DexEXO: A Wearability-First Dexterous Exoskeleton for Operator-Agnostic Demonstration and Learning

***Abstract*— Scaling dexterous robot learning is constrained by the difficulty of collecting high-quality demonstrations across diverse operators. Existing wearable interfaces often trade comfort and cross-user adaptability for kinematic fidelity, while embodiment mismatch between demonstration and deployment requires visual post-processing before policy training. We present DexEXO, a wearability-first hand exoskeleton that aligns visual appearance, contact geometry, and kinematics at the hardware level. DexEXO features a pose-tolerant thumb mechanism and a slider-based finger interface analytically modeled to support hand lengths from 140 mm to 217 mm, reducing operator-specific fitting and enabling scalable cross-operator data collection. A passive hand visually matches the deployed robot, allowing direct policy training from raw wrist-mounted RGB observations. User studies demonstrate improved comfort and usability compared to prior wearable systems. Using visually aligned observations alone, we train diffusion policies that achieve competitive performance while substantially simplifying the end-to-end pipeline. These results show that prioritizing wearability and hardware-level embodiment alignment reduces both human and algorithmic bottlenecks without sacrificing task performance. dexexo-research.github.io**

I. Introduction

Learning robust dexterous manipulation remains fundamentally limited by the availability of scalable, high-fidelity demonstrations that capture the closed-loop, contact-rich strategies humans employ in daily tasks [1–5]. Although recent advances in robot learning have shown strong gains from larger and more diverse human datasets, collecting such data for multi-finger hands remains particularly difficult due to high-dimensional kinematics, frequent occlusions, and complex hand–object contact dynamics [6–13]. In contrast to parallel-jaw grippers, where portable teaching interfaces scale effectively [14], high-DoF hands continue to rely on interfaces that trade off naturalness, wearability, and motion fidelity, especially in the thumb, whose abduction, adduction, and opposition enable complex in-hand manipulation.

Existing sources of dexterous demonstrations generally fall into three categories: (i) **simulation and videos**, (ii) **robot teleoperation**, and (iii) **wearable interfaces** such as gloves and exoskeletons. Video and simulation data scale efficiently and broadly [15–21], yet accurately capturing contact forces, fine hand–object interactions, and transferring them to hardware remains challenging [18, 22, 23]. Teleoperation provides demonstrations directly in the robot control space [24–27], but dexterous hand teleoperation is often slow, unintuitive, costly to scale, and limited by insufficient haptic feedback for contact-rich manipulation [24]. Wearable devices improve embodiment by mechanically coupling human motion to the robot, reducing retargeting ambiguity and enabling more natural demonstrations [28–30]. However, prior work shows that wearables can introduce a visual embodiment gap during data collection, requiring additional post-processing before training [28]. These systems also frequently sacrifice comfort for fidelity and can be difficult to fit across users due to anthropometric variation [31, 32].

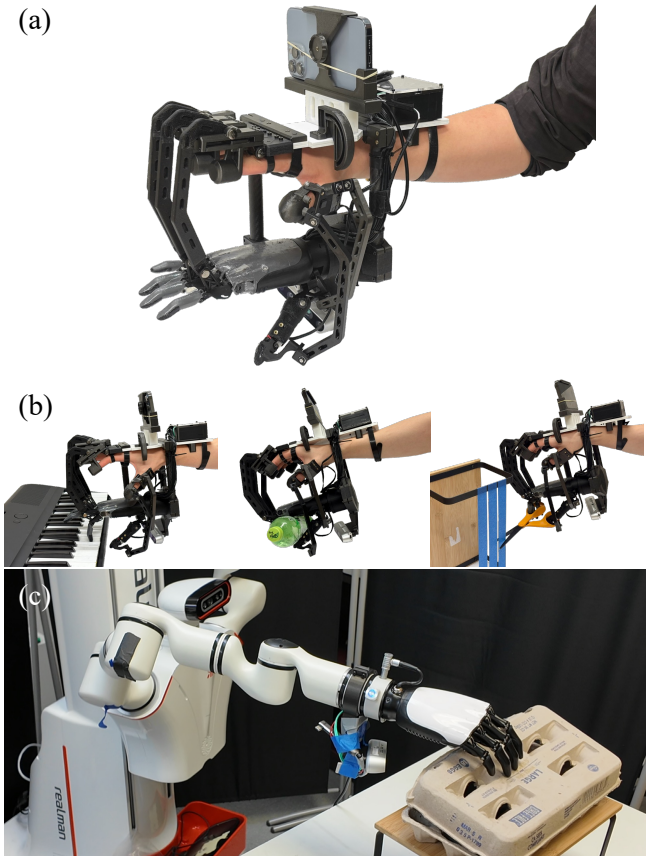

Fig. 1: **System overview of DexEXO.** (a) Full device worn on a user's hand. (b) Demonstrations of piano playing, full-hand grasping, and scissors cutting. (c) Policy deployment on the robot.

Motivated by these limitations, we present DexEXO, a wearable hand exoskeleton for data collection designed around two principles: **wearability and cross-user adaptability** to enable scalable, sustained demonstration collection, and **aligned visual and kinematic embodiment** for an efficient end-to-end pipeline from demonstration to policy training. Our approach preserves natural thumb and finger behaviors during data collection while maintaining a consistent, learnable mapping to the target robot hand. By incorporating a passive hand, the wrist-mounted camera view aligns with

that of the physical robotic hand, eliminating visual discrepancies between collection and inference. The design targets practical deployment scenarios in which operators perform repeated tasks across diverse environments with minimal setup, while retaining the motion and visual fidelity required for learning manipulation skills [29, 30, 33]. In summary, our contributions are:

- A wearability-first hand exoskeleton with analytically validated anthropometric compatibility, enabling cross-user operation without rigid alignment or calibration.
- A pose-tolerant thumb mechanism that preserves the natural human thumb workspace while maintaining a consistent, controllable mapping to robot thumb DOFs.
- An embodiment-aligned data collection and policy training pipeline that eliminates segmentation and visual post-processing, enabling direct learning from only raw wrist-mounted RGB observations.

