EgoThumb: A Vision-Driven Supernumerary Robotic Thumb for Autonomous Grasp and Holding Assistance

Jianyu Hou[1*], Jiawei Fu[1*], Weikang Wan[1], Minghui Jiang[1], Jiajun Li[1], Hao Su[2]

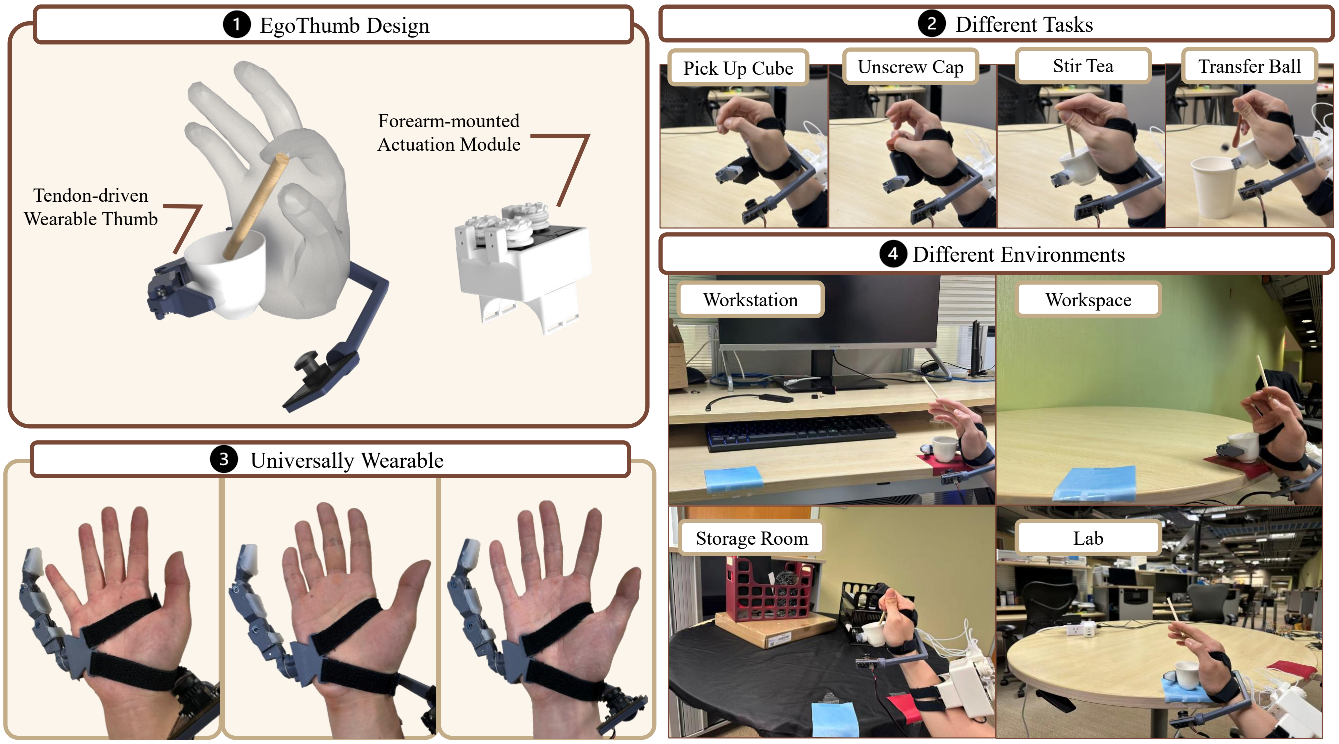

Fig. 1: **Overview of our EgoThumb System. (1) EgoThumb Design**: We present a fully open-source, vision-driven supernumerary robotic thumb, featuring a compact, tendon-driven mechanism with forearm-mounted actuation. **(2) Different Tasks**: EgoThumb enables a single hand to independently execute complex tasks that typically require bimanual coordination. **(3) Universally Wearable**: The device is secured via adjustable hook-and-loop straps, ensuring seamless physical adaptation across diverse hand morphologies. **(4) Different Environments**: The system achieves strong generalization performance across varied, unstructured environments.

***Abstract*— Supernumerary robotic fingers augment human dexterity by enabling a single hand to perform complex "hold-and-manipulate" tasks that typically require bimanual coordination. However, existing devices are often hindered by bulky hardware that limits interaction to large objects and rely on explicit hand motion triggers that lack environmental context. To address these challenges, we present EgoThumb, a fully open-source, vision-driven supernumerary robotic thumb. EgoThumb features a compact, dual tendon-driven finger whose actuation is relocated to the forearm, reducing distal weight while enabling precise active joint control. The system utilizes a visual imitation learning policy that maps egocentric images from a palm-facing camera directly to continuous joint positions. By implicitly decoding human intent from visual cues capturing both the hand and the object of interest, EgoThumb achieves robust autonomous manipulation and strong generalization across diverse users and environments. All hardware designs and software implementations are fully open-sourced at `https://ego-thumb.github.io`.**

I. Introduction

The enhancement of human physical capabilities through robotic systems has long been a central goal of robotics research. Existing literature in this domain typically falls into three branches: exoskeletons, prostheses, and supernumerary limbs. Exoskeletons attach motors to human joints to improve existing capabilities, such as playing instruments [1] and walking [2]. Prosthetic devices replace missing limbs to restore lost functions, including hand manipulation [3] and locomotion [4]. In contrast, supernumerary robotic (SR) limbs attach additional robotic arms and grippers to the human body to augment functional capacity [5]. For instance, body-mounted robotic arms can support heavy loads during overhead assembly tasks [6].

