A Benchmark of Dexterity for Anthropomorphic Robotic Hands

Davide Liconti[1*], Yuning Zhou[1*], Yasunori Toshimitsu[1], Ronan Hinchet[1] and Robert K. Katzschmann[1]

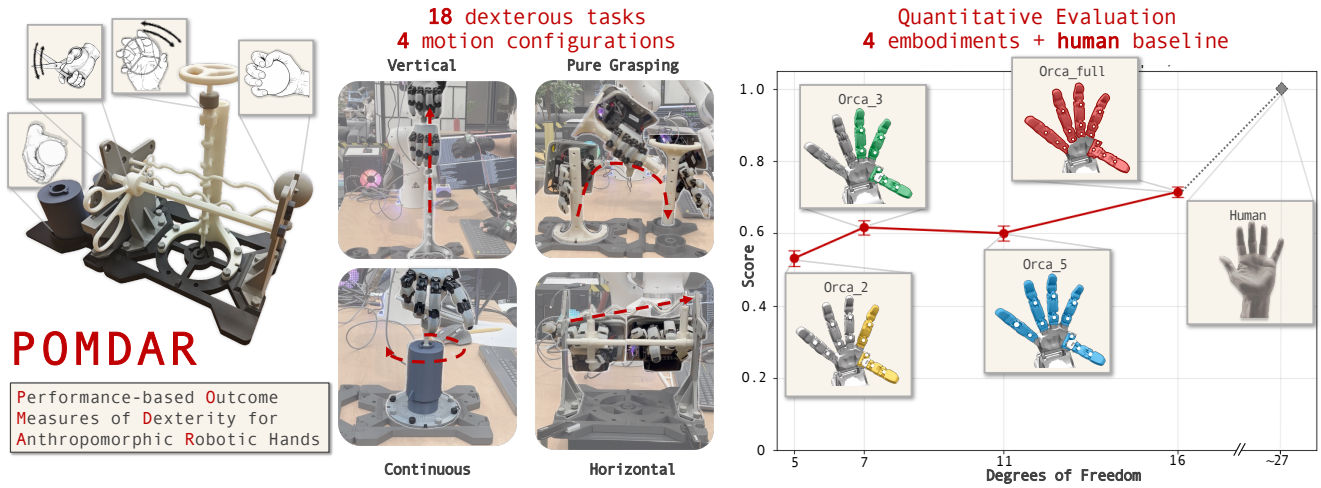

Fig. 1. POMDAR (Performance-based Outcome Measures of Dexterity for Anthropomorphic Robot Hands): a compact, fully 3D-printable benchmark for quantitative evaluation of robotic hand dexterity. The setup includes four manipulation configurations: vertical and horizontal scaffolded configurations, continuous rotation, and pure grasping, covering a wide range of dexterous skills. Quantitative results obtained via teleoperation with ORCA hands of increasing DoF (2 to 16) show improved performance with higher embodiment complexity, demonstrating the benchmark's sensitivity to dexterity variations.

***Abstract*— Dexterity is a central yet ambiguously defined concept in the design and evaluation of anthropomorphic robotic hands. In practice, the term is often used inconsistently, with different systems evaluated under disparate criteria, making meaningful comparisons across designs difficult. This highlights the need for a unified, performance-based definition of dexterity grounded in measurable outcomes rather than proxy metrics. In this work, we introduce POMDAR, a comprehensive dexterity benchmark that formalizes dexterity as task performance across a structured set of manipulation and grasping motions. The benchmark was systematically derived from established taxonomies in human motor control. It is implemented in both real-world and simulation and includes four manipulation configurations: vertical and horizontal configurations, continuous rotation, and pure grasping. The task designs contain mechanical scaffolding to constrain task motion, suppress compensatory strategies, and enable metrics to be measured unambiguously. We define a quantitative scoring metric combining task correctness and execution speed, effectively measuring dexterity as throughput. This enables objective, reproducible, and interpretable evaluation across different hand designs. POMDAR provides an open-source, standardized, and taxonomy-grounded benchmark for consistent comparison and evaluation of anthropomorphic robot hands to facilitate a systematic advancement of dexterous manipulation platforms. CAD, simulation files, and evaluation videos are publicly available at https://srl-ethz.github.io/POMDAR/.**

*These authors contributed equally to this work.
[1]Soft Robotics Lab, D-MAVT, ETH Zurich rkk@ethz.ch

I. Introduction

For anthropomorphic robotic hands, *dexterity* is not only a central design objective but also the defining attribute that distinguishes them from traditional grippers. Yet a major research gap persists: there is no unified definition of dexterity for anthropomorphic hands, nor a standardized framework for evaluating it across designs. This limits meaningful comparison across systems and hinders the community's ability to systematically design, optimize, and select hands for specific manipulation tasks.

A key challenge is that dexterity cannot be captured through kinematic properties alone. Measures such as degrees of freedom, joint limits, or manipulability indices reflect a system's *potential* but not how effectively a hand performs real manipulation under contact-rich interactions. Dexterity must therefore be evaluated in a *performance-based* manner, which requires specifying a representative set of tasks grounded in established studies of hand motion, grasping, and manipulation primitives.

The consequences of this dexterity definition and evaluation gap are visible even in top venues. Two anthropomorphic hands recently published in *Nature Communications*, the *SMA Hand* [1] and the *ILDA Hand* [2], share only one common metric: finger reachable workspace. All other assessments rely on ad-hoc measures such as specific grasp postures or custom-designed tasks.

More broadly, the lack of a standardized benchmarking framework [3] means that researchers, developers, and end

users have no reliable method to compare dexterity across hand designs [4], [5], [3]. This complicates the assessment of design improvements and limits the ability to match hand designs to specific task requirements.

To address these gaps, we introduce POMDAR, a systematic dexterity benchmarking framework applicable to both physical and simulated environments. Our benchmark is guided by the following design principles:

1) **Representative of real hand use**: Every benchmark task can be traced to an existing manipulation or grasp taxonomy, ensuring a representative choice of tasks.
2) **Reproducible across laboratories**: The benchmark design is open source and can be fully 3D printed, ensuring easy access without external procurement. Many components are reused across tasks, further facilitating fabrication. In addition, the benchmark is also available in simulation, allowing for pre-fabrication dexterity evaluation and design optimization.
3) **Directly observable, standardized motions**: The benchmark uses mechanical scaffolds to constrain task motion to the intended degrees of freedom to suppress compensatory strategies (e.g., gravity assistance, palm support, or excessive arm/wrist involvement), standardizing movements and enabling direct observation of the task completion level.
4) **Quantitative, throughput-based evaluation**: Dexterity is measured through a unified score that combines task success and execution speed, effectively capturing performance as task throughput. This enables objective, continuous, and easily comparable evaluation across different hand designs.

II. Approaches to define and measure dexterity

A. Definition of Dexterity

Early medical literature distinguished between *fine* (finger) dexterity, involving coordinated finger movements on smaller objects, and *gross* (manual) dexterity, involving whole-hand and arm movements on larger objects [9]. Later rehabilitation definitions increasingly characterized dexterity as coordinated voluntary movement for functional object manipulation, emphasizing speed and task completion [10].

More recent formulations extend the concept to include adaptation to environmental changes and task demands, framing dexterity as an interaction between motor control and external constraints [11]. This implies that dexterity is inherently tied to dynamic behavior and cannot be measured from static grasps alone.