II. Related Work

A. Teleoperation for Dexterous Manipulation

Teleoperation remains the dominant approach for collecting high-quality dexterous demonstrations, but existing interfaces involve inherent trade-offs. Vision-based systems provide an unencumbered user experience, yet they are fundamentally limited by line-of-sight occlusion and tracking instability during contact-rich interactions [15, 24, 34]. Data gloves alleviate these tracking challenges and can provide tactile feedback [35], but they introduce the "correspondence problem" [36]. Without physical constraints enforcing robot kinematics on the human hand, glove-based demonstrations may generate trajectories that are kinematically infeasible for the target robot, despite recent advances in retargeting algorithms [37, 38].

B. Learning from Human Hand Videos

To circumvent the hardware limitations of teleoperation, recent work has explored learning directly from large-scale human video data [15, 16, 39–41]. These approaches exploit the scale of internet videos to acquire rich visual and geometric priors for hand–object interaction [42–45]. However, video-based learning faces a "physicality gap," as it lacks explicit information about contact forces and closed-loop interaction dynamics [18, 22]. As a result, policies trained purely from video often struggle to transfer directly to physical systems and typically require additional fine-tuning on contact-rich demonstrations via teleoperation [7, 46].

C. Exoskeleton and Mechanically Coupled Interfaces

Exoskeleton and mechanically coupled interfaces aim to reduce embodiment mismatch by physically linking human motion to robot kinematics, improving controllability compared to loosely coupled systems [30, 33, 47]. Recent systems such as **DexUMI** [28] pursue scalable in-the-wild data collection through a lightweight wearable and vision-based reconstruction. However, their rigid exoskeleton geometry, derived from non-anthropomorphic robotic hand proportions, provides limited adaptation to diverse human hand sizes, increasing joint-alignment sensitivity across operators and potentially constraining ergonomic tolerance during sustained use [48]. Additionally, visual embodiment mismatch requires segmentation and inpainting prior to policy training. Devices such as **DexOP** [29] address embodiment alignment through hardware–robot co-design, employing linkage-driven passive mechanisms to enforce strong kinematic correspondence between the operator and robot hand. While this tight coupling improves demonstration fidelity, it binds the interface to specific robot geometries, limiting adaptability to diverse hand designs already used at scale. Moreover, rigid linkage constraints reduce tolerance to anthropometric variation and restrict residual thumb motion, constraining the natural abduction, adduction, and opposition workspace required for complex in-hand manipulation [31, 32, 49]. In contrast, our approach targets the intersection of wearability-first extended use and robust mapping by incorporating a pose-tolerant thumb mechanism that preserves natural thumb motion without sacrificing controllability.

III. Hardware Design

A. Hardware Overview

DexEXO comprises (i) a linkage-driven wearable exoskeleton, (ii) a passive demonstration hand, and (iii) an onboard sensing and power module for untethered operation. The passive hand follows the geometry of the 6-DoF OYMotion ROH-AP001 (ROHand) [50], featuring a 2-DoF thumb (IP flexion/extension and TM abduction/adduction) and single-DoF flexion for each finger. As shown in Fig. 2, the exoskeleton transmits operator motion through two coupling architectures. The four fingers use parallel linkage mechanisms that provide identical flexion/extension correspondence while accommodating inter-user variation. The thumb employs a multi-DoF coupling that allows the exoskeleton structure to translate and rotate relative to the palm while still transferring the key thumb motions to the passive hand, improving comfort and adaptability across hand sizes. A dorsal-mounted electronics module supplies onboard power and data logging, while a wrist-mounted iPhone provides pose capture for in-the-wild data collection.

B. Passive Hand

To ensure high-fidelity proprioception, DexEXO integrates six joint encoders within a rigid, rib-reinforced mounting structure that prevents misalignment, backlash, and sensor drift, maintaining stable joint-angle measurements during dynamic manipulation. In parallel, we performed kinematic identification using a URDF to design a custom linkage-slide mechanism that matches the actual hand kinematics, ensuring consistent and physically accurate trajectory mapping.

C. Wearability-First Design for Cross-Operator Deployment

DexEXO is designed for deployment without rigid joint alignment or per-user calibration. Instead of enforcing strict geometric coincidence between human and exoskeleton joints, passive tolerance mechanisms absorb anthropometric variation locally while preserving structured motion transfer.