While much of this prior work focuses on proximal limb augmentation, there is growing interest in mounting robotic fingers directly onto the hand or forearm to extend the

*These authors contributed equally.
[1]The authors are with the Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA 92093, USA. {jeh089, jif017, w2wan, mijiang, jil186}@ucsd.edu
[2]Hao Su is with Sudo AI GmbH, 8050 Zurich, Switzerland. academic@haosu.ai

degrees of freedom of the human hand [7]. These additional digits can actively collaborate with biological fingers, enabling a single hand to perform tasks that typically require bimanual coordination. They have been explored across contexts ranging from assistive and therapeutic use for motor impairments [8] to able-bodied augmentation for skilled tasks [9]. Specifically, we focus on "hold-and-manipulate" tasks [10], where the SR finger must: 1) autonomously decide when to grasp and release an object, and 2) stably hold the object while the human hand interacts with it.

Despite their potential, existing SR fingers face two key limitations. First, their bulky hardware [8, 11, 12] suits large objects but hinders interaction with the smaller, everyday objects central to daily manipulation, such as the teacup in Fig. 1.2 (Stir Tea). Second, they are typically driven by teleoperation [13, 14] or low-dimensional hand-motion triggers [10, 11], which lack the autonomy and environmental awareness needed to infer intent from context, such as grasping the bottle in Fig. 1.2 (Unscrew Cap) as the hand approaches, without an explicit trigger motion.

To address these limitations, we present **EgoThumb**, a fully open-source, vision-driven supernumerary robotic thumb system. EgoThumb actively perceives hand motion and the environment through vision to autonomously determine the optimal timing for grasping, holding, and releasing objects. The key contributions of our work are as follows:

- **Hardware Design**: Our system utilizes a dual-tendon actuation strategy for each joint, enabling active control of both flexion and extension to ensure precise position control. Thus, the actuators are located on the forearm, resulting in a lightweight, wearable finger design.
- **Egocentric Perception**: To achieve autonomous control with environmental awareness, we integrate a palm-facing camera that captures the egocentric visual context of both the hand motion and the surroundings.
- **Intelligent Autonomy**: We formulate the control problem as a visual imitation learning task, training a policy to map egocentric images directly to joint positions.

Through these hardware and algorithmic advancements, EgoThumb achieves robust performance across four distinct real-world tasks without explicitly measuring hand motion. It also demonstrates strong generalization capabilities across different users and varied environmental backgrounds.

II. Related Work

A. Design of Supernumerary Robotic Fingers

While direct-driven mechanisms provide the high torque needed for large-scale supernumerary robotic (SR) limbs [5, 6], applying this architecture to SR fingers [8, 9, 11, 12, 15] results in bulky devices. This bulk limits interaction to large objects and causes user fatigue due to significant distal mass. Similarly, pneumatic designs suffer from substantial volumetric overhead [16]. Although compact tendon-driven alternatives exist [13, 14], they remain predominantly closed-source and rely on passive, spring-based extension, which degrades joint positioning accuracy. In contrast, EgoThumb is an open-source, compact SR finger whose dual-tendon drive actively controls flexion and extension for precise joint positioning. Placing the actuators on the forearm decouples the finger dimensions from motor size, unlike joint-mounted designs, minimizing distal weight and enabling a compact form suited to everyday objects.

B. Control of Supernumerary Robotic Fingers

Existing control strategies for SR fingers largely rely on teleoperation [9, 13, 14, 16–19], which inherently lacks autonomy. Alternatively, some systems guide finger motion by mapping explicitly measured, low-dimensional hand kinematics [8, 10, 11, 15]. However, these reactive approaches are insufficient when grasping must be initiated based on environmental context rather than explicit hand modulation. In contrast, EgoThumb infers grasp timing from visual context via imitation learning, rather than relying on teleoperation or explicit hand-motion triggers.

C. Imitation Learning for Robot Hands

Imitation learning is a promising paradigm for learning robot manipulation policies from expert demonstrations. Recently, wearable data collection systems, such as UMI [20], DexUMI [21], and DexWild [22], have highlighted the efficacy of end-to-end vision-motor policies for dexterous manipulation by leveraging diverse human data. However, these frameworks target standalone robots, overlooking the physical human-robot interaction inherent to supernumerary devices. EgoThumb instead frames wearable-finger control as visual imitation learning in a shared workspace, directly mapping egocentric images to joint positions.

