# RLHND: VIDEO FOUNDATION MODELS AS PHYSICALLY GROUNDED HAND TRACKERS FOR ROBOT LEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Recently, approaches that leverage human video datasets for robot policy training have become increasingly prevalent. However, most existing hand trackers regress pose from cropped frames with limited priors on hand motion and object interaction, resulting in inaccurate and physically inconsistent estimates. Moreover, the lack of physical cues, *e.g.*, contact and force, limits the use of human videos for robot policy training. To this end, we propose RLHND, a video foundation model-based hand tracking model that jointly estimates hand pose and realistic tactile information from monocular egocentric videos. RLHND turns the pre-trained Cosmos 3 video diffusion backbone into a deterministic clip-level feature extractor via clean-latent conditioning, carrying its learned priors on hand motion and hand-object interaction into tracking. For pose estimation, RLHND (i) predicts hand poses with anatomically plausible joint angles and (ii) enables optional conditioning on the shape parameter to maintain consistent hand shape within the same video and even across videos recorded by the same actor. For tactile estimation, a separate tactile expert stream, trained with the pose stream frozen, predicts dense contact and force over the hand surface. We further adopt LBS-based feature spreading to enable vertex-wise feature extraction without costly per-vertex attention. RLHND achieves state-of-the-art performance across various benchmark datasets for pose estimation, while also achieving state-of-the-art performance in contact and force estimation. Moreover, we demonstrate the utility of RLHND for robot learning through retargeting results and real-world robot experiments. The code will be publicly available at https://anonymous4898.github.io/rlhnd/.

## 1 INTRODUCTION

Along with the rapid advancement of vision-language models (OpenAI, 2023; Liu et al., 2023; Gemini Team, 2023; Bai et al., 2025), robotics foundation models have also made significant progress in recent years (Brohan et al., 2023; Octo Model Team et al., 2024; Kim et al., 2024; Black et al., 2024; Physical Intelligence et al., 2025; Bjorck et al., 2025; Kim et al., 2026b). However, robotics datasets remain substantially more limited in scale than vision-language datasets and are considerably more labor-intensive to collect. Recently, researchers have recognized the potential of human egocentric videos, which are relatively easy to collect and can capture a wide range of tasks that are difficult to demonstrate through teleoperation. Consequently, human datasets (Grauman et al., 2022; Banerjee et al., 2024; Hoque et al., 2026) have been increasingly leveraged to train robotics foundation models (Ye et al., 2025a; Kareer et al., 2024; Yang et al., 2025; Luo et al., 2025; Kim et al., 2026b).

However, utilizing human videos for robot training requires bridging two major gaps between human videos and real-world robot episodes. The first is the *action-space discrepancy* between humans and robots. Existing works address this discrepancy by retargeting human hand motion to a specific robot embodiment (Handa et al., 2020; Qin et al., 2023; Shaw et al., 2022; Lepert et al., 2025), defining a co-action space that encompasses both human and robot joint actions (Kareer et al., 2024; Luo et al., 2025; Li et al., 2025), or adopting an end-effector-based action space consisting of wrist poses and keypoints (Haldar & Pinto, 2025; Liu et al., 2025; Kim et al., 2026a; Cai et al., 2025).



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

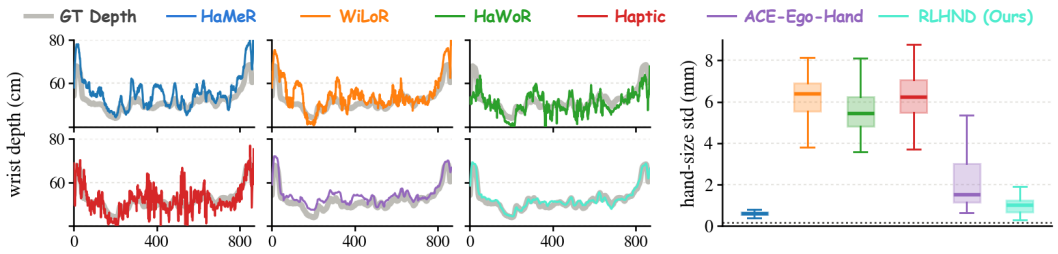



Figure 1: **3D instability of hand motion reconstruction methods.** *Left:* Wrist depth over 850 frames from an ARCTIC test-set clip. Baselines exhibit severe oscillation or drift relative to the ground-truth depth. *Right:* Per-video standard deviation of the estimated hand size. Even within a single video, existing hand motion estimators exhibit ∼8 mm variation in estimated hand size.

