YUBI: Yielding Universal Bidigital Interface for Bimanual Dexterous Manipulation at Scale

Anonymous Submission

***Abstract*—We introduce *Yielding Universal Bidigital Interface (YUBI)*, a finger-aligned gripper designed to enable intuitive, ergonomic, and scalable data curation for bimanual dexterous manipulation. While handheld data collection systems such as Universal Manipulation Interface (UMI) have lowered the barrier for in-the-wild data collection, their bulky pistol-grip designs can present ergonomic and usability challenges for fine-grained, dexterous manipulation tasks. To address this limitation, YUBI presents a distinct design principle: yielding, finger-driven actuation that directly maps human finger movements to gripper jaw motion, allowing the jaws to naturally follow the operator's grip. This intuitive interface bridges the gap between human intent and robotic execution, facilitating more precise fingertip motor control. Furthermore, by integrating VR-based 6 DoF tracking into a rig-based operation setup, our system produces accurate gripper trajectories suitable for large-scale, high-quality data acquisition in tabletop scenarios. Leveraging this capability, we curate an unprecedented UMI-based dataset for bimanual dexterous manipulation, comprising 2730 hours of data across 300K episodes and 40 distinct tasks. Our experiments demonstrate that YUBI offers advantages over the original UMI gripper in versatility for complex bimanual tasks, dexterity, and operational efficiency. YUBI delivers an end-to-end framework spanning ergonomic gripper design and large-scale dataset curation, which advances research on robotic foundation models.**

I. Introduction

Scaling data collection is fundamental to the progress of robotic foundation models. As demonstrated by the success of large multimodal models, the performance and generalization of Vision-Language-Action (VLA) policies are bounded by the volume and diversity of their training data [1], [2], [14]. This requirement is particularly evident in complex, dexterous, real-world manipulation scenarios such as assembly, which demand massive amounts of high-quality demonstrations.

However, conventional data collection systems struggle to keep pace. Traditional leader-follower teleoperation, while precise, is cost-prohibitive and requires significant operator expertise, resulting in low data throughput [3], [17], [21]. Conversely, imitation learning from human demonstrations offers a more scalable alternative. While raw human demonstration data (*e.g.*, via web videos or wearable sensors) are available at scale [6], [13], [15], they introduce the challenge of a fundamental embodiment gap between human hands and robotic hardware. Recently, handheld interfaces like Universal Manipulation Interface (UMI) [5], [11], [22] have emerged as a pivotal medium between humans and robots. These UMI systems enable bridging the embodiment gap by having operators guide a gripper mechanically identical to the robot's end-effector, eliminating the need for a physical robot during data collection. The resulting data enable direct policy learning from the gripper trajectories and wrist camera streams. Such UMI-based systems provide a cost-effective and embodiment-agnostic pathway for large-scale data collection.

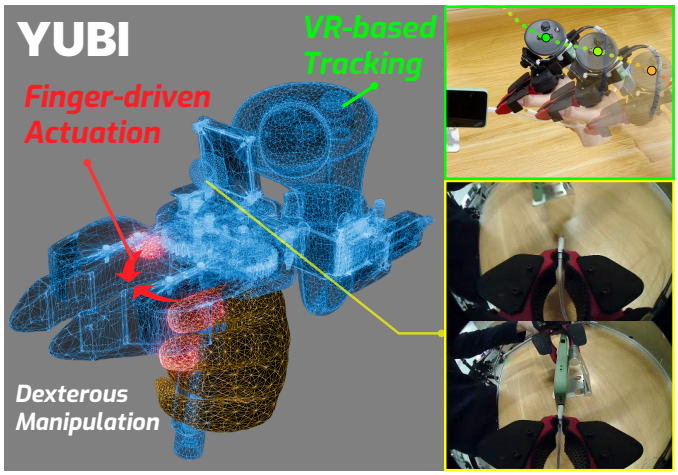

Fig. 1. **Yielding Universal Bidigital Interface (YUBI):** Our lightweight, finger-aligned gripper offers intuitive control by mirroring human digital kinematics for dexterous manipulation. Leveraging high-precision VR-based tracking, YUBI facilitates the curation of a large-scale, high-quality bimanual dataset to advance robotic foundation models.

