EgoPHI: Estimating 3D Hand-Object Contact and Force from Egocentric Vision

Andela Ilic, Rachel Schuchert, Yijing Jiang, and Christian Holz

Department of Computer Science, ETH Zürich, Zürich, Switzerland

https://siplab.org/projects/EgoPHI

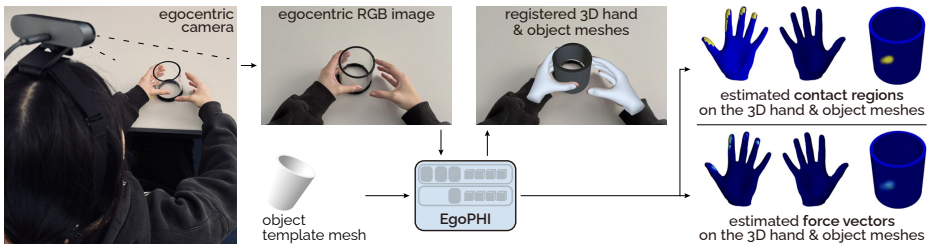

Fig. 1: EgoPHI is the first vision-based model that estimates 3D forces and contact during interaction with articulated rigid objects. Given a single egocentric monocular RGB image and the object geometry, EgoPHI registers hand meshes to refine the object pose in a shared camera coordinate frame, and predicts dense per-vertex contact and force distributions over all meshes for physically grounded interaction reasoning.

Abstract. Understanding hand-object interaction from egocentric vision is essential for modeling how people physically engage with the surrounding world. Yet reasoning about physically grounded interaction requires estimating the forces acting on hands and objects, beyond localizing contact. We present *EgoPHI*, the first method that jointly estimates dense contact maps and 3D force distributions on hand and object meshes from a single monocular RGB image and object geometry. To address the lack of scalable ground-truth force annotations, we introduce a physics-based simulation pipeline that augments existing hand-object datasets with dense per-vertex force supervision. EgoPHI then learns dense 3D contact and force on interacting hand and articulated object meshes, extending vision-based force estimation beyond image-space or planar settings. Our evaluation on in-distribution and out-of-distribution benchmarks shows that EgoPHI improves force estimation over existing approaches while generalizing to unseen datasets. To evaluate sim-to-real transfer, we constructed two physical objects that capture dense object contact and force magnitude and used them to record a dataset of interactions from eight participants across diverse touch and grasp types. Our results demonstrate that EgoPHI recovers meaningful 3D contact and force distributions in simulated, out-of-distribution, and real-world settings, advancing egocentric hand-object understanding from contact localization toward physically grounded interaction reasoning.

Code: `https://github.com/eth-siplab/EgoPHI`

1 Introduction

Understanding hand–object interaction is fundamental to modeling how humans engage with the physical world. Detecting contact and estimating the magnitude and spatial distribution of forces during interaction reveal intent [30], stability, and control [20] in object manipulation. For example, in robotics, recovering such information can enable robots to infer human manipulation strategies and learn stable grasping or object handling from observation.

Recent work has demonstrated predicting *contact* from visual input, addressing both human–scene contact [4,18,22,38,40] and fine-grained hand–object contact [1,14,16,17,21,39], where contact areas are small and vary during manipulation. Aided by datasets [36,38] and benchmarks [3,14], methods now reliably detect contact points and regions even in challenging interactions [16].

However, contact alone is insufficient to fully capture the physical dynamics of hand–object interaction, which are ultimately governed by the *forces* underlying these contacts. Nevertheless, inferring forces from egocentric RGB is difficult due to occlusion, weak visual cues, and uncertain 3D geometry. Subtle visual changes offer limited evidence of intensity, and single-view images provide little depth information to disambiguate 3D force direction. Consequently, prior approaches have estimated forces only in the image space [11,12] or flat surfaces [47].

In this paper, we present EgoPHI, a vision model that jointly estimates 3D forces and contact on both hands and the manipulated object from an egocentric view. Taking a single egocentric image and the object geometry as input, our model leverages estimated hand poses to refine the object's 3D pose in a shared camera frame and infers dense contact and force distributions across all meshes. For training EgoPHI, we augment a large dataset of hand interactions with articulated rigid objects (ARCTIC [9]) with physically simulated contact forces during the recorded hand-object interactions.

Our evaluations on ARCTIC (in distribution) and H2O [23] (out of distribution) show that EgoPHI outperforms existing baselines in force accuracy and reduces the mean error by approximately 2 N while generalizing to unseen objects. To evaluate sim-to-real transfer, we captured a real-world hand–object interaction dataset with two fabricated objects (a cube and a cylinder) from eight participants. We instrumented both objects to measure dense spatial contact and force distributions during touch and grasp. Our evaluation provides the first experimental evidence that a vision-only system can estimate contact and 3D force distributions on non-planar real-world objects.

Contributions. We make the following contributions in this paper:

1. EgoPHI, the first vision-based method that estimates 3D force distributions and contact regions over both hand and object meshes during manipulation,
2. a physics-based force simulation module that generates realistic contact and force data from hand and object meshes, and
3. an instrumented real-world dataset of ground-truth hand-object contact and 3D forces, captured via two augmented objects from eight participants during touch and grasp interaction to validate EgoPHI's sim-to-real transfer.

EgoPHI establishes a basis for inferring physically grounded hand-object interaction dynamics from vision alone. Our model performs robustly on unseen objects, diverse datasets, and real-world scenarios. We believe this opens a new direction for studying hand-object manipulation in everyday settings.

2 Related Work

2.1 Vision-based contact area estimation

Several projects have explored visual estimation of human–object and human–scene contact. Previous work has targeted whole-body contact using parametric models (e.g., SMPL [26], SMPL-X [27]) for dense 3D contact predictions across the full body, such as DECO [38], PICO [7], and GRACE [40]. Others have aimed at fine-grained hand–object contact via MANO [31], such as S2Contact [39], DyTact [6], and Ti3D-contact [5]. While single-image approaches can capture plausible contact for static frames, video-based methods such as VisTracker [43] exploit motion cues to reason about contact dynamics. HOT [4] and PIHOT [41] predict 2D contact heatmaps in image space, which are useful for visual recognition but lack detailed geometric information. Short of estimating contact masks, egocentric vision-based approaches have detected touch events on passive surfaces [34, 35]. Outside camera sensors, high-resolution contact masks have been estimated from low-resolution capacitive touch images using generative super-resolution approaches [33] and foundation-model-based discriminative approaches [32]. Conversely, methods that operate on meshes estimate dense 3D contact locations directly on the human or object surface (e.g., DECO [38], PICO [7], GRACE [40]).

In contrast to whole-body contact estimation, hand contact estimation is particularly challenging due to dual imbalance [21]. Most hand vertices are not in contact during interaction, which creates a class imbalance. In addition, spatial imbalance results from contact being predominantly concentrated on the fingertips, which limits accurate predictions in other hand regions. The state-of-the-art (SOTA) model in this area, HACO [21], addresses these challenges by introducing Balanced Contact Sampling (BCS) to mitigate class imbalance, and the Vertex-Level Class-Balanced (VCB) loss to alleviate spatial imbalance. Existing datasets vary in the types of objects they contain. They range from rigid objects [2, 3, 13, 16, 23] to articulated objects [9, 25] and deformable objects [42].

2.2 Vision-based force estimation

Early work by [8], [24], and [29] demonstrated that forces and impulses can be inferred from object and human motion cues in RGB/RGB-D video. These methods operate purely in image space without explicit 3D geometry [8, 24, 29]. Similar ideas have been explored in robotics, where vision-based estimators inferred grasp or contact forces from RGB observations of robotic hands [48]. [44] showed the feasibility of force estimation from visual deformation cues in minimally invasive surgery. Recently, FORCE introduced a physics-guided dataset

and model capturing human-applied forces and object resistance, which emphasized intuitive-physics priors for human–object interaction [46]. ViTaM-D is a visual-tactile framework to reconstruct hand-object interaction via a force-aware contact representation [45]. Force information acts as a physical constraint to fine-tune the reconstruction derived from the visual input.

Closely related to our work are PressureVision [12], PressureVision++ [11], and EgoPressure [47], which estimate pressure of hand contact with a flat surface from a single RGB image. While PressureVision and PressureVision++ predict pressure as a 2D heatmap in the image space, EgoPressure shifts the problem to the egocentric perspective and focuses on estimating the pressure distribution directly on the 3D hand mesh. Our work extends contact and force estimation from egocentric RGB images to both hands and articulated non-planar objects during interaction. To our knowledge, we are the first to generalize vision-based force estimation beyond flat surfaces to include rigid and articulated objects.

3 Method

3.1 Problem statement

Our objective is to estimate per-vertex contact points and interaction forces on the left and right hand meshes, denoted as $\mathcal{H}_L$ and $\mathcal{H}_R$, and on the object mesh $\mathcal{O}$ during hand–object interaction. To achieve this, EgoPHI leverages both visual and geometric information: it takes as input a single egocentric RGB image $\mathcal{I}$, the 3D coordinates of the left and right hands $\mathcal{P}_L, \mathcal{P}_R \in \mathbb{R}^{V_H \times 3}$, and the 3D coordinates of the object $\mathcal{P}_O \in \mathbb{R}^{V_O \times 3}$, where V_H and V_O are the numbers of hand and object vertices, respectively.