In robotics, Cutkosky [12] argued that dexterity depends on two parameters: manipulability (the ability of a grasp to impart arbitrary motions to an object) and kinematic workspace. Bicchi [13] later emphasized in-hand object repositioning and reorientation which are enabled by kinematic redundancy.

B. Classification of Dexterous Tasks

Researchers in both medicine and robotics have proposed taxonomies of human hand activities. Bullock et al. [14] provide a compact, motion-centric map of manipulation using five binary descriptors (contact, prehensile/non-prehensile, motion, within-hand motion, motion-at-contact), yielding 15 mutually exclusive classes. From these, grasping (static contact to constrain an object) and in-hand manipulation (dynamic contact for repositioning or reorienting an object) emerge as two principal categories.

For *manipulation*, Elliott and Connolly [6] contribute a human-derived set of 13 coordination patterns (synergy- and sequence-based), forming a minimal library of *dexterity primitives* for task selection. Ma and Dollar [7] augment this set with robotics-relevant categories (regrasping, in-grasp manipulation, finger gaiting, rolling, sliding, and finger pivoting/tracking), explicitly covering pivoting/tracking behaviors that are otherwise underrepresented.

For *grasping*, Napier's power–precision split [15] distinguishes force-dominant from dexterity-dominant grasps. Cutkosky [12] adds a task-driven hierarchy based on object geometry, clamping, and force. Feix et al.'s GRASP Taxonomy [8] provides a comprehensive set of 33 stable, one-handed prehensile grasp types, classified by opposition type, virtual-finger roles, and thumb position, forming a practical catalog for selecting standardized grasping postures.

III. Existing Dexterity Benchmarks

In the literature, existing evaluation practices broadly fall into two families: *parameter-based* (proxy) evaluation and *task-performance-based* evaluation, the latter aligning with *performance-based outcome measures of dexterity* (PBOMD) used in clinical assessment [16].

A. Parameter-based (proxy) evaluation paradigm

A common practice in both academic reporting and commercial specifications is to infer "dexterity" from design parameters such as the number of degrees of freedom, joint ranges of motion, actuator count, speed, and kinematic indices (e.g., Jacobian-based manipulability and conditioning) [11]. Such measures implicitly target *potential* capability rather than *realized* capability in contact-rich manipulation. Although such evaluation is fast to compute and inexpensive, it does not correspond to a concrete task set and does not directly evaluate grasp stability, contact transitions, or control in actual physical interaction.

B. Task-performance-based evaluation (PBOMD-style) paradigm

PBOMDs, widely adopted in clinical settings [16], evaluate dexterity by measuring how effectively a subject (human or robot) performs physical tasks. Rather than relying on abstract mathematical constructs, PBOMDs use quantitative performance metrics such as task completion time, error rates, or motion efficiency to provide practical, outcome-driven assessments of manual dexterity.

In the robotics context, task-performance-based benchmarks assess dexterity by measuring whether and how well a hand accomplishes a set of physical tasks (e.g.,

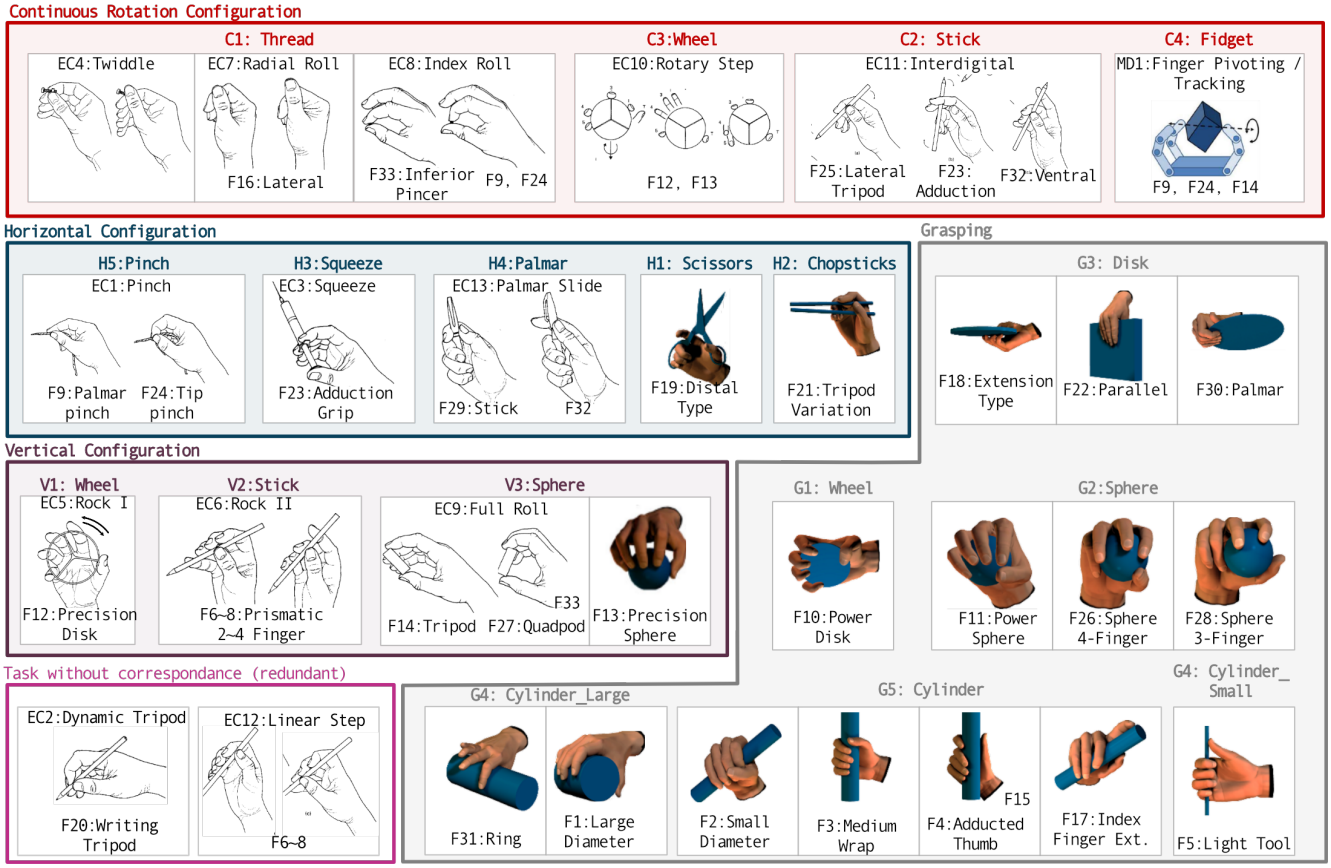

Fig. 2. Overview of hand tasks based on those identified in Elliott & Connolly's taxonomy for manipulation [6], Ma & Dollar's extension [7], and Feix's GRASP taxonomy [8], respectively identified with the symbols EC, MD, and F. Many of the GRASP taxonomy postures are inherently contained in each of the manipulation patterns, providing an opportunity to make the benchmark more efficient. The taxonomies are grouped in the POMDAR benchmark task. Refer to Fig. 3 for the actual implementation of the tasks.

stable grasping, in-hand object reorientation, sequential finger gaiting), using outcome metrics such as success rate, achieved displacement/rotation, or completion time [4], [5], [17]. Compared to parameter-based proxies, PBOMD-style benchmarking improves face validity and typically yields more actionable comparisons for end users.