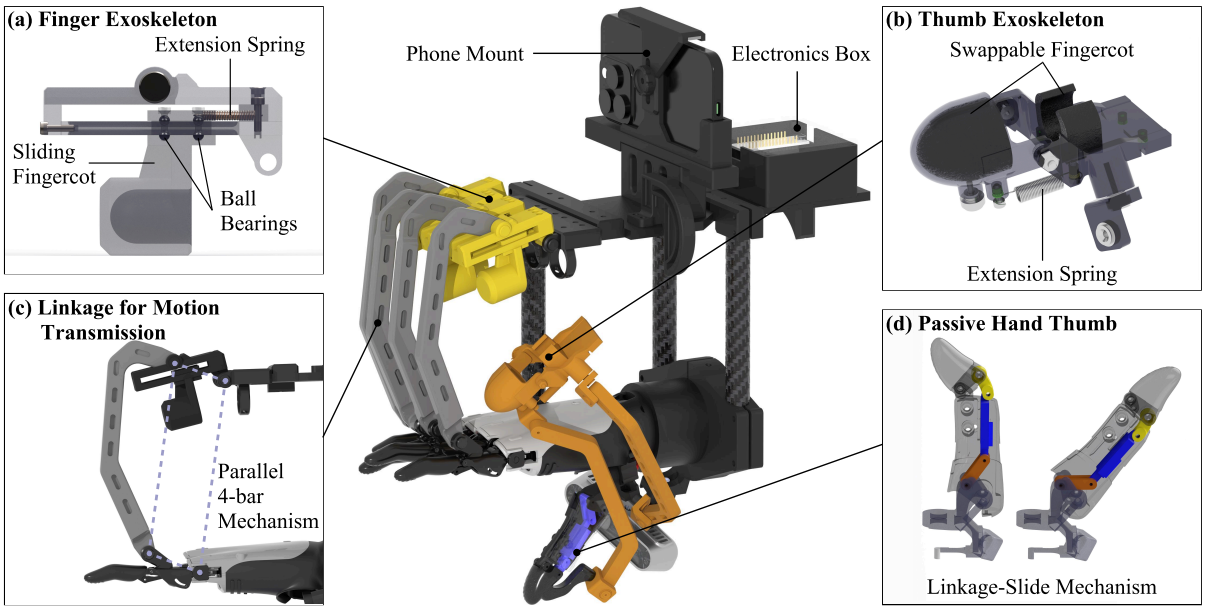

Fig. 2: Mechanical overview of DexEXO. DexEXO integrates a linkage-driven wearable exoskeleton, a passive data-capture hand, and an onboard sensing/power module. Insets highlight key subsystems: (a) passive finger slider for cross-user fit, (b) pose-tolerant thumb coupling interface, (c) parallel four-bar finger linkage for motion transmission, and (d) passive hand thumb that reproduces the intended thumb DOF

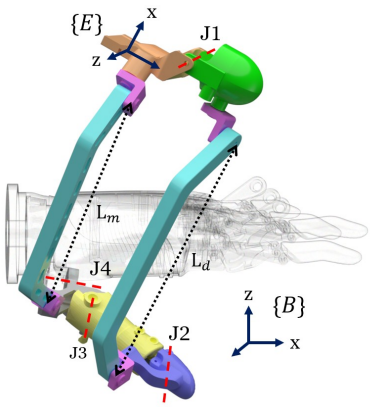

Fig. 3: Kinematic schematic of the exoskeleton thumb and its distal and metacarpal linkages. The four swivel joints (shown in purple) allow self-alignment between the exoskeleton frame E and the palm base frame B while maintaining fixed linkage lengths L_d and L_m.

1) Slider-Based Finger Interface: Each finger employs a passive spring-loaded linear slider coupled to a compliant fingercot. The slider fits variation in finger length while preserving sufficient curl displacement, decoupling insertion depth from joint-axis alignment and improving fit robustness.

To estimate the range of compatible hand sizes, we analyze the mechanical limits of the slider using the middle finger as the governing digit. Let L_{min} denote the minimum TPU ring–to–fingercot distance at rest, d_{max} the maximum distance permitted by slider travel, d_{curl} the minimum free slider length required for full finger flexion, δ the maximum allowable offset of the TPU ring above the webbing, and r the middle-finger-to-hand-length ratio ($r \approx 0.39$–0.40) reported in [51]. The compatible bounds are:

$$L_{max} = d_{max} - d_{curl}, \qquad MFL_{max} = L_{max} + \delta \quad (1)$$

$$H_{min} = \frac{L_{min}}{r}, \qquad H_{max} = \frac{MFL_{max}}{r} \quad (2)$$

Placement variability is incorporated by allowing the TPU ring to sit up to $\delta = 17\,\text{mm}$ above the webbing before restricting PIP flexion. Even accounting for the ring width (8 mm), sufficient clearance remains for natural finger curl.

Substituting $L_{min} = 56\,\text{mm}$, $d_{max} = 86\,\text{mm}$ (30 mm travel), $d_{curl} = 16\,\text{mm}$, $\delta = 17\,\text{mm}$, and $r = 0.40$ yields $L_{max} = 70\,\text{mm}$, $MFL_{max} = 87\,\text{mm}$, and a compatible hand-length range of $H_{min} = 140\,\text{mm}$ to $H_{max} = 217\,\text{mm}$. All participants in our user study ($n = 14$, 165–195 mm) fall within this range, validating cross-user compatibility.

2) Swappable Compliant Thumb Interface: Complementing the pose-tolerant linkage, the thumb employs a swappable TPU fingercot coupled to the distal linkage through compliant elements. Unlike rigid shells requiring precise axis alignment, the soft fingercot accommodates variation in thumb length and joint-center location while maintaining stable contact during pinch and grasp.

D. Pose-tolerant Thumb Mechanism

The thumb is challenging for wearable interfaces due to inter-user anatomical variability, where rigid axis alignment can cause discomfort and restrict motion. DexEXO addresses this with a pose-tolerant thumb coupling that preserves wearability while remaining functional with the robotic hand's IP flexion/extension and TM ab/ad configuration.

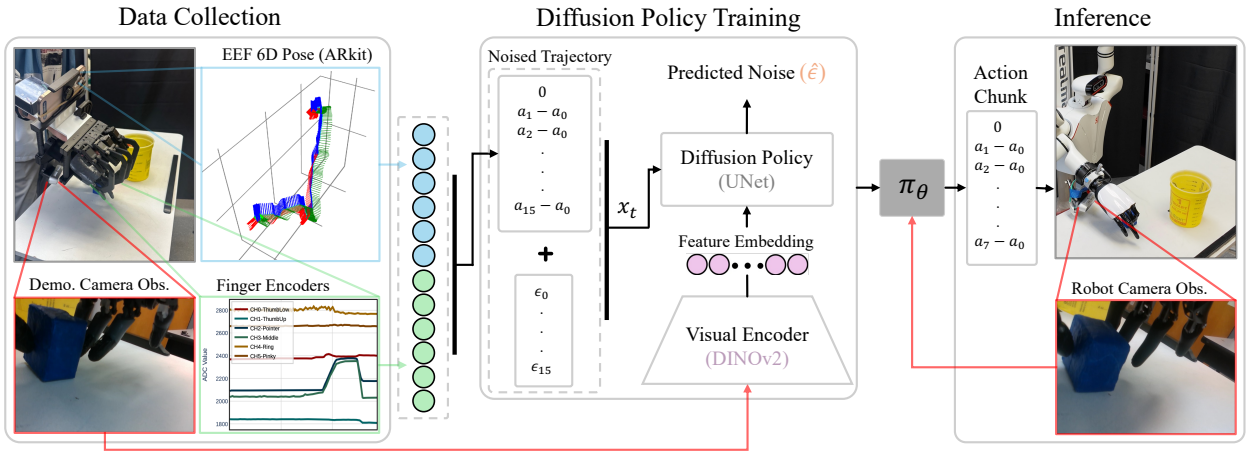

Fig. 4: An overview of the full demonstration data modalities, policy training, and inference with visual-aligned observations.