Despite the success of initial handheld designs, scaling to increasingly complex tasks reveals ergonomic and functional bottlenecks, as shown in Table I. Most existing UMI frameworks use a pistol-grip design, which introduces a mechanical offset between the operator's fingers and the gripper's pinch-point. This offset, combined with the heavy and bulky shell, often obstructs natural hand movements, limiting the manipulability and dexterity of the gripper. Furthermore, the gripper's tracking fidelity remains a hurdle. Unlike SLAM-based tracking in [5], [9]–[11], recent systems [19], [20] have incorporated Virtual Reality (VR) systems for high-fidelity gripper tracking by mounting VR controllers. However, reliance on wearing the headset introduces substantial physical strain on the neck, which limits the session length and operational efficiency (up to 30 min of continuous use [12]).

To address these limitations, we propose **YUBI (Yielding Universal Bidigital Interface)**, a novel finger-aligned gripper for intuitive and sustained data collection in bimanual dexterous manipulation. As shown in Figure 1, YUBI introduces *yielding, finger-driven actuation*, where the gripper aperture follows the operator's natural pinch motion, eliminating the mechanical offset caused by pistol-grip UMI designs and offering intuitive haptic feedback to the fingers.

Combined with the gripper redesign, the YUBI system in-

TABLE I
COMPARATIVE ANALYSIS OF UMI-BASED RESEARCH

Method	UMI [5] FastUMI [22]	ActiveUMI [20] exUMI [19]	YUBI (Ours)
Gripper	Pistol-grip	Pistol-grip	*Finger-driven*
Weight	780 g	($\geqslant$ 905 g)	319 g
Ergonomics	◑ (bulky)	○ (heavy HMD)	● (haptic ↑)
Dexterity	◑	◑	●
Tracking	SLAM	VR	VR
Tracking prec.	◑	●	●
Data size (hrs)	12 / (60)	– / 5	2730
#Tasks	4 / 22	6 / 9	40

●: Superior, ◑: Moderate, ○: Limited. / Parenthesized values are estimates.

tegrates high-frequency VR sensors directly into each gripper. This ensures high-fidelity gripper trajectory tracking while mitigating the tracking drift common in SLAM-based systems. We further improve data collection throughput via a decoupled configuration that mounts the VR headset on a stationary camera rig rather than mounting it on the operator's head, thereby eliminating physical load. This camera rig is additionally equipped with a stereo camera to provide a top-view visual stream for stable workspace monitoring. Collectively, our system enables the sustained, high-fidelity data collection sessions necessary for training robotic foundation models.

Leveraging this framework, we provide the largest UMI-based dataset for bimanual dexterous manipulation, comprising 2730 hours of data across 300K episodes and 40 tasks. Our user study finds YUBI adaptable to diverse tasks from daily to industrial scenarios, while exhibiting measurable gains in both operational efficiency and task success rates. We will open-source the YUBI hardware and the dataset upon publication.

Our contributions are summarized as follows:

- We propose YUBI, a novel finger-aligned gripper with yielding jaw actuation designed to collect massive, high-quality data for bimanual dexterous tasks.
- We present a decoupled operation setup that mounts the VR headset on the rig and the VR controller on the gripper, enabling high-fidelity and long data collection.
- We offer a large-scale dataset for diverse bimanual dexterous manipulation tasks, comprising 2730 hours of interaction data with 300K episodes and 40 tasks.
- We demonstrate the system's efficacy through a comparative user study with the original UMI device, suggesting improved dexterity and operational efficiency.

II. METHOD

YUBI is designed to curate massive, high-quality UMI-based data for bimanual dexterous manipulation tasks. This section presents the design principles of YUBI (Section II-A), our operation setup with an integrated camera and VR rig (Section II-B), and data processing details (Section A).

A. YUBI Design

The design of YUBI aims to shift from the conventional pistol-grip interaction used in prior UMI works [5], [9]–[11] to yielding, finger-aligned actuation, enabling more intuitive operation and supporting data collection for dexterous manipulation.