Hand pose estimation from a single image has become increasingly reliable in recent years [28]. In contrast, object pose is often partially occluded, particularly in egocentric views where the hands frequently obstruct large portions of the object. To handle this, EgoPHI estimates the object pose in the camera frame from a single egocentric RGB image $\mathcal{I}$, the object template mesh $\mathcal{O}_{\text{template}}$, and auxiliary cues derived from the image (object segmentation mask, hand centroids). Given the inputs

$$\{\mathcal{I}, \mathcal{O}_{\text{template}}\},$$

our goal is to predict per-vertex contacts

$$\mathcal{C}_L, \mathcal{C}_R, \mathcal{C}_O \in \mathbb{R}^{V_H \times 1}, \mathbb{R}^{V_H \times 1}, \mathbb{R}^{V_O \times 1}$$

and interaction forces

$$\mathcal{F}_L, \mathcal{F}_R, \mathcal{F}_O \in \mathbb{R}^{V_H \times 3}, \mathbb{R}^{V_H \times 3}, \mathbb{R}^{V_O \times 3}.$$

3.2 Method overview

Figure 2 presents an overview of the EgoPHI pipeline, which jointly estimates the object pose and per-vertex interaction forces and contacts. These force and

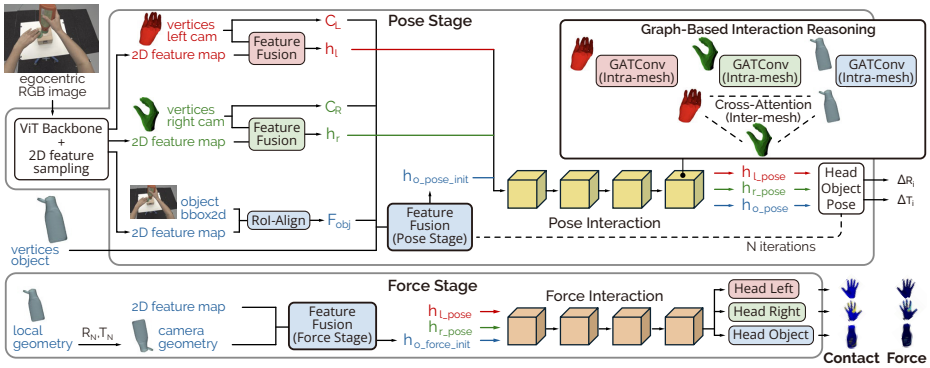

Fig. 2: EgoPHI's 3-stage pipeline: (1) visual & geometric feature extraction with cross-modal fusion, (2) object pose estimation, and (3) contact and force estimation. Our novel *Graph-Based Interaction Blocks* perform the core 3D reasoning by encoding intra-mesh structure and inter-mesh relationships between the hands and object. These blocks represent the 3D pose and spatial configuration of each mesh relative to the others. We thus explicitly model proximity and geometric interaction, which is critical for inferring physically grounded forces.

contact values are obtained through physics-based simulation using the hand and object meshes available in the dataset. Our pipeline consists of three main stages: **Visual and Geometric Feature Extraction**, **Iterative Pose Refinement**, and **Force Estimation**. The first stage maps visual and geometric inputs into per-vertex features. The second stage refines the object pose through successive updates. The third stage then predicts per-vertex forces and contact areas on the final 3D object mesh.

3.3 Physically Simulating Contact and Force

We supervise our model using physics-simulated, per-vertex contact and force distributions generated from existing hand–object datasets.

Given a dataset $\mathcal{D}$ comprising hand and object meshes and corresponding motion sequences, each frame includes the left and right hand meshes $\mathcal{H}_L$, $\mathcal{H}_R$, and the object mesh $\mathcal{O}$, all expressed in camera coordinates with vertex positions $\mathbf{v}_i \in \mathbb{R}^3$. Our simulation then estimates a force vector $\mathbf{f}_i \in \mathbb{R}^3$ at every vertex i on $\mathcal{H}_L$, $\mathcal{H}_R$, and $\mathcal{O}$. This produces a dense per-vertex distribution of 3D forces

$$\mathcal{F} = \{\, \mathbf{f}_i \mid \mathbf{v}_i \in \mathcal{H}_L \cup \mathcal{H}_R \cup \mathcal{O} \,\}.$$

We provide $\mathcal{H}_L$, $\mathcal{H}_R$, and $\mathcal{O}$ aligned with their recorded poses to SOFA simulator [10]. To estimate forces consistent with the observed hand-object configurations, we employ a penalty-based formulation that treats the meshes as quasi-rigid bodies. We model interactions via soft virtual springs that activate when vertices enter a predefined contact proximity zone. This formulation generates a restoring force $\mathbf{f}_i$ proportional to the collision displacement d_i,

$$\mathbf{f}_i = -k\, d_i\, \mathbf{n}_i$$

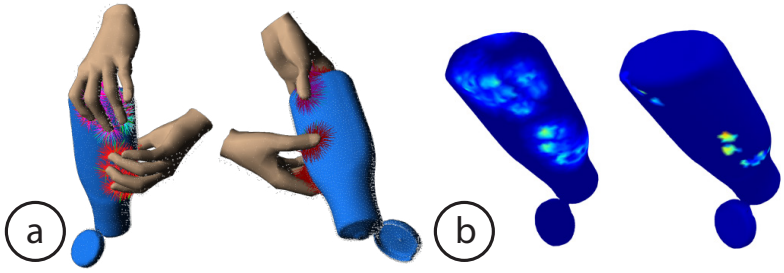

Fig. 3: (a) Activation of collision detection: pink lines show line-line contact, blue lines show point-line contact, yellow lines show point-point contact, and red lines represent constraints. (b) Simulated forces before and after applying contact masks.

where k represents the stiffness constant and $\mathbf{n}_i$ is the surface normal at vertex i. Here, d_i refers to the depth of the vertex within the contact threshold rather than a physical interpenetration of the underlying geometries. Each mesh is imported as a triangle-based surface model with a uniform mass m. We maintain a constant stiffness $k = 10$ across all simulations, which provides stable and physically-consistent supervision for the rigid and articulated rigid objects. Section A in the Appendix describes our implementation in more detail.

To derive contact areas on the meshes, we filter the raw forces $\mathcal{F}$ by a binary contact mask $\mathcal{M}$ that retains only physically valid contact regions. The final ground-truth forces are therefore

$$\tilde{\mathcal{F}} = \mathcal{M} \odot \mathcal{F},$$

where $\odot$ denotes element-wise masking. This step removes spurious non-contact forces and yields spatially consistent distributions over $\mathcal{H}_L$, $\mathcal{H}_R$, and $\mathcal{O}$ (Figure 3). The simulated forces and contact regions serve as dense, physically grounded supervision for EgoPHI. Combined with the ground-truth hand and object poses in the dataset, our augmentations provide realistic annotations to train our model's per-vertex contact and force prediction heads.

3.4 Data preparation

Before training, we normalize all meshes in both local and camera coordinates by centering them at the origin and scaling them to unit magnitude. We then map the articulated object mesh from its local coordinate system into the camera coordinate frame through a similarity transformation with rotation R and translation t. This aligns the template object to its ground-truth pose, and our pose-estimation model is later trained to predict these parameters (R, t).

For the hand meshes $\mathcal{H}_L$ and $\mathcal{H}_R$ and the object mesh $\mathcal{O}$, vertex positions are expressed as $\mathbf{v}_i \in \mathbb{R}^3$ in camera space. For force supervision, we normalize the simulated forces $\tilde{\mathcal{F}}$ by the maximum force magnitude observed across the dataset for the left hand, right hand, and object. To improve numerical stability and training efficiency, we decompose each force vector $\mathbf{f}_i$ into its magnitude $|\mathbf{f}_i|$

and unit direction $\mathbf{u}_i = \mathbf{f}_i/|\mathbf{f}_i|$, rather than predicting the full 3D vector directly. This representation enables cleaner loss formulations for force magnitude and direction and reduces sensitivity to extreme values in the raw force simulation.

3.5 Model architecture

Visual and Geometric Feature Extraction Our model integrates a Vision Transformer (ViT) backbone to extract patch-level features from the input image. The backbone returns a 2D feature map $F \in \mathbb{R}^{B\times C\times H_p\times W_p}$, where B is the batch size, C is the embedding dimension, and H_p, W_p are the spatial dimensions of the patch grid. We remove the classification head so that the feature map retains spatial detail needed for sampling features at projected mesh vertices.

To link visual cues with the 3D vertices of the hand meshes $\mathcal{H}_L$, $\mathcal{H}_R$ and the object mesh $\mathcal{O}$, we project each vertex $\mathbf{v}_i$ into the image using camera intrinsics. We sample F at the projected coordinates via bilinear interpolation and linearly embed the sampled vectors to the model's internal dimensionality D. Geometric and visual information jointly initialize the representation of each mesh vertex. **Hands:** For each vertex $\mathbf{v}_i \in \mathcal{H}_L \cup \mathcal{H}_R$, the 3D camera-space coordinate is linearly projected into a D-dimensional space and concatenated with the corresponding sampled visual feature. This concatenation is passed through a small multilayer perceptron (MLP) to produce the initial hand vertex embeddings. **Object (Pose Stage):** Our method dynamically computes the object's representation within an iterative pose refinement loop. Each iteration integrates four distinct feature vectors: (1) **visual features** sampled from a RoI-aligned feature map F_{obj} using the current 2D projection; (2) **local geometry** from the template mesh coordinates (x, y, z); (3) **relative geometry** to the left hand center; (4) **relative geometry** to the right hand center. All four are embedded and combined to form the object vertex embeddings for the current iteration.

