DexHLM: An Egocentric Dexterous Human Loco-Manipulation Dataset for Humanoids with Tactile Prediction

Zhenyu Wei[1,2,*] **Haoyang Luo**[1,3,*] **Chixuan Zhang**[1,3] **Guo Chen**[1,3] **Suting Ni**[1]
Siyuan Huang[2] **Baoxiong Jia**[2,†] **Jingya Wang**[1,3,†]

[1]ShanghaiTech University [2]State Key Laboratory of General Artificial Intelligence, BIGAI
[3]InstAdapt

{weizhy2024, luohy12024, zhangchx12024}@shanghaitech.edu.cn
{chenguo2024, nist2024, wangjingya}@shanghaitech.edu.cn
huangsiyuan@ucla.edu baoxiongjia@g.ucla.edu



Figure 1: **Overview of DexHLM.** DexHLM is a multimodal egocentric HOI benchmark for humanoid whole-body dexterous manipulation, featuring synchronized tactile sensing, egocentric RGB-D, body motion, object trajectories, and hand motion across diverse spatial layouts and long-horizon activities.

Abstract

Humanoid whole-body loco-manipulation requires scalable and physically reliable demonstrations, yet existing human-object interaction datasets lack direct contact supervision and egocentric observability, often resulting in physically inconsistent interactions. We present DexHLM, a multimodal egocentric HOI benchmark for humanoid whole-body dexterous manipulation, featuring synchronized tactile sensing, egocentric RGB-D, body motion, object trajectories, and hand motion across diverse spatial layouts and long-horizon activities. To improve physical consistency, we introduce a tactile-guided refinement pipeline that leverages measured pressure to correct hand-object interactions while preserving global motion fidelity. We further establish an Egocentric HOI Recovery method for tactile-guided reconstruction from first-person observations, along with an egocentric tactile prediction model that supplies contact evidence at recovery time. Together, our work provides a scalable and physically grounded foundation for learning humanoid whole-body dexterous interaction, pointing toward low-cost egocentric data collection for embodied intelligence. Project page is available at: `https://dexhlm.github.io/`.

*Equal contribution. †Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Humanoid whole-body loco-manipulation requires the tight integration of locomotion, postural balance, and fine-grained dexterous interaction. Learning such coordinated behaviors typically demands large-scale, high-quality demonstrations [33], yet collecting them on humanoid robots is costly and difficult to scale due to hardware, safety, and teleoperation constraints [8]. Human-object interaction (HOI) data offers a promising alternative, since everyday human activities naturally contain rich whole-body interaction relevant to humanoid skill acquisition. To serve as physically grounded supervision for robot learning, however, such data must capture contact- and force-aware hand-object interactions, span diverse object layouts and spatial configurations, include long-horizon activities, and preserve egocentric observations aligned with robot perception. Due to these rigorous requirements on data quality, this type of large-scale, interaction-centric HOI data remains largely underexplored.

Existing HOI datasets only partially satisfy these requirements. Hand-object datasets [4, 6, 7] provide fine-grained manipulation data, primarily focusing on hand-object interaction. Whole-body HOI datasets [27, 3, 12, 1, 19, 14, 36, 15, 20] extend to body-scale interactions but typically infer contact from geometric proximity rather than direct sensing [38, 32], leading to problems like hand-object penetration, floating fingers, and physically implausible grasps. Beyond pose capturing, most existing datasets cluster around fixed table-top setups or task-specific scenes, rarely revisiting the same interaction across heights, postures, and viewpoints that everyday activity actually spans.

Based on these observations, we introduce DexHLM, a multimodal HOI dataset that jointly addresses the aforementioned requirements by *(i)* capturing tactile information for contact- and force-aware interaction; *(ii)* systematically varying interaction height, posture, and view point for diverse spatial configurations; *(iii)* including long-horizon activities with transitions across atomic interactions; and *(iv)* providing egocentric observations and camera poses for future robot learning. Synchronized capture alone, however, does not guarantee physical reliability: body and object trajectories are stable, but finger-level contact remains fragile, and geometric proximity cannot distinguish fingers that truly carry contact from those merely close to the surface. We therefore introduce a tactile-guided refinement pipeline anchored by measured pressure. We further train RL policies on the refined demonstrations as a humanoid usability check.

To reduce the data acquisition cost for humanoid whole-body dexterous manipulation, we explore a low-cost egocentric data collection pipeline designed for human-to-humanoid transfer. The core challenge lies in obtaining high-quality 3D human-object interaction (HOI) data, particularly high-precision hand motion, from portable egocentric recordings, a setting where traditional motion capture and tactile gloves are unavailable. To address this, we build upon our DexHLM benchmark and propose an Egocentric HOI Recovery method with Tactile Refinement. Specifically, we first recover hand articulation, object pose, and body motion from RGB streams using existing methods, then refine the physically implausible parts (e.g., hand-object penetration) via a tactile-guided optimization. Since tactile sensing is not accessible during deployment, we further introduce a VLM-based tactile predictor that estimates contact pressure from egocentric video clips. Experiments on DexHLM demonstrate that our approach significantly improves the physical fidelity of reconstructed interactions, validating the effectiveness of leveraging VLM-guided tactile prediction for low-cost, high-precision HOI data acquisition. In summary, our contributions are:

- We present DexHLM, a multimodal egocentric HOI benchmark for humanoid whole-body dexterous manipulation. It provides synchronized whole-body motion, object trajectories, hand motion, real tactile sensing, and egocentric RGB-D across diverse spatial layouts and long-horizon activities.
- We introduce a tactile-guided refinement pipeline that leverages measured pressure to correct hand-object interactions (e.g., penetration and unsupported contacts) while preserving global motion fidelity, significantly improving the physical reliability of dexterous whole-body motion captures.
- We establish an Egocentric HOI Recovery method on DexHLM for tactile-guided HOI reconstruction from first-person observations, together with a VLM based tactile prediction model that supplies contact evidence at recovery time, enabling downstream physical refinement without requiring measured tactile during deployment.

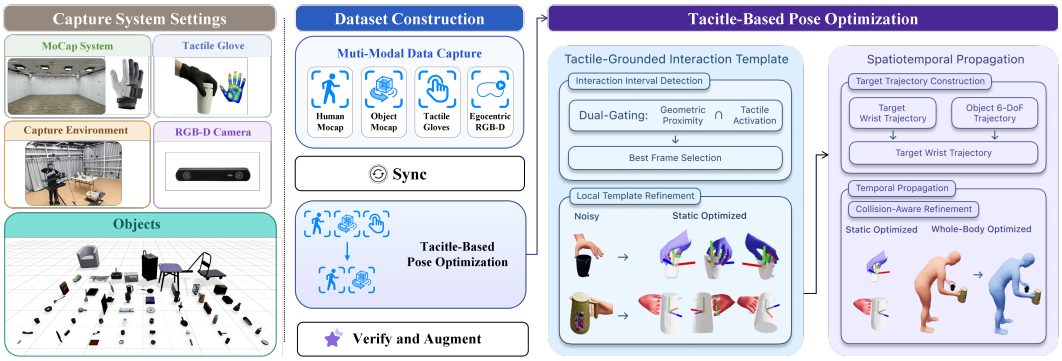

Figure 2: **Overview of DexHLM and tactile-guided post-optimization.** DexHLM aligns whole-body motion, object motion, egocentric RGB-D, real tactile sensing, and controlled spatial layouts; the refinement uses measured tactile to convert raw capture into physically grounded demonstrations.

