-inch cube at a time from a pile of cubes in a box. In [2], a benchmark was developed to assess object rearrangement efficiency of picking and placing one item at a time. In [3], a benchmark was defined to measure the efficiency of picking and transferring a list of items from a tote, one item at a time. Several competitions have also been organized based on those benchmarks to measure and showcase the research development of single-object grasping in those scenarios [2], [4], [5], [6]. The protocols and rules in those benchmarkers particularly forbid picking up multiple items at a time.

In concept, many single-object-oriented grasping benchmarks could be expanded to measure the performance of grasping multiple objects as they may share the same goal: accurately picking a number of objects efficiently. Allowing robots to pick up multiple items at once opens up many opportunities for efficiency improvement. They naturally introduce many factors that are not counted for in those single-object-grasping benchmarks. For example, strategizing the grasping sequence becomes critically important, and grasping accuracy in one round may not be as critical. There is a clear need for a set of

Received 14 March 2025; accepted 7 July 2025. Date of publication 4 August 2025; date of current version 12 August 2025. This article was recommended for publication by Associate Editor N. Rojas and Editor J. Borràs Sol upon evaluation of the reviewers' comments. This work was supported in part by USF CREATE AWARD and in part by the Italian Ministry of Education and Research (MUR) within the framework of the FoReLab Project (Departments of Excellence) . *(Corresponding author: Tianze Chen.)*

Tianze Chen, Ricardo Frumento, Villa Keth, Shahaddin Gafarov, Jian Gong, and Yu Sun are with the Robot Perception and Action Lab (RPAL) of Computer Science and Engineering Department, University of South Florida, Tampa, FL 33620 USA (e-mail: tianzechen@usf.edu; ricardod@usf.edu; villak@usf.edu; gafarov@usf.edu; gongj@usf.edu; yusun@usf.edu).

Giulia Pagnanelli, Gianmarco Cei, Marco Baracca, and Matteo Bianchi are with the Research Center "E. Piaggio", and the Department of Information Engineering, University of Pisa, 56126 Pisa, Italy (e-mail: giulia.pagnanelli@phd.unipi.it; gianmarco.cei@phd.unipi.it; marco.baracca@phd.unipi.it; matteo.bianchi@unipi.it).

Zihe Ye is with the Rutgers University, New Brunswick, NJ 08901 USA (e-mail: zihe.ye@rutgers.edu).

Salvatore D'Avella is with the Department of Excellence in Robotics & AI, Institute of Mechanical Intelligence, Scuola Superiore Sant'Anna, 56127 Pisa, Italy (e-mail: salvatore.davella@santannapisa.it).

This article has supplementary downloadable material available at https://doi.org/10.1109/LRA.2025.3595043, provided by the authors.

Digital Object Identifier 10.1109/LRA.2025.3595043

2377-3766 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Scuola Superio Sant'Anna di Pisa. Downloaded on November 26,2026 at 11:17:24 UTC from IEEE Xplore. Restrictions apply.

new benchmarks to unpack the influence of the factors and properly measure the progress of robotics research for multi-object grasping.

II. Related Work

In addition to the benchmarking works on single-object grasping mentioned in the Introduction, we introduce here the related advancements in Multi-Object Grasping (MOG). Harada and Kaneko [7], [8] explored multi-object grasping using robotic hands, focusing on enveloping grasp and active force closure (AFC). They analyzed kinematic conditions for stable grasping, ensuring rolling contacts at interfaces, and derived mathematical models to validate AFC feasibility. Later, [9] optimized power grasps by maximizing stability while minimizing gripping force, validated via simulations. [10] derived stability conditions for grasping two objects in 2D based on geometry and contact points, while [11] extended this to 3D, modeling robotic fingers as linear springs and incorporating contact surface geometry. These studies collectively provide theoretical and simulation-based foundations for stable multi-object grasping.

Recent methods extend these ideas to more complex scenarios. [12] proposed a multi-object push-grasping approach for rigid convex polygonal objects on a planar surface. They established the necessary conditions for frictionless push-grasps and developed a grasp planner to filter inadmissible grasps. Expanding on this, [13] introduced Push-MOG, which used "fork pushing" actions with a parallel-jaw gripper to cluster objects, facilitating multi-object grasping. Push-MOG employs hierarchical clustering to identify clusters and executes stable pushes to optimize grasp configurations. Other methods incorporate learning-based approaches. MOG-Net [14] plans robust grasps for multiple convex polygonal objects by integrating frictional interactions. [15] proposed a method for multi-fingered robotic hands to grasp multiple objects by leveraging kinematic redundancy. Their algorithm uses pairwise contacts and a reachability map to optimize stable grasps, validated on a robotic hand. The work in [16] tackled the problem of estimating the number of objects within a robotic grasp when the hand is submerged in a pile, highlighting the importance of precise perception. Similarly, [17] analyzed human and robotic motion data to define 12 distinct grasping types, providing a structured foundation for improving grasp strategies based on real-world insights.

For scenarios involving transferring objects, the transferring algorithm based on Markov decision processes (MDP) [18] ensures consistent and efficient execution. The Experience Forest algorithm [19] enhanced robotic precision by organizing finger movements into structured patterns, enabling more effective interactions with objects. The Only-Pick-Once (OPO) system [20] demonstrated significant improvements in efficiency through optimized grasp-and-transfer processes. [21] introduces an imitation learning-based approach using a diffusion policy network to enable robots to dynamically generate action sequences for pushing, grouping, and grasping multiple objects simultaneously. Advanced systems and designs have further elevated MOG performance. A multimodal gripper system integrating a suction cup was proposed by [22], allowing for enhanced flexibility in handling different types of objects. Human-inspired designs like MOGrip [23] emulate natural grasping abilities, enabling sequential object handling, secure storage, and smooth transfer in a single operation, setting new standards for robotic dexterity and adaptability.