Iterative Pose Refinement We model interactions between the left hand, right hand, and object using a stack of Graph-Based Interaction Blocks. Each block contains three components:

1. Intra-mesh Graph Attention Networks (GAT): These layers capture geometric relationships within each mesh ($\mathcal{H}_L$, $\mathcal{H}_R$, and $\mathcal{O}$). Edges follow the mesh connectivity and the outputs of the GAT use residual connections and layer normalization.

2. Inter-mesh Cross-Attention: Each mesh attends to the vertex features of the other two meshes (e.g., $\mathcal{H}_L$ attends to $\mathcal{H}_R$ and $\mathcal{O}$). This way our model can capture inter-object interaction cues.

3. Feed-forward Network (FFN): An MLP with a nonlinearity, residual connection, and layer normalization refines the vertex features.

We chain multiple Interaction Blocks to infer the object pose in two stages:

Stage 1. Initialization: The object pose is initialized with rotation R_0 as the identity and translation t_0 as zero.

Stage 2. Refinement Loop (for $i = 1$ to N):
- **Feature Update:** Given the current pose estimate (R_{i-1}, t_{i-1}), we update the object's vertex features while keeping the hand features fixed.
- **Interaction:** Static hand features and dynamic object features pass through the Interaction Blocks.
- **Global Pooling:** The resulting vertex features $(h_{l\_pose},\ h_{r\_pose},\ h_{o\_pose})$ are globally average-pooled and concatenated.
- **Incremental Prediction:** From the fused vector, a pose decoder predicts an incremental rotation $\Delta R_i \in \mathbb{R}^6$ (6D representation) and translation $\Delta t_i \in \mathbb{R}^3$.
- **Pose Composition:** The pose is updated by composing the increment: $R_i = \Delta R_i \cdot R_{i-1}$ and $t_i = \Delta t_i + t_{i-1}$.

Through this iterative process, the model refines the object's pose by repeatedly re-evaluating the visual and geometric relationships across the three meshes.

Force Estimation We estimate per-vertex contact probability, normalized force magnitude, and direction in a second stage. This stage is conditioned on the final predicted object pose (R_N, t_N) from our refinement process. We reuse the final hand features (h_{l_pose}, h_{r_pose}) from the last refinement iteration. Object features are recomputed by transforming each object vertex $\mathbf{v}_i \in \mathcal{O}$ into camera space using (R_N, t_N) and fusing two components: (1) visual features sampled from the full (non-RoI) backbone feature map, and (2) the final 3D camera-space coordinates of the object vertices. These updated hand and object features are processed by a second and separate stack of Interaction Blocks dedicated to contact and force prediction. The output heads for the left hand, right hand, and object predict per-vertex contact probabilities and force vectors. Predicted forces are multiplied by the predicted contact probabilities to enforce physical consistency and to suppress non-contact forces.

3.6 Loss Functions

We train our model with multiple objectives that jointly supervise contact prediction, force estimation, and object pose refinement. Contact maps for $\mathcal{H}_L$, $\mathcal{H}_R$, and $\mathcal{O}$ are supervised with a weighted binary cross-entropy (BCE) loss. We add three regularization terms to encourage spatial and semantic consistency [21]: (i) a vertex-level class-balanced (VCB) loss, (ii) a dataset-level contact regularization term, and (iii) a smoothness loss for locally coherent contact regions.

For every mesh vertex $\mathbf{v}_i$, we supervise both force magnitude and force direction. The magnitude is trained with a contact-weighted L2 loss together with a regularization term that encourages physically reasonable mean magnitudes per mesh. The direction is trained with a normalized L2 loss between predicted and ground-truth force vectors, weighted by the corresponding contact probabilities.

For pose estimation, we supervise the predicted pose using three terms: (i) a vertex reconstruction loss $\mathcal{L}_{\text{verts}}$, (ii) a geodesic rotation loss $\mathcal{L}_{\text{rot}}$, and (iii) a translation loss $\mathcal{L}_{\text{trans}}$. The full training objective combines all components:

$$\begin{aligned}\mathcal{L}_{\text{total}} = \lambda_c \mathcal{L}_{\text{contact}} + \lambda_m \mathcal{L}_{\text{force-mag}} + \lambda_v \mathcal{L}_{\text{force-vec}} \\ + \lambda_t \mathcal{L}_{\text{trans}} + \lambda_r \mathcal{L}_{\text{rot}} + \lambda_p \mathcal{L}_{\text{verts}}\end{aligned}$$

$\lambda_c, \lambda_m, \lambda_v, \lambda_r, \lambda_t, \lambda_p$ control the relative weight of each loss term.

4 Implementation

EgoPHI uses a ViT-B/16 backbone and our graph-based Interaction Blocks ($d = 512$, 4 heads). Our Iterative Refinement Module performs object pose refinement over three iterations. During training, the model uses ground-truth hand poses as input; at inference time, these are replaced with hand poses estimated by HAMER [28]. We train using Adam with a learning rate of 1×10^{-4} and a batch size of 8. The appendix provides further training and ablation details.

5 Experiments

Baselines We evaluate EgoPHI against existing methods, noting that prior work focuses exclusively on the hand; to our knowledge, no existing model estimates contact or force on the manipulated object. While EgoPressure [47] addresses 3D hand pressure, its code is not publicly available. Therefore, we establish two baseline strategies. **3D Mesh-Based Comparison:** We utilize the state-of-the-art contact estimation model HACO [21]. For contact, we use the publicly available version trained on 14 datasets, following their evaluation protocol on ARCTIC subject `s05` and H2O `subject4_ego`. To compare force estimation in 3D, we fine-tune the HACO architecture on our augmented ARCTIC training set for 10 epochs, maintaining the original training hyperparameters. **2D Image-Based Comparison:** We benchmark against PressureVision [12], a model specialized for force that we fine-tune on our training data. Since it estimates 2D hand pressure maps in image space, we project our 3D per-vertex forces onto the 2D image plane and discretize them into eight pressure bins. This allows for a fair comparison of hand force localization accuracy within the image domain.

Evaluation Protocol We train exclusively on the ARCTIC dataset, whose objects are strictly articulated and rigid. We evaluate on both the ARCTIC test set and H2O. H2O is excluded from training because its hand and object meshes are less accurate. Also, many sequences involve interactions with other hands or objects rather than the main object, which could confuse the model. To evaluate sim-to-real performance, we additionally collected a real-world dataset that captures forces exerted by hands on objects, because such data is highly challenging to obtain and rarely available in existing benchmarks.

Table 1: Comparison to PressureVision [12], a specialized 2D hand force estimation model, using 2D metrics computed on our projected 3D hand predictions.

	ARCTIC					**H2O**				
Model	**Prec.**	**Rec.**	**F1**	**IoU**	**Vol. IoU**	**Prec.**	**Rec.**	**F1**	**IoU**	**Vol. IoU**
PressureVision	.101	.175	.128	.068	6.6	.004	.008	.005	.003	0.2
EgoPHI w/o IRM	.186	.023	.041	.021	1.6	**.041**	.013	.020	.010	0.8
EgoPHI	**.245**	**.304**	**.271**	**.157**	**15.4**	.029	**.331**	**.053**	**.027**	**2.7**

Table 2: Contact and force estimation for HACO[1] and our models[2] (Full vs. without Iterative Refinement Module (IRM)). [1]Pretrained on 14 datasets (incl. ARCTIC, H2O). [2]Trained only on ARCTIC train set (excl. s05).

		HAND force			OBJECT force			HAND contact				OBJECT contact			
Model	**Train/Test**	**MAE [N]**	**RMSE [N]**	**vIoU**	**MAE [N]**	**RMSE [N]**	**vIoU**	**Prec.**	**Rec.**	**F1**	**IoU**	**Prec.**	**Rec.**	**F1**	**IoU**
evaluation on: ARCTIC s05															
HACO	**in-distribution**	6.62	7.29	1.0	*not supported*			.120	**.886**	**.196**	**.113**	*not supported*			
EgoPHI w/o IRM	**in-distribution**	4.80	5.94	0.7	5.01	6.03	0.7	.061	.092	.057	.033	.024	.229	.038	.020
EgoPHI	**in-distribution**	**4.03**	**5.06**	**2.2**	**4.42**	**5.28**	**2.0**	**.136**	.500	.190	.112	**.033**	**.517**	**.060**	**.032**
evaluation on: H2O s4_ego															
HACO	**in-distribution**	6.37	7.13	0.7	*not supported*			**.127**	**.831**	**.198**	**.117**	*not supported*			
EgoPHI w/o IRM	**out-of-distribution**	**4.65**	**5.87**	**0.8**	4.70	5.35	0.4	.038	.040	.026	.016	.020	.153	.030	.016
EgoPHI	**out-of-distribution**	5.16	6.62	0.6	**3.88**	**4.18**	**0.7**	.064	.350	.097	.055	**.020**	**.476**	**.037**	**.019**

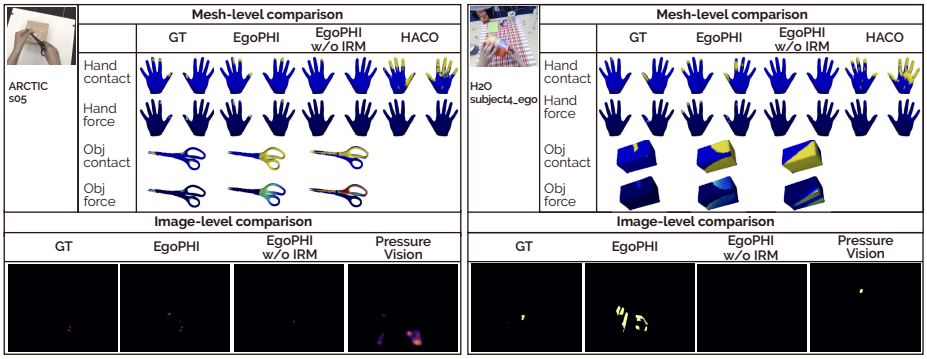

Fig. 4: EgoPHI estimates dense per-vertex force and contact for both hands and objects and outperforms the hand-only HACO baseline. EgoPHI's projected 2D force distributions outperform PressureVision and align better with the egocentric image.