1) Mechanism overview: As illustrated in Fig. 3, the exoskeleton thumb contains an instrumented IP joint J_1 with angle θ_1. The passive thumb includes the IP joint J_2 with angle θ_2 and the TM ab/ad joint J_4 with angle θ_4, where J_3 is mechanically coupled to J_2. The exoskeleton thumb is connected to the passive thumb through two rigid linkages: a distal linkage and a metacarpal linkage. This architecture avoids enforcing rigid orientation alignment between the exoskeleton and the human thumb. Instead, only geometric distance constraints are imposed, enabling the exoskeleton to translate and rotate relative to the palm while remaining mechanically coupled.

2) Simplified kinematic model: Let $\{B\}$ and $\{E\}$ denote a palm-base frame and an exoskeleton frame. The relative pose of the exoskeleton with respect to the palm is

$$ {}^{B}T_{E} = \begin{bmatrix} {}^{B}R_{E} & {}^{B}p_{E} \\ 0 & 1 \end{bmatrix} \in SE(3), \tag{3} $$

.

Denote the passive thumb configuration as

$$ q_p \triangleq [\theta_2 \quad \theta_4]^{\top}, \qquad \theta_3 = f(\theta_2), \tag{4} $$

and let ${}^{B}r_d(q_p)$ and ${}^{B}r_m(q_p)$ be the distal and metacarpal attachment points expressed in $\{B\}$, computed from the passive hand kinematics.

The corresponding attachment points on the exoskeleton are constant vectors ${}^{E}\bar{r}_d$ and ${}^{E}\bar{r}_m$ expressed in $\{E\}$. Their positions in $\{B\}$ are

$$ {}^{B}r_i^E = {}^{B}R_E \, {}^{E}\bar{r}_i + {}^{B}p_E, \qquad i \in \{d, m\}. \tag{5} $$

The two-link coupling imposes holonomic distance constraints

$$ \|{}^{B}r_d^E - {}^{B}r_d(q_p)\| = L_d, \qquad \|{}^{B}r_m^E - {}^{B}r_m(q_p)\| = L_m, \tag{6} $$

where L_d and L_m are the distal and metacarpal link lengths.

3) Residual pose freedom and self-alignment.: The exoskeleton pose ${}^{B}T_E$ has six degrees of freedom, while Eqs. (6) impose two independent scalar holonomic constraints under typical thumb configurations. Consequently, for a fixed passive thumb posture q_p, the coupled system generically admits a four-dimensional self-motion manifold in the exoskeleton pose space. This residual freedom corresponds to the experimentally observed "wiggle space," in which the exoskeleton body can translate and rotate relative to the palm without altering the passive thumb posture.

IV. Data Collection and Policy Training

A. Data Collection

1) Finger Position Data: Finger joint positions are measured using six analog encoders embedded within the exoskeleton mechanism. Encoder values are sampled by an onboard microcontroller at 1 kHz and streamed to a host computer using a lightweight binary protocol. We account for the target hand's non-linear actuation kinematics by mapping the exoskeleton encoder data to actuator commands using piecewise linear interpolation across waypoints sampled at identical physical postures on both DexEXO and ROHand.

2) End-Effector Pose: The 6-DOF end-effector pose is captured using an iPhone-based AR tracking system through the TeleDex application [52]. Pose data, consisting of position and orientation, is streamed to the host computer and resampled to a fixed 60 Hz rate to ensure consistent timing.

3) Visual Observations: A wrist-mounted Intel RealSense camera records RGB images at 640×480 resolution and 30 Hz. Each frame is timestamped and stored for downstream policy training.

4) Time Synchronization: All sensor modalities are temporally aligned using video timestamps as the master reference. Asynchronous encoder and pose measurements are matched via nearest-neighbor association.

B. Policy Architecture and Training Setup

Our aligned embodiment enables an efficient pipeline from demonstration to policy training. In particular, the passive hand ensures that the wrist-mounted camera observes a

TABLE I: Summary of quantitative results in user study (mean ± SEM). Bold indicates best performance per metric.

	Scissors Cutting		Page Flipping		Cup Stacking		Piano Playing	
Method	Success Rate	Time (s)†	Success Rate	Time (s)†	Success Rate	Time (s)†	Success Rate	Time (s)†
DexEXO	$\mathbf{0.79 \pm 0.10}$	$\mathbf{11.7 \pm 1.4}$	$\mathbf{0.88 \pm 0.03}$	5.4 ± 0.6	$\mathbf{0.82 \pm 0.07}$	12.0 ± 1.1	$\mathbf{0.96 \pm 0.02}$	$\mathbf{21.6 \pm 1.8}$
DexUMI	0.00 ± 0.00	—	0.86 ± 0.04	$\mathbf{4.7 \pm 0.7}$	0.80 ± 0.07	$\mathbf{8.9 \pm 1.0}$	0.62 ± 0.13	25.9 ± 2.5
Teleoperation	0.00 ± 0.00	—	0.51 ± 0.06	18.0 ± 2.1	0.33 ± 0.09	68.6 ± 13.1	0.60 ± 0.09	97.4 ± 7.8

† Completion time was defined as the average time from picking up scissors to finishing the cut, time for 5 page flips, average time for a successful 3-cup stack, and time to play 16 piano notes, respectively.