1) The Elliott & Connolly Benchmark (E&C): Coulson *et al.* [4] proposed using the E&C taxonomy [6] to specifically isolate *in-hand dexterity* of humanoid robot hands.

E&C focuses on *within-hand* object motions (translations and rotations relative to the hand) achieved through digit motion alone, without the arm, wrist, or external support [4]. The benchmark targets core in-hand primitives such as rolling, sliding, and sequential repositioning (finger-gaiting-like patterns). Object selection largely leverages the YCB set, supporting standardization, although its procurement has become increasingly difficult. However, it does not include pure grasping tasks (static grasp formation and stability), and it typically relies on visual tracking (e.g., AprilTags), which adds instrumentation burden and can suffer from occlusions during manipulation.

2) 50 Hand Dexterity Benchmarks (HD-marks): Zhou *et al.* introduced a broad checklist-style benchmark containing 50 tasks spanning grasping, thumb dexterity, and in-hand manipulation [5]. The benchmark includes: (i) a large set of static grasps drawn from the GRASP taxonomy [8], (ii) thumb postures inspired by the Kapandji test [18], and (iii) a small set of in-hand tasks parameterized by translations/rotations along Cartesian axes [5]. Its primary strength is breadth: it covers grasp diversity, thumb mobility, and basic in-hand motion directions within a single framework [5]. However, evaluation is largely binary (success/failure), and the benchmark does not enforce standardized objects across tasks, reducing cross-lab comparability.

3) Dexterity Test Board: Elangovan *et al.* proposed a modular dexterity test board comprising 24 tasks across multiple categories (simple manipulation, reorientation, fine manipulation, fastening tasks, and puzzles) [17]. The motorized board includes pick-and-place tasks, object reorientation, threaded insertions/turning, nut/bolt tool interactions, and puzzle-like force/control tasks, with an open, modular construction that promotes replication [17]. Although the tasks were adapted from existing dexterity tests, they are not mapped one-to-one to prior tests or to a formal taxonomy, making the theoretical foundation weaker than that of the other benchmarks.

IV. Construction of benchmark tasks for POMDAR

To cover the full breadth of hand capabilities, we draw on 14 manipulation patterns for in-hand manipulation and 33 grasp types from the GRASP Taxonomy [8] as the basis for constructing benchmark tasks. The 14 manipulation patterns are derived from Elliott and Connolly's taxonomy of 13 patterns [6] (where "Rock" is split into two variants [4]), plus Finger Pivoting/Tracking from Ma and Dollar [7]. We translate these taxonomies into a set of self-contained, concrete tasks that constitute the benchmark.

As shown in the upper portion of Fig. 2, many grasp patterns are inherently contained within the manipulation patterns: 16 of the 33 grasp types already appear. This provides an opportunity to streamline the benchmark by evaluating manipulation and grasping simultaneously. The remaining grasp patterns, shown in the lower portion of Fig. 2, do not overlap with any manipulation motion.

Based on this observation, a dual-structured task categorization is proposed for the comprehensive dexterity benchmark:

- **Manipulation Tasks**: Motion-based tasks that primarily measure in-hand dexterity based on Elliott and Connolly's manipulation taxonomy. Because these tasks involve both object interaction and configuration transitions, they are well suited for simultaneously evaluating grasping and in-hand manipulation performance.
- **Pure Grasping Tasks**: These tasks specifically evaluate the grasp types that do not overlap with in-hand manipulation patterns (shown in the lower part of Fig. 2). They primarily assess the hand's ability to form and maintain stable grasp configurations.

This task division aligns closely with the observation of dexterity seen in medical literature [9], which distinguishes between:

- *Gross Dexterity*, defined as the coordinated action between fingers and palm, and primarily reflected in **pure grasping tasks**;
- *Fine Dexterity*, defined as inter-digit coordination during object interaction, and best assessed through **manipulation tasks**.

Such a structured task design ensures both comprehensiveness and efficiency, and reflects the multidimensional nature of dexterity in anthropomorphic robotic hands.

A. Consolidating manipulation tasks

The proposed benchmarking system is designed to be performance-based and outcome-oriented, and should avoid direct penalization of robotic hands with unconventional structures, such as those with fewer fingers. Thus, the prismatic 2–4 finger patterns used for precision grasping (F6 to F8) were consolidated into a single representative manipulation task *V2: Stick* (Fig. 3), which rotates a thin object in its fingers.

The grasping patterns not assigned to manipulation tasks in Fig. 2 are mostly focused on firmly securing objects in the hand, with two notable exceptions: F19 (Distal Type) and F21 (Tripod Variation), typically demonstrated through the use of scissors and chopsticks, respectively. Because these patterns entail functional within-hand motions that resemble in-hand manipulation rather than static grasping, they were incorporated into the **Manipulation Tasks** category.

Tripod Variation (F21) is highly similar to Dynamic Tripod (EC2, F20), which describes writing with a pen. In both patterns, the index finger rests on the top surface of the pen or upper chopstick, while the middle finger and thumb press against the side surfaces, forming a stable three-point contact. The key distinction is that F21 requires simultaneously holding an additional chopstick. Given this similarity, F21 was integrated into *H2: Chopsticks*, while G19 (Distal Type) was renamed *H1: Scissors*. EC12 (linear step) is a variation of EC13, but using the grap of EC6, so it was considered redundant and not implemented in the benchmark.

The tasks EC4, EC7, and EC8 all require using a combination of thumb and index flexion to rotate a very small object. Since it is difficult to implement tasks that test all these different manipulation strategies, and we aim to obtain quantitative and objective performance measures (rather than subjective evaluation of how closely the intended motion is followed), all these tasks were incorporated into the *C1: Thread* task. The motion of EC9 (full roll), using the F13 precision sphere grasp, was grouped into a vertical configuration task *V3: Sphere*, with the goal of reusing as many components as possible while maintaining a compact setup.

This results in a total of 12 manipulation tasks (Fig. 3).

B. Consolidating grasping tasks

To make the benchmark fairer to robot hands with fewer fingers and as objective as possible, a similar rationale applies to the grasping patterns. For the grasping tasks, since we aim to avoid non-objective scoring and instead rely purely on performance-based evaluation, we cluster the grasps based on the manipulated object. Specifically, we group F2, F3, F4, F15, and F17 into cylindrical grasps, with variations including F1 and F31 for large-diameter cylinders, and F5 for small-diameter cylinders.

We then consider grasping of spherical objects (F26, F28) and disk-shaped objects (F18, F22, F30). In total, six grasp tasks are evaluated. An overview of the final grasp tasks is shown in Fig. 3B.