Evaluation metrics We report precision, recall, and F1 score for contact prediction [21] along with contact intersection over union (IoU) [12, 47]. For force prediction, we report Masked Mean Absolute Error (Masked MAE), Masked Root Mean Squared Error (Masked RMSE), computed within ground-truth contact regions, and volumetric Intersection over Union (vIoU) [12, 47].

Ablation Study To isolate the contribution of the Iterative Refinement Module, we evaluate a baseline without the recursive object pose update. Contact and force distributions are predicted directly from the initial template mesh features without further geometric alignment.

5.1 Results on ARCTIC and H2O

Tables 1 and 2 compare EgoPHI's quantitative results with HACO and PressureVision. We also include an ablated version of EgoPHI without the Iterative Refinement Module. Evaluation is performed on the ARCTIC and H2O test splits used by the pretrained HACO model. Section C.1 in the Appendix reports results on the full H2O dataset. Figure 4 shows qualitative comparisons between EgoPHI, its ablation, HACO, and PressureVision.

Comparison against specialized force baseline EgoPHI consistently outperforms PressureVision across both contact and force evaluation metrics. While 2D models rely on local appearance cues such as skin blanching, they struggle under the severe occlusions typical of egocentric manipulation. EgoPHI leverages 3D geometric reasoning to anchor force predictions to the mesh topology and maintains high localization accuracy even when visual cues are obscured.

Testing on ARCTIC (in-distribution for both) EgoPHI surpasses the HACO baseline on force metrics with lower MAE and RMSE and higher Volumetric IoU. We note that absolute contact scores (F1 < 0.2, IoU < 0.15) reflect the inherent difficulty of the task where ground-truth contact regions are sparse. For contact regions, EgoPHI achieves higher precision with comparable F1 score and IoU, while HACO attains higher hand-contact recall, likely due to its training on a large, diverse set of 14 hand contact datasets. For object contact and force estimation, EgoPHI provides the first results of this kind; performance reflects the increased difficulty of localizing interaction on small, occluded object regions, especially when both hand and object poses are estimated. Overall, EgoPHI produces more physically meaningful force estimates and extends reasoning to object surfaces, a capability not supported by prior methods.

Testing on H2O (out-of-distribution for EgoPHI) H2O poses a challenging out-of-distribution setting (unseen objects, scenes, and capture conditions). Here, EgoPHI remains the only method that produces object contact and force estimates at all, which we consider the primary result. Our model produces hand forces with lower MAE and RMSE and comparable Volumetric IoU, even though we train it only on ARCTIC. HACO achieves higher hand contact recall, as expected from its large multi-dataset training.

Ablation of the Iterative Refinement Module (IRM) As shown in Tables 1 and 2, IRM consistently improves both contact and force estimation across in-distribution (ARCTIC) and out-of-distribution (H2O) datasets. In 2D image-level force comparisons, IRM substantially enhances contact localization. On ARCTIC, IoU increases from 0.021 to 0.157 and volumetric IoU from 1.6 to 15.4 compared to EgoPHI without IRM. On H2O, IoU rises from 0.010 to 0.027 and volumetric IoU from 0.8 to 2.7. At the mesh level, the full model achieves consistently better hand and object contact and more accurate object force estimation (IoU and vIoU improve up to $3\times$). The only minor exception is H2O hand forces, where the ablated model performs slightly better ($\sim 10\%$). These results demonstrate that IRM enforces tighter geometric alignment between hand and object meshes and produces more physically meaningful force predictions.

Pose Refinement Analysis Beyond contact metrics, we observe that our iterative refinement significantly improves 3D geometric understanding. Vertex-level alignment error consistently decreases across iterations, with Mean Per-Vertex Error (MPV) dropping by over 70% (ARCTIC: $64 \rightarrow 19$ cm, H2O: $48 \rightarrow 10$ cm). Pose correction allows the interaction module to operate more effectively and yields more accurate force and contact predictions even on imperfect initial poses.

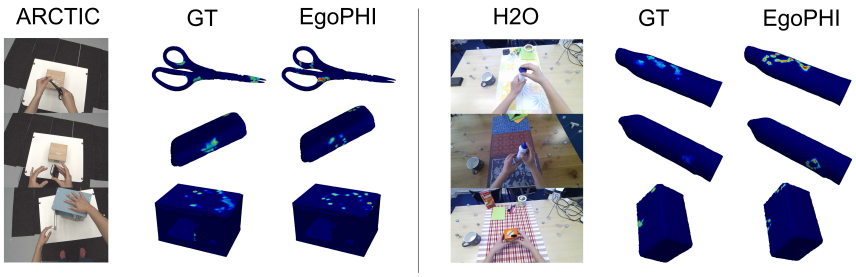

Fig. 5: Error-isolation analysis with ground-truth poses.

Error-Isolation Analysis with Ground-Truth Poses To isolate the source of misalignment observed in Figure 4, we repeat evaluation with ground-truth hand and object poses and bypass pose refinement entirely. This removes pose-related errors and yields accurate mesh alignment (Figure 5). Results confirm that residual misalignment stems from the difficulty of egocentric object pose estimation under occlusion, rather than from our simulated force labels.

Hand pose domain gap EgoPHI is trained with ground-truth hand poses. At inference time, it relies on HAMER-estimated hand poses [28]. To quantify this train-test domain gap, we fine-tuned EgoPHI using HAMER-estimated hand poses and observed a tradeoff: hand force prediction improves (MAE: 4.03 $\rightarrow$ 3.32 N, RMSE: 5.06 $\rightarrow$ 4.08 N, vIoU: 2.2 $\rightarrow$ 2.9), since the model adapts to the noisier input distribution at test time. However, hand contact prediction degrades (F1: 0.190 $\rightarrow$ 0.128, Recall: 0.500 $\rightarrow$ 0.212), and object force estimation worsens slightly (MAE: 4.42 $\rightarrow$ 4.73 N, vIoU: 2.0 $\rightarrow$ 1.8), while object contact remains comparable (F1: 0.060 $\rightarrow$ 0.061). We attribute this to a mismatch in our supervision: SOFA supervision is anchored to ground-truth mesh geometry, which becomes inconsistent with HAMER-predicted vertex positions. Contact labels computed from GT geometry no longer align precisely with the HAMER mesh used at inference, even though the overall force distribution remains learnable. This highlights an interplay between distribution matching for force regression and geometry-dependent contact supervision.

5.2 Evaluation on real object contact and force

To evaluate the sim-to-real performance of EgoPHI and validate our force simulation for objects, we designed two touch- and force-sensitive objects: a cube and a cylinder, made from 8 mm acrylic with polished edges (Figure 6). Along the bottom of each object we mounted an LED strip (TOPAI 12V COB) and covered it with opaque black tape. Light is reflected within the acrylic until a finger touches the surface, causing light to escape and illuminate the skin [15]. The refractive-index similarity between skin and acrylic causes light to escape into the finger, where it diffuses and reflects. Because surface residues and fin-

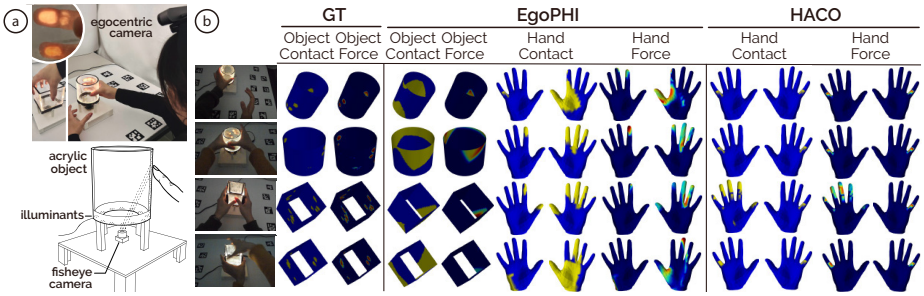

Fig. 6: (a) Apparatus of our real-world contact and force evaluation. Participants touched and grasped these two objects, which we constructed such that touch force corresponds to light intensity on the fingers and hands. (b) Qualitative results on real-world force recordings. While our experimental apparatus recorded ground-truth contact and force on the object only, we additionally show the estimated contact and force maps on the hands for visual comparison.

gerprint ridges naturally perturb the optical path, stronger pressure produces brighter reflections and provides an established proxy for force magnitude.

To capture contact areas, we mounted a fisheye camera (8 MP Sony IMX179 with a near-180° lens) below the object and calibrated camera intrinsics with a checkerboard. As acrylic has insufficient visual texture for reliable pose detection, we placed ArUco markers on the surface to track the poses of our objects.