2 Related Work

Human Object Interaction Datasets. Hand-object datasets such as ContactPose [4], DexYCB [6], H2O [17], and ARCTIC [7] provide fine hand articulation under close-range observation—some with first-person views—and primarily focus on hand-object interaction. Whole-body HOI datasets such as BEHAVE [3], InterCap [12], OMOMO [19], and HOI-M^3 [36] capture body-scale interactions, yet typically retain coarse hand articulation and infer contact from geometric proximity rather than direct measurement [38, 32]. Datasets such as GRAB [27], HUMOTO [20], and ParaHome [15] approach whole-body dexterity, but remain limited by missing tactile evidence, missing synchronized egocentric observation, or insufficient systematic spatial variation. To bridge this gap, to our knowledge, DexHLM is the first whole-body HOI corpus to unify measured tactile contact, egocentric RGB-D, systematic spatial layouts, and long-horizon activity at dexterous whole-body resolution.

Interaction Optimization and Refinement. Hand level methods such as ContactOpt [9], reconstruction pipelines [5], GraspTTA [13], and TOCH [38] achieve fine grained hand-object refinement; whole-body pipelines such as CONTHO [21], InterDiff [31], and HOI-FHLI [30] extend contact reasoning to full-body motion yet rely on inferred geometric proximity or learned priors rather than measured contact; and dataset level methods such as InterAct [32] unify geometric correction across corpora yet lack direct tactile supervision. To bridge this gap, our approach grounds correction in measured tactile pressure, modulating contact attraction while softly preserving captured motion.

Egocentric Interaction Recovery. Egocentric datasets [17, 11] provide rich first person observability yet lack synchronized tactile measurements. H+O [28] estimates hand-object poses. ContactHands [22] studies contact-state recognition, PressureVision [10] estimates hand pressure, and OpenTouch [26] provides synchronized egocentric video, tactile signals, and hand pose. To bridge egocentric observation and physical contact, our approach integrates tactile estimation into whole-body HOI recovery, where predicted contact evidence guides finger refinement and object-trajectory correction rather than remaining a standalone perception baseline.

3 The DexHLM Dataset

3.1 Data Collection

DexHLM synchronizes whole-body SMPL-X [24] motion, rigid-body 6-DoF object trajectories, camera poses, head-mounted egocentric RGB-D, and per-frame tactile pressure from a pair of custom gloves, each with 256 sensing points, for humanoid whole-body loco-manipulation. The corpus spans 20.1 hours over 52 objects and six everyday scenes, with both single-handed and bimanual activities covering prehensile, in-hand, and non-prehensile behaviors. It includes six participants, four male and two female, with varying heights and body proportions. The modalities are temporally aligned at 30 Hz, with human and object motion expressed in a shared coordinate system.

Two protocol-level choices distinguish DexHLM from existing HOI corpora. First, same-level interactions sample a body-calibrated grid of five height levels (L1 floor to L5 neck) and three horizontal variations (V1–V3), so posture, reach, and viewpoint vary independently of task semantics.

Table 1: Representative HOI datasets grouped by interaction scope. Columns summarize available sensing modalities and protocol properties, including tactile sensing, egocentric observation, spatial diversity, long-horizon activity, and dexterous-level hand interaction.

Dataset	Tactile	Egocentric	Spatial Diversity	Long Horizon	Hand Interaction	Duration (h)	# of Objects
Hand-object interaction							
ContactPose [4]	–	–	–	–	✓	2.0	25
DexYCB [6]	–	–	–	–	✓	1.0	20
H2O [17]	–	✓	–	–	✓	1.0	8
ARCTIC [7]	–	✓	–	–	✓	2.1	11
Whole-body HOI with coarse-hand annotation							
BEHAVE [3]	–	–	–	–	–	1.4	20
InterCap [12]	–	–	–	–	–	0.6	10
OMOMO [19]	–	–	–	–	–	10.0	15
HOI-M^3 [36]	–	–	–	✓	–	20.0	90
Whole-body HOI with dexterous hand articulation							
GRAB [27]	–	–	–	–	✓	3.8	51
ParaHome [15]	–	–	–	✓	✓	8.1	22
HUMOTO [20]	–	–	–	✓	✓	2.2	63
DexHLM (Ours)	✓	✓	✓	✓	✓	20.1	52

Heights are calibrated to each participant's body proportions. Second, capture proceeds through three temporal stages: same-level interactions (S1) populate grid cells, cross-level transitions (S2) chain locomotion with manipulation, and long-horizon scene-level activity (S3) combines interactions in freely arranged layouts. S3 also provides synchronized exocentric views as a qualitative reference, not as recovery inputs. Table 1 compares DexHLM with representative HOI datasets along the modality and protocol axes most relevant to egocentric whole-body recovery; further dataset details are in Section A. In raw DexHLM capture the dominant residual error lies in finger-level contact rather than in body or object motion. Section 3.2 describes the tactile-guided post-optimization applied to these captures, anchored in measured pressure rather than inferred geometric proximity.

3.2 Measured-Tactile Human-Motion Refinement

Body and object trajectories from MoCap are reliable, while local hand-object contact is the most fragile part of whole-body interaction data. Methods that infer contact from learned predictors or geometric proximity [9, 38, 31, 32] can confuse fingers that truly carry contact with those merely close to the surface, and over-correcting the latter disturbs the captured motion. We therefore anchor correction in measured tactile pressure, modulating contact attraction while softly constraining pose deviations. During dataset construction, the pipeline refines hand articulation and wrist pose against the reliable MoCap object trajectory, then propagates the correction to body posture. We describe these steps next, with additional details in Section B.1; object-trajectory correction at egocentric recovery is described separately in Section 4.3.

Interval detection. A dual tactile-geometric gate flags frame t when the maximum aggregated finger pressure exceeds τ_p and the hand-mesh-to-object distance falls below τ_d. Within each smoothed interval, a representative template frame t^* is chosen by maximizing a weighted score over tactile confidence, kinematic stability, and geometric proximity, defining the wrist-relative reference for the subsequent steps.

Pressure-modulated local refinement. At the template frame t^*, hand articulation $\boldsymbol{\theta}_{t^*}$ and wrist pose $\mathbf{T}_w$ are optimized under a per-finger objective whose weights derive from tactile evidence. We split the articulation into per-finger axis-angle blocks, $\boldsymbol{\theta}_{t^*} = \big(\boldsymbol{\theta}_{t^*}^{(1)}, \ldots, \boldsymbol{\theta}_{t^*}^{(F)}\big)$ with $\boldsymbol{\theta}_{t^*}^{(f)} \in \mathbb{R}^{3J_f}$ stacking the axis-angle parameters of finger f's J_f joints. For each finger f, we form a tactile weight $w_f \in [0, 1]$ by averaging measured fingertip pressure over a temporal window centered at t^*, normalizing to a global pressure scale, and clipping to $[0, 1]$; the following simplified objective illustrates the balance between contact attraction and pose preservation:

$$\min_{\boldsymbol{\theta}_{t^*},\, \mathbf{T}_w} \sum_{f=1}^{F} \Big[w_f \sum_{v \in \mathcal{V}_f} \big|\mathrm{SDF}_{\mathcal{O}}(\mathbf{v}_{t^*})\big| \;+\; (1 - w_f)\, \big\|\boldsymbol{\theta}_{t^*}^{(f)} - \boldsymbol{\theta}_{t^*}^{(f),0}\big\|^2 \Big], \tag{1}$$

where $\mathcal{V}_f$ is the index set of hand vertices on finger f, $\mathbf{v}_{t^*}$ denotes a hand vertex at frame t^* expressed in the object's local frame, $\mathrm{SDF}_{\mathcal{O}}$ is the object signed distance field, and $\boldsymbol{\theta}_{t^*}^{(f),0}$ is the captured pose of finger f. The implementation uses Euclidean distances from sparse contact points to targets on the object surface, with attraction introduced in stages and additional regularizers (Section B.1). Continuous weights balance attraction and soft pose preservation rather than freezing individual fingers.