III. Benchmarking Protocols

Standardized benchmarking is essential for evaluating and comparing robotic systems in multi-object grasping, helping to identify strengths, address limitations, and set performance goals. Consistent metrics and protocols improve reproducibility and enable meaningful comparisons, advancing robust and adaptable real-world systems. This paper introduces three protocols for benchmarking MOG in surface and pile scenarios with identical objects. The only-pick-once (OPO) protocol measures performance in grasping multiple objects in a single attempt. The accurate pick-transferring (APT) protocol builds on OPO to evaluate the efficiency of picking and transferring a specific number of objects. The pick-transferring-all (PTA) protocol assesses the efficiency of transferring all items from one container to another. All protocols involve a robotic manipulator operating in both surface and pile conditions.

a) Robot Setup: The experimental setup consists of a robotic manipulator positioned in front of a support surface, with an RGB-D camera mounted either above or on the robotic arm to enable real-time object detection and pose estimation. The camera captures depth and color data, which vision algorithms process to identify object locations and orientations within the source bin, facilitating grasp planning and execution. In some experiments, a simulated counterpart replicates the real-world setup, maintaining identical relative positions and dimensions to ensure consistency in benchmarking.

b) Object Setup and Scenarios: Both the source and target bins are placed within the robot's reachable workspace, with their exact positioning left to the user's discretion. The source bin contains objects arranged in either piles of identical items or a single-layer grouping of identical objects. This physical setup remains consistent across all three experimental protocols to ensure uniform evaluation. Further details on the baseline implementation and specific setup configurations will be provided in subsequent sections.

A. OPO Protocol

Only-pick-once (OPO) protocol is designed to reflect and standardize the grasping process of picking tasks where picking up an exact number of items is the goal. For example, in warehouses, workers need to fulfill an order of multiple identical items, such as cosmetic jars, USB drives, or similar small components.

1) Goal: The objective of the OPO protocol is to measure the accuracy and efficiency of the evaluated approach while only allowing the robot to perform one round of picking, which includes approaching, forming pre-grasp, grasping/re-grasping, and lifting. It directly measures how well a robotic system can grasp a specific number of objects in a single round of picking.

2) Procedure:

- Provide a target number p to the robotic system, start the timer, and record t_0.
- When the approaching ends, record t_1, and report the approaching time $t_a = t_1 - t_0$.
- When the robotic arm starts to rise, record t_2, and report the grasping/regrasping time $t_g = t_2 - t_1$.
- When the robotic arm stops rising, record t_3, and report the lifting time $t_l = t_3 - t_2$.[1]

[1]For our setup, the robot hand is 70cm above the objects and $t_a = t_l = 3.5$s

Authorized licensed use limited to: Scuola Superio Sant'Anna di Pisa. Downloaded on November 26,2026 at 11:17:24 UTC from IEEE Xplore. Restrictions apply.

- Wait three seconds.
- Exam how many objects are in the robotic hand, report it as q.
- Put objects back into the container and shuffle.
- Repeat 100 times.
- Perform statistical analysis with baseline results.

The total time cost of finishing one OPO procedure is $t_{OPO} = t_a + t_g + t_l$.

B. APT Protocol

Accurate pick-transferring (APT) protocol represents a frequent and practical task in warehouse automation and logistics environments, designed to fulfill requirements where the target quantity of objects exceeds the maximum capacity of a robot's gripper in a single grasp. For instance, imagine a situation where a robot is assigned to move a certain number of apples into a crate for shipping. Since the robotic hand can only grasp several apples at once, the robot must repeat the process a few times to reach the desired amount.

1) Goal: The APT protocol measures the efficiency of a robotic system in grasping and transferring objects across multiple attempts, allowing for sequential picks when the object count exceeds the gripper's capacity. It evaluates the system's ability to plan and execute actions for each grasp, ensuring efficiency. Essentially, it extends the OPO protocol by incorporating multiple rounds of picking to achieve the goal.

2) Procedure:

- Provide a target number N_{target} (usually more than 5) to the robotic system.
- Select a target number p of objects to grasp for the current round.
- Perform an OPO round without shuffling and record the t_{OPO} and the number of objects grasped q.
- Transfer the objects to the target bin and record the transfer time $t_{transfer}$ (including the return to the start position for the next round).
- Update the target with N_{target} - q.
- Repeat this process until the target is zero. Record the number of performed OPO procedures m_{OPO}
- Put objects back into the container and shuffle.
- Repeat 100 times.
- Perform statistical analysis with baseline results.

In this procedure, if the number of objects in the target bin exceeds the intended amount, the excess objects must be picked up from the target bin and returned to their original bin. The additional time taken for this step should also be included. The total time cost of finishing one APT procedure is $t_{APT} = m_{OPO}(t_{OPO} + t_{transfer})$. If, at a particular round, the $p = 1$, an SOP procedure will be performed and the time cost will be assigned to t_{OPO} for that round.

C. PTA Protocol

Pick-transferring-all (PTA) protocol represents a highly challenging task that demands precise planning and adaptability to handle difficult grasping scenarios. This task is particularly complex when the system must clear objects with similar shapes and sizes, which can create additional challenges such as dense packing, limited distinguishability, or unpredictable object interactions. For example, consider a scenario in a warehouse where a robot is tasked with clearing a bin filled with uniformly shaped and sized plastic cups stacked tightly together. The robot must plan its actions to efficiently remove all the cups while addressing potential difficulties such as cups slipping, deforming, or sticking to one another due to friction or static.