We recruited eight participants for the data-capture study. Each wore a Logitech C920 webcam (78° diagonal FOV) to record egocentric views and interacted with each instrumented object using a set of predefined poses. These included two-finger grips at multiple locations and force levels; three-, four-, and five-finger configurations; full-hand grasps; and individual touches and presses, with multiple captures per interaction. Participants performed about ten repetitions per gesture for the left hand, right hand, and both hands. A typical session lasted under 20 minutes. Both cameras operated at fixed exposure and 30 fps. In total, we collected ∼2000 synchronized pairs of egocentric and contact+force frames.

Calibration of light intensity to force Before the experiment, we captured calibration data to map light intensity to physical force. We instrumented a small 8 mm acrylic panel with LEDs and tape, placed it on the non-illuminated cube above the fisheye camera, and mounted the assembly on a Beurer scale with 1 g resolution. An experimenter applied pressure to reach a target weight of 1000 g and adjusted the fisheye exposure so that the sensor did not saturate.

We then recorded light intensities for a series of controlled force levels. The experimenter pressed on the panel in 50 g increments up to 1000 g (i.e., 9.81 N) while the fisheye camera captured the reflected light pattern. For each step, we recorded ten frames, averaged the light intensity over the contact region (excluding a small margin), and converted the measured weights to Newtons. This lookup table maps optical intensity to force, through which we converted measured optical responses to per-vertex force measurements.

Table 3: Real-world evaluation for object contact and force.

		Object Contact				Object Force [N]	
Model	**Object**	**Precision**	**Recall**	**F1**	**IoU**	**MAE**	**RMSE**
EgoPHI	cylinder	.114	.305	.121	.067	1.35	3.24
EgoPHI	cube	.115	.316	.151	.085	0.48	1.54

Results Table 3 shows that EgoPHI can recover real-world object contact and force even though we trained it only on simulated forces (Figure 6). Force errors are lower than on ARCTIC, which indicates that the model captures object force magnitudes accurately. This may partly reflect the limited object variety. Contact prediction is challenging due to small, sparse real contact regions, yet EgoPHI successfully recovers the main contact areas. These results demonstrate that egocentric vision can infer physically meaningful 3D contact and force distributions on real, non-planar objects.

5.3 Validation of Simulated Force Supervision

To verify the fidelity of our SOFA-generated supervision, we compare simulated outputs against our real-world FTIR-instrumented object measurements (Table 4). SOFA simulation produces mean forces close to real-world FTIR measurements, with low per-frame MAE and strong Spearman correlations. These results demonstrate that the simulations provide a reliable and consistent source of force supervision for the rigid objects used in our training and evaluation.

6 Discussion and Limitations

Our evaluation shows that egocentric vision, when combined with known object geometry, can effectively infer dense 3D force fields. EgoPHI generalizes from simulated supervision to unseen interactions and, to some extent, real-world objects, showing that visual cues support physically meaningful force estimation.

Several limitations remain. First, EgoPHI processes frames independently and discards temporal structure. Since forces are inherently dynamic, a sequence-aware model could enforce temporal consistency and recover force trajectories that a per-frame approach cannot. Second, the model requires the object mesh template as input. Integrating feed-forward mesh reconstruction models such as SAM3D [37] could make EgoPHI applicable to fully unconstrained egocentric video. Third, our force simulation applies uniform stiffness across all objects. In our setting, force ratios between vertices are k-independent under a linear penalty, so the spatial distribution of forces is preserved regardless of the chosen

Table 4: Validation of SOFA simulation against real-world FTIR-based measurements.

Object	**SOFA Force (N)**	**FTIR Force (N)**	**Per-frame MAE (N)**	**Spearman ρ**
Cylinder	$\mu_F = 1.02 \pm 0.59$	$\mu_F = 1.21 \pm 0.50$	0.36 ± 0.18	.760
Cube	$\mu_F = 1.61 \pm 0.39$	$\mu_F = 1.67 \pm 0.39$	0.25 ± 0.16	.604

constant. However, this limits accuracy for objects with heterogeneous or compliant material properties. Fourth, EgoPHI predicts only normal contact forces, consistent with our simulation and FTIR sensing setup. Extending to tangential forces would require additional sensing, friction-aware simulation, and learning from weaker visual cues. A promising direction is post-hoc inference that combines predicted normal forces with object motion via Newton's second law, without extra hardware. Furthermore, because our model takes only geometry and a single RGB image as input, it lacks object properties (e.g., mass) needed to enforce global force equilibrium. Finally, the two real-world objects that resolved contact and force in our evaluation were a cube, which is piecewise planar, and a cylinder, which has a developable lateral surface. Future work should evaluate vision-based contact and force estimation on truly curved 3D objects. To obtain ground truth on such real-world geometries, our recent fabrication-based approach for retrofitting existing 3D objects with surface-conforming capacitive sensing could be a first step [19].

7 Conclusion

We have presented EgoPHI, a method for estimating physically grounded hand-object interaction from egocentric vision. EgoPHI estimates dense contact maps and 3D force distributions on both hands and manipulated objects by combining image evidence, object geometry, graph-based interaction reasoning, and iterative pose refinement. Since dense force annotations are difficult to obtain in natural interaction, we introduced a physics-based simulation pipeline that turns existing hand-object datasets into per-vertex contact and force supervision.

Across simulated, out-of-distribution, and real-world evaluations, EgoPHI shows that egocentric RGB can support meaningful 3D force estimation without tactile sensing at test time. EgoPHI improves force prediction over prior hand-only and image-space baselines, generalizes to unseen interaction data, and remains the only evaluated approach that estimates force and contact on object surfaces. Our instrumented-object benchmark further provides dense physical measurements on real 3D objects, offering a step toward validating sim-to-real transfer for force-aware egocentric perception. At the same time, the remaining errors highlight open challenges in egocentric physical perception, including robust object pose recovery under occlusion, temporal force consistency, and modeling material-dependent interaction.

Our results contribute to a broader shift in egocentric perception from recognizing contact toward estimating the physical interaction it represents. By predicting forces jointly on hands and objects, EgoPHI adds a force-aware layer to visual hand-object understanding, complementing existing work on contact, pose, and manipulation. This capability is important because force distributions provide cues about how an object is being held, pressed, stabilized, or controlled, which can support downstream applications such as robot learning from human demonstration, augmented assistance, and embodied scene understanding.

References

1. Brahmbhatt, S., Tang, C., Twigg, C.D., Kemp, C.C., Hays, J.: Contactpose: A dataset of grasps with object contact and hand pose (2020),
2. Cao, Z., Radosavovic, I., Kanazawa, A., Malik, J.: Reconstructing hand-object interactions in the wild. In: ICCV (2021)
3. Chao, Y.W., Yang, W., Xiang, Y., Molchanov, P., Handa, A., Tremblay, J., Narang, Y.S., Wyk, K.V., Iqbal, U., Birchfield, S., Kautz, J., Fox, D.: Dexycb: A benchmark for capturing hand grasping of objects (2021),
4. Chen, Y., Dwivedi, S.K., Black, M.J., Tzionas, D.: Detecting human-object contact in images (2023),
5. Chen, Z., Cao, C., Ouyang, Y., Chen, H., Jin, H., Zhang, S.: Ti3d-contact, a high-resolution and whole-body dataset of hand-object contact area based on 3d scanning method (2024). ,
6. Cong, X., Xing, A., Pokhariya, C., Fu, R., Sridhar, S.: Dytact: Capturing dynamic contacts in hand-object manipulation (2025),
7. Cseke, A., Tripathi, S., Dwivedi, S.K., Lakshmipathy, A., Chatterjee, A., Black, M.J., Tzionas, D.: Pico: Reconstructing 3d people in contact with objects (2025),
8. Ehsani, K., Tulsiani, S., Gupta, S., Farhadi, A., Gupta, A.: Use the force, luke! learning to predict physical forces by simulating effects (2020),
9. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: Arctic: A dataset for dexterous bimanual hand-object manipulation (2023),
10. Faure, F., Duriez, C., Delingette, H., Allard, J., Gilles, B., Marchesseau, S., Talbot, H., Courtecuisse, H., Bousquet, G., Peterlik, I., Cotin, S.: SOFA: A Multi-Model Framework for Interactive Physical Simulation. In: Payan, Y. (ed.) Soft Tissue Biomechanical Modeling for Computer Assisted Surgery, Studies in Mechanobiology, Tissue Engineering and Biomaterials, vol. 11, pp. 283–321. Springer (Jun 2012). ,
11. Grady, P., Collins, J.A., Tang, C., Twigg, C.D., Aneja, K., Hays, J., Kemp, C.C.: Pressurevision++: Estimating fingertip pressure from diverse rgb images (2024),
12. Grady, P., Tang, C., Brahmbhatt, S., Twigg, C.D., Wan, C., Hays, J., Kemp, C.C.: Pressurevision: Estimating hand pressure from a single rgb image (2022),
13. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: Honnotate: A method for 3d annotation of hand and object poses. In: CVPR (2020)
14. Hampali, S., Sarkar, S.D., Lepetit, V.: Ho-3d_v3: Improving the accuracy of hand-object annotations of the ho-3d dataset (2021),
15. Han, J.Y.: Low-cost multi-touch sensing through frustrated total internal reflection. In: Proceedings of the 18th annual ACM symposium on User interface software and technology. pp. 115–118 (2005)
16. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects (2019),
17. Hu, J., Zhang, H., Chen, Z., Li, M., Wang, Y., Liu, Y., Sun, Z.: Learning explicit contact for implicit reconstruction of hand-held objects from monocular images (2024),
18. Huang, C.H.P., Yi, H., Höschle, M., Safroshkin, M., Alexiadis, T., Polikovsky, S., Scharstein, D., Black, M.J.: Capturing and inferring dense full-body human-scene contact (2022),
19. Ilic, A., Gao, J., Li, Z., Jiang, Y., Schuchert, R., Meier, M., Herholz, P., Holz, C.: Retrofitting existing 3d objects with surface-conforming capacitive sensing. In: ACM SIGGRAPH 2026 Conference Papers. SIGGRAPH '26, Association for Computing Machinery, New York, NY, USA (2026)