Spatiotemporal propagation. The refined wrist-relative transform is composed with the object trajectory to obtain a target wrist trajectory $\hat{\mathbf{T}}_t$, blended with the captured wrist position $\mathbf{p}_t^{w,0}$ and finger pose $\boldsymbol{\theta}_t^0$ through smooth weights $\alpha_t^w, \alpha_t^f \in [0,1]$ that ramp inside each interval and decay at its boundaries:

$$\mathbf{p}_t^w \;=\; (1-\alpha_t^w)\,\mathbf{p}_t^{w,0} + \alpha_t^w\,\hat{\mathbf{p}}_t, \qquad \boldsymbol{\theta}_t^{\text{blend}} \;=\; (1-\alpha_t^f)\,\boldsymbol{\theta}_t^0 + \alpha_t^f\,\boldsymbol{\theta}^*, \tag{2}$$

with $\boldsymbol{\theta}^*$ the locally refined hand articulation. Body posture is propagated through the arm chain by bounded inverse kinematics with the root pose fixed, so contact frames track the refined wrist target while frames outside the interval stay close to the captured motion.

Collision-aware finger refinement. With the wrist trajectory fixed, finger pose on near-contact frames is refined by

$$\min_{\boldsymbol{\theta}_t} \;\; w_{\text{pen}} \sum_v \kappa_v\,\mathrm{ReLU}\big(\mathrm{SDF}_{\mathcal{O}}(\mathbf{v}_t) + \delta\big) \;+\; w_{\text{close}}\,\big\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^{\text{blend}}\big\|^2, \tag{3}$$

where $\mathbf{v}_t$ ranges over hand vertices in the object's local frame, $\mathrm{SDF}_{\mathcal{O}}$ is the object signed distance field positive inside the object, $\delta \geq 0$ is a clearance margin, and κ_v favors fingertip and palmar vertices. The closeness term anchors the result to the blended pose $\boldsymbol{\theta}_t^{\text{blend}}$, limiting drift from the tactile-guided correction.

4 Egocentric HOI Recovery

4.1 Desiderata and Setup

The Egocentric HOI Recovery Benchmark turns the question raised in Section 1 into a concrete evaluation: which whole-body HOI states can be recovered from egocentric observations given GT camera poses, object meshes, and depth? We decompose whole-body HOI into three coupled states—hand articulation, object 6-DoF motion, and body pose—and let each egocentric method fill the column it can provide. Camera trajectory and object mesh are held at GT values, with wrist positions lifted using GT depth, so that metric variation reflects the component being evaluated rather than scene-reconstruction noise; tactile, by contrast, is treated as a derived signal predicted by Section 4.2 and fed into the correction of Section 4.3, so the benchmark probes *tactile-guided* egocentric recovery without requiring measured tactile at evaluation time. Hand and body errors use MoCap motion refined using measured tactile signals as reference; object errors use MoCap object trajectories.

The component baselines pair FoundationPose [29] with HaPTIC [34] for hand articulation or EgoEgo [18] for body pose. Our pipeline uses WiLoR [25] for hands and EgoAllo [35] for body pose (Section 4.3).

4.2 Tactile Prediction from Egocentric Video

Problem formulation. The tactile predictor supplies the contact mask used by the object-trajectory correction of Section 4.3, so that recovery operates without a tactile glove at evaluation time. Given a short egocentric video clip and a fixed task-conditioning prompt, the model predicts a non-negative 32-dimensional tactile representation for the target frame, associated with selected SMPL-X hand keypoints. Framewise predictions are thresholded into a binary contact mask for downstream object-trajectory correction.

Architecture. We adopt Qwen3-VL [2] at 2B and 8B scales as the multimodal backbone. Because our curated egocentric–tactile corpus is orders of magnitude smaller than the backbone's web-scale

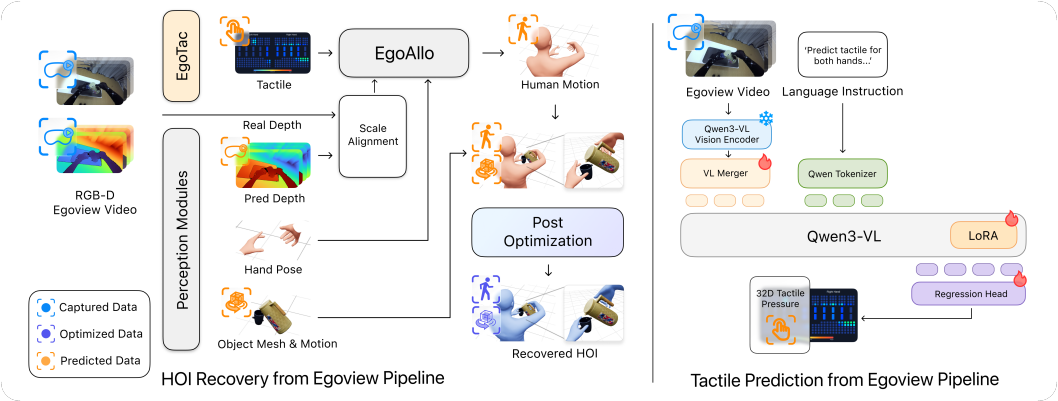

Figure 3: **Egocentric HOI Recovery Benchmark and tactile prediction.** The reported evaluation uses GT camera poses, GT object meshes, and wrists lifted using GT depth; predicted depth and mesh modules shown in the schematic are not evaluated here. Predicted tactile signals guide recovery-time post-processing, rather than body estimation.

pre-training data, we fine-tune through lightweight LoRA adapters on the language-layer attention projections instead of full-parameter updates, and attach a small two-layer MLP regression head inspired by OpenVLA [16]. The head maps the backbone features to target-frame tactile values using synchronized glove measurements as supervision. We also train the visual-language merger and regression head while freezing the remaining backbone weights. Denoting a pre-trained weight matrix by $\mathbf{W}_0 \in \mathbb{R}^{d\times k}$, the adapted forward pass is

$$\mathbf{h} \;=\; \mathbf{W}_0\,\mathbf{x} \;+\; \frac{\alpha}{r}\,\mathbf{B}\mathbf{A}\,\mathbf{x}, \qquad \mathbf{A}\in\mathbb{R}^{r\times k},\;\; \mathbf{B}\in\mathbb{R}^{d\times r}, \tag{4}$$

with the adapter settings given in Section B.2.

Training objective. The predictor is trained on DexHLM using synchronized egocentric–tactile pairs. In the following general loss form, the temporal-smoothing weight is set to zero in our experiments:

$$\mathcal{L} \;=\; \mathcal{L}_{\text{MSE}} \;+\; \lambda_{\text{bce}}\,\mathcal{L}_{\text{BCE}} \;+\; \lambda_{\text{sm}}\,\mathcal{L}_{\text{smooth}}, \tag{5}$$

The active terms combine weighted mean-squared error on tactile values, with non-zero targets up-weighted to counter contact sparsity, and binary cross-entropy on contact masks derived from a fixed pressure threshold.

Pipeline integration. Once trained, the predictor operates as an integrated component whose contact predictions feed the object-trajectory correction of Section 4.3, turning egocentric tactile estimation from an isolated perception baseline into a usable signal for downstream HOI recovery. As validated in Section 5.2, the 8B fine-tuned model reaches a Pearson correlation of 0.812 and a contact F1 of 0.816, while the 2B model reaches 0.790 and 0.789. Trainable modules, supervision, and baseline settings are described in Section B.2.

4.3 Recovery with Predicted Tactile

Hands and objects often appear in close proximity even when no interaction occurs, so geometric proximity is unreliable for correcting object trajectories. We therefore separate component-level recovery from tactile-guided post-processing: egocentric perception produces initial hand, body, and object estimates, and predicted contact intervals guide object-trajectory correction.