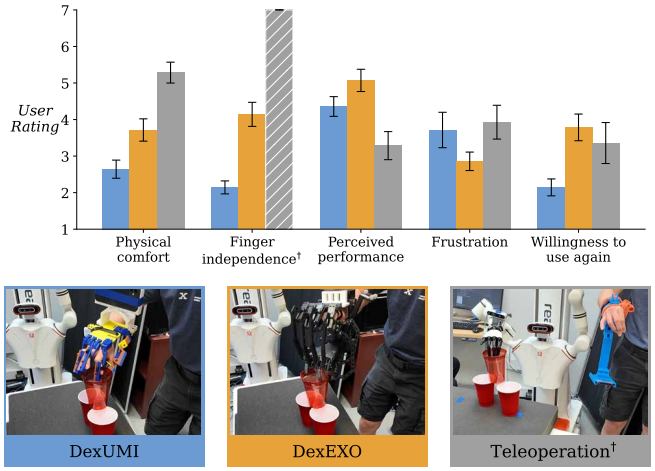

Fig. 5: Subjective feedback results in user study (mean ± std.). † Finger independence is not applicable to teleoperation, as the user's natural hand motion is unconstrained.

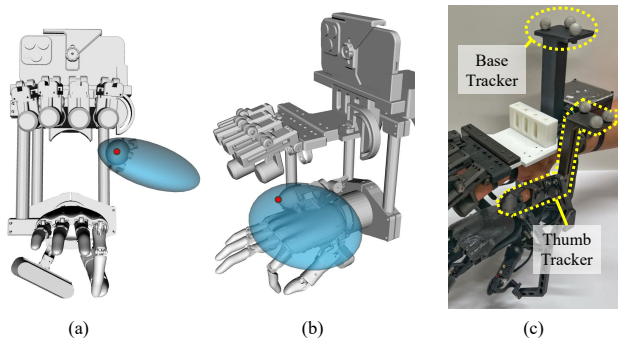

Fig. 6: Wiggle-space evaluation of the thumb interface. (a) and (b) Two views of the measured wiggle space. The fitted ellipsoid represents the covariance envelope of the sampled trajectory. The red point indicates the nominal fingertip position within the ellipsoid. (c) Experimental setup with reflective markers attached for motion-capture measurement.

hand geometry consistent with the deployed robotic hand, eliminating the visual embodiment gap that typically necessitates segmentation, masking, or inpainting [28]. As a result, policies are trained directly from raw wrist RGB observations paired with synchronized end-effector and finger signals.

a) Observations: Each training sample includes an RGB frame from the wrist-mounted camera and (optionally) a low-dimensional hand state. The RGB image is resized to 240×240, randomly cropped to 224×224, and augmented with color jitter during training. Visual features are extracted using a DINOv2 ViT-S/14 encoder [53], and the resulting embedding is used as the primary conditioning signal for the policy. When used, the hand state is the 6D absolute finger pose.

b) Actions: The policy outputs a 12D action consisting of a 6-DoF end-effector command and 6 finger commands. We train a diffusion policy [54] to predict an action horizon of 16 steps and execute the first 8 actions in a receding-horizon manner at inference time. Actions are expressed relative to the initial state of the horizon: the k-th predicted action corresponds to $T_k - T_0$. This representation supports reactive closed-loop control while retaining multi-step prediction capability.

All policies in this work use the same diffusion policy backbone and vision encoder; differences between action parameterizations and conditioning signals are evaluated in Sec. V.

V. Results

A. Experimental Validation of Thumb Wiggle Space

The residual pose freedom predicted in Sec. III-D manifests as a self-motion manifold of the exoskeleton relative to the base. As shown in Fig. 6 (c), to experimentally characterize the extent of this allowable motion, we conducted a wiggle-space experiment with the hand maintained in a pinch configuration. Reflective markers were attached to the base and the exoskeleton thumb linkage, and their relative pose was recorded using a motion-capture system for approximately 25 s while the user performed small natural adjustments within the interface.

The sampled marker positions were expressed in the base frame to obtain the relative motion between the hand and the exoskeleton. The resulting point cloud represents the allowable configuration space ("wiggle space") during pinch interaction. To summarize the spatial distribution of this motion, we fitted a 3-D ellipsoid to the sampled points using the covariance of the trajectory:

$$\Sigma = \frac{1}{N-1} \sum_{i=1}^{N} (p_i - \bar{p})(p_i - \bar{p})^\top,$$

where $p_i \in \mathbb{R}^3$ are the measured positions and $\bar{p}$ is the mean. The ellipsoid axes are obtained from the eigenvalues λ_i of Σ as

$$a_i = k\sqrt{\lambda_i},$$

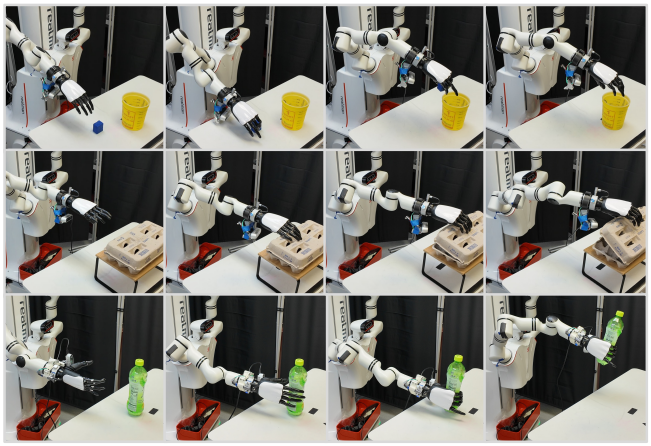

Fig. 7: Policy rollouts for the block pick and place, egg carton, and bottle tasks.

where $k = 2$ corresponds to approximately 95% coverage.