20. Jiang, Y., Yu, M., Zhu, X., Tomizuka, M., Li, X.: Contact-implicit model predictive control for dexterous in-hand manipulation: A long-horizon and robust approach (2024),
21. Jung, D.S., Lee, K.M.: Learning dense hand contact estimation from imbalanced data (2025),
22. Kim, H., Han, S., Kwon, P., Joo, H.: Beyond the contact: Discovering comprehensive affordance for 3d objects from pre-trained 2d diffusion models (2024),
23. Kwon, T., Tekin, B., Stuhmer, J., Bogo, F., Pollefeys, M.: H2o: Two hands manipulating objects for first person interaction recognition (2021),
24. Li, Z., Sedlar, J., Carpentier, J., Laptev, I., Mansard, N., Sivic, J.: Estimating 3d motion and forces of person-object interactions from monocular video. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 8632–8641. IEEE (Jun 2019). ,
25. Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., Yi, L.: Hoi4d: A 4d egocentric dataset for category-level human-object interaction (2024),
26. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015)
27. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 10975–10985 (2019)
28. Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., Malik, J.: Reconstructing hands in 3d with transformers (2023),
29. Pham, T.H., Kheddar, A., Qammaz, A., Argyros, A.A.: Towards force sensing from vision: Observing hand-object interactions to infer manipulation forces. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2810–2819 (2015).
30. Pham, T.H., Kyriazis, N., Argyros, A.A., Kheddar, A.: Hand-object contact force estimation from markerless visual tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(12), 2883–2896 (2018).
31. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6) (Nov 2017)
32. Schuchert, R., Holz, C.: CapFM: A foundation model for capacitive touch sensing. In: Proceedings of the ACM Symposium on User Interface Software and Technology (UIST). Association for Computing Machinery, New York, NY, USA (2026)
33. Streli, P., Holz, C.: CapContact: Super-resolution contact areas from capacitive touchscreens. In: Proceedings of the ACM CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA (2021)
34. Streli, P., Jiang, J., Rossie, J., Holz, C.: Structured light speckle: Joint egocentric depth estimation and low-latency contact detection via remote vibrometry. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST). Association for Computing Machinery, New York, NY, USA (2023)
35. Streli, P., Richardson, M., Botros, F., Ma, S., Wang, R., Holz, C.: TouchInsight: Uncertainty-aware rapid touch and text input for mixed reality from egocentric vision. In: Proceedings of the 37th Annual ACM Symposium on User Interface Soft-

ware and Technology (UIST). Association for Computing Machinery, New York, NY, USA (2024).
36. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: GRAB: A Dataset of Whole-Body Human Grasping of Objects, p. 581–600. Springer International Publishing (2020). ,
37. Team, S.D., Chen, X., Chu, F.J., Gleize, P., Liang, K.J., Sax, A., Tang, H., Wang, W., Guo, M., Hardin, T., Li, X., Lin, A., Liu, J., Ma, Z., Sagar, A., Song, B., Wang, X., Yang, J., Zhang, B., Dollár, P., Gkioxari, G., Feiszli, M., Malik, J.: Sam 3d: 3dfy anything in images (2025),
38. Tripathi, S., Chatterjee, A., Passy, J.C., Yi, H., Tzionas, D., Black, M.J.: Deco: Dense estimation of 3d human-scene contact in the wild (2023),
39. Tse, T.H.E., Zhang, Z., Kim, K.I., Leonardis, A., Zheng, F., Chang, H.J.: S^2contact: Graph-based network for 3d hand-object contact estimation with semi-supervised learning (2023),
40. Wang, C., Zhai, W., Yang, Y., Cao, Y., Zha, Z.: Grace: Estimating geometry-level 3d human-scene contact from 2d images (2025),
41. Wang, Y., Neng, W., Wei, Z., Lei, Y., Xue, W., Zhuang, N., Xu, Y., Jiang, X., Liu, Q.: Precision-enhanced human-object contact detection via depth-aware perspective interaction and object texture restoration (2024),
42. Xie, W., Yu, Z., Zhao, Z., Zuo, B., Wang, Y.: Hmdo: Markerless multi-view hand manipulation capture with deformable objects (2023),
43. Xie, X., Bhatnagar, B.L., Pons-Moll, G.: Visibility aware human-object interaction tracking from single rgb camera (2023),
44. Yang, S., Le, M.H., Golobish, K.R., Beaver, J.C., Chua, Z.: Vision-based force estimation for minimally invasive telesurgery through contact detection and local stiffness models (2024),
45. Yu, Z., Xu, W., Xie, P., Li, Y., Anthony, B.W., Zhang, Z., Lu, C.: Dynamic reconstruction of hand-object interaction with distributed force-aware contact representation (2025),
46. Zhang, X., Bhatnagar, B.L., Starke, S., Petrov, I., Guzov, V., Dhamo, H., Pérez-Pellitero, E., Pons-Moll, G.: Force: Physics-aware human-object interaction (2024),
47. Zhao, Y., Kwon, T., Streli, P., Pollefeys, M., Holz, C.: Egopressure: A dataset for hand pressure and pose estimation in egocentric vision (2024),
48. Zhu, Y., Hao, M., Zhu, X., Bateux, Q., Wong, A., Dollar, A.M.: Forces for free: Vision-based contact force estimation with a compliant hand. Science robotics **10 103**, eadq5046 (2025),

A Physical contact and force simulation

In this section, we provide the implementation details for our physical simulation of touch contact and forces on the object and hand meshes for our two datasets (Section 3.3).

A.1 SOFA Physics Simulation Setup

We created our environment in the physics simulator SOFA [10]. While physics simulators are typically used to predict how objects move in response to applied forces, we use it in the opposite direction: we take the hand and object poses already provided by the dataset for each frame, and simulate the contact forces that arise from those poses.

To this end, we import the left hand, right hand, and object meshes as triangle-based surface models and assign them a `UniformMass` to capture their inertial properties. For the ARCTIC objects, we select the following masses: laptop (0.8 kg), phone (0.1 kg), ketchup bottle (0.05 kg), mixer (0.85 kg), waffle iron (0.8 kg), scissors (0.09 kg), capsule machine (0.4 kg), box (0.75 kg), notebook (0.5 kg), espresso machine (0.8 kg), and microwave (1.1 kg). For the H2O dataset, we use the following object masses: book (0.5 kg), espresso cup (0.2 kg), lotion bottle (0.3 kg), deodorant (0.25 kg), milk carton (1.0 kg), cocoa container (0.4 kg), chips bag (0.1 kg), and cappuccino cup (0.2 kg). For all simulations, we assign each hand a uniform mass of 0.5 kg.

A.2 Contact Modeling and Force Estimation

Each body is equipped with collision models (point, line, and triangle) mapped to the surface geometry through `BarycentricMapping`, enabling SOFA to compute contact at the mesh vertices consistently. We employ the `PenaltyContactForceField` to handle collisions. This penalty-based method is well-suited to our setting because it generates contact forces based on proximity: once two surfaces come within a predefined contact distance, a restoring force is applied proportional to how much that distance threshold is exceeded. This naturally handles both cases that arise from imperfect mesh reconstructions - surfaces that are very close but not quite touching, and those that slightly overlap - without requiring the meshes to be in perfect geometric contact.

For handling collisions, we rely on SOFA's `CollisionPipeline`, which integrates different stages of the collision detection process (Figure 3 (a)). Specifically, a BruteForce broad-phase detector is used to initially identify potential colliding pairs, followed by a BVH (Bounding Volume Hierarchy) narrow phase to refine the detection with higher geometric accuracy. Finally, the `DefaultContactManager` coordinates how detected contacts are resolved, passing the collision information to the selected contact response model, which determines the appropriate reactive forces.

A.3 Simulation Parameters and Solver Configuration

We set the simulation timestep to 0.01 s and employ an implicit Euler solver with Rayleigh damping coefficients of 0.1 for both mass and stiffness. For the resulting linear systems at each timestep, we use a Conjugate Gradient solver with a maximum of 200 iterations and a tolerance of 10^{-9}. The collision detection pipeline includes a `MinProximityIntersection` component, with dataset-specific alarm and contact distances of 0.02 m for the ARCTIC dataset and 0.04 m for the H2O dataset. Contact constraints are resolved using a `GenericConstraintSolver` with a maximum of 1000 iterations and a tolerance of 10^{-5}, combined with a penalty-based contact force field. Forces at each mesh vertex are stored in dedicated `MechanicalObjects`, while `ContactListeners` capture hand-object interactions and save the resulting forces frame-by-frame.