III. System Design

In this section, we present the architectural design of the EgoThumb system, the teleoperation framework utilized for data collection, and the formulation of the vision-driven imitation learning policy. To facilitate low-latency, cross-platform communication between these subsystems, we developed a communication middleware leveraging `Zenoh` [23] and `Protocol Buffers` [24]. We use Network Time Protocol (NTP) [25] to synchronize all the Raspberry Pis and PCs involved in the system.

A. EgoThumb Design

The EgoThumb system comprises two primary components: 1) a compact, tendon-driven wearable thumb with a wide field-of-view (FOV) palm-facing camera; and 2) a forearm-mounted actuation module.

1) Wearable Thumb and Palm-Facing Camera: The wearable thumb's kinematic structure mimics human anatomy (Fig. 2.2), incorporating three joints: the carpometacarpal (CMC), metacarpophalangeal (MCP), and interphalangeal (IP). The MCP and IP joints are modeled as 1-degree-of-freedom (DoF) revolute joints, consistent with the biological thumb. While the biological CMC joint exhibits complex, multi-DoF kinematics [26, 27], our system focuses primarily on grasping and holding tasks, making a high-DoF design

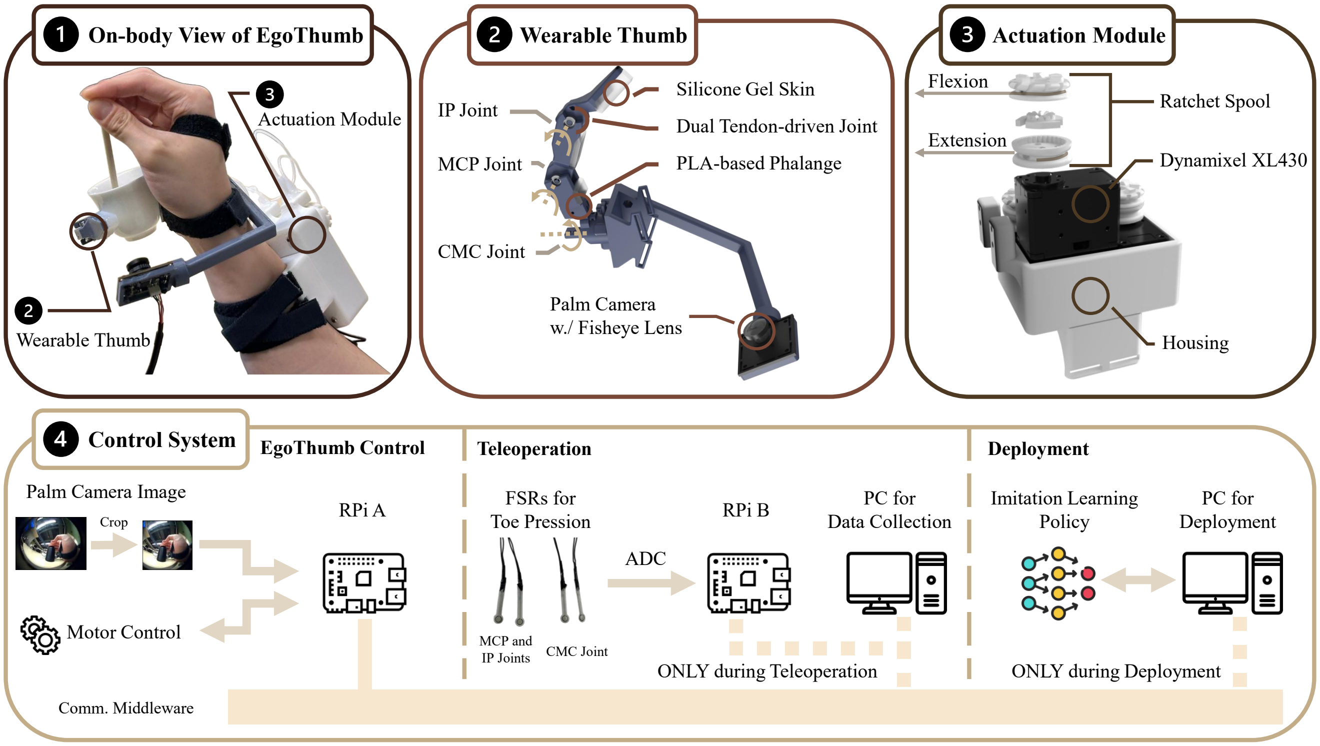

Fig. 2: **Overview of the EgoThumb System Design. 1) On-body View**: The system comprises a wearable thumb and a forearm-mounted actuation module. **2) Wearable Thumb**: The thumb features dual tendon-driven joints for precise control and PLA phalanges with a silicone-gel coating for stable grasping. A palm-facing fisheye camera provides egocentric visual feedback. **3) Actuation Module**: The actuation module houses three Dynamixel XL430-driven ratchet spools. This forearm-mounted configuration significantly reduces distal weight. **4) Control System**: Our EgoThumb interfaces with teleoperation and deployment systems via custom communication middleware.

unnecessary. Consequently, we simplify the CMC into a single-DoF revolute joint that rotates perpendicular to the palm plane [13, 14]. To maintain terminological consistency across all of the device's joints, we refer to the CMC joint's abduction and adduction as flexion and extension throughout this paper.