Limitation of prior UMI grippers: Although the existing UMI gripper provides a portable platform for in-the-wild data collection [5], it has critical limitations in real-world scenarios: *limited adaptability to dexterous manipulation tasks and long-duration sessions* due to its bulky and heavy interface.

The conventional pistol-grip interface introduces a mechanical offset between the operator's fingers and the gripper's pinch-point, significantly reducing haptic transparency [16]. This lack of tactile feedback causes operators to apply excessive gripping force (overcompensation), while mechanical backlash in the gears further hinders fine motor control.

For the gripper tip design, previous works adopt Fin-Ray type designs [5], [8], [10], enabling soft, compliant grasping; however, their deformation characteristics result in poor positional repeatability and insufficient gripping force for heavy-object handling ($\geqslant$ 2 kg) [7], [18]. These limitations are particularly critical for real-world assembly tasks, which demand both precise handling of small components (*e.g.*, nuts) and the ability to handle heavy industrial parts.

Lastly, the device weight induces early fatigue on the operator's wrist; $\approx$ 780 g in the original UMI gripper and $>$900 g when integrated with the VR controller [19], [20]. Such physical stress introduces trajectory noise and drift in demonstrations, undermining the dataset quality.

Shift to a yielding finger-aligned design: To address the above limitations, we introduce a yielding, finger-driven actuation design into YUBI (see Figure 3). Specifically, one gripper jaw is actuated by the thumb, while the opposing jaw is driven by the coordinated motion of the index and middle fingers. Each jaw yields directly to its driving finger, so the gripper aperture follows the operator's natural pinch motion without motor-driven resistance. This configuration mitigates control mismatch and improves haptic transparency, enabling operators to directly leverage their inherent dexterity in manipulation tasks.

To preserve the precision of finger-driven actuation while supporting loads of up to 2 kg, YUBI employs a dimensionally stable gripper design. An integrated support grip serves as a mechanical fulcrum; while the index and middle fingers actuate the jaws, the remaining fingers stabilize the grip to distribute loads across the entire hand. Additionally, finger geometry is optimized to balance reach against load capacity. By minimizing the moment arm, YUBI ensures the structural stiffness required for heavy objects without sacrificing accessibility to confined spaces.

To mitigate wrist fatigue caused by previous heavy designs (780–900 g), YUBI adopts a miniaturized gripper architecture and a lightweight camera module, reducing the handheld mass to approximately 319 g (200 g gripper + 119 g controller). This