C. Design of the POMDAR benchmark

We designed the POMDAR benchmark tasks to fulfill the design principles introduced in Section I, as illustrated in Fig. 3. The manipulation tasks (Fig. 3A) are further organized into three configurations:

- **Vertical (scaffolded)** configuration, where the hand grasps an object and wiggles it while pulling a rod upward through a sequence of notches.
- **Continuous Rotation** configuration, where the hand grasps and continuously rotates an object through stepping sequences. A gravity-based clutch engages the

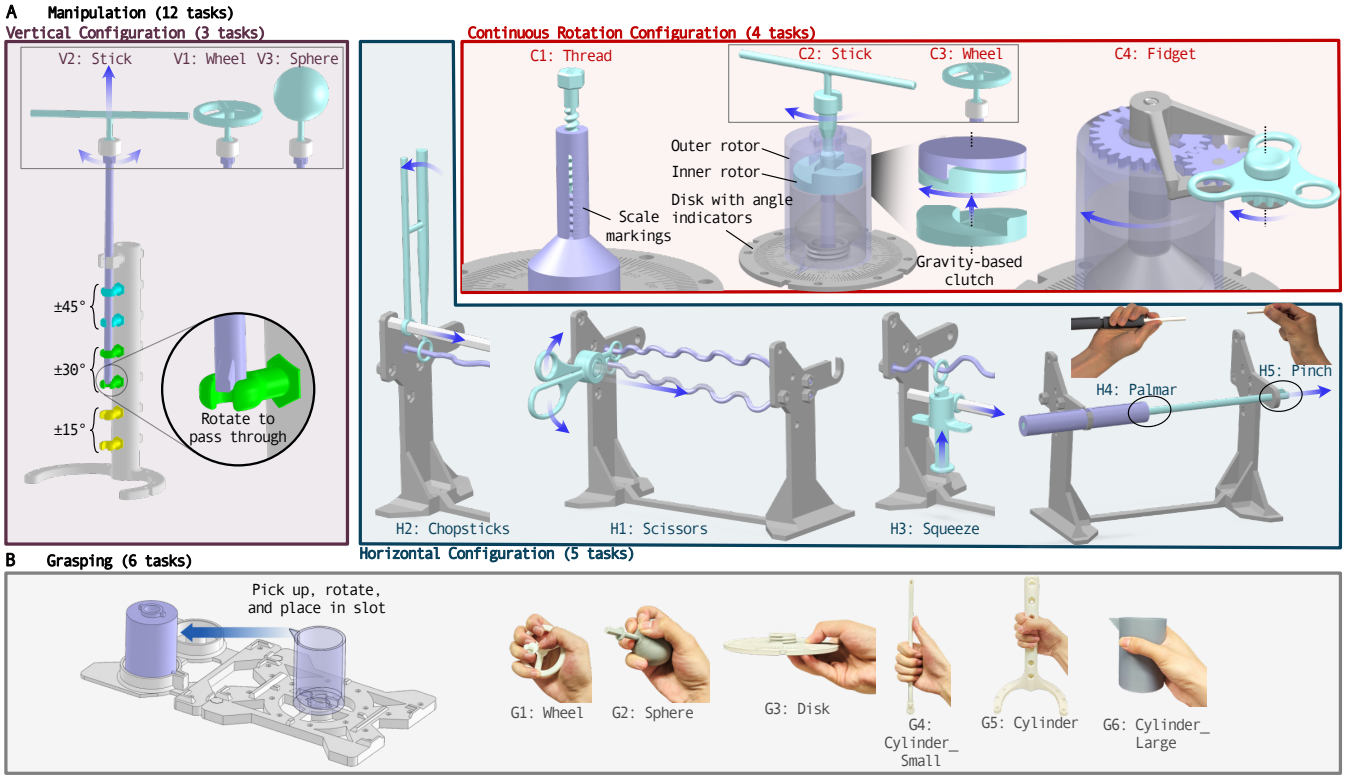

Fig. 3. The benchmark comprises four task configurations: two scaffolded manipulation setups (vertical and horizontal), a continuous rotation configuration, and a set of pure grasping tasks. In the vertical scaffolded configuration (V), the hand grasps an object and moves it upward along a rod with discrete notches of increasing angle (from $\pm 15°$ to $\pm 45°$), requiring coordinated in-hand adjustments to pass each constraint. In the horizontal scaffolded configuration (H), the object is translated along a curved rail with progressively increasing curvature, enforcing controlled in-hand manipulation to clear each section. The number of curves depends on the task: three for the chopsticks task (H5), four for the squeeze task (H2), and six for the scissors task (H1). The palmar (H3) and pinch (H4) tasks share the same object but involve manipulating different parts of it. The setup uses a horizontal hexagonal rod to constrain motion along the horizontal axis. For the chopsticks task, the two sticks undergo a constrained motion of approximately $\pm 20°$. The continuous rotation configuration (C) features a gravity-based clutch mechanism that engages the outer rotor to maintain object suspension, enabling continuous, stepwise rotation without resetting (e.g., C2: stick, C3: wheel). The fidget task uses a simple geared transmission to transfer motion to a rotating element, while the thread task consists of a plastic screw that must be removed from a threaded hole. The pure grasping tasks (G) isolate grasp quality by requiring the hand to pick up objects from inserts and relocate them in free space, without external scaffolding. This avoids artificially stabilizing suboptimal grasps, enforcing more robust grasping. A total of six grasp objects are used. Many objects are reused across configurations to ensure consistency and reduce redundancy in the benchmark design. The setup is compact (see Fig. 1), fully 3D printable, requires no additional components beyond printing material, and can be fixed to a table using clamps.

outer rotor to keep the object suspended. This platform can also be adapted for tasks such as fidget spinning and threading.

- **Horizontal (scaffolded)** configuration, where the object is manipulated in-hand while being translated horizontally. The required range of motion increases progressively to clear a curved rail, thereby increasing task difficulty. This setup can also be used for manipulating thin rods.

The pure grasping tasks (Fig. 3B) reuse components from the manipulation setups and measure grasps not inherently covered by the manipulation tasks. In these tasks, objects are not constrained by external structures and must be stably held while undergoing rotations in free space, enabling evaluation of grasp robustness and quality.

D. Benchmark protocol and scoring

The POMDAR benchmark quantifies dexterity as the *throughput* over a set of tasks, capturing both how well and how fast each task is executed. Accordingly, each task is evaluated through two components: a *correctness score* and a *speed score*.

The correctness component reflects task completion quality and is normalized between 0 and 1. For the scaffolded manipulation tasks, correctness is defined as the fraction of progress achieved relative to the task goal. In the vertical configuration, this corresponds to the number of notches successfully traversed divided by the total number of notches. Similarly, in the horizontal configuration, correctness is the number of curves cleared divided by the total. For the continuous rotation configuration, correctness is computed as the achieved rotation normalized by a full $360°$ rotation.

Pure grasping tasks use a discrete scoring scheme: a score of 0 is assigned if the object is not lifted, 0.5 if the object is lifted but dropped during relocation, and 1 if the object is successfully grasped and relocated.

The speed component captures execution efficiency and is defined relative to a human baseline. Specifically, it is

computed as the ratio between a baseline time and the time taken by the evaluated system to complete the task. The baseline time is obtained as the average completion time of human participants, as measured in the user study described in Section V-A.

The overall task score is computed as a weighted combination of correctness and speed:

$$\text{Score} = 0.8 \cdot \text{Correctness} + 0.2 \cdot \text{Speed}. \quad (1)$$

This weighting reflects the current state of the field: as in-hand manipulation remains a challenging frontier in robotic hand development, reliable task completion is prioritized over execution speed. Consequently, a higher weight is assigned to correctness (0.8) to emphasize robustness, while still accounting for efficiency through a smaller contribution from speed (0.2). Scores greater than 1 indicate superhuman performance, which may arise in future scenarios such as learning-based controllers achieving faster-than-human execution; therefore, the speed score is intentionally left unbounded.