Hand articulation uses WiLoR [25], with each per-frame wrist lifted by the GT depth map; this depth-assisted lifting is a privileged diagnostic input rather than a deployable solution (Section 6). Body pose is recovered with EgoAllo [35], conditioned on GT camera poses and the depth-lifted wrist constraints. Object 6-DoF is recovered by FoundationPose [29], complemented by four post-processing steps targeting the residual errors of the tracker: symmetric-orientation disambiguation, adaptive rotation-outlier rejection, sliding-window quaternion-mean smoothing, and multi-object secondary alignment.

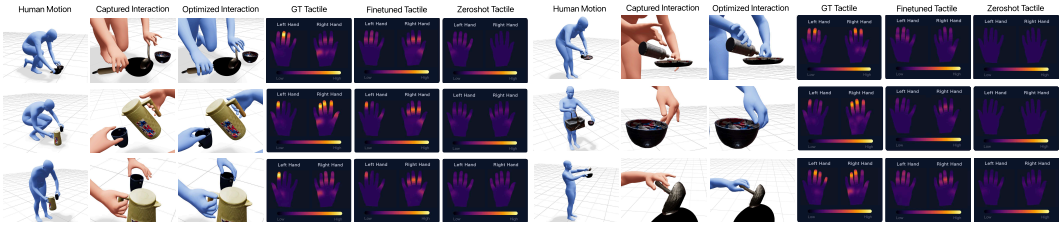

Figure 4: **Qualitative results of tactile-guided refinement and tactile prediction.** Tactile-guided refinement improves hand-object alignment while preserving natural articulation. The fine-tuned tactile predictor produces outputs closely matching measured tactile signals.

Predicted-tactile object-trajectory correction. The recovered body and wrist trajectories are held fixed, while finger poses and the object trajectory are adjusted. The perception-time object estimate at the template frame directly anchors the correction; no iterative object-pose optimization is performed. For each detected contact interval $\mathcal{C}$, we record the wrist-relative object pose $T^{\mathrm{rel}}_{\mathrm{obj}|\mathcal{C}} \in SE(3)$ at the template frame and propagate it through the interval by composing it with the recovery-time wrist trajectory. Under approximately stable contact, this relation is expressed as $T_{\mathrm{obj}}(t) = T_{\mathrm{wrist}}(t)\, T^{\mathrm{rel}}_{\mathrm{obj}|\mathcal{C}}$ for $t \in \mathcal{C}$. The object trajectory can also change outside $\mathcal{C}$ through anchoring; finger poses remain subject to refinement. Predicted tactile signals identify contact intervals, rather than supplying the measured-pressure local weights of Section 3.2. Recovery thus operates without measured tactile at evaluation time (Section B.2).

5 Experiments

Section 5.1 validates the tactile-guided refinement on raw DexHLM captures, Section 5.2 evaluates the egocentric tactile predictor, Section 5.3 reports the headline egocentric HOI recovery results, and Section 5.4 provides a preliminary humanoid usability check.

5.1 Measured-Tactile Human-Motion Refinement

Table 2: Refinement ablation on raw DexHLM captures. FP-C, HPD, NTF-D, and MPJPE are not reported for the Raw row. Ours Full applies tactile-guided refinement combined with collision-aware finger refinement (Section 3.2). Metric definitions: Section C.

Condition	PR↓(%)	FP-C↓(%)	HPD↓(rad)	NTF-D↓(mm)	MPJPE↓(mm)
Raw MoCap	83.5	—	—	—	—
Geometry only	5.1	72.5	1.294	31.6	43.7
Tactile only	71.3	72.8	1.143	30.4	40.9
Ours Full	**4.2**	**71.9**	**1.086**	**29.9**	**39.4**

Table 2 validates the refinement of Section 3.2 on raw DexHLM captures along two axes: physical validity (PR, FP-C) and motion fidelity (HPD, NTF-D, MPJPE). Full refinement reduces the penetration-frame rate from 83.5% to 4.2%, slightly below the geometry-only baseline (5.1%), while also lowering FP-C (71.9% vs 72.5%) and keeping HPD, NTF-D, and MPJPE below the geometry-only baseline. The two ablated variants illustrate the balance between geometric correction and motion preservation: the geometry-only baseline reaches comparable penetration only at the cost of the largest motion deviation in the table (HPD 1.294, NTF-D 31.6 mm, MPJPE 43.7 mm), and the Tactile-only variant without collision refinement leaves substantial penetration (71.3%) despite using the same tactile-guided weighting. The full pipeline balances contact alignment with soft preservation of captured motion, consistent with Section 3.2; representative cases are shown in Fig. 4.

5.2 Tactile Prediction from Egocentric Video

Table 3 measures whether the synchronized tactile of DexHLM makes egocentric tactile prediction a viable learning problem. The main evaluation allows seen participants and objects, including

Table 3: Egocentric tactile prediction. CF1: element-level contact F1; FC@k: frame-level contact F1. The zero-shot generation and invalid-output handling are described in Section B.2.

	Model	Pearson↑	CF1↑	FC@3↑	FC@5↑	MAE↓	MSE↓	R^2 ↑
Zero-shot	2B	0.000	0.000	0.000	0.000	3.006	33.22	−0.42
	8B	0.320	0.435	0.431	0.468	2.776	25.34	−0.18
Ours	2B	0.790	0.789	0.910	0.921	1.614	8.16	0.50
	8B	**0.812**	**0.816**	**0.925**	**0.922**	**1.449**	**7.12**	**0.60**

Table 4: Egocentric HOI recovery on DexHLM using GT camera poses, GT object meshes, and wrists lifted using GT depth (Section 4.1). "–" marks states a method does not estimate; best per-column values are bolded. The Tactile column repeats the Ours-8B prediction scores from Table 3; its predictions supply the contact mask used by Ours Full. †FP: FoundationPose [29]. Metric definitions: Section C.

Method	Object		Hand				Body			Tactile	
	OTE↓	ORE↓	Rel-MPJPE↓	PA-MPJPE↓	GA-MPJPE↓	Rel-ACC↓	MPJPE↓	PA-MPJPE↓	Wrist↓	Pearson↑	CF1↑
HaPTIC [34] + FP†	286.8	120.7	38.2	33.5	132.3	0.48	–	–	–	–	–
EgoEgo [18] + FP†	286.8	120.7	–	–	–	–	232.0	144.5	609.8	–	–
Ours w/o tactile	286.8	120.7	**21.5**	**19.7**	**62.0**	**0.34**	**115.2**	**92.9**	**202.3**	–	–
Ours Full	**171.4**	**57.8**	**21.5**	**19.7**	**62.0**	**0.34**	**115.2**	**92.9**	**202.3**	**0.812**	**0.816**

the same interactions at different heights. Under the zero-shot generation and output-handling protocol (Section B.2), the 2B model scores zero contact F1, and the 8B model yields only a weak signal (Pearson 0.32, contact F1 0.44). Fine-tuning on DexHLM brings both scales into a usable regime (Pearson 0.79/0.81, contact F1 0.79/0.82 for 2B/8B), and the 8B backbone delivers only marginal additional headroom over 2B. In the main evaluation, Ours-8B outperforms our supervised DINOv2 [23] ViT-B/14 baseline, adapted from the design of TouchAnything [37] (baseline Pearson 0.674, MAE 1.994, MSE 12.45). On jointly unseen participants and objects, Table 5 compares both predictors without adaptation across all seven metrics, with Ours-8B reaching Pearson 0.795. These results show that DexHLM supplies useful supervision for turning egocentric observations into contact evidence for the object-trajectory correction of Section 4.3. Baseline settings and additional results are in Sections B.2 and D.

5.3 Recovery with Predicted Tactile

Following Sections 4.1 and 4.3, Table 4 reports egocentric HOI recovery on DexHLM under the fixed GT inputs. The first two rows are component-decoupled pipelines built from publicly available perception components: HaPTIC + FoundationPose covers hand and object but leaves body unmodeled, while EgoEgo + FoundationPose covers body and object but leaves hand unmodeled. Each baseline is missing one of the three whole-body components and cannot deliver a complete HOI recovery on its own. Our pipeline (Section 4.3) supplies all three states simultaneously, with hand GA-MPJPE at 62.0 mm, body MPJPE at 115.2 mm, and wrist error at 202.3 mm.