To avoid the bulk associated with direct-drive mechanisms, we adopt a tendon-driven architecture inspired by recent anthropomorphic robotic hands [28–31]. Specifically, we leverage the dual-tendon actuation design proposed by [28], where each joint is controlled by two tendons: one for flexion and one for extension. This dual-tendon mechanism enables active, precise position control in both directions. This stands in contrast to conventional tendon-driven SR fingers, which often rely on passive spring-based mechanisms for extension, leading to reduced control authority [10, 13, 14, 18]. The mechanical rotation ranges are 180° for the CMC joint, 84.14° for the MCP joint, and 141.34° for the IP joint. These ranges are determined by the joints' mechanical design and are sufficient to cover the grasping and holding postures required by our tasks. A calibration procedure is conducted to map the physical joint positions to the corresponding motor angles in the actuation module [28].

The phalanges are fabricated using Fused Deposition Modeling (FDM) with Polylactic Acid (PLA). To enhance grasp stability and friction, a custom-molded layer of Shore 10A silicone (`Dragon Skin 10`) is bonded to the palmar surface. The device is secured to the user's hand via two adjustable hook-and-loop straps—one traversing the purlicue and the other encircling the wrist—ensuring a secure fit across varying hand sizes and preventing slippage during operation.

The system perceives hand motion and environmental context via a global-shutter camera (Fig. 2.2) that is fixed to the wearable thumb on the palmar side of the hand and oriented from the hand root toward the fingers (i.e., palm-facing). Equipped with a fisheye lens (220° FOV) [22], the camera captures comprehensive visual data encompassing the hand, the robotic thumb, and the target object. Unlike DexUMI [21], which mounts the camera on the ulnar side, we position our sensor on the radial side to maintain an unobstructed view of the interaction zone. We favor this egocentric, palm-facing placement over a third-person camera to avoid the occlusions that arise as the wearable finger and hand move through the workspace. The raw 240×320 RGB frames are asymmetrically cropped to 224×224 pixels, prioritizing the bottom-right quadrant. This cropping strategy serves as a spatial prior, centering the human-robot interaction workspace while filtering out extraneous background noise. The policy consumes these raw cropped frames without geometric undistortion [20, 21] and requires no intermediate hand-pose estimation, avoiding on-device vision overhead.

2) Actuation Module: The actuation module (Fig. 2.3) houses three actuator-spool units, each comprising a ratchet spool driven by a `Dynamixel XL430-W250-T` servo motor. The ratchet spool anchors the two tendons, converting motor rotation into flexion/extension movements. A key feature of this design is the ability to rapidly re-tension the tendons to mitigate slack accumulation [28]. The `Dynamixel` motors provide a stall torque of 1.4 N·m, ensuring stable grasping force, while their 12-bit encoders (0.088° resolution) enable precise proprioceptive feedback.

Tendons are routed from the wearable unit to the actuation module through low-friction Polytetrafluoroethylene (PTFE) tubes. The high axial stiffness of the PTFE tubing maintains invariant tendon path lengths during wrist articulation. This effectively decouples the robotic finger's motion from the user's wrist movements, ensuring consistent control regardless of hand pose.

A `Raspberry Pi 3B+` (RPi A) serves as the embedded controller (Fig. 2.4, Left), managing motor control loops at 60 Hz and image acquisition at 20 Hz. It interfaces with the teleoperation and deployment systems via the aforementioned middleware. The RPi runs only these control and acquisition loops; policy inference is offloaded to the external PC, so it performs no vision processing. The 20 Hz acquisition rate is sufficient for the low-speed dynamics of our hold-and-manipulate tasks.

B. Teleoperation System

To facilitate data collection for imitation learning, we designed a customized foot-operated teleoperation interface [13, 14] (Fig. 2.4, Middle). The system utilizes four Force Sensitive Resistors (FSRs), each integrated into a voltage divider with a 10$k\Omega$ resistor and sampled by a 16-bit ADC (`TI ADS1115`) at 50 Hz. Signal processing is handled by a secondary `Raspberry Pi 3B+` (RPi B).

The control mapping is intuitive: the user modulates the left two FSRs with their left toe to control the flexion/extension velocity of the MCP and IP joints. Simultaneously, the right toe presses the right two FSRs to control the CMC joint. To reduce the control dimensionality for the operator, we enforce a linear kinematic synergy between the MCP and IP joints, defined as $\theta_{\text{IP}} = \frac{2}{3}\theta_{\text{MCP}}$.

During data collection, the palm-facing camera images, joint positions of the wearable thumb, and target actions from teleoperation are captured by RPi B and logged to a PC. The teleoperation hardware is disconnected during autonomous deployment.

C. Imitation Policy Learning

We formulate the control of the EgoThumb as a vision-based imitation learning problem. We adopt the widely used Diffusion Policy framework [20, 32], which models multimodal action distributions and generates smooth action sequences suited to continuous joint control; the control policy is defined as $a_{t+1:t+H} = \pi(I_{t-T:t}, q_{t-T:t})$. At time step t, a_t represents the target action, I_t denotes the cropped palm-facing camera image, and q_t represents the current proprioceptive joint positions. Following the short observation history of UMI [20], we set $T = 2$ and use a prediction horizon of $H = 8$, which we found to work well on our tasks.