The fitted ellipsoid has semi-axis lengths of $66.12\,\mathrm{mm}$, $49.19\,\mathrm{mm}$, and $21.14\,\mathrm{mm}$. The large ellipsoidal volume indicates that the interface tolerates substantial variation in thumb placement while maintaining stable kinematic coupling. This tolerance enables the mechanism to accommodate inter-user variation in thumb morphology and placement without requiring precise anatomical alignment.

B. Demonstration User Studies

A user study was conducted to compare experience and performance across different demonstration methods. We recruited 14 university students (7 female, 7 male; aged 18–27) with hand sizes ranging from 165 mm to 195 mm. Participants engaged with with 3 demonstration devices: DexEXO, DexUMI, and vision-based teleoperation [52]. Teleoperation served as a baseline, as it is the most commonly used approach in prior work. For each device, participants were asked to perform the following tasks:

Scissors cutting: Pick up scissors and cut a strip of tape.
Page flipping: Use the fingertip to flip a notebook page.
Cup stacking: Stack 3 cups facing up.
Piano playing: Play 16 notes on a piano using 4 fingers.

For each task, we recorded success rate and completion time as quantitative metrics. All tasks were performed under a 120-second time limit, and completion time was capped at 120 seconds if unfinished. In addition to objective metrics, subjective feedback was collected using Likert-scale questions adapted from NASA-TLX dimensions [55]. We tested whether DexEXO receives higher subjective ratings than DexUMI using a Wilcoxon signed-rank test.

The quantitative results from the user study are shown in Table I. DexEXO was the only device capable of performing the scissors cutting task. DexUMI failed as its added exoskeleton geometry, absent in the original robot hand, prevented the fingers from fitting within the handles. Teleoperation failed at the same task due to a lack of precision, responsiveness, and force feedback. While DexUMI outperforms DexEXO in page flipping and cup stacking in terms of completion time (13.0% and 25.8% faster, respectively),

TABLE II: Policy evaluation comparisons across methods and tasks.

Method	**Tasks**		
Finger Condition	Block	Carton	Bottle
No	**0.90**	0.90	**0.85**
Yes	0.85	**0.95**	0.80

DexEXO achieves higher success rate in both tasks. DexEXO outperforms DexUMI significantly in the piano task, with 54.5% higher success and 16.6% faster completion time. It is also worth noting that teleoperation performs worse overall compared to both exoskeleton methods.

The subjective feedback results from the user study is shown in Fig. 5. Participants reported greater finger independence for the exoskeleton design ($p \ll 0.01$), which is consistent with DexEXO's superior performance in the piano task under the quantitative evaluation. DexEXO also received higher ratings in physical comfort ($p = 0.0127$) and lower frustration ($p = 0.0219$) compared to DexUMI, which can be attributed to the analytical design considerations to accommodate a wider range of hand sizes, as well as better dexterity from finger independence. Additionally, participants expressed greater willingness to use DexEXO again in future sessions ($p \ll 0.01$), and slightly higher perceived performance rating compared to DexUMI ($p = 0.0544$), supporting our overall hypothesis.

Notably, despite its lowest quantitative performance, teleoperation received the highest ratings in physical comfort and finger independence, attributable to the user's hand remaining unconstrained during operation. Overall, across both quantitative and subjective metrics, DexEXO demonstrated the strongest performance among the three devices, while offering improved physical comfort over DexUMI and greater efficiency over teleoperation.

C. Policy Evaluation

We evaluate whether aligned visual and contact geometry embodiment enables effective end-to-end policy learning without visual post-processing, and whether explicit hand-state conditioning remains necessary under this setting.

a) Experimental Setup: Policies are trained on demonstrations collected using DexEXO as described in Sec. IV. The Block task is trained on 200 demonstrations, while Carton and Bottle are trained on 150 demonstrations each. All policies are trained for 300–500 epochs until convergence under identical data splits and augmentation settings. We evaluate three representative manipulation tasks:

Block: Grasp a block and place it into a cup, testing precision and fingertip alignment.
Carton: Open an egg carton lid using coordinated multi-finger interaction and distributed contact.
Bottle: Grasp a bottle and lift it above 50 mm, highlighting whole-hand grasping with a palm-assisted enclosure.

For each trained policy, we conduct 20 evaluation trials with randomized object initial poses. Success is defined as complete placement into the cup (Block), opening the lid

beyond 30° (Carton), and lifting the bottle by at least 50 mm and holding it stably for 2 s (Bottle).

b) Ablation Study: We ablate the use of explicit hand-state conditioning by comparing policies trained with and without absolute finger pose inputs. All policies share the same diffusion architecture, visual encoder, and training configuration; only the observation inputs differ.

c) Quantitative Results: Policy success rates are reported in Table II, with representative policy rollouts shown in Fig. 7. For the **Block** task, success primarily depends on precise fingertip alignment and stable grasp closure during placement. Under wrist-aligned embodiment, finger configuration remains visually observable from RGB input, allowing the policy to infer grasp posture directly from image features.

For the **Carton** task, which requires coordinated multi-finger interaction and distributed contact, visual cues such as lid deformation and relative hand pose provide sufficient information for closed-loop adjustment, resulting in similar performance with and without explicit finger-state inputs.

For the **Bottle** task, which emphasizes whole-hand grasping and stable lifting, performance remains similar with and without explicit finger-state conditioning. Because the task primarily relies on gross hand-object alignment and power grasp formation rather than precise fingertip articulation, the policy can recover sufficient hand configuration even under major occlusion.

These results suggest that when geometric and visual alignment are addressed at the hardware level, RGB observations alone provide sufficient information for recovering hand configuration, reducing the benefit of explicit finger-state conditioning.

d) Comparison to Prior Wearable-Based Pipelines: We adopt block placement and carton opening tasks to enable comparison with DexUMI. DexUMI reports success rates of 1.00 (Cube) and 0.85 (Carton) under their best configuration using relative actions, image and tactile conditioning, and segmentation with inpainting to mitigate visual embodiment mismatch. Under our aligned pseudo-hand embodiment, we achieve 0.90 success on both tasks without segmentation, masking, inpainting, or any tactile feedback as conditioning.