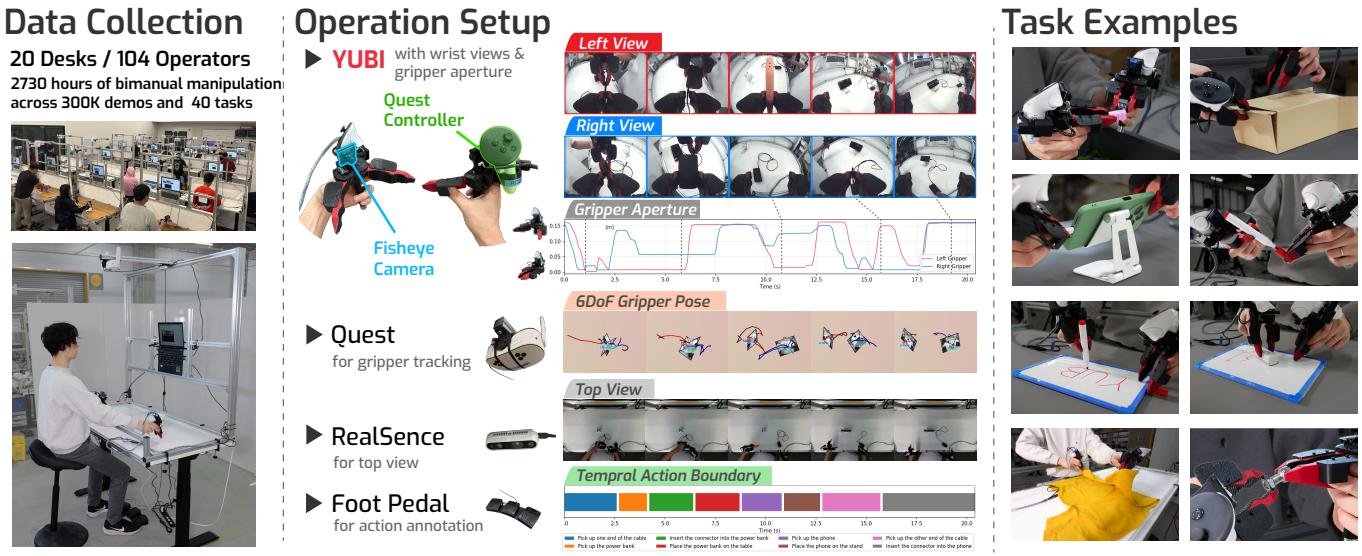

Fig. 2. **Overview:** To ensure massive, high-quality robot manipulation data for scaling robot learning, we collect bimanual YUBI-based demonstrations at 20 desks in parallel from 104 operators. Our operation setups feature (i) a stereo top-view camera for stable workspace observation, (ii) a rig-mounted VR system for 6 DoF gripper tracking where the VR controllers are attached to the YUBI gripper, and (iii) foot pedal-based action segmentation for hands-free control. The collected dataset features diverse 40 tasks that require various skills including assembling bricks and paper boxes, insertion, writing and erasing, folding clothes, and tightening nuts and bolts.

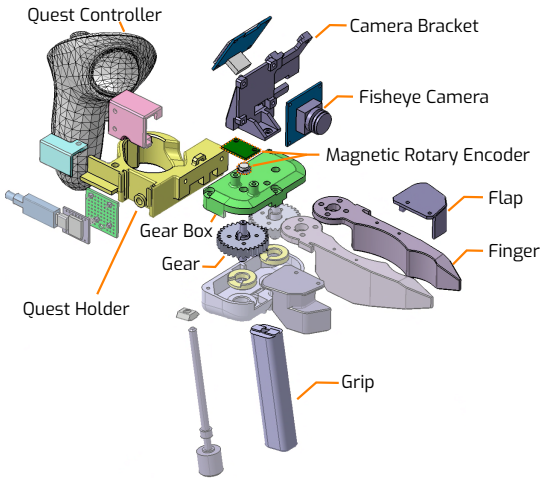

Fig. 3. **Exploded view of the YUBI gripper:** The bidigital mechanism utilizes internal gears to actuate the jaws, all supported by an ergonomic grip and flap. A fisheye camera is attached for task observation. The Quest controller is integrated to provide high-frequency 6-DoF trajectory tracking.

reduction in weight burden enables sustained demonstrations over long-duration sessions.

Overall, YUBI's finger-aligned design resolves the trade-off between dexterous manipulation and heavy-object handling while reducing operator fatigue for sustained data collection.

B. Operation Setup

As shown in Figure 2, we design an operation setup tailored for sustained, bimanual manipulation tasks where the operator grasps one YUBI device in each hand. For the sake of data quality, we adopt a fixed desktop setup rather than a fully portable setup. The stationary setup is integrated with VR systems (Quest) for 6 DoF controller tracking, a top-view stereo camera (RealSense) for workspace monitoring, a task user interface (UI) on a laptop and a foot pedal. These visual observations, trajectories, and gripper signals are transmitted to the laptop in real time.

YUBI: Two YUBI devices are installed on the desk to be used with each hand. Each YUBI device is equipped with an onboard wrist camera with 180° FOV, a Quest controller, and a magnetic encoder to measure the gripper aperture, tracked at 100, 80, and 100 Hz, respectively. Design details are found in Section II-A.

VR-based gripper tracking: We use the Quest 3S to track the 6 DoF trajectory of the controller mounted on YUBI, yielding higher tracking fidelity than SLAM-based systems, which are prone to drift, scale ambiguity, and failure under fast motion or textureless environments. Unlike the head-worn VR-based UMI systems (*e.g.*, ActiveUMI [20] and exUMI [19]), our setup mounts the heavy headset on the fixed rig. This reduces neck fatigue while ensuring tracking coverage of the controllers from the mounted headset.