The visual input I_t is processed by a CLIP-pretrained [33] Vision Transformer (ViT) [34] to extract robust latent representations, which are then concatenated with the proprioceptive states. The core generative backbone is a ConditionalUnet1D, which maps the concatenated features to a sequence of future actions through iterative denoising. To ensure precise synchronization between the asynchronous visual stream and hardware states, we implement continuous latency matching [20]. During deployment, as shown in Fig. 2.4 (Right), the policy runs on an external PC, communicating real-time inference commands to the wearable system via the middleware.

IV. Experiments

Autonomous supernumerary fingers must dynamically adapt to human hand motions, accommodate various hand shapes, and function reliably across diverse environments. To evaluate our proposed system, we investigate the following research questions: **Q1**: Can our EgoThumb assist grasping and free human hands for other motion? **Q2**: Can our system generalize to *different hands*? **Q3**: Can our system generalize to *"in-the-wild" environments*?

A. Evaluatoin Setup

We evaluate our wearable finger across four different real-world tasks that involve human-robot collaboration, adapted from prior hand-augmentation studies [13, 14], as shown in Fig. 3:

- **Pick Up Cube**: The user reaches toward a rigid cube placed in the initial area with their right hand. The wearable thumb grasps the cube and then the user lifts the cube off the table.
- **Unscrew Cap**: The user approaches a capped bottle in an initial area. The wearable thumb secures the bottle's body, enabling the user to lift it. While the device maintains a stable grasp, the user employs their biological fingers to pinch and rotate the cap. Finally, the user returns the bottle, and the wearable thumb disengages.
- **Stir Tea**: The user, holding a stirring stick, approaches a cup positioned in a red initial area. The wearable thumb secures and stabilizes the cup, allowing the user to lift it and stir inside. The user then transports the cup to a target blue area and places it down, at which point the wearable thumb disengages.
- **Transfer Ball**: A teacup containing a small ball is positioned at a red initial area. The user approaches the cup with a spoon. The wearable thumb secures the cup, allowing the user to lift and tilt it. The user then employs the spoon to guide the ball out of the tilted teacup and into a secondary target receptacle. Following the transfer, the user transports the teacup to a blue target region and places it down, prompting the wearable thumb to disengage.

A demonstrator collects 50 episodes of training dataset for each task using our teleoperation system (Sec. III-B). The objects are randomly placed at the initial area and moved to

Fig. 3: **Execution Sequences for the Four Evaluation Tasks.** Policy rollouts demonstrating collaborative human-robot manipulation across four tasks: **Pick Up Cube**, **Unscrew Cap**, **Stir Tea**, **Transfer Ball**. For each task, the top row displays the user's first-person view of the interaction, while the bottom row shows the corresponding egocentric palm-facing camera view used by the policy. The timelines below outline the coordinated sequence of autonomous EgoThumb actions and human motions.

	Pick Up Cube	Unscrew Cap	Stir Tea	Transfer Ball	Avg.
Demonstrator	1.0	0.7	0.9	1.0	0.90
User 1	1.0	0.9	0.8	1.0	0.925
User 2	1.0	0.8	0.8	0.9	0.875

TABLE I: **Success Rates across Different Hands.** We report the performance when objects are placed within initial and target areas in the training data (Fig. 4, Top, darker regions). EgoThumb system achieves strong performance across all four tasks when operated by the demonstrator, and it generalizes effectively to novel participants (User 1 and User 2).

	Pick Up Cube	Unscrew Cap	Stir Tea	Transfer Ball
Demonstrator	1.0	0.8	0.6	0.2
User 1	1.0	0.7	0.3	0.1
User 2	1.0	0.7	0.7	0.2
Avg.	1.00	0.73	0.53	0.17

TABLE II: **Success Rates in Expanded Initial and Target Areas.** We evaluate our system on an expanded initial and target areas (Fig. 4, Top, lighter regions) compared to the training data. Results show strong robustness on **Pick Up Cube** and **Unscrew Cap**, moderate robustness on **Stir Tea**, and a clear drop on **Transfer Ball**, likely due to increased foreground visual variation and its longer-horizon manipulation requirements.

a random location of the target area for each task, as shown in Fig. 4 (Top, marked with darker regions). The imitation learning policy is trained on Intel I9-9960X and NVIDIA TITAN RTX for 120 epochs.

We conduct 10 evaluation episodes per task. For each evaluation episode, we use the same initial and target areas as the training data. To assess the generalization of our system to different hands, two other human operators (whose hands are not used in training data collection, User 1 and User 2) participate in this study. We maintain a consistent initial object configuration across different users. The evaluation is conducted on Intel I9-9960X and NVIDIA RTX 2080 Ti. The policy rollouts for four tasks are shown in Fig. 3.