1) Goal: The primary goal of the PTA protocol is to measure the efficiency of grasping and transferring all objects from the source bin to the target bin through a series of picks and transfers, aiming to optimize operational efficiency and minimize handling time. This protocol evaluates the system's ability to manage a complete clearing task, requiring effective planning and execution across multiple grasping attempts. By integrating aspects of both OPO (focusing on precise object counts) and APT (handling larger quantities across several attempts), the PTA protocol challenges the robot's capacity for sequential task planning, adaptability, and consistent performance in dynamic environments.

2) Procedure:

- Select a target number p of objects to grasp for the current round.
- Perform an OPO round without shuffling and record the t_{OPO} and the number of objects grasped q.
- Transfer the objects to the target bin and record the transfer time $t_{transfer}$.
- Repeat this process until the original bin is empty. Record the number of performed OPO procedures m_{OPO}
- Put objects back into the container and shuffle.
- Repeat 100 times.
- Perform statistical analysis with baseline results.

The total time cost of finishing one PTA procedure is $t_{PTA} = m_{OPO}(t_{OPO} + t_{transfer})$. If, at a particular round, the $p = 1$, an SOP procedure will be performed and the time cost will be assigned to t_{OPO} for that round.

D. Evaluation Metrics

We propose several evaluation metrics using the recorded values in the protocol procedures to comprehensively assess the accuracy, reliability, and efficiency of MOG techniques. They are picking accuracy (PA), overall success rate (OSR), availability rate (AR), and cost of grasping per unit (CGPU).

Picking Accuracy (PA) is the normalized root mean square error (RMSE). It measures the accuracy of a picking outcome compared to the desired, making it particularly suitable for evaluating grasping performance in OPO Protocol. We define PA in (1), where N_{total} is the total number of grasp trials performed, $target$ is the desired number of objects to be grasped, and o_i is the number of objects grasped in the i_{th} grasping trial. It serves as a performance indicator, particularly for tasks that can accommodate minor errors—such as grabbing a few ice cubes to toss into a drink. It can also be applied to analyze how performance impacts subsequent tasks, as measured by APT.

$$PA(object, target) = \sqrt{\frac{\sum_{n=1}^{N_{total}} (o_i - target)^2}{N_{total}}}, \quad (1)$$

Overall Success Rate (OSR) is an intuitive and widely adopted metric in robotic grasping evaluations, representing the proportion of successful grasps relative to the total number of attempts. It is defined in (2), where $N_{success}$ is the number of grasp trials achieving the target. It is also a good measurement for evaluating grasping performance in OPO Protocol. It serves as a performance indicator, particularly for tasks that can high degree

Authorized licensed use limited to: Scuola Superio Sant'Anna di Pisa. Downloaded on November 26,2026 at 11:17:24 UTC from IEEE Xplore. Restrictions apply.

of accuracy is critical - such as order fulfillment. As PA, it can also be applied to analyze how performance impacts subsequent tasks, as measured by APT.

$$OSR(object, target) = \frac{N_{success}}{N_{total}}, \tag{2}$$

Cost of Grasping per Unit (CGPU) is designed to measure the efficiency of an MOG process. It directly compares the MOG process with the SOG process. In this metric, we assume the SOG process is always perfect, and it takes exactly the number of picking-transferring as the targeted number of objects, k. CGPU can be calculated using the average t_{PTA} or t_{APT} normalized by the time of applying single-object picking and transferring (t_{sopt}).

$$CGPU = \frac{t_{PTA/APT}}{kt_{sopt}} \tag{3}$$

If we assume each round of OPO procedure uses the same time as a round of SOP, the CGPU can be simplied as

$$CGPU_s = \frac{N_{OPO}}{k}, \tag{4}$$

where N_{OPO} represents the total OPO rounds the system executes for PTA or APT, including both successful and failed grasping trials, to obtain the target number of k objects. For example, for picking and transferring five apples using MOG, the PTA procedure has two rounds of OPOs. s is short for simplified. The $CGPU_s = \frac{2}{5}$.

Since there could be overhead in each OPO process caused by extra steps such as regrasping introduced in [19] to improve PA. Therefore, the CGPU contains

$$CGPU_o = \lambda_o CGPU_s \tag{5}$$

$CGPU_o$ represents the overhead in the MOG process that is not in the SOP process. The average extra cost for each picking-transferring measured t_{PTA} or t_{PTA} compared to SOP is normalized by the time SOP and defined as λ_o.

However, since t_{PTA} and t_{PTA} contain transferring time $t_{transfer}$, it could be affected by the setup. To remove the variances in the setup, we define the adjusted CGPU as

$$CGPU_{ao} = \lambda_{ao} CGPU_s \tag{6}$$

where $\lambda_{ao} = \frac{t_{OPO} - t_{SOG}}{t_{SOG}}$. Thus the total CGPU can be represented as:

$$CGPU_{total} = (1 + \lambda_o + \lambda_{ao}) CGPU_s \tag{7}$$

For CGPU and its variations, a score closer to 0 indicates better performance because it means the MOG process is more efficient relative to the theoretical SOG baseline. CGPU aims to quantify how efficient multi-object grasping is compared to standard one-by-one grasping, considering both the number of attempts needed and the time spent in each attempt and subsequent transfer.