Fixed stereo camera: We use the RealSense D435, rigidly mounted on the forward-extending frame, to capture the task workspace. This camera provides a stable top-down view of the workspace at 30 Hz (RGB + depth). The observations can be used for additional supervision signals for VLA training, such as object tracking, fine-grained action annotation, etc.

Task UI: A laptop is also mounted at the center of the frame wall, serving as the central hub of the system and the user interface for task instruction. It aggregates all sensor streams, while displaying real-time observations, tracking signals, and task information to the operator.

A major challenge in bimanual operation is that both hands are occupied by the grippers, leaving no free hand to annotate task transitions and sub-action boundaries through conventional inputs (*e.g.*, buttons, touchscreens, keyboards).

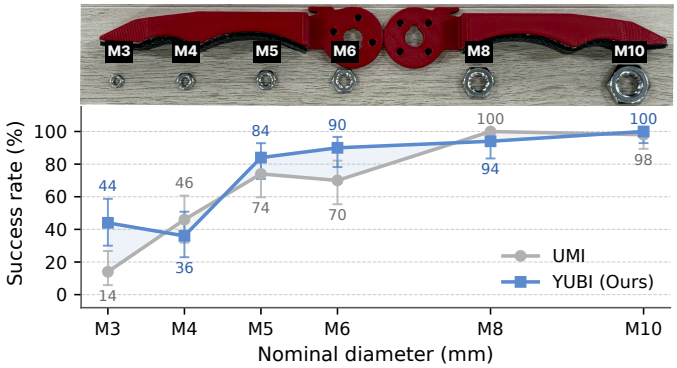

Fig. 4. **Dexterity test:** Pick-and-place success rate on hex nuts M10–M3 for UMI and YUBI. Error bars show 95% binomial confidence intervals (Clopper–Pearson, n=50). YUBI's advantage widens as nut size decreases.

To address this, we install a *foot pedal* under the desk to start/stop action segments and to accept an episode for storage. Operators press the pedal to indicate boundaries between predefined sub-action steps.

III. Results

A. Data Statistics

As shown in Figure 2, we collected YUBI-based manipulation data at scale across 20 desks. The resulting data consist of 2730 hours across 300K episodes and 40 tasks. Data collection was conducted 24/7 over one month by 104 operators (73 male, 31 female). Our dataset is significantly larger than previous UMI-based datasets, such as the Fast-UMI data (≈ 60 hours and 22 tasks) [22] and the original UMI data (12 hours and 4 tasks) [5]. Detailed statistical analysis is found in the appendix.

B. Dexterity Test

To quantitatively evaluate this advantage, we conducted a user study assessing dexterity in grasping objects of varying sizes using nuts. Here, operators were instructed to pick up a nut from a table and place it into a tray. Six standardized hex nut sizes (M10–M3) were prepared, and operators performed the pick-and-place task sequentially from the largest to the smallest. Note that operators had only one attempt per nut. If they failed to grasp or place it, it was counted as a failure, and they proceeded to the next size. A total of 10 operators were recruited with gender balanced, each performing five attempts.

Figure 4 compares the success rates of UMI and YUBI. On the larger nuts (M8 and M10), both devices approach the ceiling ($\geqslant 94\%$). As the diameter decreases, however, the two methods diverge. YUBI maintains a clear margin over UMI at M6 ($+20$ pp) and M5 ($+10$ pp), and the gap widens most markedly at the smallest size: on M3 nuts, YUBI achieves a 44% success rate while UMI drops to 14%, a roughly $3\times$ improvement. The sole exception to this trend appears at M4, where YUBI's success rate dips below its own M3 rate. We hypothesize that a geometric mismatch exists between the nut diameter and the fingertip curvature, *i.e.*, a size-specific contact artifact rather than a loss of precision. Indeed, we find that UMI also exhibits the same dip from M6→M5, supporting this interpretation. Overall, these results suggest that YUBI offers more robust performance for precision tasks than UMI.