B. Results and Analysis

Our wearable finger enables the human hand to perform concurrent tasks by providing stable grasp assistance (Q1): As detailed in Tab. I (Demonstrator row), our system achieves an average success rate of 0.9 across all four tasks when evaluated on the demonstrator's hand (the in-distribution hand). In every scenario, the system demonstrates the ability to: 1) precisely determine the optimal timing for grasping and releasing objects; and 2) maintain a stable grasp during dynamic hand interactions. We attribute this to our occlusion-free camera placement and the dual-tendon, silicone-skinned finger that ensures secure grasping and efficient release.

Our finger system demonstrates robustness across varying hand morphologies (Q2): Adjustable hook-and-loop fasteners allow the wearable thumb to physically adapt across users. However, novel hands introduce out-of-distribution (OOD) visual scenarios, since training data was collected solely from the demonstrator's hand. Specifically, anatomical variations alter the human hand's appearance and relative pose within the camera's field of view. Despite these visual discrepancies, Tab. I (User 1 and User 2 rows) illustrates that our system maintains average success rates exceeding 0.875 for both User 1 and User 2 across all four tasks. This sustained success suggests strong general-

Envs	Seen #1	Seen #2	Unseen #1	Unseen #2
Demonstrator	0.8	1.0	0.7	0.9
User 1	0.9	1.0	0.7	0.6
User 2	0.7	1.0	0.6	0.7
Avg. User	0.80	1.00	0.67	0.73

TABLE III: **Task Success Rates Across Different Backgrounds.** We evaluate our policy on two seen and two unseen environments using the Stir Tea Task. The policy maintains robust performance across both seen and entirely novel (unseen) backgrounds for all users.

ization to novel users; the system effectively learns to attend to human-object interactions and accurately identifies the optimal timing and regions for grasping. We attribute this robustness to the rigidly mounted camera, which keeps the robotic thumb's pose constant in the visual frame regardless of hand size while centering the view on the interaction and its environmental context.

Our system is robust to variations in object distributions (Q3): To assess resilience against spatial variations, we evaluated the system by enlarging the initial and target areas for the objects compared to training data, as depicted in Fig. 4 (Top, lighter regions). As shown in Tab. II, the system consistently achieves average success rates over 0.73 in the **Pick Up Cube** and **Unscrew Cap** tasks across all users. Performance degrades slightly to an average success rate of 0.53 for the **Stir Tea** task and drops to 0.17 for the **Transfer Ball** task. We hypothesize that these variances stem from differing visual and temporal complexities. The **Pick Up Cube** task serves as a robust baseline despite spatial variation. Similarly, in the **Unscrew Cap** task, although the initialization area is expanded, the primary background remains static and distant, preserving critical visual features. In contrast, the **Stir Tea** task suffers because variations in the initial and target zones introduce significant foreground color and texture shifts, creating substantial visual distractors. Finally, the **Transfer Ball** task exhibits the lowest performance due to its long-horizon nature, which demands precise sequential planning for identifying grasp regions, stabilizing the object during interaction, and executing timely releases.

Our finger system demonstrates robustness to background variations (Q3): To evaluate robustness against diverse visual backgrounds and lighting conditions, we conducted an extended "in-the-wild" study on the **Stir Tea** task. The control policy was trained on a broadened dataset of 105 demonstrations collected across 9 distinct environmental settings (Fig. 4, Seen Environments). We evaluated the policy in two in-distribution (seen) environments and two completely novel (unseen) environments (marked #1 and #2 in Fig. 4, Bottom). Tab. III shows that our system maintains average success rates above 0.8 across three users in the seen environments. In unseen environments, performance declines slightly but remains above 0.67. To investigate the drivers of this generalization, we varied the training dataset size and evaluated performance on environments Seen #2 and Unseen #2. As illustrated in Fig. 5, scaling the number of training demonstrations yields a corresponding increase in success rates across both settings for all users. This confirms that the system effectively learns generalizable visual representations invariant to environmental distractors, demonstrating that

Fig. 4: **Task Area and "In-the-Wild" Environments. Top (Task Area):** Darker areas represent the initial and target areas for training data, while the larger, lighter areas (including the original darker areas) represent the expanded areas used for the robustness test of object distributions. **Pick Up Cube** and **Unscrew Cap** utilize the green areas, expanded from 10 cm × 10 cm to 50 cm × 10 cm. **Transfer Ball** and **Stir Tea** use red for pick-up and blue for drop-off, expanded from 10 cm × 10 cm to 30 cm × 25 cm. For the Transfer Ball task, expanding the evaluation area shifts the target receptacle's position ("Target Change"). **Bottom ("In-the-wild" Environments):** The "Seen Environments" panel displays the diverse environments included in the broadened training dataset. The "Unseen Environments" panel displays completely novel backgrounds used strictly to test out-of-distribution generalization. Labels #1 and #2 denote the specific seen and unseen test environments.

data scaling successfully drives out-of-distribution robustness rather than simply overfitting to specific backgrounds.

C. Ablation Study

We conducted an ablation study on the latency matching and image cropping modules to isolate their contributions. Both ablations were trained and evaluated on the **Unscrew Cap** task using the in-distribution demonstrator. To ensure a fair comparison, the training dataset, policy architecture, and hyperparameters remained identical to the main experiments. Quantitative results are detailed in Tab. IV.