The object metrics are identical for the first three rows because the same FoundationPose-based object stack is shared across all rows; predicted-tactile object-trajectory correction distinguishes Ours Full. With predicted tactile signals, OTE drops by 115.4 mm (-40%) and ORE by 62.9° (-52%). The reported hand and body errors remain unchanged at the displayed precision. The Tactile column repeats the Ours-8B prediction scores from Table 3.

5.4 Humanoid Policy Learning

We further train whole-body dexterous manipulation policies for a simulated SMPL-X-based humanoid following InterMimic [33], using refined DexHLM demonstrations as supervision. During training, each policy is optimized using physics-based rewards that encourage both task completion and realistic hand-object contacts. In a paired comparison with matched object trajectories, simulator, reward, and PPO configurations, policy success improves from 28.0% with raw MoCap demonstrations to 62.0% with demonstrations refined using measured tactile signals. The policies

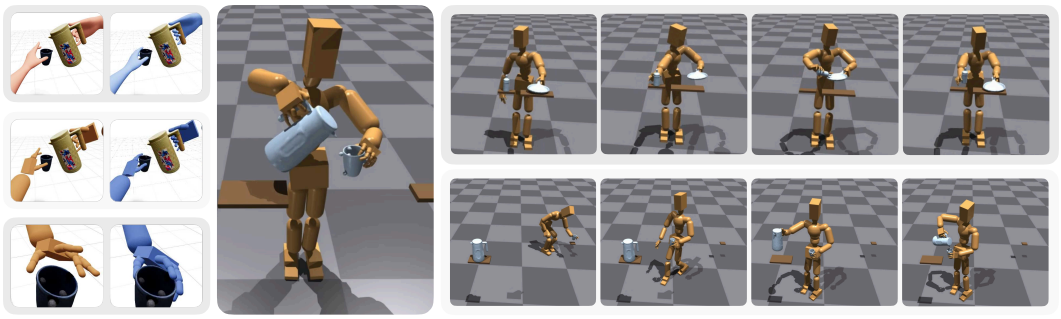

Figure 5: **Qualitative Humanoid Policy Rollouts.**

consistently reproduce key interaction patterns observed in the demonstrations, such as grasping, lifting, and manipulating objects at varying heights (Fig. 5). These results demonstrate that DexHLM can effectively support downstream humanoid policy learning. Task examples and further discussion are in Section D.

6 Conclusion

In this work, we introduce **DexHLM**, a multimodal egocentric HOI benchmark designed to advance humanoid whole-body loco-manipulation from purely kinematic tracking toward physically grounded interaction. DexHLM provides synchronized whole-body motion, object trajectories, real tactile sensing, and egocentric RGB-D across diverse spatial layouts, enabling contact-aware evaluation and learning. Building upon this benchmark, we propose a tactile-guided refinement pipeline that leverages measured pressure as a grounding constraint to correct hand-object interactions while preserving global motion fidelity. This pipeline explicitly improves physical consistency, reducing the penetration-frame rate from 83.5% to 4.2%. Furthermore, we establish the Egocentric HOI Recovery Benchmark, together with an egocentric tactile prediction model, demonstrating that contact evidence can be recovered from first-person observations and used to guide object-trajectory correction at recovery time. Overall, DexHLM provides a scalable and physically grounded foundation for studying whole-body dexterous interaction from egocentric human demonstrations. More importantly, it opens a practical path toward low-cost data collection for embodied intelligence: future systems may use lightweight egocentric sensors and predicted tactile evidence to recover contact-consistent demonstrations at scale, reducing reliance on expensive robot teleoperation or fully instrumented capture setups. We believe this benchmark can serve as a long-term testbed for connecting egocentric perception, tactile reasoning, and humanoid policy learning.

Limitations & Future Work While DexHLM provides rich demonstrations for whole-body loco-manipulation, our current humanoid policy evaluation remains in simulation. Discrepancies in sensor noise, actuator dynamics, and physical compliance between simulation and real hardware can degrade the accuracy of learned interactions when deploying policies directly. To mitigate this, future work will explore on-robot fine-tuning with real tactile feedback, shifting from offline data curation to active, closed-loop adaptation. Recovery currently uses GT camera poses, GT object meshes, and GT depth; estimating these inputs and learning policies directly from egocentrically recovered demonstrations remain future work. More broadly, integrating vision, proprioception, and real-time tactile sensing within a unified learning framework represents a promising and critical direction toward achieving robust, real-world humanoid manipulation.

Acknowledgments and Disclosure of Funding

This work was supported by the State Key Laboratory of General Artificial Intelligence, the National Natural Science Foundation of China (62406195), the HPC Platform of ShanghaiTech University, the Core Facility Platform of Computer Science and Communication of ShanghaiTech University, the Key Laboratory of Intelligent Perception and Human-Machine Collaboration (Shanghaitech University), Ministry of Education, and the Shanghai Engineering Research Center of Intelligent Vision and Imaging.

References

[1] Joao Pedro Araújo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Jiajun Wu, Deepak Gopinath, Alexander William Clegg, and Karen Liu. Circle: Capture in rich contextual environments. In *CVPR*, pages 21211–21221, 2023.

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[3] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15935–15946, 2022.

[4] Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. ContactPose: A dataset of grasps with object contact and hand pose. In *European Conference on Computer Vision (ECCV)*, volume 12358, pages 361–378, 2020.

[5] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *International Conference on Computer Vision (ICCV)*, pages 12417–12426, 2021.

[6] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9044–9053, 2021.

[7] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12943–12954, 2023.

[8] Zipeng Fu, Tony Zhao, and Chelsea Finn. Mobile ALOHA: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv preprint arXiv:2401.02117*, 2024.

[9] Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. ContactOpt: Optimizing contact to improve grasps. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1471–1481, 2021.

[10] Patrick Grady, Chengcheng Tang, Samarth Brahmbhatt, Christopher D. Twigg, Chengde Wan, James Hays, and Charles C. Kemp. PressureVision: Estimating hand pressure from a single rgb image. In *European Conference on Computer Vision (ECCV)*, volume 13666, pages 328–345, 2022.

[11] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the World in 3,000 Hours of Egocentric Video. In *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2022.

[12] Yinghao Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. Intercap: Joint markerless 3d tracking of humans and objects in interaction. In *DAGM German Conference on Pattern Recognition*, pages 281–299. Springer, 2022.

[13] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11107–11116, 2021.

[14] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[15] Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. Parahome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions. *arXiv preprint arXiv:2401.10232*, 2024.

[16] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[17] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2O: Two hands manipulating objects for first person interaction recognition. In *International Conference on Computer Vision (ICCV)*, pages 10138–10148, 2021.

[18] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17142–17151, 2023.

[19] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023.

[20] Jiaxin Lu, Chun-Hao Paul Huang, Uttaran Bhattacharya, Qixing Huang, and Yi Zhou. Humoto: A 4d dataset of mocap human object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10886–10897, 2025.

[21] Hyeongjin Nam, Daniel Sungho Jung, Gyeongsik Moon, and Kyoung Mu Lee. Joint reconstruction of 3d human and object via contact-based refinement transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10218–10227, 2024.

[22] Supreeth Narasimhaswamy, Trung Nguyen, and Minh Hoai Nguyen. Detecting hands and recognizing physical contact in the wild. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

[23] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL `https://arxiv.org/abs/2304.07193`.

[24] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[25] Rolandos Alexandros Potamias et al. WiLoR: End-to-end 3d hand localization and reconstruction in-the-wild. In *CVPR*, 2025.