E. Real-World Benchmark Design

A central design goal of the POMDAR benchmark is reproducibility across laboratories. To this end, all physical components of the benchmark apparatus are manufactured using consumer-grade 3D printing, eliminating the need for specialized machining or procurement of proprietary parts.

The system is secured to the table with clamps, while protrusions, bolt holes, and a custom bolt–nut retention mechanism ensure repeatable mounting and easy replacement of damaged parts without reprinting the full base. This makes the real-world setup compact, maintainable, and practical for repeated benchmarking.

F. Simulation Benchmark Design

The benchmark was implemented in MuJoCo with a task-oriented modeling strategy that preserves only the geometries essential for contact and scoring, while simplifying or discarding non-functional parts to keep the simulation efficient and stable. Because MuJoCo does not natively handle non-convex collision meshes well, complex components were either approximated with primitive shapes or preprocessed using CoACD for convex decomposition before import. Additional task-specific workarounds were required for mechanisms that are hard to reproduce directly, such as manually enforcing gear couplings and using a feedforward PID controller to emulate screw-thread motion. The simulated benchmark supports teleoperation through a variety of interfaces, including motion capture gloves and VR systems (see Fig. 4), which also provides an overview of all tasks implemented in MuJoCo. Beyond teleoperation, the simulation environment enables rapid evaluation of hand designs prior to fabrication and establishes a foundation for future learning-based approaches, where policies can be trained directly on the tasks, decoupling performance from operator skill and teleoperation setup.

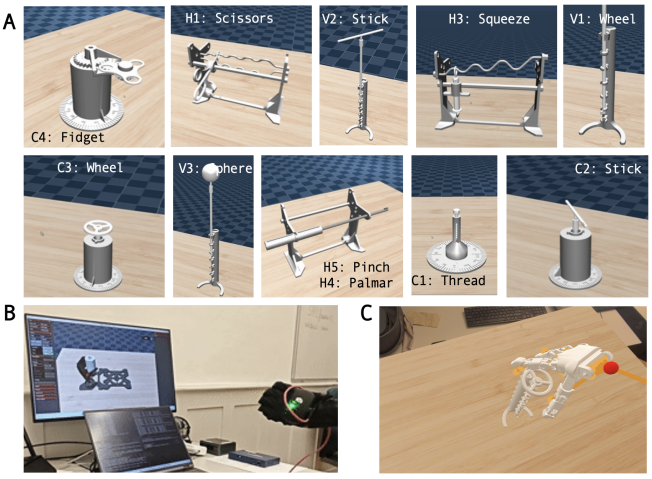

Fig. 4. (A) Examples of the POMDAR tasks implemented in MuJoCo across different manipulation configurations. (B) Teleoperation using motion-capture gloves for direct control of the simulated hand. (C) Teleoperation in virtual reality using Apple Vision Pro, where the simulation is streamed to the user for an immersive and interactive experience; the tracked hand keypoints are visible as orange overlays.

V. Results

A. User Study

We conducted a user study to validate the benchmark and establish human baseline performance. Six participants were asked to solve all tasks given only the start and goal configurations, without an explicit strategy. Five out of six participants were right-handed, and all performed the tasks with their right hands. Each participant completed three trials per task, yielding a total of 18 trajectories.

Participants wore a motion-capture glove tracking 22 hand keypoints at 100 Hz, enabling detailed analysis of hand motion (Fig. 5A).

We verify that participants adopt motion and grasp patterns consistent with the taxonomies introduced in Section IV-A and Section IV-B. The low variability in strategies suggests that the scaffolded design effectively constrains interactions to the intended motion primitives (Fig. 5B).

Baseline completion times are computed as the mean across all participants and trials, and they define the speed component of the benchmark score.

Finally, recorded hand trajectories are visualized in Fig. 5, showing consistent motion patterns across participants.

B. Benchmark results

We evaluate the POMDAR benchmark through teleoperation experiments using progressively more dexterous embodiments of the ORCA hand [19].

Specifically, we consider four configurations: (i) a 2-finger setup (thumb and index, no abduction, 5 DoF), (ii) a 3-finger setup (no abduction), (iii) a 5-finger setup without abduction, and (iv) a full 5-finger configuration with 16 DoF (Fig. 6A). The hand is mounted on a 7-DoF Franka Emika arm.

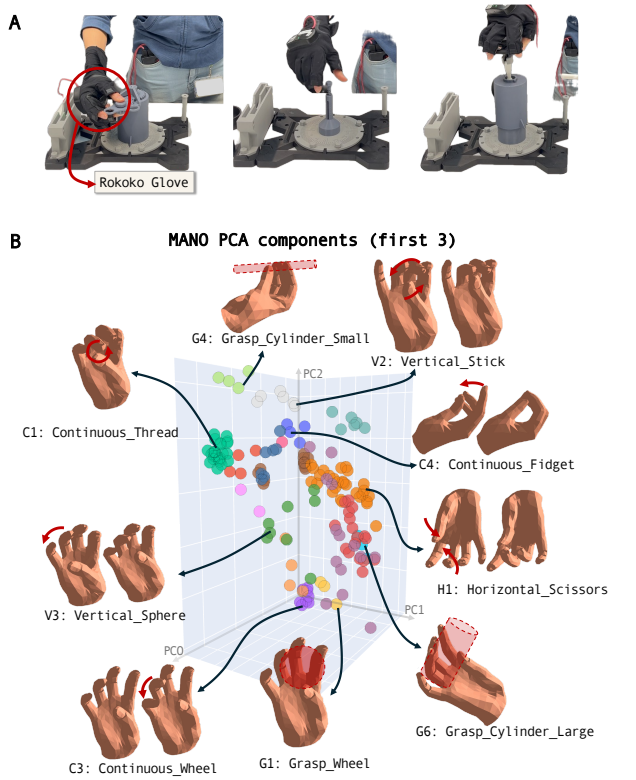

Fig. 5. Human study and motion analysis. (A) Example snapshots of the human data collection for three continuous rotation tasks. Participants perform the benchmark using a motion-capture Rokoko glove (circled), which tracks hand keypoints during task execution. (B) Principal component analysis (PCA) of the recorded hand motions using the MANO representation. The first three principal components (out of six total) are shown. Each point corresponds to a 1.5 s trajectory segment, and colors indicate different tasks. The clustering of points by color suggests that participants, both across and within subjects, adopt similar strategies for each task, indicating that the tasks are intuitive and well-constrained. Example MANO reconstructions are shown around the plot, illustrating representative hand configurations. These are consistent with the intended manipulation patterns and grasp types derived from the taxonomy in Fig. 2.

a) Robot Setup: All experiments are conducted using the ORCA hand mounted with a 90° adapter to a Franka Emika arm. The teleoperator wears Rokoko motion-capture gloves, which track 22 keypoints on the hand and forearm, as well as the 6-DoF wrist pose via a base station providing absolute tracking. The captured keypoints are mapped to the robot through a global scaling and rotation transformation to account for user-specific hand morphology. These parameters are automatically tuned via a Bayesian optimization procedure based on a small calibration dataset of seven poses (four pinch configurations, hand open with and without abduction, and closed hand), where glove keypoints are paired with ground-truth robot joint configurations.