This configuration results in objects that behave as rigid bodies with soft penalty-based contact, which is sufficient for our application, where the primary goal is to measure contact forces at the hand–object interface rather than simulate detailed material deformations. After computing the simulated forces, we apply a contact mask to restrict them to the actual contact regions. Since the simulation uses a `Vec3d` representation in SOFA (typically used for meshes), the resulting forces initially resemble those on a deformable object. The contact mask removes these non-physical forces outside contact areas, yielding forces consistent with rigid-body behavior (Figure 3 (b)).

B Implementation

B.1 Training details

For training, our method requires only the input image and the template object mesh. All additional supervisory cues (hand poses, object segmentation mask, hand centroids) are extracted from the input image. Specifically, we apply the HAMER model to estimate hand poses [28]. As described in Section 3.4, we compute the ground-truth (GT) rotation and translation required to transform the normalized object template mesh into the object mesh in the camera coordinate system. During training, GT hand poses are used as input, while at evaluation time we replace them with the poses estimated by HAMER. The model is supervised using the GT object vertices expressed in the camera coordinate frame.

Our architecture integrates a Vision Transformer backbone (ViT-B/16) pretrained on ImageNet, with a hidden dimension of 768, together with a graph-based module using a latent feature size of 512 and four attention heads. A key component of our pipeline is the set of Interaction Blocks: a stack of four blocks is used for pose estimation, followed by another stack of four blocks for contact and force estimation. Pose refinement is performed over the three iterative steps within the Iterative Refinement Module.

The model is trained using the Adam optimizer with a learning rate of 1×10^{-4}, weight decay of 1×10^{-4}, and gradient clipping with a maximum norm of 1.0. We use a batch size of 8.

The weights of the loss terms defined in Section 3.6 are set as follows:

$$\lambda_c = 0.1, \quad \lambda_m = 0.6, \quad \lambda_v = 0.25, \quad \lambda_r = 0.7, \quad \lambda_t = 120, \quad \lambda_p = 200.$$

B.2 Evaluation details: Hand Force Projection to the Image Space

To compare 3D per-vertex hand forces with 2D pressure maps from PressureVision, it is necessary to project the sparse 3D contact signal into a dense, camera-aligned representation. This section details the rasterization-based projection pipeline used to transform MANO-based vertex pressures into pixel-wise pressure class maps. This process maps the discrete 3D physical quantities into a continuous 2D pressure map, $P \in \mathbb{R}^{H \times W}$, where H, W represent the target resolution.

Pressure Conversion and Normalization Given the per-vertex masked force magnitudes F_v (in Newtons), we define the expected vertex-wise pressure P_v (in kPa) as $P_v = F_v \cdot \Phi$. Here, Φ is a conversion constant determined by the reference sensor's pixel area (1.25×1.25 mm^2), yielding a scale factor of $\Phi = 640.0$.

Mesh Rasterization and Occlusion Handling To ensure a dense and topologically consistent 2D mapping, we employ a triangle rasterization approach. Using the camera intrinsics K, vertices are projected from camera space to the image grid. For every pixel $\mathbf{p}$ enclosed by a projected MANO face $f = \{i_0, i_1, i_2\}$, the pressure value $P(\mathbf{p})$ is calculated via barycentric interpolation of the vertex pressures $P_{i_{0,1,2}}$. To properly handle self-occlusions (e.g., a fingertip occluded by the palm relative to the camera) and ensure a fair comparison with PressureVision, a Z-buffer is maintained so that only the pressure values of the visible hand surface contribute to the final 2D map.

Quantization into Pressure Classes To align with the evaluation protocol of PressureVision, the continuous pressure map is discretized into N bins using a non-linear thresholding scale:

$$\mathcal{T} = \{0.0, 0.5, 1.0, 2.0, 4.0, 8.0, 16.0, 32.0, 64.0\} \text{ kPa}.$$

The final output is a categorical class map where each pixel value corresponds to the index of its pressure magnitude within $\mathcal{T}$. This quantization effectively accounts for the high dynamic range of human grasp forces while providing a robust format for image-space comparison.

C Additional experiments

C.1 Evaluation on the whole H2O dataset

In Section 5.1, we compared EgoPHI's performance to HACO on the H2O `subject4_ego` split, following the evaluation protocol used in the HACO paper.

Table 5: Extension of Table 1 to the full H2O dataset. We report the performance of EgoPHI, its ablated version (without IRM), and PressureVision in image space. The models are trained only on ARCTIC and evaluated out-of-distribution on H2O.

	H2O s4_ego					H2O full				
Model	**Prec.**	**Rec.**	**F1**	**IoU**	**Vol. IoU**	**Prec.**	**Rec.**	**F1**	**IoU**	**Vol. IoU**
PressureVision	.004	.008	.005	.003	0.2	.008	.012	.010	.005	0.4
EgoPHI w/o IRM	**.041**	.013	.020	.010	0.8	**.147**	.018	.032	.016	1.2
EgoPHI	.029	**.331**	**.053**	**.027**	**2.7**	.087	**.400**	**.143**	**.077**	**7.2**

Table 6: Extension of Table 2 from the main paper to the full H2O dataset. EgoPHI is trained only on ARCTIC and evaluated out-of-distribution on H2O.

	HAND force			OBJECT force			HAND contact				OBJECT contact			
Model	**MAE [N]**	**RMSE [N]**	**vIoU**	**MAE [N]**	**RMSE [N]**	**vIoU**	**Prec.**	**Rec.**	**F1**	**IoU**	**Prec.**	**Rec.**	**F1**	**IoU**
evaluation on H2O s4_ego														
EgoPHI w/o IRM	**4.65**	**5.87**	**0.8**	4.70	5.35	0.4	.038	.040	.026	.016	.020	.153	.030	.016
EgoPHI	5.16	6.62	0.6	**3.88**	**4.18**	**0.7**	**.064**	**.350**	**.097**	**.055**	.020	**.476**	**.037**	**.019**
evaluation on full H2O														
EgoPHI w/o IRM	**4.96**	**6.19**	0.7	4.40	5.09	0.4	.039	.038	.026	.016	.020	.098	.025	.014
EgoPHI	5.30	6.76	0.7	**3.76**	**4.05**	**0.7**	**.074**	**.233**	**.094**	**.055**	**.021**	**.211**	**.033**	**.018**

This setting is out-of-distribution for EgoPHI, but in-distribution for HACO, since H2O was part of HACO's training data. Here, we extend that evaluation in two complementary ways: we report image-space contact and force results alongside PressureVision (Table 5), and we report 3D mesh-space force and contact estimates for EgoPHI and its ablation across both the `subject4_ego` split and the full H2O dataset (Table 6).

Image-space evaluation. We observe that all models demonstrate consistently higher performance on the full H2O dataset compared to the `subject4_ego` test split, suggesting that the proposed method generalizes effectively across a broader range of participants and manipulation styles. Table 5 shows that EgoPHI substantially outperforms PressureVision across both evaluation splits, with particularly large gaps in recall, F1, and volumetric IoU. This is notable given that PressureVision is a supervised method trained on image data, while our approach operates entirely in 3D and projects to image space only at evaluation time.

The ablated variant (EgoPHI w/o IRM) achieves higher precision in isolation, but at the cost of drastically lower recall, resulting in worse F1 and IoU overall. This confirms that IRM plays a critical role in producing spatially spread and well-calibrated contact predictions, rather than overly conservative ones.

3D mesh-space evaluation. Table 6 reports per-vertex force and contact metrics directly on the mesh. Results on the full H2O dataset are consistent with those on `subject4_ego`: metric values shift only slightly, and the relative ranking

between EgoPHI and its ablation remains stable. The inclusion of IRM particularly benefits object-force estimation and contact recall, while hand-force MAE and RMSE remain comparable between the two variants. The consistent performance across splits suggests that EgoPHI captures physical cues that generalize across the broader H2O distribution.

C.2 Model Performance on ARCTIC vs. H2O

While Tables 1 and 2 in the main paper report absolute performance per dataset, here we focus specifically on how performance changes when moving from the in-distribution ARCTIC split to the out-of-distribution H2O dataset. The most consistent degradation is observed in mesh contact localization: hand contact F1 drops from 0.190 to 0.097, and object contact F1 from 0.060 to 0.037. In image space, volumetric IoU drops from 15.4 to 2.7, reflecting sparser and less spatially accurate contact predictions on unseen objects. Notably, force magnitude estimation is considerably more robust to this domain shift: object force MAE remains comparable (4.42 N on ARCTIC vs. 3.88 N on H2O), suggesting that the model learns transferable physical priors for force intensity, while struggling more with precisely localizing where contact occurs.

C.3 Per-Object Performance Analysis

Model performance is significantly influenced by the geometric and physical properties of the manipulated objects. While our proposed simulation framework assumes rigid-body interactions, which aligns well with objects in the ARCTIC dataset that exhibit minimal surface deformation, the shape and affordances of the objects also play a crucial role. Consequently, contact regions for many ARCTIC objects are typically localized and stable, which facilitates both contact detection and force estimation. This is reflected in stronger performance on objects with well-defined grasp structures and predictable contact zones; the best-performing objects are those that force the human hand into a specific, highly constrained pose. For example, the notebook achieves the lowest force errors overall (hand force MAE: 2.73 N, object force MAE: 3.15 N). A notebook is a simple, flat, rigid plane. When a hand presses against it, the force is distributed evenly and the rigid 3D template perfectly matches reality. Similarly, scissors yield the strongest object contact localization (P: 0.13, R: 0.79, F1: 0.21, IoU: 0.13), reflecting its highly constrained finger placement (Table 7). Conversely, objects like the ketchup bottle or a generic box can be grabbed from almost any angle, with the whole hand or just the fingertips. This high variance in interaction styles makes predicting exact contact patches and force vectors much harder.