Regardless of the choice of action space, these approaches require human action labels that provide 3D hand joint locations. However, existing hand motion reconstruction methods are often optimized primarily for 2D reprojection or image-space alignment, which can make their 3D predictions unreliable for robot action labeling. As shown in Figure 1, existing hand reconstructions exhibit severe jitter along the depth axis (left), while the estimated size of the same hand can fluctuate substantially across frames (right). The latter arises from the inherent scale-depth ambiguity under perspective projection, where depth errors can be absorbed into the estimated hand shape. Such temporally inconsistent and physically implausible 3D trajectories are problematic for robot learning, where small errors in motion estimation can lead to substantially different or infeasible actions.

The second gap is *missing physical information*. Recent vision-language-action models increasingly incorporate tactile input alongside vision (Zhang et al., 2026a; Niu et al., 2026; Chen et al., 2026; Zhang et al., 2026b), as tactile information can distinguish a successful grasp from merely hovering over an object. However, human datasets generally cannot provide tactile information, except for those collected with pressure boards (Grady et al., 2022) or wearable tactile gloves (Song et al., 2025). To this end, several vision-based contact and force estimation models (Jeon et al., 2026; Jung & Lee, 2025) have emerged, but these remain underexplored compared to hand pose estimation.

In this paper, we present RLHND, which produces both metrically accurate, anatomically realistic 3D bimanual hand motion and dense contact and force estimates over the hand surface from a monocular human video. Since recent works (Wang et al., 2026; Liu et al., 2026) show that pre-trained video diffusion models internalize strong priors on hand-object interactions, RLHND uses Cosmos 3 Nano (NVIDIA, 2026) as its video feature encoder. Unlike Liu et al. (2026), which feeds clean videos directly into a denoising backbone, we instead route them through the conditioning-frame interface of Cosmos 3, which is designed to receive clean context frames during pre-training. This design keeps our fine-tuning aligned with the pre-trained interface of the backbone.

On top of the Cosmos 3 encoder, we make three design choices motivated by real-world deployment. *(i) Shape caching.* RLHND caches the shape parameter and uses a single value throughout the entire video. This not only eliminates the hand-size fluctuations shown in Figure 1, but also allows direct injection of a pre-calibrated $\boldsymbol{\beta}$, which is particularly useful in real-world data collection settings where the hand shape of the operator is known. *(ii) Anatomical pose parameterization.* Rather than regressing the 45 rotational DoF of MANO, we align each joint with its anatomical axes and freeze the axes about which a human finger cannot rotate, regressing 29 DoF. *(iii) A decoupled tactile expert.* Contact and force are predicted by a second expert stream that is trained with the pose stream frozen, so tactile supervision never perturbs the pose. To obtain vertex-wise features, we introduce LBS-based feature spreading, which uses the linear blend skinning weights of MANO to spread bone-level features to the 778 MANO vertices, avoiding the need for costly vertex-wise attention.

For motion reconstruction, we evaluate RLHND on HOT3D and ARCTIC, as well as EgoDex, which is held out during training. RLHND consistently outperforms all baselines across the three benchmark datasets. For tactile estimation, RLHND achieves the best performance in both contact and force estimation among various benchmarks. Finally, we demonstrate the utility of RLHND for robot learning by retargeting its human demonstrations to five dexterous hands and training Dexterous Point Policy (Kim et al., 2026a). RLHND produces more accurate and less jittery joint actions than the baselines, and training with keypoints from RLHND improves Dexterous Point Policy notably.