The keypoints are then retargeted online to the robot joint space using an energy minimization-based algorithm (similar to [20]), running at approximately 25 Hz. The relative wrist pose is sent to the robot, which is controlled through a low-level impedance controller.

Depending on the task, the arm motion is either fixed (e.g., C4 fidget, H4 pinch, H5 palmar), constrained along

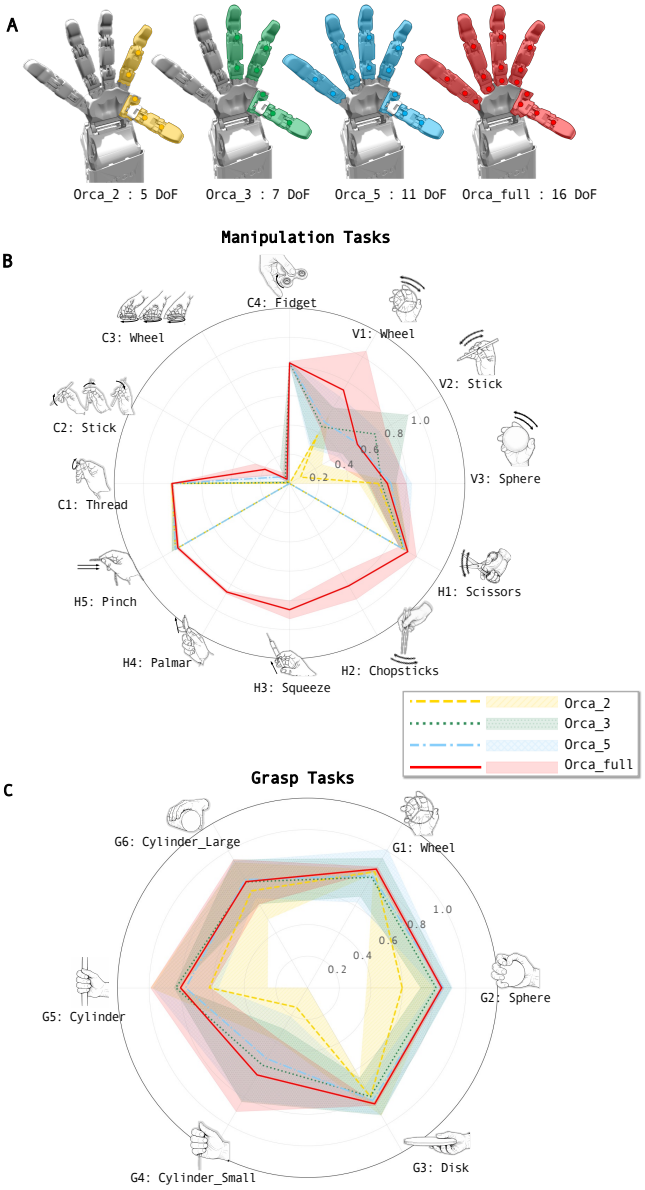

Fig. 6. Results of the POMDAR benchmark across different ORCA hand embodiments. (A) The evaluated embodiments (Orca_2, Orca_3, Orca_5, Orca_full) are shown, where highlighted joints indicate the unlocked degrees of freedom. This selection of DoFs is specific to the ORCA hand and should not be interpreted as a general mapping between number of DoFs and achievable capabilities, but rather as a controlled study of embodiment variations within the same platform. All embodiments use the same controller and teleoperation interface; differences arise solely from locking selected DoFs at zero position in the hand controller. (B) Radar plots for the manipulation tasks (12 total), and (C) radar plots for the grasping tasks (6 total). Each task is repeated 20 times. The shaded regions represent the standard deviation across trials. Icons around the plots provide a visual reference for each task to aid interpretation. The color of each plot corresponds to the embodiment shown in (A); an explicit legend is omitted for clarity.

the vertical axis (V1–V3, C1–C3), or constrained along the horizontal axis (H1–H3). The initial pose of the robot hand is selected by the operator and may vary across tasks and embodiments to compensate for alignment differences.

b) Experimental protocol.: All experiments are performed by the same operator to ensure consistency. For each task and embodiment, the operator performs five practice trials to determine an effective strategy and a start pose for the robot, followed by 20 recorded trials.

c) Disclaimer.: The reported results reflect the performance of a *combined system* comprising the robotic hand, the teleoperation interface, and the operator's skill. While the results are directly comparable across the tested embodiments, they are not necessarily comparable to results obtained with different teleoperation systems. They nonetheless provide meaningful insights into the role of embodiment in dexterous manipulation. Moreover, the object dimensions are designed for anthropomorphic hands, which may disadvantage embodiments whose morphology deviates from human hand proportions, potentially leading to lower benchmark scores.

d) Teleoperation Results.: To test whether the benchmark provides useful insights into the dexterity of robotic hands and whether it correlates with intuitive measures such as the number of degrees of freedom, we evaluate four different embodiments (Fig. 6A). These correspond to variants of the ORCA hand with selected degrees of freedom locked or unlocked. The configurations span from 5 to 16 DoF (full hand without wrist), where in the full configuration all finger abduction DoFs are also enabled.

Fig. 6B reports results for vertical, horizontal, and continuous manipulation tasks. Performance differences are most pronounced in tasks requiring coordinated multi-finger interactions and finger abduction.

Certain tasks are only feasible with specific embodiments, particularly those requiring thumb or multi-finger abduction (e.g., squeeze, palmar, and chopsticks). Conversely, several tasks exhibit identical execution strategies across embodiments (e.g., scissors, thread, and pinch), resulting in comparable performance and reduced sensitivity to embodiment differences. This highlights that dexterity gains are strongly task-dependent.

Fig. 6C shows radar plots for the grasping tasks. All embodiments are able to successfully grasp the objects; however, increased dexterity improves both grasp stability and execution speed. This effect is most pronounced for smaller objects, where additional fingers significantly enhance robustness, while for larger objects the main benefit is faster relocation. Notably, increasing the number of fingers from three to five without abduction yields limited improvement, whereas adding a third finger (from two to three) provides a substantial performance gain, primarily due to improved grasp stability.

Task difficulty varies across configurations. Vertical tasks exhibit higher variance due to finer-grained correctness metrics and increased coordination requirements. In contrast, continuous rotation tasks are generally the most challenging, particularly those involving the gravity-based clutch, which

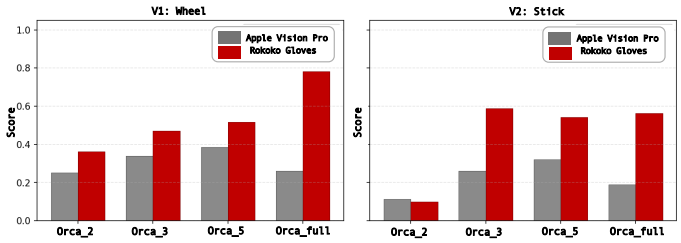

Fig. 7. Comparison of teleoperation methods using motion-capture gloves (Rokoko) and Apple Vision Pro (AVP) across different ORCA embodiments for representative tasks (V1: Wheel, V2: Stick). Results are averaged over 20 trials per task. The teleoperation stack, controller, and robot setup are identical in both cases; the only difference lies in the source of hand keypoints. Performance with AVP is consistently lower, primarily due to occlusions in egocentric perception, which limit accurate tracking of finger motions during manipulation.

cannot be reliably solved with the current hand and teleoperation system. Simpler tasks such as fidget and thread are easier, as they require fewer degrees of freedom and do not penalize temporary loss of contact.