In certain scenarios, it is important to evaluate the system's capability to effectively identify and group objects for MOG. To address this, we introduce the *Availability Rate (AR)* as a performance metric. AR measures the percentage of arbitrary scenes where the system determines it is feasible to grasp k objects simultaneously. This metric provides insight into the system's ability to locate and group objects based on their spatial arrangement and graspability. A higher AR indicates that the system can efficiently identify opportunities for MOG.

TABLE I
SUMMARY OF MOG BENCHMARKS

MOG Protocol	Goal	Evaluation Metrics
OPO	Grasp a small, specific target number of objects in a single attempt.	OSR, RMSE, AR, CGPU
APT	Grasp and transfer a large, specific target number of objects in multiple attempts.	$CGPU_{adjusted}$
PTA	Clear all objects from the environment in multiple grasps.	$CGPU_{adjusted}$

Our chosen metrics build upon and extend those commonly found in SoA benchmarks. Metrics like RMSE and OSR are rooted in traditional single-object grasping evaluations, but their adaptation to the multi-object context sets them apart. For example, benchmarks such as the YCB Object [24] and the ACRV Picking Benchmark [25] prioritize success rates and object-handling efficiency but often lack metrics that address the cumulative challenges of multi-object scenarios. By introducing $CGPU$ and $CGPU_a$, we address a critical gap in SoA benchmarks, enabling a comprehensive assessment of both effectiveness and efficiency in MOG tasks. Summary of MOG benchmarks mentioned above can be found in Table I.

E. Environment

For MOG benchmarking, we use six fundamental convex shapes—sphere, cube, cuboid, cylinder, cone, and ellipsoid, which are shown in Table II. These shapes are common in grasp datasets such as YCB [24] and frequently appear in industrial applications across manufacturing, logistics, and assembly. This selection ensures alignment with standard benchmarks and relevance to real-world robotic tasks.

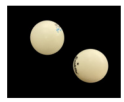
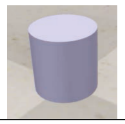
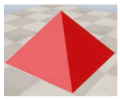
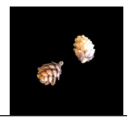
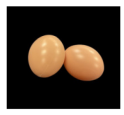
We carefully considered object size variations in both simulation and real-world experiments to ensure comprehensive evaluation and practical relevance. In simulation, each basic shape was tested in three sizes—small, medium, and large—allowing us to assess performance across a broad range of dimensions. For real-world testing, we selected two representative sizes per shape, balancing variability with practical constraints. Object dimensions were inspired by the YCB dataset [24], ensuring consistency across simulated and physical benchmarks. Simulated object sizes ranged from approximately 30mm to 40mm, while real-world counterparts extended slightly beyond (e.g., spheres from D40 to D52mm) to reflect realistic handling scenarios. This sizing strategy helped bridge the gap between controlled simulations and real-world applications, enhancing the benchmark's robustness and applicability.

IV. BASELINES

To illustrate how to use the protocols and facilitate comparison, we present the implementation of the protocols on three robotic systems that represent the mainstream pick-and-place

Authorized licensed use limited to: Scuola Superio Sant'Anna di Pisa. Downloaded on November 26,2026 at 11:17:24 UTC from IEEE Xplore. Restrictions apply.

TABLE II
OBJECTS USED FOR GRASPING

Basic Shape	picture	size 1(mm)	size 2(mm)	size 3(mm)	real object 1	size(mm)	real object 2	size(mm)
Cube		cube_s_s: 30	cube_m_s: 35	cube_l_s: 40		cube_l_r: 55 weight: 9g		cube_s_r: 25 weight: 9.5g
Sphere		sphere_s_s: D30	sphere_m_s: D35	sphere_l_s: D40		sphere_l_r: D52 weight: 24g		sphere_s_r: D40 weight: 2.7g
Cylinder		cylin_s_s: D30 H30	cylin_m_s: D35 H35	cylin_l_s: D40 H40		cylin_l_r: D26 H63 weight: 5.2g		cylin_s_r: D36 H30 weight: 11g
Cuboid		cuboid_s_s: L30 W10 H30	cuboid_m_s: L40 W20 H40	cuboid_l_s: L50 W30 H50		cuboid_l_r: L107 W37 H29 weight: 34.2g		cuboid_s_r: L71 W16 H44 weight: 37g
Cone		cone_s_s: D10 H30	cone_m_s: D15 H35	cone_l_s: D20 H40		cone_l_r: D8 H17 weight: 1.1g		cone_s_r: D40 H56 weight: 21.6g
Ellipsoid		ellip_s_s: L30 S20	ellip_m_s: L35 S25	ellip_l_s: L40 S30		ellip_l_r: L60 S45 weight: 3.2g		ellip_s_r: L33 S21 weight: 9g

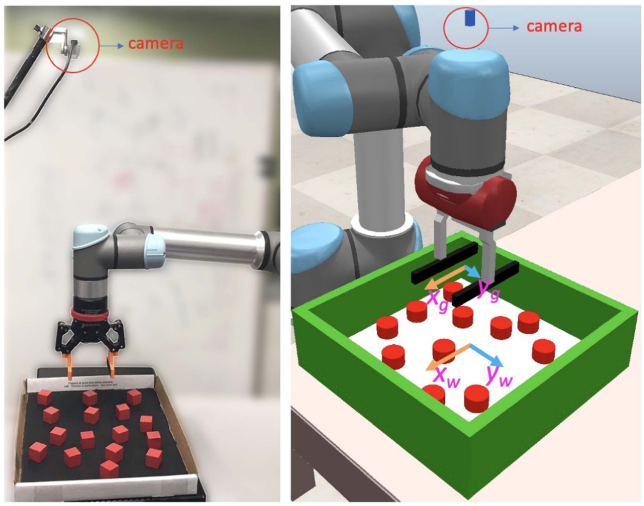


Fig. 1. Set up for the Protocol 1 MOG experiments with the Robotiq.