In contrast, several objects in the H2O dataset are less rigid or partially deformable, such as milk carton or chips bag. Such objects can deform under grasping forces, producing more distributed and less consistent contact regions compared to rigid objects. As a result, the rigid-body assumption underlying the force simulation introduces a mismatch between predicted contact-force patterns

Table 7: Per-object results on ARCTIC s05.

	Hand Force [N]		Object Force [N]		Hand Contact				Object Contact			
Object	MAE	RMSE	MAE	RMSE	P	R	F1	IoU	P	R	F1	IoU
Box	2.94	3.91	6.64	9.14	.074	.288	.106	.060	.011	.494	.021	.011
Capsule machine	4.97	5.92	3.61	4.25	.138	.703	.208	.122	.019	.681	.036	.019
Espresso machine	3.69	4.76	5.12	6.17	.152	.513	.186	.107	.013	.589	.025	.013
Ketchup	4.21	5.29	4.22	4.58	.259	.526	.317	.199	.025	.254	.044	.023
Laptop	3.64	4.70	5.48	6.72	.131	.376	.168	.098	.014	.444	.026	.013
Microwave	3.56	4.44	5.74	6.84	.114	.447	.162	.093	.015	.510	.028	.015
Mixer	4.24	5.30	4.69	5.52	.125	.586	.191	.111	.017	.396	.032	.016
Notebook	2.73	3.64	3.15	3.96	.144	.384	.191	.113	.040	.492	.073	.039
Phone	4.49	5.50	4.04	4.37	.137	.611	.210	.123	.049	.463	.086	.047
Scissors	6.59	8.01	2.11	2.40	.140	.513	.202	.118	.127	.789	.214	.126
Waffleiron	3.83	4.93	4.46	5.28	.129	.505	.186	.107	.023	.523	.043	.022

Table 8: Per-object results on H2O (out-of-distribution for EgoPHI).

	Hand Force [N]		Object Force [N]		Hand Contact				Object Contact			
Object	MAE	RMSE	MAE	RMSE	P	R	F1	IoU	P	R	F1	IoU
Milk carton	3.20	4.00	4.93	5.66	.057	.378	.091	.052	.013	.159	.023	.012
Book	7.92	8.99	3.49	3.81	.071	.360	.105	.059	.018	.146	.026	.015
Coffee box	4.57	6.03	3.62	3.80	.067	.383	.105	.060	.011	.225	.021	.011
Tea box	5.67	7.10	3.68	3.84	.039	.367	.066	.036	.014	.224	.025	.013
Lotion bottle	7.07	9.49	3.58	3.72	.038	.479	.067	.037	.018	.225	.030	.017
Ovomaltine box	4.61	5.98	3.44	3.61	.083	.512	.133	.075	.024	.289	.042	.022
Deodorant	6.31	8.40	3.36	3.49	.080	.514	.128	.071	.061	.314	.090	.050
Chips bag	4.32	5.85	3.34	3.57	.064	.512	.107	.059	.019	.159	.031	.017

and the actual interactions. This effect is reflected in lower recall and IoU values for several H2O objects, particularly for the milk carton (P: 0.013, R: 0.16, F1: 0.02, IoU: 0.012) and chips bag (P: 0.019, R: 0.16, F1: 0.03, IoU: 0.017). Rigid objects, however, achieve relatively better performance; the best results are observed for the deodorant, which reaches P: 0.06, R: 0.31, F1: 0.09, and IoU: 0.05 (Table 8). Ultimately, for partially deformable objects in H2O, we observe an accumulated error stemming from both the out of distribution (OOD) scenario and the rigidity mismatch. Meanwhile, for rigid objects, the better overall performance suggests that the error can be attributed primarily to the OOD nature of the data.

D Additional qualitative results

Below, we provide additional qualitative comparisons of our method (EgoPHI) and the ablated version of our method (EgoPHI without Iterative Refinement Module-IRM) against the HACO baseline, on the ARCTIC test split (Figure 7) and the H2O test split (Figure 8).

We further present additional qualitative results on our real-world setup, using the instrumented cube and cylinder objects introduced in Section 5.2. Figure 9 shows ground-truth and EgoPHI force predictions overlaid on the egocentric RGB images, illustrating that our method produces localized and physically plausible force estimates despite residual object pose errors.

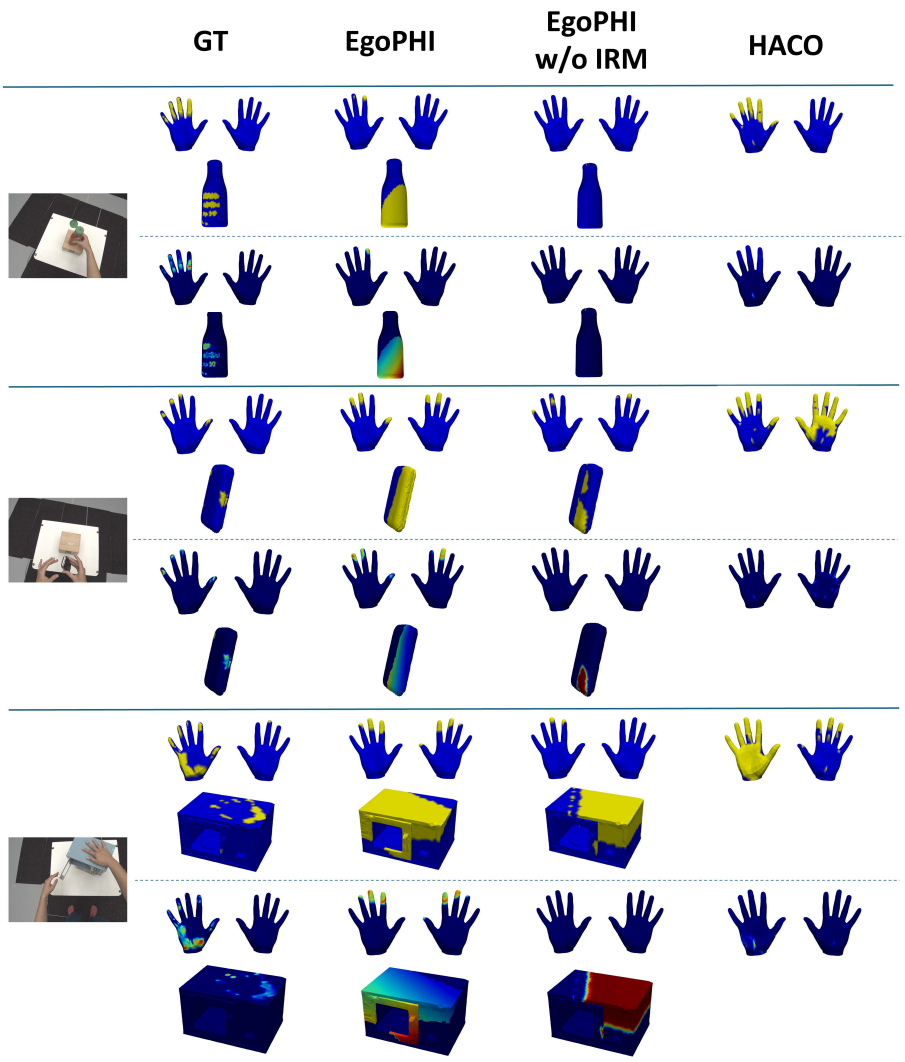

Fig. 7: Additional qualitative results on the **ARCTIC dataset** (participant s05). We compare our model (EgoPHI) with the HACO baseline and with an ablated version of our method (EgoPHI without the Iterative Refinement Module–IRM) for contact and force prediction. The comparison with HACO is limited to hand contact estimation, as the method does not provide object contact predictions. Our full model produces more localized and physically consistent contact estimates, while HACO tends to overestimate contact regions. Furthermore, we note our steadily better force intensity predictions.

Fig. 8: Additional qualitative results on the **H2O dataset** (participant subject4_ego). We compare our model (EgoPHI) with the HACO baseline and with an ablated version of our method (EgoPHI without the Iterative Refinement Module–IRM) for contact and force prediction. Consistent with the ARCTIC results in Figure 7, our approach provides more localized estimates, in contrast to HACO, which tends to overestimate the contact regions. Furthermore, we note our steadily better force intensity predictions.

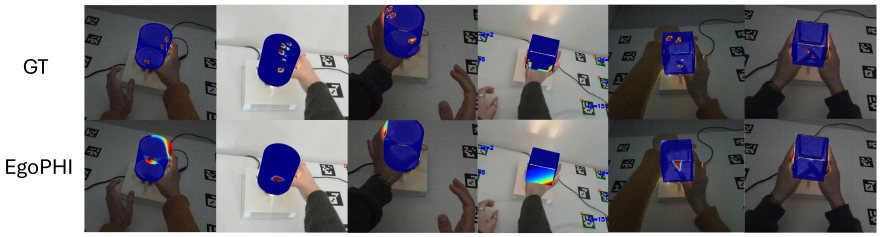

Fig. 9: Additional in-the-wild qualitative results on our real-world cube and cylinder objects, with GT and EgoPHI predictions overlaid on the egocentric RGB frame. Predicted forces remain well-localized and follow the observed interactions despite residual object pose errors.