DexUMI's raw image baseline without tactile conditioning or visual post-processing achieves substantially lower success rates, indicating that segmentation and inpainting play a critical role in compensating for embodiment mismatch. In contrast, our hardware-level alignment enables strong performance directly from raw RGB observations while maintaining a substantially simpler end-to-end pipeline.

e) Discussion: Overall, these results suggest that hardware-level alignment of geometry, appearance, and viewpoint reduces both human and algorithmic bottlenecks in dexterous learning. By eliminating segmentation and inpainting and reducing reliance on explicit hand-state conditioning, DexEXO enables a streamlined demonstration-to-policy pipeline while retaining competitive task performance.

VI. Limitations

DexEXO has several limitations worth noting. First, finger visibility from a top-down viewpoint is partially occluded by the exoskeleton structure, which may affect visual monitoring during certain tasks. Second, the linkage architecture introduces mechanical interference that can limit the range of motion, particularly when interacting with objects on flat surfaces. Third, the pseudo-hand embodiment introduces a slight spatial offset between the operator's natural hand and the passive hand, which may reduce intuitiveness for first-time users. Finally, although the hardware-level embodiment alignment improves policy transfer, adapting the system to different robot hand form factors requires non-trivial mechanical redesign and integration effort.

In addition, the current system primarily targets demonstration collection for visually guided manipulation with a wrist-mounted camera. Tasks that require significant occlusion handling, multi-view perception, or rich tactile feedback may still benefit from additional sensing modalities such as tactile arrays or depth sensing. Future work will explore integrating multi-modal sensing and improving the mechanical modularity of the system to support rapid adaptation to a broader range of robotic hands and manipulation scenarios.

VII. Conclusion

We presented **DexEXO**, a wearability-first dexterous exoskeleton designed to enable scalable, cross-operator demonstration collection while preserving structured kinematic correspondence with a target robotic hand. Through analytically modeled finger interfaces and a pose-tolerant thumb mechanism, DexEXO accommodates anthropometric variation without rigid joint alignment or per-user calibration. Experimental validation confirmed consistent IP transmission, structured TM coupling under low-dimensional pose modeling, and substantial residual self-alignment. User studies demonstrated improved comfort and usability relative to prior wearable systems, and policy experiments showed that hardware-level embodiment alignment enables effective end-to-end learning directly from raw wrist-mounted RGB observations. Together, these results suggest that prioritizing wearability and geometric alignment at the hardware level can reduce both human and algorithmic bottlenecks in dexterous robot learning without sacrificing task performance.

References

[1] A. Rajeswaran et al., "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," in *Proceedings of Robotics: Science and Systems (RSS)*, 2018.

[2] M. Andrychowicz et al., "Learning dexterous in-hand manipulation," *The International Journal of Robotics Research*, vol. 39, no. 1, pp. 3–20, 2020.

[3] Y. Jiang, A. Stone, Z. Tan, et al., "Vima: General robot manipulation with multimodal prompts," in *International Conference on Machine Learning (ICML)*, 2023.

[4] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," in *International Symposium on Experimental Robotics (ISER)*, 2016.

[5] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Robotics: Science and Systems (RSS)*, 2023.

[6] V. Kumar et al., "Robohive: A unified framework for robot learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[7] Q. Ye et al., "Visual-tactile pretraining and online multitask learning for humanlike manipulation dexterity," *Science Robotics*, vol. 11, no. 110, eady2869, 2026.

[8] Y. Liu, Y. Yang, Y. Wang, et al., "Realdex: Towards human-like grasping for robotic dexterous hand," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

[9] Y. Huang, D. Fan, H. Duan, et al., "Human-like dexterous manipulation for anthropomorphic five-fingered hands: A focused review," *Biomimetic Intelligence and Robotics*, vol. 5, no. 1, p. 100 212, 2025.

[10] Y. Tanaka, Y. Shirai, A. Schperberg, X. Lin, and D. Hong, "Scaler: Versatile multi-limbed robot for free-climbing in extreme terrains," *IEEE Transactions on Robotics*, 2025.

[11] E. Welte et al., "Interactive imitation learning for dexterous robotic manipulation: Challenges and perspectives," *Frontiers in Robotics and AI*, 2025.

[12] A. O'Neill, A. Rehman, A. Gupta, et al., *Open X-Embodiment: Robotic learning datasets and RT-X models*, https://arxiv.org/abs/2310.08864, 2023.

[13] A. Brohan et al., "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *arXiv preprint arXiv:2307.15818*, 2023.

[14] C. Chi et al., "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," in *Robotics: Science and Systems (RSS)*, 2024.

[15] K. Shaw, S. Bahl, A. Sivakumar, A. Kannan, and D. Pathak, "Learning dexterity from human hand motion in internet videos," *The International Journal of Robotics Research*, vol. 43, no. 4, pp. 513–532, 2024.

[16] H. G. Singh et al., *Hand-object interaction pretraining from videos*, 2024. arXiv: 2409.08273 [cs.RO].

[17] A. Zhu, Y. Tanaka, A. Goldberg, and D. Hong, *Aura: Autonomous upskilling with retrieval-augmented agents*, 2025. arXiv: 2506.02507 [cs.RO].

[18] Y. Li et al., *Taccel: Scaling up vision-based tactile robotics via high-performance gpu simulation*, 2025. arXiv: 2504.12908 [cs.RO].

[19] J. Yu et al., *Real2render2real: Scaling robot data without dynamics simulation or robot hardware*, 2025. arXiv: 2505.09601 [cs.RO].

[20] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. K. Liu, "Dexcap: Scalable and portable mocap data collection system for dexterous manipulation," *arXiv preprint arXiv:2403.07788*, 2024.

[21] M. Torne et al., *Reconciling reality through simulation: A real-to-sim-to-real approach for robust manipulation*, 2024. arXiv: 2403.03949 [cs.RO].