In total, the benchmark results are based on 1140 recorded trajectories (18 tasks, 4 embodiments, 20 trials per task, excluding non-repeated cases), corresponding to approximately 25 hours of real-world testing. All experiments were conducted with the same robot, operator, and teleoperation interface, ensuring consistent comparison across embodiments.

Finally, Fig. 1 (right) presents aggregated scores as a function of the number of DoF. While performance generally increases with dexterity, the results confirm that improvements are task-dependent, with some tasks benefiting from additional fingers and others from specific kinematic capabilities such as abduction. The error bars represent variability across trials: for each trial, we compute the average score over all tasks and report the mean and standard deviation across the two trials.

e) Ablations: We demonstrate that the benchmark can also be used to compare different teleoperation methods and equipment. Fig. 7 shows the results obtained using motion capture gloves and Apple Vision Pro (AVP) across all embodiments. Overall, the scores achieved with AVP are lower, mainly due to occlusions in egocentric perception. This is particularly evident in tasks such as the vertical configurations, where finger movements are partially occluded by the thumb and therefore not fully captured, leading to reduced control accuracy.

VI. Discussion

The results demonstrate that POMDAR provides a quantitative, taxonomy-grounded framework for evaluating dexterous manipulation across robot hand embodiments. We outline key limitations and future directions.

a) Teleoperation-only evaluation.: The current results reflect a combined system of the hand, teleoperation interface, and human operator, and therefore do not isolate mechanical dexterity. Evaluating autonomous policies, whether learned or programmed, would decouple hand capability

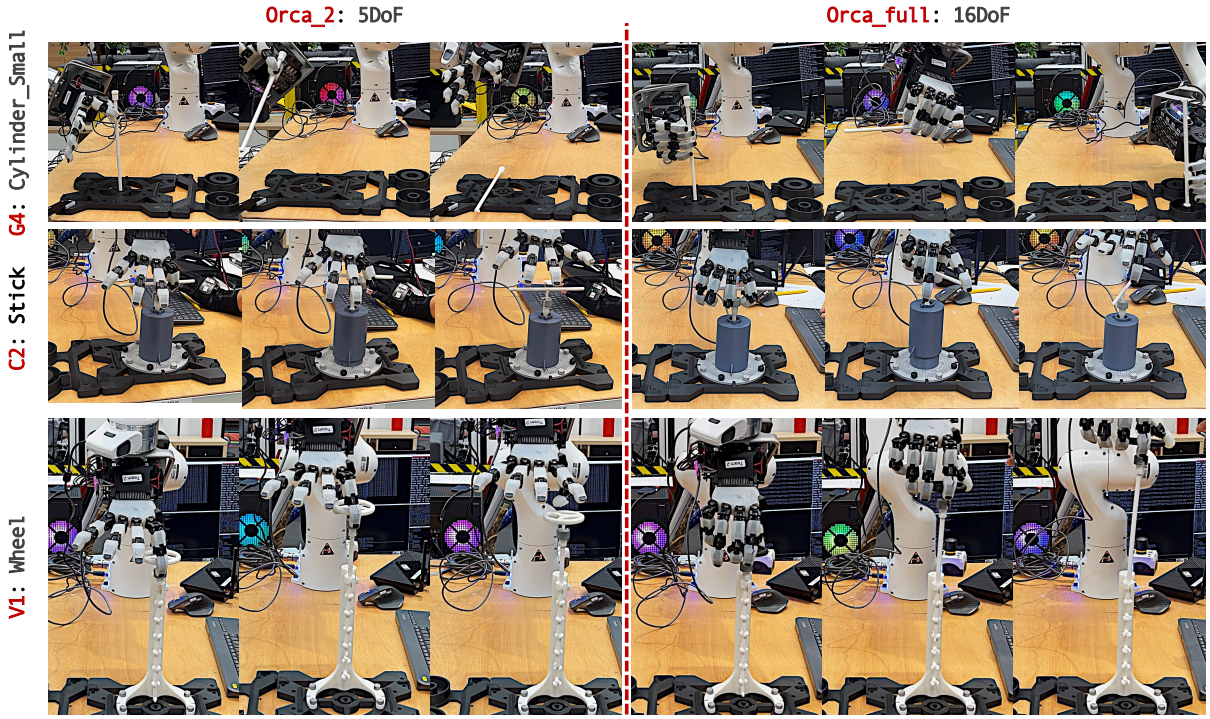

Fig. 8. Qualitative example task execution sequences for the 5Dof (left) and 16 Dof (right) versions of the ORCA hand. The shown tasks where selected to highlight differences in the performance due to the added degrees of freedom.

from operator skill and enable more direct embodiment and control method comparisons.

b) Object size and anthropomorphic bias.: The benchmark objects are designed for anthropomorphic hands, which may disadvantage embodiments with different morphologies. Future work could explore scalable or parameterized object sets to enable fairer comparisons.

c) Limited interaction dynamics.: The benchmark primarily evaluates kinematic dexterity and controlled contact interactions, and does not capture dynamic behaviors relying on inertial or forceful interactions (e.g., pushing, tossing, flipping). Extending the task set in this direction would broaden the notion of dexterity.

d) Dependence on external robotic arm.: The benchmark requires mounting the hand on an external robotic arm, which may introduce variability across systems. This is partially mitigated by constraining end-effector motion (fixed or single-axis) in most tasks. Future designs could explore standardized mounting or arm-independent setups.

e) Toward automated benchmarking.: A key direction is the development of learning-based methods for automatic evaluation of robotic hands in simulation and reality. Recent work in dexterous retargeting, including sampling-based approaches [21] and reinforcement learning-based methods [22], suggests that human motion can be transferred across embodiments. Such systems could generate task policies and compute benchmark scores without teleoperation. The simulation version of POMDAR provides a foundation for scalable and reproducible evaluation framework.

VII. Conclusion

We presented POMDAR, a performance-based benchmark for evaluating dexterity in anthropomorphic robot hands. The benchmark is grounded in established manipulation and grasp taxonomies, translating 14 manipulation patterns and 33 grasp types into a structured set of 18 physical tasks covering both manipulation (12 tasks) and pure grasping (6 tasks).

POMDAR provides a quantitative dexterity score based on task correctness and execution speed relative to a human baseline, enabling direct and interpretable comparisons across embodiments.

A key design principle is reproducibility: all components are fully 3D-printable and open-source, and a simulation counterpart enables evaluation without physical hardware.

We validated the benchmark through a user study and experiments across multiple ORCA hand embodiments (5–16 DoF), showing that POMDAR scores track embodiment dexterity and reveal task-dependent performance differences.

Acknowledgment

D.L is supported by Swiss National Science Foundation (SNSF) Project Grant No. 200021 215489. Y.T. received support from the Takenaka Scholarship Foundation, Max Planck ETH Center for Learning Systems, and the Swiss Government Excellence Scholarship. The authors thank Harry Durham for assistance with the hardware design of the benchmark.