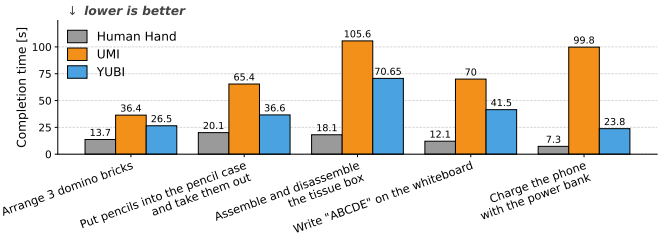

Fig. 5. **Operation efficiency.** Mean completion time (seconds) on five tasks for Hand, UMI, and YUBI. YUBI is significantly faster than UMI and narrows the gap to direct hand operation.

C. Operational Efficiency Test

We further demonstrate that the intuitive design and improved manipulability of YUBI enable faster data collection compared to existing interfaces. In our experiment, we measured the time required to complete a task for five distinct tasks. Each participant was assigned two tasks and performed them under three manipulation conditions: manual execution (Hand), UMI [5], and YUBI. For each task-device combination, operators completed five trials, and the average completion time was computed across trials. A total of 10 operators were recruited, with gender balanced, resulting in four operators assigned to each task. To mitigate order effects, operators were counterbalanced such that half performed the conditions in the order Hand→UMI→YUBI, while the remaining half followed Hand→YUBI→UMI.

Figure 5 shows the average task completion time of each device on the five evaluation tasks. Across all tasks, YUBI consistently outperforms UMI, with speed-up ratios ranging from $1.37\times$ on the domino arrangement task to $4.19\times$ on the phone-charging task. The results demonstrate that YUBI substantially narrows the efficiency gap toward the human hand compared with UMI, with clear advantages on tasks that require fine and dexterous manipulation, such as connector insertion in the phone-charging task and precise tool control in the whiteboard-writing task.

IV. Conclusion

We present **Yielding Universal Bidigital Interface (YUBI)**, a novel finger-aligned gripper with yielding jaw actuation designed to collect massive, high-quality data for bimanual dexterous tasks. The ergonomic gripper design enables fine control and heavy-object handling while reducing fatigue for sustained data collection. Our operation setup also supports high-fidelity gripper tracking with a decoupled VR system and stable visual observations from the top-view camera. The resulting dataset is the largest to date, comprising 2730 hours of interaction data across 300K episodes and 40 distinct tasks. Our user study demonstrates the gripper's advantages in precise grasping and operational efficiency. By open-sourcing the hardware configurations and dataset, our work offers a scalable and reproducible path toward large-scale, high-fidelity data collection for robotic foundation models.

REFERENCES

[1] Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_0.5$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

[2] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[3] Thanpimon Buamanee, Masato Kobayashi, Yuki Uranishi, and Haruo Takemura. Bi-act: Bilateral control-based imitation learning via action chunking with transformer. In *2024 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*, pages 410–415. IEEE, 2024.

[4] Remi Cadene, Simon Alibert, Alexander Soare, Quentin Gallouedec, Adil Zouitine, Steven Palma, Pepijn Kooijmans, Michel Aractingi, Mustafa Shukor, Dana Aubakirova, Martino Russi, Francesco Capuano, Caroline Pascal, Jade Choghari, Jess Moss, and Thomas Wolf. Lerobot: State-of-the-art machine learning for real-world robotics in pytorch. https://github.com/huggingface/lerobot, 2024.

[5] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *Robotics: Science and Systems XX*, 2024.

[6] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022.

[7] Chuanyu Li, Chaoyi Liu, Daotan Wang, Shuyu Zhang, Lusong Li, Zecui Zeng, Fangchen Liu, Jing Xu, and Rui Chen. Vitamin-b: A reliable and efficient visuo-tactile bimanual manipulation interface. *arXiv preprint arXiv:2511.05858*, 2025.

[8] Siwei Liang, Yixuan Guan, Jing Xu, Hongyu Qian, Xiangjun Zhang, Dan Wu, Wenbo Ding, and Rui Chen. Alltact fin ray: A compliant robot gripper with omni-directional tactile sensing. *IEEE Transactions on Automation Science and Engineering*, 2025.