[26] Yuxin Ray Song, Jinzhou Li, Rao Fu, Devin Murphy, Kaichen Zhou, Rishi Shiv, Yaqi Li, Haoyu Xiong, Crystal Elaine Owens, Yilun Du, Yiyue Luo, Xianyi Cheng, Antonio Torralba, Wojciech Matusik, and Paul Pu Liang. Opentouch: Bringing full-hand touch to real-world interaction, 2025. URL `https://arxiv.org/abs/2512.16842`.

[27] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pages 581–600. Springer, 2020.

[28] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+O: Unified egocentric recognition of 3d hand-object poses and interactions. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4511–4520, 2019.

[29] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: Unified 6d pose estimation and tracking of novel objects. In *CVPR*, 2024.

[30] Zhen Wu, Jiaman Li, Pei Xu, and C. Karen Liu. Human-object interaction from human-level instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2025.

[31] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. InterDiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, 2023.

[32] Sirui Xu, Dongting Li, Yucheng Zhang, Xiyan Xu, Qi Long, Ziyin Wang, Yunzhi Lu, Shuchang Dong, Hezi Jiang, Akshat Gupta, Yu-Xiong Wang, and Liang-Yan Gui. Interact: Advancing large-scale versatile 3d human-object interaction generation. In *CVPR*, 2025.

[33] Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. InterMimic: Towards universal whole-body control for physics-based human-object interactions. In *CVPR*, 2025.

[34] Yufei Ye, Yao Feng, Omid Taheri, Haiwen Feng, Shubham Tulsiani, and Michael J Black. Predicting 4d hand trajectory from monocular videos. *arXiv preprint arXiv:2501.08329*, 2025.

[35] Brent Yi, Vickie Ye, Maya Zheng, Yunqi Li, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7072–7084, 2025.

[36] Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. HOI-M^3: Capture multiple humans and objects interaction within contextual environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 516–526, 2024.

[37] Jianyi Zhou, Ziteng Gao, Feiyang Hong, Zirui Liu, Guannan Zhang, Weisheng Dai, Ruichen Zhen, Chuqiao Lyu, Haotian Wu, Yinian Mao, Xushi Wang, Yuxiang Jiang, Wenbo Ding, and Shuo Yang. TouchAnything: A dataset and framework for bimanual tactile estimation from egocentric video, 2026. URL `https://arxiv.org/abs/2605.13083`.

[38] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. TOCH: Spatio-temporal object-to-hand correspondence for motion refinement. In *European Conference on Computer Vision (ECCV)*, volume 13663, pages 1–19, 2022.

A DexHLM Dataset Details

This appendix expands the dataset construction summarized in Section 3. We describe the capture hardware, tactile representation, participants, and spatial protocol. The provided annotations include synchronized SMPL-X body and hand motion, rigid-object 6-DoF trajectories, camera poses, egocentric RGB-D, and per-frame measured tactile signals.

Capture hardware. DexHLM is captured with 22 OptiTrack Prime X 22 motion-capture cameras, a pair of Manus MetaGloves for fine hand motion, a pair of custom tactile gloves, each with 256 sensing points, and an Orbbec Quad-EGO four-camera headband. The exported modalities are temporally aligned at 30 Hz, with human and object motion expressed in a shared coordinate system.

Tactile signals and SMPL-X hand keypoints. The tactile measurements are mapped to a 32-dimensional representation associated with selected SMPL-X hand keypoints. Measured tactile signals guide human-motion refinement during dataset construction (Section 3.2) and provide supervision for tactile prediction (Section 4.2).

Participants and spatial protocol. The original dataset contains data from six participants, four male and two female, with varying heights and body proportions. The five interaction levels—Floor, Knee, Waist, Chest, and Neck—are calibrated to each participant's height rather than fixed at the same absolute heights for all participants. Together with three horizontal variations, the body-calibrated grid exposes changes in posture, reach, and viewpoint across interaction types.

Collection stages. The three collection stages of Section 3.1 are deployed as follows. **S1 (same-level interactions)** records each object–action pairing within a single layout cell and is the primary source of controlled variation across the body-calibrated grid. **S2 (cross-level transitions)** records HOI motion that includes locomotion, by carrying objects across multiple height levels. **S3 (long-horizon scene-level activity)** chains multiple interactions and transitions into continuous activities across freely arranged scene layouts. S3 additionally provides synchronized exocentric views recorded with four Z CAM cameras for independent qualitative inspection of body motion, object trajectories, and interaction transitions; these views are not recovery inputs.

Interaction coverage. DexHLM covers six everyday scenes and 52 objects, with both single-handed and bimanual activities. It includes prehensile, in-hand, and non-prehensile behaviors, such as grasping and transport, object reorientation during use, pushing, pouring, and bimanual cooking activities, embedded within long-horizon whole-body interactions. The object annotations describe rigid-body motion.

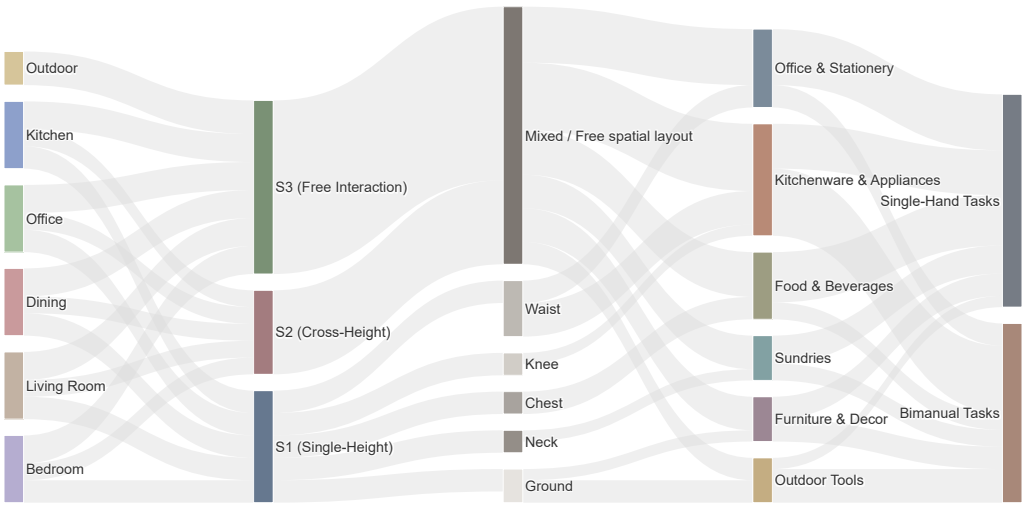

Figure 6: Scenes, collection stages, spatial layouts, object categories, and hand modes in DexHLM.

Synchronized examples. Figures 7 and 8 show synchronized egocentric and exocentric views, SMPL-X HOI, and tactile visualizations, with the complete object collection displayed in the background.