REFERENCES

[1] H. Yang, Z. Tao, J. Yang, W. Ma, H. Zhang, M. Xu, M. Wu, S. Sun, H. Jin, W. Li, L. Wang, and S. Zhang, "A lightweight prosthetic hand with 19-DOF dexterity and human-level functions," *Nature Communications*, vol. 16, no. 1, p. 955, Jan. 2025, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41467-025-56352-5

[2] U. Kim, D. Jung, H. Jeong, J. Park, H.-M. Jung, J. Cheong, H. R. Choi, H. Do, and C. Park, "Integrated linkage-driven dexterous anthropomorphic robotic hand," *Nature Communications*, vol. 12, no. 1, p. 7177, Dec. 2021. [Online]. Available: https://www.nature.com/articles/s41467-021-27261-0

[3] J. Falco, K. Van Wyk, S. Liu, and S. Carpin, "Grasping the Performance: Facilitating Replicable Performance Measures via Benchmarking and Standardized Methodologies," *IEEE Robotics & Automation Magazine*, vol. 22, no. 4, pp. 125–136, Dec. 2015, conference Name: IEEE Robotics & Automation Magazine. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7349337

[4] R. Coulson, C. Li, C. Majidi, and N. S. Pollard, "The Elliott and Connolly Benchmark: A Test for Evaluating the In-Hand Dexterity of Robot Hands," in *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*, July 2021, pp. 238–245, iSSN: 2164-0580. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9555798?casa_token=NQCUaC07HeIAAAAA:qpEUeW3ykRbli2XDIXrSANepzcWPek6o09_AytJ7pIFsDxZ8Jv22zAsoK6UVvvv9TOCHLcjXCZAN9w

[5] J. Zhou, Y. Chen, D. C. F. Li, Y. Gao, Y. Li, S. S. Cheng, F. Chen, and Y. Liu, "50 Benchmarks for Anthropomorphic Hand Function-based Dexterity Classification and Kinematics-based Hand Design," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2020, pp. 9159–9165, iSSN: 2153-0866. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9340982?casa_token=pPb9yabmEwAAAAAA:oc8g5XYQ3OsFBjjK6yBJM7NnZ_jrnIe9wFOj5e2Lw8QDAuGuUNGysahIsCgAadJ0_7JB8z-ZP9lv4A

[6] J. M. Elliott and K. J. Connolly, "A CLASSIFICATION OF MANIPULATIVE HAND MOVEMENTS," *Developmental Medicine & Child Neurology*, vol. 26, no. 3, pp. 283–296, June 1984. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/j.1469-8749.1984.tb04445.x

[7] R. R. Ma and A. M. Dollar, "On dexterity and dexterous manipulation," in *2011 15th International Conference on Advanced Robotics (ICAR)*. Tallinn, Estonia: IEEE, June 2011, pp. 1–7. [Online]. Available: http://ieeexplore.ieee.org/document/6088576/

[8] 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, Feb. 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7243327/

[9] E. A. Fleishman and W. E. Hempel, "A Factor Analysis of Dexterity Tests," *Personnel Psychology*, vol. 7, no. 1, pp. 15–32, Mar. 1954. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/j.1744-6570.1954.tb02254.x

[10] C. Backman, S. C. D. Gibson, and J. Parsons, "Assessment of Hand Function: The Relationship between Pegboard Dexterity and Applied Dexterity," *Canadian Journal of Occupational Therapy*, vol. 59, no. 4, pp. 208–213, Oct. 1992. [Online]. Available: https://journals.sagepub.com/doi/10.1177/000841749205900406

[11] N. Elangovan, "Analysis and Evaluation of the Dexterity, Grasping, and Manipulation Capabilities of Human and Robot Hands," Doctoral Thesis, ResearchSpace at Auckland, 2022. [Online]. Available: https://hdl.handle.net/2292/62956

[12] M. Cutkosky, "On grasp choice, grasp models, and the design of hands for manufacturing tasks," *IEEE Transactions on Robotics and Automation*, vol. 5, no. 3, pp. 269–279, June 1989. [Online]. Available: http://ieeexplore.ieee.org/document/34763/

[13] A. Bicchi, "Hands for dexterous manipulation and robust grasping: a difficult road toward simplicity," *IEEE Transactions on Robotics and Automation*, vol. 16, no. 6, pp. 652–662, Dec. 2000. [Online]. Available: http://ieeexplore.ieee.org/document/897777/

[14] I. M. Bullock, R. R. Ma, and A. M. Dollar, "A Hand-Centric Classification of Human and Robot Dexterous Manipulation," *IEEE Transactions on Haptics*, vol. 6, no. 2, pp. 129–144, Apr. 2013, conference Name: IEEE Transactions on Haptics. [Online]. Available: https://ieeexplore.ieee.org/document/6298887

[15] J. R. Napier, "THE PREHENSILE MOVEMENTS OF THE HUMAN HAND," *The Journal of Bone & Joint Surgery British Volume*, vol. 38-B, no. 4, pp. 902–913, Nov. 1956, publisher: Bone & Joint. [Online]. Available: https://boneandjoint.org.uk/Article/10.1302/0301-620X.38B4.902

[16] J. Yong, J. C. MacDermid, T. Packham, P. Bobos, J. Richardson, and S. Moll, "Performance-based outcome measures of dexterity and hand function in person with hands and wrist injuries: A scoping review of measured constructs," *Journal of Hand Therapy*, vol. 35, no. 2, pp. 200–214, Apr. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0894113021000648

[17] N. Elangovan, C.-M. Chang, G. Gao, and M. Liarokapis, "An Accessible, Open-Source Dexterity Test: Evaluating the Grasping and Dexterous Manipulation Capabilities of Humans and Robots," *Frontiers in Robotics and AI*, vol. 9, Apr. 2022, publisher: Frontiers. [Online]. Available: https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2022.808154/full

[18] A. Kapandji, "Cotation clinique de l'opposition et de la contre-opposition du pouce," *Annales de Chirurgie de la Main*, vol. 5, no. 1, pp. 67–73, Jan. 1986. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0753905386800539

[19] C. C. Christoph, M. Eberlein, F. Katsimalis, A. Roberti, A. Sympetheros, M. R. Vogt, D. Liconti, C. Yang, B. G. Cangan, R. J. Hinchet, and R. K. Katzschmann, "Orca: An open-source, reliable, cost-effective, anthropomorphic robotic hand for uninterrupted dexterous task learning," 2025. [Online]. Available: https://arxiv.org/abs/2504.04259

[20] A. Sivakumar, K. Shaw, and D. Pathak, "Robotic telekinesis: Learning a robotic hand imitator by watching humans on youtube," 2022. [Online]. Available: https://arxiv.org/abs/2202.10448

[21] C. Pan, C. Wang, H. Qi, Z. Liu, H. Bharadhwaj, A. Sharma, T. Wu, G. Shi, J. Malik, and F. Hogan, "Spider: Scalable physics-informed dexterous retargeting," 2026. [Online]. Available: https://arxiv.org/abs/2511.09484

[22] Z. Mandi, Y. Hou, D. Fox, Y. Narang, A. Mandlekar, and S. Song, "Dexmachina: Functional retargeting for bimanual dexterous manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2505.24853