[9] Fanqi Lin, Yingdong Hu, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. Data scaling laws in imitation learning for robotic manipulation. In *The Thirteenth International Conference on Learning Representations*, 2024.

[10] Fangchen Liu, Chuanyu Li, Yihua Qin, Jing Xu, Pieter Abbeel, and Rui Chen. Vitamin: Learning contact-rich tasks through robot-free visuo-tactile manipulation interface. *arXiv preprint arXiv:2504.06156*, 2025.

[11] Kehui Liu, Zhongjie Jia, Yang Li, Pengan Chen, Song Liu, Xin Liu, Pingrui Zhang, Haoming Song, Xinyi Ye, Nieqing Cao, et al. Fastumi-100k: Advancing data-driven robotic manipulation with a large-scale umi-style dataset. *arXiv preprint arXiv:2510.08022*, 2025.

[12] Meta Platforms, Inc. Health and safety warnings – Meta Quest 3S. https://www.meta.com/legal/quest/health-and-safety-warnings/quest-3s/, 2026. Accessed: 2026-04-23.

[13] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12999–13008, 2023.

[14] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.

[15] Ryan Punamiya, Simar Kareer, Zeyi Liu, Josh Citron, Ri-Zhao Qiu, Xiongyi Cai, Alexey Gavryushin, Jiaqi Chen, Davide Liconti, Lawrence Y Zhu, et al. Egoverse: An egocentric human dataset for robot learning from around the world. *arXiv preprint arXiv:2604.07607*, 2026.

[16] Longyan Wu, Checheng Yu, Jieji Ren, Li Chen, Yufei Jiang, Ran Huang, Guoying Gu, and Hongyang Li. Freetacman: Robot-free visuo-tactile data collection system for contact-rich manipulation. *arXiv preprint arXiv:2506.01941*, 2025.

[17] Philipp Wu, Yide Shentu, Zhongke Yi, Xingyu Lin, and Pieter Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12156–12163. IEEE, 2024.

[18] Jingyi Xu, Tamay Aykut, Daolin Ma, and Eckehard Steinbach. 6dls: Modeling nonplanar frictional surface contacts for grasping using 6-d limit surfaces. *IEEE Transactions on Robotics*, 37(6):2099–2116, 2021.

[19] Yue Xu, Litao Wei, Pengyu An, Qingyu Zhang, and Yong-Lu Li. exumi: Extensible robot teaching system with action-aware task-agnostic tactile representation. In *Conference on Robot Learning*, pages 2536–2554. PMLR, 2025.

[20] Qiyuan Zeng, Chengmeng Li, Jude St John, Zhongyi Zhou, Junjie Wen, Guorui Feng, Yichen Zhu, and Yi Xu. Activeumi: Robotic manipulation with active perception from robot-free human demonstrations. *arXiv preprint arXiv:2510.01607*, 2025.

[21] Tony Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *Robotics: Science and Systems XIX*, 2023.

[22] Zhaxizhuom Zhaxizhuoma, Kehui Liu, Chuyue Guan, Zhongjie Jia, Ziniu Wu, Xin Liu, Tianyu Wang, Shuai Liang, Pengan Chen, Pingrui Zhang, et al. Fastumi: A scalable and hardware-independent universal manipulation interface with dataset. In *Conference on Robot Learning*, pages 3069–3093. PMLR, 2025.

TABLE II
STATISTICS OF THE YUBI BIMANUAL MANIPULATION DATASET.