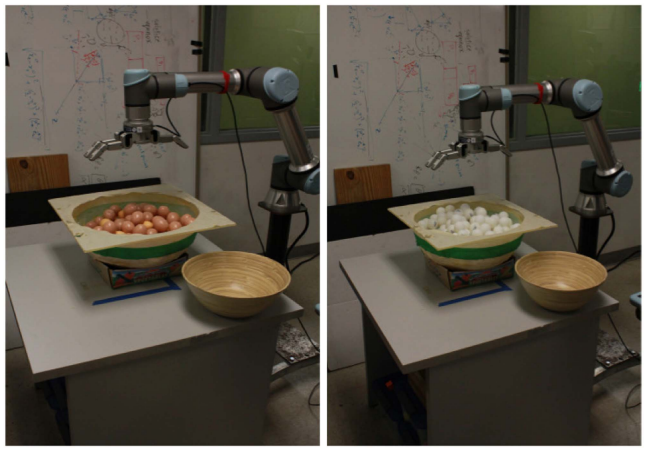


Fig. 2. Set up for the Protocols 1 & 2 MOG experiments with the Barretthand.

setups: parallel gripper, rigid dexterous hand, and soft hand. We carry out the protocols with those robotic systems and our latest algorithms and present the results as baselines. We also carry out the protocols with 5 human participants and present the human performances and a baseline.

a) Setup-Robotiq: We set up a typical hand-eye system with a Robotiq 2F-85 gripper, a UR5e arm, and an Intel RealSense Depth Camera D435, widely used in single-object pick and place applications. It is shown in Fig. 1. We also duplicated the setup in CoppeliaSim for simulation.

b) Setup-Barrett: With the same UR5e robotic arm and the camera, we replace the gripper with a Barrett hand as shown in Fig. 2. It represents a traditional grasping and manipulation system with dexterous in-hand manipulation capability. The hand has three fingers and four DoFs. It has tactile sensors on the fingertips and palm to perceive complex grasping patterns in real-time.

c) Setup-SoftHand We also implement a robotic system with a PISA/IIT SoftHand-2 mounted on a Franka Emika Robotic, with

Authorized licensed use limited to: Scuola Superio Sant'Anna di Pisa. Downloaded on November 26,2026 at 11:17:24 UTC from IEEE Xplore. Restrictions apply.

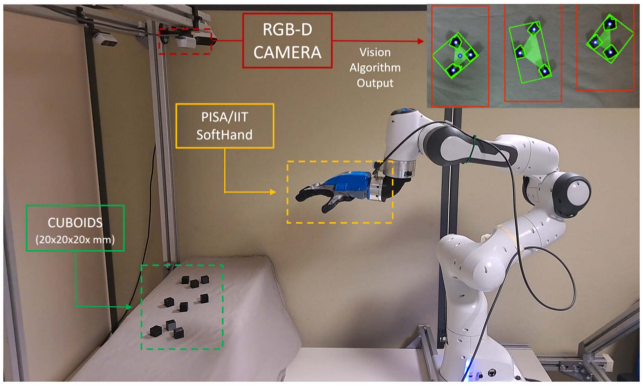


Fig. 3. Set up for the Protocol 1 MOG experiments with the PISA/IIT Softhand-2. Objects are placed in clusters on the table in front of the robot, to test the feasibility and repeatability of grasping 3 at a time.

an Intel RealSense Depth Camera D455, as shown in Fig. 3. The SoftHand-2 is an anthropomorphic, underactuated soft gripper, featuring 19 DoFs and 2 degrees of actuation, coordinated finger motions called "synergies" enabled by the elastic ligaments between the phalanxes [26]. These synergies are designed to imitate the most commonly observed behaviors in human hands: the first synergy realizes the opening and closing of the thumb and fingers, while the second operates a relative rotation between the fingers [27].

For the setups employing Robotiq and Barretthand, we directly apply the grasping algorithms proposed in our previous works [19], [20] [2]. For Setup-SoftHand, we developed the following grasping strategy[3]. Since the objects rest on a table, the strategy begins by clustering cubes using depth data from the camera. Constrained K-Means groups K objects based on planar proximity. Each cluster undergoes a feasibility check: using hand dimensions and camera intrinsics, a rectangle is projected over the cluster to verify if the SoftHand-2 can cover all objects without collision. If a valid cluster exists, it is selected; otherwise, the robot manually forms a feasible cluster by identifying and grouping the K closest cubes through three planned movements. Once a valid group is formed, the robot approaches with the hand open, parallel to the table, and properly oriented for maximum coverage. It then configures the hand using its second synergy to enclose the group, descends while avoiding collisions, and drags the objects to the grasping pose before closing the hand—mimicking how a person sweeps crumbs off a table. Key parameters—like grasp poses, object spacing, and motion paths—must be adapted to the specific hand's morphology and kinematics. While the clustering algorithm is generalizable, actions like selecting feasible clusters or planning rearrangement are tailored to the end-effector's capabilities.