*1) **Latency Matching:*** Precise temporal alignment is critical for continuous control; latency matching synchronizes mul-

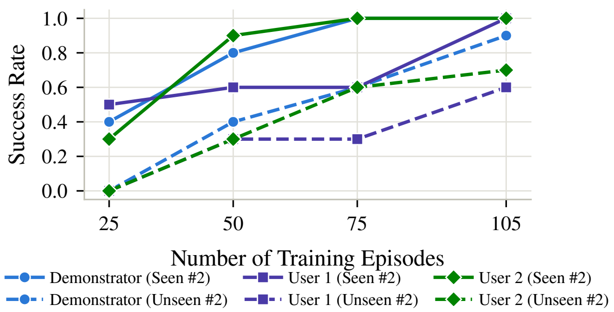

Fig. 5: **Impact of Training Data Scale on Generalization.** Scaling the demonstrations from 25 to 105 generally improves policy performance in both seen (solid lines) and unseen (dashed lines) environments across all users, proving that more data drives robust out-of-distribution generalization.

System Configuration	Success Rate
Full System	0.7
w/o Latency Matching	0.0
w/o Image Cropping	0.0

TABLE IV: **Ablation Study Results.** Disabling either latency matching or image cropping results in catastrophic policy failure, confirming their necessity for robust visuo-motor control.

timodal observation streams and mitigates hardware pipeline delays [20].

Ablated Setup: We disabled all temporal compensation. During training, we replaced latency-compensated continuous interpolation with a naive nearest-neighbor approach, matching images to the chronologically closest joint state. During inference, we queried the most recent joint position instead of interpolating to the image timestamp. Additionally, we disabled the forward predictive action delay—which normally offsets the computational latency between policy querying and physical execution—and executed predictions immediately.

Results: The ablated policy failed completely (0 success rate vs. 0.7 for the full system). This catastrophic drop confirms that temporal alignment and pipeline delay compensation are necessary for dynamic, high-precision human-robot coordination.

*2) **Image Cropping:*** Explicit spatial priors are crucial for filtering spurious environmental distractors and learning robust visuo-motor policies.

Ablated Setup: We bypassed our standard image cropping—which normally isolates the primary region of interest (the hand and target object), as shown in Fig. 2.4 (Left). Instead, uncropped camera frames were fed directly into an ImageNet-pretrained [35] ResNet34 [36] encoder during both training and inference. We use ResNet34 because it can directly process the uncropped 240×320 input resolution, whereas CLIP does not support this input size.

Results: The policy with uncropped image input also failed completely (0 success rate vs. 0.7 for the full system). Without this spatial bottleneck, the policy cannot isolate fine-grained manipulation features and likely overfits to background noise, proving that image cropping is an indispensable preprocessing step for our vision pipeline.

V. Conclusion and Limitation

We presented EgoThumb, an open-source, vision-driven supernumerary thumb that augments human dexterity in "hold-and-manipulate" tasks. By pairing a lightweight, compact, tendon-driven joint design with egocentric visual imitation learning, the system autonomously decodes human intent and generalizes robustly across diverse users and environments, validating our palm-facing camera design for capturing invariant visual representations. Several limitations remain and motivate future work. Performance degrades on long-horizon tasks with large spatial variation in the initial and target areas (**Stir Tea** and **Transfer Ball**), motivating hierarchical learning frameworks. Rapid hand motions can induce motion blur; this does not affect our low-speed hold-and-manipulate tasks, but high-speed manipulation remains unevaluated. The palm-facing camera protrudes from the hand and may reduce wearing comfort, motivating a more compact, integrated sensor. Our data were collected under relatively consistent hand approach angles, leaving robustness to varied approach angles untested. Finally, our evaluation relies on binary task success over 10 trials per condition; finer measures of coordination quality (e.g., grasp timing and motion smoothness) and larger-scale evaluation with statistical analysis are also needed.

References

[1] Shinichi Furuya et al. "Surmounting the ceiling effect of motor expertise by novel sensory experience with a hand exoskeleton". In: *Science Robotics* 10.98 (2025), eadn3802.

[2] Tianchi Chen et al. "Biarticular Rigid Powered Lower Extremity Exoskeleton Robot". In: *IEEE Robotics and Automation Letters* (2025).

[3] Rebecca J Greene et al. "Evaluating the impact of a semi-autonomous interface on configuration space accessibility for multi-dof upper limb prostheses". In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2024, pp. 154–161.

[4] Jairo Maldonado-Contreras et al. "Transfer Learning for Walking Speed Estimation Across Novel Prosthetic Devices and Populations". In: *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2025, pp. 12019–12025.

[5] Federico Parietti and Harry Asada. "Supernumerary robotic limbs for human body support". In: *IEEE Transactions on Robotics* 32.2 (2016), pp. 301–311.

[6] Federico Parietti, Kameron Chan, and H Harry Asada. "Bracing the human body with supernumerary robotic limbs for physical assistance and load reduction". In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2014, pp. 141–148.