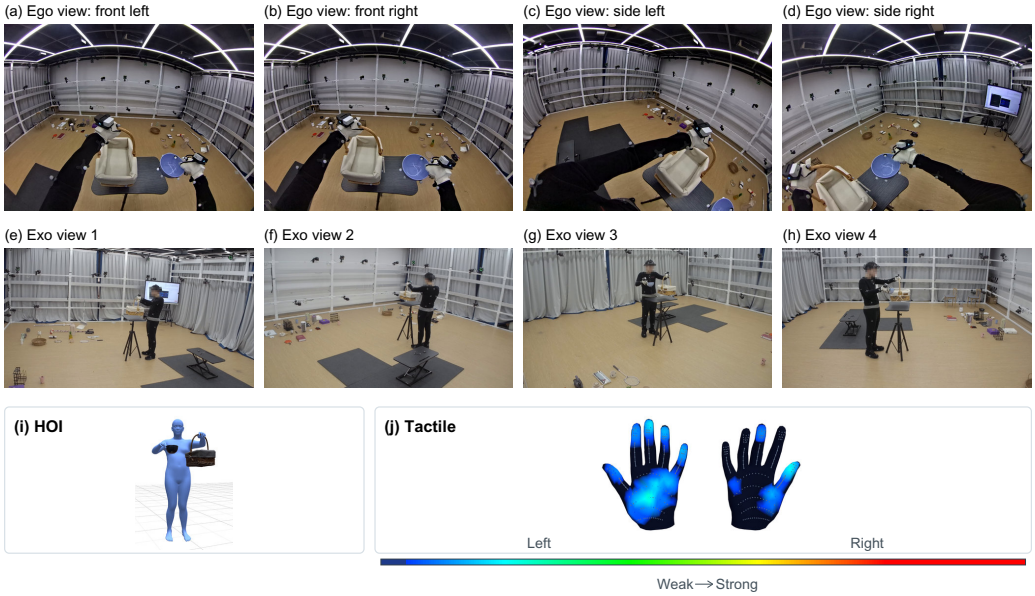

Figure 7: Synchronized egocentric and exocentric views, SMPL-X HOI, and tactile visualizations in DexHLM.

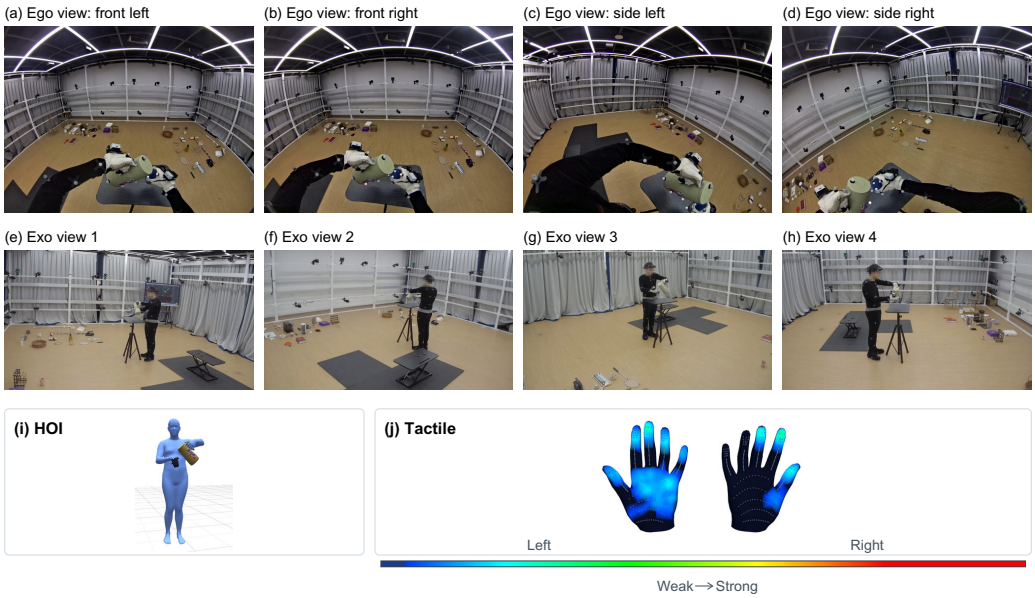

Figure 8: An additional synchronized multimodal example from DexHLM.

B Tactile Refinement and Prediction

B.1 Measured-Tactile Human-Motion Refinement

Body and object trajectories from MoCap are reliable, while local hand-object contact is the most fragile part of whole-body interaction data. We therefore anchor human-motion refinement in

measured tactile pressure. During dataset construction, human motion is refined while the MoCap object trajectory remains fixed.

Interaction interval detection. The dual tactile-geometric gate of Section 3.2 flags a frame when the maximum aggregated finger pressure exceeds τ_p and the hand-mesh-to-object distance falls below τ_d. Short isolated intervals are removed, and nearby intervals are merged before template selection. A representative template frame is selected using tactile confidence, kinematic stability, and geometric proximity, defining the wrist-relative reference for the subsequent steps.

Pressure-modulated local refinement. At the template frame, hand articulation and wrist pose are optimized under an objective whose weights derive from measured fingertip tactile signals. For each finger, pressure readings are aggregated over a temporal neighborhood and normalized to form a continuous tactile weight. These weights modulate attraction between sparse hand contact points and object-surface targets, while a soft pose-preservation term penalizes deviations from the captured hand articulation. The attraction is introduced in stages.

Spatiotemporal propagation and collision refinement. The refined wrist-relative transform is composed with the object trajectory to obtain a target wrist trajectory. The wrist target and hand-pose template are blended with the captured motion through smooth weights that ramp inside each interaction interval and decay at its boundaries. Body posture is adjusted through bounded inverse kinematics to follow the refined wrist target, with the root pose fixed. With the wrist trajectory fixed, the collision-aware step refines finger pose on near-contact frames. The penetration term penalizes hand vertices inside the object, while the closeness term anchors the result to the blended pose (Eq. (3)). The wrist-relative propagation assumes approximately stable contact within each interval.

B.2 Tactile Prediction from Egocentric Video

Input and backbone. The predictor consumes a short egocentric video clip together with a fixed task-conditioning prompt. We use Qwen3-VL [2] 2B and 8B as the multimodal backbone. LoRA adapters are injected into the language-layer attention projections, with rank r=16 and scaling α=32. The adapters, visual-language merger, and regression head are trained, while the remaining backbone weights are frozen.

Regression output. Inspired by OpenVLA [16], we attach a small two-layer MLP regression head to the backbone. The head is supervised by synchronized measured tactile signals and predicts non-negative values for the target-frame representation defined in Section A. The predictor learns visual interaction cues for human-demonstration collection; it does not explicitly estimate physical properties such as mass, friction, or fill level. Robots with reliable tactile sensors can use measured feedback during execution.

Regression and contact supervision. Weighted mean-squared error supervises the tactile-signal magnitude. Non-zero targets are up-weighted to counter contact sparsity. Binary cross-entropy supervises contact masks derived from a fixed pressure threshold.

Supervised baseline. The DINOv2 [23] ViT-B/14 baseline uses a frozen visual encoder with a learned temporal prediction head. The baseline and Ours use the same training data, egocentric RGB inputs, measured tactile supervision, and evaluation metrics. This comparison complements the zero-shot baselines with a learned predictor of the tactile signal.

Zero-shot baselines. The zero-shot Qwen3-VL baselines use native autoregressive text generation with fully frozen models, without LoRA adapters or a learned regression head. A structured prompt produces ordinal contact scores, which are calibrated and mapped to the same tactile representation for evaluation. All samples are retained: missing or invalid fields are set to zero, and outputs with no valid fields are scored as all-zero predictions. The zero-shot baselines and Ours are evaluated against the same measured signals using the same metrics.

Use of predicted tactile signals. At recovery time, predicted tactile signals identify contact intervals for post-processing. The recovered body and wrist trajectories are held fixed, while finger poses and

the object trajectory are adjusted. The measured-pressure weighting described in Section B.1 is used during dataset construction and is not applied to predicted tactile values.

C Evaluation Metric Definitions

Recovery is evaluated using GT camera poses and GT object meshes, with wrist positions lifted using GT depth. Hand and body errors are measured against the measured-tactile-refined MoCap motion, while object errors are measured against the MoCap object trajectories. The following definitions distinguish recovery accuracy from changes introduced during dataset refinement.

Hand metrics. **Rel-MPJPE** subtracts the wrist root per frame before averaging the per-joint Euclidean error. **PA-MPJPE** applies per-frame Procrustes alignment between predicted and GT joints before computing MPJPE. **GA-MPJPE** aligns each evaluated trajectory once via a global rigid transform and reports MPJPE on the aligned joints. **Rel-ACC** reports the mean root-relative joint-acceleration error.