A. OPO Baselines

The results of MOG under Protocol OPO are shown in Fig. 4. In the simulated setup, the Robotiq hand achieved high OSR values when grasping two objects across different object shapes, such as 97.5% for $cube_m_s$, 96.5% for $cyli_m_s$, and 98.5% for $sphere_s_s$. However, OSR values decreased significantly as the target number increased. For example, OSR for $cube_m_s$ dropped from 97.5% when grasping 2 objects to 6.5% when grasping 4 objects, reflecting the challenges in maintaining grasp success as the number of objects increases. The results for the Robotiq hand in the real system highlight differences in performance across object numbers. For $cylin_l_r$, with OSRs of 81.25% for 2 objects, 40.00% for 3 objects, and only 10.00% for 4 objects. Overall, the Robotiq hand shows strong grasping capabilities for a smaller number of objects but struggles as the task complexity and target numbers increase.

The real setup evaluated OSR and RMSE with the PISA/IIT SoftHand-2 grasping 2 and 3 objects of cube and cylindrical shapes. OSR with 3 objects was greater with $cube_s_r$ and equal with $cylin_s_r$ with respect to 2 objects. RMSE with 3 $cube_s_r$ was higher than using 2 $cube_s_r$, while with $cylin_s_r$ RMSE achieved a bit less score with 3 objects than considering 2 objects. This can be explained by accounting for the higher intra-forces between the objects in the case of the cubes with respect to the cyilinders when grasped by the compliant hand, which favors successful grasping of a higher object count in the first case with respect to the second. Although its compliant behavior can prevent the execution of precisely repeatable tasks, SoftHand intrinsic compliance allows it to adapt to various items and leverage environmental constraints, fostering MOG applications in heavily unstructured environments.

In the simulated setup, the Barrett Hand demonstrated varying levels of OSR and RMSE, with OSR decreasing as the target number increased. For example, the $sphere_m_s$ object type achieved an OSR of 63% when grasping 2 objects, which dropped to 55% for 3 objects, with corresponding RMSE values of 0.7746 and 0.9798, respectively. Similar trends were observed with other shapes, such as $cuboid_m_s$.

In the real setup, OSR values were generally higher for tasks involving 2 objects. For instance, the $sphere_m_r$ object achieved an OSR of 71% with an RMSE of 0.6232 for 2 objects, compared to 60% with an RMSE of 0.8521 for 3 objects. These results highlight the system's effectiveness at handling smaller target numbers, with relatively consistent accuracy and precision.

For the task of grasping two $sphere_m_s$ objects under the OPO procedure, the proposed MOG algorithm in [19] achieves a mean performance of 1.74 with a variance of 0.518. In comparison, the naive baseline algorithm in [19] yields a lower mean of 1.43 and a higher variance of 0.712 for the same task. The Student's t-test reveals a p-value of 0.0057, indicating that the improvement in performance of the proposed algorithm over the baseline is statistically significant. This example demonstrates how statistical analysis can be used to validate algorithmic performance. When comparing more than two groups, ANOVA provides an effective alternative to assess significance across multiple conditions.

These results show that all three robotic hands perform well on simpler MOG tasks but struggle as object count increases. Robotiq excels in simulation with fewer objects, SoftHand-2 maintains accuracy but lower OSR in real settings, and Barrett Hand performs reliably at low counts but loses precision with more objects. This highlights the need to match system capabilities to task complexity.

[2]Starter code for the Barrett hand scene can be found at https://github.com/RPAL-USF/Benchmarking-Starter-Code.git here

[3]Toolkit for the SoftHand controller can be found at https://github.com/CentroEPiaggio/panda-softhand.git here and https://wiki.ros.org/qb_hand_control here

Authorized licensed use limited to: Scuola Superio Sant'Anna di Pisa. Downloaded on November 26,2026 at 11:17:24 UTC from IEEE Xplore. Restrictions apply.

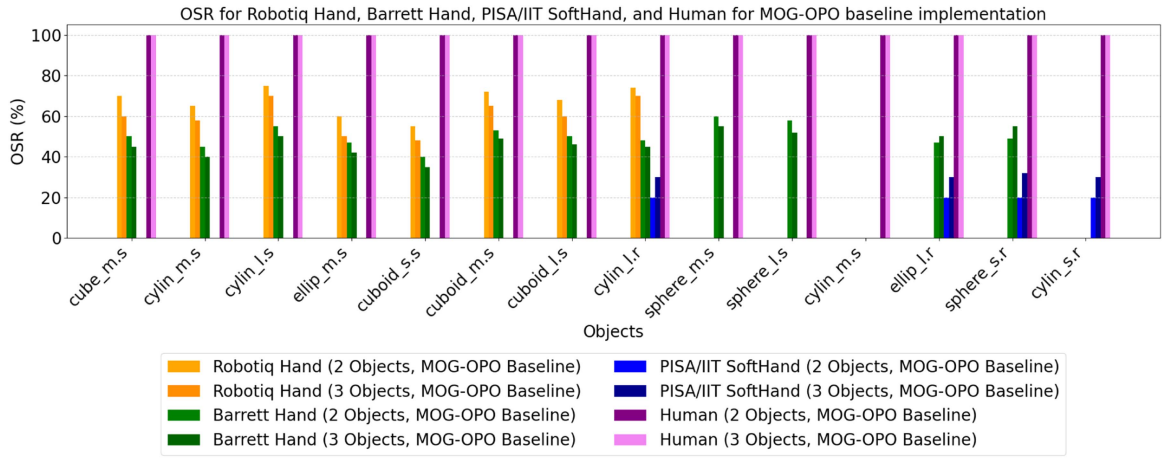


Fig. 4. OSR for robotiq hand, barrett hand, PISA/IIT softhand-2, and human performance when targeting 2 and 3 objects.