[22] Z. Zhao, H. Dong, Z. He, Y. Li, X. Yi, and Z. Li, *Closing the reality gap: Zero-shot sim-to-real deployment for dexterous force-based grasping and manipulation*, 2026. arXiv: 2601.02778 [cs.RO].

[23] J. Zhao, N. Kuppuswamy, S. Feng, B. Burchfiel, and E. Adelson, *Polytouch: A robust multi-modal tactile sensor for contact-rich manipulation using tactile-diffusion policies*, 2025. arXiv: 2504.19341 [cs.RO].

[24] A. Handa et al., "Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[25] T. He et al., "Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning," *arXiv preprint arXiv:2406.08858*, 2024.

[26] Z. Fu, T. Z. Zhao, and C. Finn, *Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation*, 2024. arXiv: 2401.02117 [cs.RO].

[27] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. 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)*, 2024, pp. 12 156–12 163.

[28] M. Xu et al., *Dexumi: Using human hand as the universal manipulation interface for dexterous manipulation*, 2025. arXiv: 2505.21864 [cs.RO].

[29] H.-S. Fang et al., *Dexop: A device for robotic transfer of dexterous human manipulation*, 2025. arXiv: 2509.04441 [cs.RO].

[30] J. Du et al., *Mile: A mechanically isomorphic exoskeleton data collection system with fingertip visuotactile sensing for dexterous manipulation*, 2025. arXiv: 2512.00324 [cs.RO].

[31] R. Ge, Y. Liu, Z. Yan, Q. Cheng, S. Qiu, and D. Ming, "Design of a self-aligning four-finger exoskeleton for finger abduction/adduction and flexion/extension motion," in *2023 International Conference on Rehabilitation Robotics (ICORR)*, 2023.

[32] C. Brogi et al., "An original hybrid-architecture finger mechanism for wearable hand exoskeletons," *Mechatronics*, vol. 98, p. 103 117, 2024.

[33] Z. Si, K. L. Zhang, Z. Temel, and O. Kroemer, *Tilde: Teleoperation for dexterous in-hand manipulation learning with a deltahand*, 2024. arXiv: 2405.18804 [cs.RO].

[34] R. Wen et al., *Dexterous teleoperation of 20-dof bytedexter hand via human motion retargeting*, 2025. arXiv: 2507.03227 [cs.RO].

[35] P. Ruppel et al., "Elastic tactile sensor glove for dexterous teaching by human demonstration," *Sensors*, vol. 24, no. 6, p. 1912, 2024.

[36] X. Lin et al., *Dexflow: A unified approach for dexterous hand pose retargeting and interaction*, 2025. arXiv: 2505.01083 [cs.RO].

[37] Z. Mandi, Y. Hou, D. Fox, Y. Narang, A. Mandlekar, and S. Song, *Dexmachina: Functional retargeting for bimanual dexterous manipulation*, 2025. arXiv: 2505.24853 [cs.RO].

[38] Z.-H. Yin et al., *Geometric retargeting: A principled, ultrafast neural hand retargeting algorithm*, 2025. arXiv: 2503.07541 [cs.RO].

[39] J. Hsieh, K.-H. Tu, K.-H. Hung, and T.-W. Ke, *Dexman: Learning bimanual dexterous manipulation from human and generated videos*, 2025. arXiv: 2510.08475 [cs.RO].

[40] Y. Liu et al., "Hoi4d: A 4d egocentric dataset for category-level human-object interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 21 013–21 022.

[41] Y.-W. Chao et al., "DexYCB: A benchmark for capturing hand grasping of objects," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[42] K. Shaw, S. Bahl, and D. Pathak, "Videodex: Learning dexterity from internet videos," in *Conference on Robot Learning*, PMLR, 2023, pp. 654–665.

[43] P. Mandikal and K. Grauman, "Dexvip: Learning dexterous grasping with human hand pose priors from video," in *Conference on Robot Learning*, PMLR, 2022, pp. 651–661.

[44] H. Bharadhwaj, R. Mottaghi, A. Gupta, and S. Tulsiani, "Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation," in *European Conference on Computer Vision*, Springer, 2024, pp. 306–324.

[45] K. Grauman et al., *Ego4d: Around the world in 3,000 hours of egocentric video*, 2022. arXiv: 2110.07058 [cs.CV].

[46] Q. Liu, Y. Cui, Z. Sun, G. Li, J. Chen, and Q. Ye, "Vtdexmanip: A dataset and benchmark for visual-tactile pretraining and dexterous manipulation with reinforcement learning," in *International Conference on Learning Representations (ICLR)*, 2025.

[47] H. Zhang, S. Hu, Z. Yuan, and H. Xu, "Doglove: Dexterous manipulation with a low-cost open-source haptic force feedback glove," *arXiv preprint arXiv:2502.07730*, 2025.

[48] H. Zhang et al., "Human-exoskeleton kinematic calibration to improve hand tracking for dexterous teleoperation," *IEEE Robotics and Automation Letters*, 2026.

[49] X. Wang et al., "Medical imaging-based kinematic modeling for biomimetic finger joints and hand exoskeleton validation," *Biomimetics*, vol. 10, no. 10, p. 652, 2025.

[50] OYMotion Technologies Co., Ltd., *Roh-ap001 dexterous robotic hand*, https://www.oymotion.com/en/product62, Five-finger robotic hand with human-like proportions and independent finger motion, 2024.

[51] A. Buryanov and V. Kotiuk, "Proportions of hand segments," *International Journal of Morphology*, vol. 28, no. 3, pp. 755–758, 2010.

[52] O. Rayyan, M. Gilles, and Y. Cui, *Teledex: Accessible dexterous teleoperation*, GitHub repository, 2026.

[53] M. Oquab et al., "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[54] C. Chi et al., "Diffusion policy: Visuomotor policy learning via action diffusion," in *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[55] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in Psychology*, vol. 52, Elsevier, 1988, pp. 139–183.