Object metrics. **OTE** is the per-frame translation error in millimeters between predicted and GT object centroids. **ORE** is the per-frame angular error in degrees between predicted and GT object orientations, measured as the geodesic distance on $SO(3)$. Both are averaged over the evaluation frames of each object.

Body metrics. **MPJPE** is the per-frame Euclidean joint error in millimeters on the common body-joint subset. **PA-MPJPE** applies per-frame Procrustes alignment before MPJPE. **Wrist** is the world-frame body-FK position error averaged over the two wrist joints.

Refinement-ablation metrics. For the refinement ablation in Table 2, **PR** is the fraction of frames containing any penetrating hand vertex. **FP-C** is computed over frame–finger pairs as the fraction where the hand is in geometric near-contact with the object but the corresponding measured tactile signal is inactive. Near-contact denotes penetration or an unsigned surface distance below 30 mm. **NTF-D** is the mean displacement in millimeters of fingers without tactile evidence relative to the raw capture. **HPD** is the mean per-joint hand-pose deviation in radians relative to the raw capture. The refinement-ablation **MPJPE** reports full-body joint displacement against the captured motion. These displacement measures quantify changes to the captured demonstration, whereas PR and FP-C describe geometric and contact consistency.

Tactile prediction metrics. **Pearson** measures correlation between predicted and measured tactile signals. **MAE** and **MSE** report absolute and squared prediction errors in the processed tactile-signal space. **CF1** is element-level contact F1, computed from predicted and measured contact masks using a threshold of 0.5. **FC@k** is frame-level contact F1: a frame is labeled as contact when any of its tactile channels exceeds pressure threshold k on the processed tactile-signal scale. $\mathbf{R}^2$ is computed after flattening the evaluated frames and tactile channels within each sequence, and the resulting scores are averaged equally across sequences. Together, these metrics assess tactile-signal agreement and contact detection.

D Additional Experimental Results

Supervised comparison. In the main tactile-prediction evaluation, Ours-8B outperforms the supervised DINOv2 ViT-B/14 baseline, improving Pearson correlation from 0.674 to 0.812 and reducing MAE from 1.994 to 1.449 and MSE from 12.45 to 7.12.

Generalization to unseen subjects and objects. We further evaluate tactile prediction on jointly unseen subjects and objects. Both models are evaluated without retraining or adaptation. Ours-8B achieves a Pearson correlation of 0.795, compared with 0.812 in the main evaluation, and outperforms the supervised DINOv2 ViT-B/14 baseline on all reported metrics (Table 5). These results concern unseen subjects and objects within our capture setting; controlled physical-property shifts and in-the-wild evaluation remain future work.

The unseen objects introduce variation in appearance, geometry, material, surface characteristics, and manipulation affordances. Examples include stone mortar-and-pestle tools, a transparent glass kettle, fabric-covered furniture, a metal cup, and wooden building blocks. These objects also involve distinct grasp/contact configurations and manipulation patterns, such as grinding and block assembly, rather than only new instances with minor visual changes.

Table 5: Tactile prediction on jointly unseen subjects and objects. Metric definitions are given in Section C.

Method	Pearson↑	CF1↑	FC@3↑	FC@5↑	MAE↓	MSE↓	R^2↑
DINOv2 ViT-B/14	0.485	0.645	0.740	0.715	3.581	23.90	0.22
Ours-8B	**0.795**	**0.757**	**0.816**	**0.936**	**1.891**	**9.42**	**0.61**

Whole-body dexterous policy learning. DexHLM targets interactions that require coordinated whole-body control across different heights and precise bimanual manipulation. For example, in a low-height kettle-pouring motion, the humanoid must maintain a squat, insert one hand into a narrow kettle handle and control the pouring motion, while using the other hand to grasp and stabilize the cup. Small errors in body balance, wrist pose, or finger placement can break the required contacts and cause failure. We follow InterMimic [33] for simulated SMPL-X-based humanoid policy learning and evaluation. In a paired comparison with matched object trajectories, simulator, reward, and PPO configurations, measured-tactile refinement improves policy success from 28.0% with raw MoCap demonstrations to 62.0% with refined demonstrations. This comparison assesses the learnability of measured-tactile-refined MoCap demonstrations; policy learning directly from egocentrically recovered demonstrations remains future work.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state the dataset, measured-tactile refinement, and egocentric recovery contributions. Section 5 reports the corresponding empirical results; the controlled GT inputs and demonstration sources are specified in Sections 4.1 and 5.4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 contains a dedicated *Limitations & Future Work* paragraph that discusses the simulation-only humanoid policy evaluation and the sim-to-real gap in sensor noise, actuator dynamics, and physical compliance, together with planned mitigations (on-robot fine-tuning with real tactile feedback, closed-loop adaptation).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper presents a dataset and empirical methods and does not include theoretical results requiring proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 3.2 specifies measured-tactile human-motion refinement, and Sections 4.2 and 4.3 describe tactile prediction and egocentric recovery. Sections B.2, C, D and 4.1 specify the controlled inputs, evaluation references, supervision, baseline settings, metric definitions, and additional comparisons underlying the main results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code for the tactile-guided refinement and egocentric HOI recovery pipelines, including the tactile predictor and evaluation scripts, has been prepared for public release together with the full DexHLM dataset and reproduction instructions. These materials will be made openly available under a research-use license with the final dataset release.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Sections 3.1 and 4.1 describe the capture protocol and evaluation settings, including GT inputs and reference motions. Sections B.2 and 4.2 specify the predictor inputs, trainable modules, LoRA settings, regression head, supervision, and supervised and zero-shot baselines. Sections C and D define the metrics and the evaluation on unseen subjects and objects.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We do not report error bars, confidence intervals, or statistical significance tests for the experiments, including policy learning.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper focuses on dataset construction, methods, and evaluation; the operational details requested here are not reported.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics: data collection involves consenting human participants performing controlled manipulation under instrumented motion capture (Section A), and third-party models and services are used under their respective terms. The dataset, code, and models are intended for release under a research-use license.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The dataset can lower the barrier to humanoid learning from human demonstrations. Human-motion recordings may expose participant information or be used beyond the intended research setting; collection uses informed consent, and the planned release is for research use.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The assets are controlled human-demonstration recordings and a tactile predictor that outputs low-dimensional contact pressure; we do not consider them to fall under the high-risk class targeted by this question. Data collection uses informed consent, and the planned release is under a research-use license.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Third-party methods and assets are credited at first use. The component licenses include Apache-2.0 (Qwen3-VL and DINOv2), MIT (EgoAllo, EgoEgo, OpenVLA code, and InterMimic), CC-BY-NC-ND-4.0 (WiLoR models), the NVIDIA Source Code License (FoundationPose code), and the non-commercial research licenses for SMPL-X, SMPL-H, and MANO. HaPTIC [34] code and pretrained weights from the official repository are used with permission for academic evaluation and publication of comparative results.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The DexHLM dataset is documented in Sections A and 3, including the capture protocol, synchronized modalities, annotations, tactile representation, and spatial coverage. Sections B.2 and 4.2 document the tactile predictor's inputs, backbone, trainable components, regression head, and supervision. Evaluation metrics and limitations are described in Sections C and 6.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The full participant instructions and compensation details have been prepared and will be made publicly available with the dataset. Participants gave informed consent, could withdraw at any time, and were compensated at or above the local minimum wage.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: Data collection was conducted under IRB approval or the institutional equivalent for low-risk human-motion-capture research. The risks of wearing capture equipment and performing the controlled manipulation tasks were disclosed before enrollment, and participants gave informed consent.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Sections B.2 and 4.2 describe the Qwen3-VL [2] tactile predictor used in recovery-time refinement. VLM-assisted visual preprocessing is also used for off-the-shelf hand estimation within the recovery pipeline. Both are methodological uses of LLMs/VLMs.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.