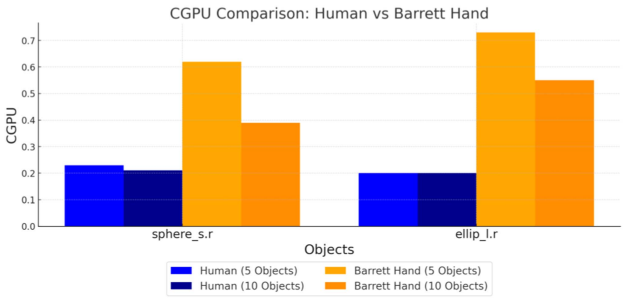


Fig. 5. CGPU comparison between human and barrett hand for grasping 5 and 10 objects.

B. APT Baselines

The results for Protocol APT, targeting 5 and 10 objects, are shown in Fig. 5. In the simulated setup, the Barrett Hand's performance improved as the target number of objects increased. For instance, the $sphere_m_s$ object type achieved $CGPU_{adjusted}$ values of 0.70 for 5 objects and 0.52 for 10 objects. Similar trends were observed with the $cuboid_m_s$ object type, which recorded CGPU values of 0.79 for 5 objects and 0.71 for 10 objects.

In the real setup, $CGPU_{adjusted}$ values also decreased with higher target numbers. For example, $sphere_m_r$ achieved a $CGPU_{adjusted}$ of 0.62 for 5 objects, which dropped to 0.39 for 10 objects. Similar patterns were observed with the $ellip_l_r$ object type, where $CGPU_{adjusted}$ values ranged from 0.61 for 5 objects to 0.51 for 10 objects.

The decreasing CGPU values show that the Barrett Hand becomes more time-efficient as the target object count increases. In APT, grasp success rate matters less—objects are transferred until the target is met, making CGPU a practical measure of overall efficiency. However, the task remains challenging, as real-world clutter, object interactions, and sensor noise still hinder performance, underscoring the need for continued advancements in system design and control.

C. PTA Baselines

For this protocol, we only have human experiments, since it is challenging for the current robotic system to achieve. PTA demands a level of robustness, adaptability, and consistent performance in cluttered, dynamic environments that exceeds the current capabilities demonstrated by robotic systems. Especially for Barrett hand, in the later stages when specific perception, planning, or control challenges arise from the reduced number of objects in potentially difficult locations. The human baseline setup uses two bowls for object placement and task execution. Thirteen objects were used, with five subjects performing multiple trials across all protocols: OPO (target 2 and 3), transferring 5 or 10 objects, and scene clearing. Subjects had an average hand length of 18cm and width of 9cm. In cases with only 10 objects, the 10-object transfer and scene-clearing tasks overlap. Results in Figs. 4 and 5 provide an upper-bound comparison for robotic performance.

V. Discussion and Conclusion

Standardized benchmarking is essential for consistent evaluation of robotic MOG systems. We propose three protocols—OPO, APT, and PTA—targeting grasp accuracy, sequential transfer, and large-scale clearance. These protocols offer a comprehensive framework for assessing performance across varied real-world scenarios, promoting reproducibility and advancement in robotic grasping.

Our protocols address key limitations of prior SOG benchmarks by introducing evaluation metrics tailored to the unique challenges of MOG. While SOG benchmarks focus on isolated object performance, our MOG framework provides a more comprehensive evaluation through four key metrics—PA, OSR, CGPU, and AR—that collectively capture the complexity, scalability, and reliability required in cluttered, high-object-count scenarios. PA evaluates accuracy, OSR and AR measure consistency and robustness across trials, and CGPU quantifies efficiency at scale under realistic conditions. These standardized protocols are essential for enabling fair, reproducible comparisons across systems and for driving progress toward algorithms capable of handling real-world, multi-object manipulation tasks. Although evaluating generalization remains challenging and is beyond the scope of this paper, we have explored this direction in prior work [19], [20] by varying object sizes and introducing novel shapes. Additional factors such as mass and friction represent further opportunities for extending the benchmark's comprehensiveness in future studies.

Authorized licensed use limited to: Scuola Superio Sant'Anna di Pisa. Downloaded on November 26,2026 at 11:17:24 UTC from IEEE Xplore. Restrictions apply.

MOG baseline results reveal that robotic hands perform well in simple tasks but struggle with clutter and irregularity, unlike humans who remain efficient. This underscores the need for progress in perception, control, and learning. Vision systems face occlusion and segmentation issues, highlighting the importance of sensor fusion. Adaptive control and compliant hands like SoftHand-2 are needed for dynamic environments. Learning must move beyond single-object tasks, incorporating hierarchical reinforcement and domain adaptation. Our benchmarks help direct these future improvements.

REFERENCES

[1] A. S. Morgan et al., "Benchmarking cluttered robot pick-and-place manipulation with the box and blocks test," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 454–461, Apr. 2020.

[2] Z. Liu et al., "OCRTOC: A cloud-based competition and benchmark for robotic grasping and manipulation," *IEEE Robot. Automat. Lett.*, vol. 7, no. 1, pp. 486–493 2022, doi: 10.1109/LRA.2021.3129136.

[3] H. Mnyusiwalla et al., "A bin-picking benchmark for systematic evaluation of robotic pick-and-place systems," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 1389–1396, Apr. 2020.

[4] N. Correll et al., "Analysis and observations from the first amazon picking challenge," *IEEE Trans. Automat. Sci. Eng.*, vol. 15, no. 1, pp. 172–188, Jan. 2018.

[5] Y. Sun, J. Falco, M. A. Roa, and B. Calli, "Research challenges and progress in robotic grasping and manipulation competitions," *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 874–881, 2022, doi: 10.1109/LRA.2021.3129134.