Statistic	Value
# Episodes	300K
Total duration (hours)	2730
# Tasks	40
Avg. sub-actions per task	8.66
Avg. episode duration (s)	32.9

APPENDIX

A. Data Processing

Dataset Structure: Each recording is stored as a single episode file that includes three synchronized camera streams (two wrist cameras and RealSense), relative poses expressed in the wrist frames, and metadata for the task and segmented sub-actions (avg. 8.66 steps per task). Specifically, we store the translational components $\mathbf{t}^{\text{right}}, \mathbf{t}^{\text{left}} \in \mathbb{R}^3$ and rotational components $\mathbf{r}^{\text{right}}, \mathbf{r}^{\text{left}} \in \mathbb{R}^3$, where rotation is represented as Euler angles. We additionally record the fingertip jaw angles for each hand, $d^{\text{right}}, d^{\text{left}} \in \mathbb{R}$. All sensor streams are initially recorded in rosbag2 format at their native frequencies and are subsequently converted to the LeRobot format [4], standardized at 30 Hz, to facilitate downstream VLA training.

Calibration: Since the gripper trajectories have independent origins across different sessions or desk setups, we align them to a shared table frame via a dedicated calibration session. We estimate a rigid transformation between the Quest and table coordinates, where a calibration board placed on the table is observed from multiple viewpoints using the wrist-mounted camera of YUBI. We then compute the relative transformation between the calibration board (defining the table coordinate system) and the Quest tracking frame. This transformation enables all tracked trajectories, including hand and controller motions, to be expressed consistently in the table-centered coordinate system.

Data Filtering: We remove defective episodes through a cascade of detectors as follows. (i) We remove episodes shorter than a minimum duration that can occur when operators accidentally press the foot pedal. (ii) Stuck-signal detectors flag three freeze patterns: *full pose freeze*—a pose signal whose standard deviation is at or below numerical precision

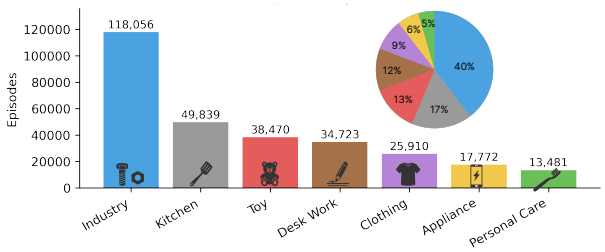

Fig. 6. **Domain distribution:** The figure presents the share of the 40 tasks across seven categories (industrial, kitchen, toy, desk work, clothing, appliance, personal care), reflecting YUBI's target scope of precise, heavy, and everyday object handling.

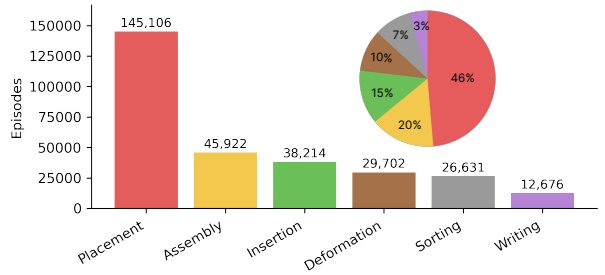

Fig. 7. **Skill distribution:** The figure indicates the share of tasks by primary skill type, such as placement, insertion, assembly, deformation, sorting, and writing. Labels reflect each task's dominant skill; most tasks combine multiple skills in practice.

over the entire episode; *partial pose freeze*—a contiguous run of frames whose per-frame translation $|\Delta \mathbf{t}|$ stays below a channel-specific noise floor for longer than a duration threshold; and *aperture freeze*—a gripper whose jaw-angle variance is near zero across the episode. (iii) Kinematic-plausibility detectors additionally drop episodes containing a single-frame translational jump exceeding an implausible velocity threshold given the recording rate, or a rotation increment greater than 90° in either roll or yaw within a single frame.

B. Data Distribution

The detailed dataset statistics are found in Table II. We show the domain and skill distribution of the recorded tasks in Figures 6 and 7. Given the high versatility of YUBI for precise and heavy object handling, we focus primarily on industrial settings along with daily scenarios, such as kitchen, toy, desk work, clothing, appliance, and personal care. The skill diversity is also crucial for general-purpose foundation models. We intend to cover various skill types including placement, insertion, assembly, deformation, sorting, and writing. While each task is categorized based on a major skill type, most tasks require a composition of different skills. For instance, a "writing on a whiteboard" task requires pick-and-place of the whiteboard marker, tactile-sensitive writing and erasing motions, and insertion of the cap onto the marker.