[7] Faye Wu and Harry Asada. "Supernumerary robotic fingers: an alternative upper-limb prosthesis". In: *Dynamic Systems and Control Conference*. Vol. 46193. American Society of Mechanical Engineers. 2014, V002T16A009.

[8] Teddy Ort et al. “Supernumerary robotic fingers as a therapeutic device for hemiparetic patients”. In: *Dynamic Systems and Control Conference*. Vol. 57250. American Society of Mechanical Engineers. 2015, V002T27A010.

[9] James Cunningham et al. “The supernumerary robotic 3 rd thumb for skilled music tasks”. In: *2018 7th IEEE International Conference on Biomedical Robotics and Biomechatronics (Biorob)*. IEEE. 2018, pp. 665–670.

[10] Faye Y Wu and H Harry Asada. ““Hold-and-manipulate” with a single hand being assisted by wearable extra fingers”. In: *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2015, pp. 6205–6212.

[11] Faye Y Wu and Haruhiko Asada. “Bio-artificial synergies for grasp posture control of supernumerary robotic fingers”. In: (2014).

[12] John Fasoulas, Michael Sfakiotakis, and Vasileios Bekas. “A mechatronic thumb for grasping and in-hand manipulation tasks”. In: *IEEE Transactions on Medical Robotics and Bionics* 6.2 (2024), pp. 733–745.

[13] Paulina Kieliba et al. “Robotic hand augmentation drives changes in neural body representation”. In: *Science robotics* 6.54 (2021), eabd7935.

[14] Dani Clode et al. “Evaluating initial usability of a hand augmentation device across a large and diverse sample”. In: *Science robotics* 9.90 (2024), eadk5183.

[15] Faye Y Wu and H Harry Asada. “Implicit and intuitive grasp posture control for wearable robotic fingers: A data-driven method using partial least squares”. In: *IEEE Transactions on Robotics* 32.1 (2016), pp. 176–186.

[16] Xinquan Liang et al. “A fabric-based wearable soft robotic limb”. In: *Journal of Mechanisms and Robotics* 11.3 (2019), p. 031003.

[17] Xinquan Liang et al. “Design and characterization of a novel fabric-based robotic arm for future wearable robot application”. In: *2017 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE. 2017, pp. 367–372.

[18] Sun Weijun et al. “Kinematic modeling and prototype experiment of flexible supernumerary robotic finger”. In: *2024 8th International Conference on Robotics, Control and Automation (ICRCA)*. IEEE. 2024, pp. 13–17.

[19] Rateb Katmah et al. “Cognitive Adaptation and Cortical Connectivity Alterations during Daily Tasks with Supernumerary Robotic Finger Assistance”. In: *Journal of Biomedical Engineering and Biosciences (JBEB)* 11.1 (2023), pp. 01–07.

[20] Cheng Chi et al. “Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots”. In: *arXiv preprint arXiv:2402.10329* (2024).

[21] Mengda Xu et al. “Dexumi: Using human hand as the universal manipulation interface for dexterous manipulation”. In: *arXiv preprint arXiv:2505.21864* (2025).

[22] Tony Tao et al. “Dexwild: Dexterous human interactions for in-the-wild robot policies”. In: *arXiv preprint arXiv:2505.07813* (2025).

[23] Angelo Corsaro et al. “Zenoh: Unifying communication, storage and computation from the cloud to the microcontroller”. In: *2023 26th Euromicro Conference on Digital System Design (DSD)*. IEEE. 2023, pp. 422–428.

[24] Google LLC. *Protocol Buffers*. https://protobuf.dev/. Accessed: 2026-02-01.

[25] David Mills et al. *Network time protocol version 4: Protocol and algorithms specification*. Tech. rep. 2010.

[26] Josephine M Kalshoven et al. “The passive biomechanics of the thumb carpometacarpal joint: an in vitro study”. In: *Journal of biomechanics* 168 (2024), p. 112129.

[27] Eni Halilaj et al. “In vivo kinematics of the thumb carpometacarpal joint during three isometric functional tasks”. In: *Clinical Orthopaedics and Related Research*® 472.4 (2014), pp. 1114–1122.

[28] Clemens C Christoph et al. “ORCA: An open-source, reliable, cost-effective, anthropomorphic robotic hand for uninterrupted dexterous task learning”. In: *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2025, pp. 8503–8510.

[29] Anya Zorin et al. “RUKA: Rethinking the design of humanoid hands with learning”. In: *arXiv preprint arXiv:2504.13165* (2025).

[30] Shadow Hand. *Shadow Hand*. https://shadowrobot.com/. [Accessed 26-02-2026].

[31] Dex Hand. *Dex Hand*. https://www.dexhand.org/. [Accessed 26-02-2026].

[32] Cheng Chi et al. “Diffusion policy: Visuomotor policy learning via action diffusion”. In: *The International Journal of Robotics Research* 44.10-11 (2025), pp. 1684–1704.

[33] Alec Radford et al. “Learning transferable visual models from natural language supervision”. In: *International conference on machine learning*. PmLR. 2021, pp. 8748–8763.

[34] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[35] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[36] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.