[6] S. D'Avella, M. Bianchi, A. M. Sundaram, C. A. Avizzano, M. A. Roa, and P. Tripicchio, "The cluttered environment picking benchmark (CEPB) for advanced warehouse automation: Evaluating the perception, planning, control, and grasping of manipulation systems," *IEEE Robot. Automat. Mag.*, vol. 31, no. 4, pp. 45–58, Dec. 2024.

[7] K. Harada and M. Kaneko, "Enveloping grasp for multiple objects—Kinematics and shovelling up condition—," *J. Robot. Soc. Jpn.*, vol. 16, no. 6, pp. 860–867, 1998.

[8] K. Harada, M. Kaneko, and T. Tsuji, "Active force closure for multiple objects," *J. Robotic Syst.*, vol. 19, no. 3, pp. 133–141, 2002.

[9] T. Yoshikawa, T. Watanabe, and M. Daito, "Optimization of power grasps for multiple objects," in *Proc. 2001 IEEE Int. Conf. Robot. Automat.*, 2001, pp. 1786–1791.

[10] T. Yamada, T. Ooba, T. Yamamoto, N. Mimura, and Y. Funahashi, "Grasp stability analysis of two objects in two dimensions," in *Proc. 2005 IEEE Int. Conf. Robot. Automat.*, 2005, pp. 760–765.

[11] T. Yamada and H. Yamamoto, "Static grasp stability analysis of multiple spatial objects," *J. Control Sci. Eng.*, vol. 3, pp. 118–139, 2015.

[12] W. C. Agboh, J. Ichnowski, K. Goldberg, and M. R. Dogar, "Multi-object grasping in the Plane," *Robot. Res.*, A. A. Billard and O. T. Khatib, Eds., Cham: Springer Nature Switzerland, pp. 222–238, 2022.

[13] S. Aeron, E. Llontop, A. Adler, W. C. Agboh, M. Dogar, and K. Goldberg, "Push-MOG: Efficient pushing to consolidate polygonal objects for multi-object grasping," in *Proc. IEEE 19th Int. Conf. Automat. Sci. Eng.*, 2023, pp. 1–6.

[14] W. C. Agboh et al., "Learning to efficiently plan robust frictional multi-object grasps," in *Proc. 2023 IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2023, pp. 10660–10667.

[15] K. Yao and A. Billard, "Exploiting kinematic redundancy for robotic grasping of multiple objects," *IEEE Trans. Robot.*, vol. 39, no. 3, pp. 1982–2002, Jun. 2023, doi: 10.1109/TRO.2023.3253249.

[16] T. Chen, A. Shenoy, A. Kolinko, S. Shah, and Y. Sun, "Multi-object grasping–estimating the number of objects in a robotic grasp," in *Proc. 2021 IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 4995–5001.

[17] Y. Sun, E. Amatova, and T. Chen, "Multi-object grasping - types and taxonomy," in *Proc. 2022 Int. Conf. Robot. Automat.*, 2022, pp. 777–783.

[18] A. Shenoy, T. Chen, and Y. Sun, "Multi-object grasping - efficient robotic picking and transferring policy for batch picking," in *Proc. 2022 IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 2741–2747.

[19] T. Chen and Y. Sun, "Multi-object grasping—Experience forest for robotic finger movement strategies," *IEEE Robot. Automat. Lett.*, vol. 9, no. 6, pp. 5222–5229, Jun. 2024.

[20] Z. Ye, R. Frumento, and Y. Sun, "Only pick once: Algorithms for efficiently picking an exact number of multiple identical objects," *IEEE Trans. Automat. Sci. Eng.*, vol. 22, pp. 14723–14737, 2025.

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

[22] S. D'Avella, A. M. Sundaram, W. Friedl, P. Tripicchio, and M. A. Roa, "Multimodal grasp planner for hybrid grippers in cluttered scenes," *IEEE Robot. Automat. Lett.*, vol. 8, no. 4, pp. 2030–2037, Apr. 2023.

[23] J. Eom, S. Y. Yu, W. Kim, C. Park, K. Y. Lee, and K.-J. Cho, "MOGrip: Gripper for multiobject grasping in pick-and-place tasks using translational movements of fingers," *Sci. Robot.*, vol. 9, no. 97, 2024, Art. no. eado3939. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.ado3939

[24] B. Calli et al., "Yale-CMU-berkeley dataset for robotic manipulation research," *Int. J. Robot. Res.*, vol. 36, no. 3, pp. 261–268, 2017. [Online]. Available: https://doi.org/10.1177/0278364917700714

[25] J. Leitner et al., "The ACRV picking benchmark (APB): A robotic shelf picking benchmark to foster reproducible research," 2016, *arXiv:1609.05258*.

[26] C. D. Santina, C. Piazza, G. Grioli, M. G. Catalano, and A. Bicchi, "Toward dexterous manipulation with augmented adaptive synergies: The Pisa/IIT SoftHand 2," *IEEE Trans. Robot.*, vol. 34, no. 5, pp. 1141–1156, Oct. 2018.

[27] M. Santello, M. Flanders, and J. F. Soechting, "Postural hand synergies for tool use," *J. Neurosci.*, vol. 18, no. 23, pp. 10105–10115, 1998. [Online]. Available: https://www.jneurosci.org/content/18/23/10105

Authorized licensed use limited to: Scuola Superio Sant'Anna di Pisa. Downloaded on November 26,2026 at 11:17:24 UTC from IEEE Xplore. Restrictions apply.