**Contributions.** We highlight the main contributions of our work as follows:

- We extend video foundation model-based hand tracking beyond pose estimation: RLHND leverages the clean-latent conditioning interface of Cosmos 3 to extract rich spatiotemporal representations from monocular video and jointly estimates metrically accurate bimanual motion with dense contact and force over the hand surface.
- We achieve state-of-the-art performance in both hand motion reconstruction and tactile estimation, enabled by physically grounded design choices including cached hand shape, anatomically constrained pose, and a decoupled tactile expert with LBS-based feature spreading.
- We demonstrate the utility of RLHND for robot learning: its outputs yield accurate and smooth retargeted commands across five dexterous hands and enable training Dexterous Point Policy from human demonstrations with automatically labeled contact.

## 2 Related Work

**Hand motion reconstruction.** Hand motion reconstruction methods have been developed to improve both pose accuracy and temporal consistency. The primary approach uses image-based methods (Pavlakos et al., 2024; Potamias et al., 2024), which regress MANO (Romero et al., 2017) parameters from a cropped image input by a hand detection model (Potamias et al., 2024). Since they treat every frame independently, they cannot consider any temporal consistency among frames. In contrast, video-based methods (Zhang et al., 2025; Ye et al., 2025b; Xu et al., 2026) add temporal context either by fusing crop features across frames (Ye et al., 2025b) or by tracking the hand and the camera in a world frame (Zhang et al., 2025; Xu et al., 2026). Despite this, their reliance on per-frame crops causes them to inherit detector box jitter and remain only weakly consistent over time. Most recently, models built on video foundation backbones take the entire clip as input and infer over both hands jointly (Wang et al., 2026; Liu et al., 2026), which notably improves temporal coherence. Even these models, however, regress a hand shape per window, leading to temporal variation in the estimated hand size, while the scale-depth ambiguity of perspective projection allows depth errors to manifest as variations in the estimated shape. RLHND keeps the clip-level formulation but fixes a single shape per video and constrains the pose to the anatomical degrees of freedom, which removes both failure modes.

**Hand contact and force estimation.** Unlike motion reconstruction, force labels require specialized hardware, *e.g.*, pressure boards (Grady et al., 2022) or tactile gloves (Song et al., 2025). PressureVisionDB (Grady et al., 2022) records RGB video of a bare hand pressing a planar capacitive pad, providing dense and finely resolved pressure maps that are spatially limited to the palm-pad contact region. OpenTouch (Song et al., 2025) instead moves the sensor onto the hand, pairing in-the-wild egocentric video with a $16\times16$ taxel grid on the palmar side of a wearable glove, enabling force measurements during natural manipulation of diverse objects. EgoPressure (Zhao et al., 2025) extends pressure sensing to egocentric video with multi-view MANO fits, while EgoTactile (Zeng et al., 2026) provides full-hand glove pressure on 63 everyday objects and a bare-hand subset. However, differences in sensor calibration and hand-surface correspondence make it difficult to pool these datasets without converting their labels to a common representation. For contact estimation, tactile hardware is not required: HACO (Jung & Lee, 2025) generates hand contact labels from the distances between hand and object meshes in HOI datasets and uses them to train contact estimation models. For force estimation, however, such geometric supervision is insufficient, requiring datasets with physical force measurements. HOPE (Jeon et al., 2026) unifies contact and force labels from pressure boards and tactile gloves on the MANO surface, enabling joint learning of contact and force estimation. RLHND follows the unified label format of HOPE while introducing a decoupled tactile expert, achieving improved tactile estimation performance.

## 3 Method

Given a monocular video, RLHND estimates both hand pose and tactile signals. The video is processed as a sequence of $W$ clips $\mathcal{C}_1, \ldots, \mathcal{C}_W$, each holding $T$ consecutive frames $\mathcal{C}_w = (\mathbf{I}_t)_{t=1}^{T}$. The pose output consists of MANO (Romero et al., 2017) parameters, *i.e.*, shape $\hat{\boldsymbol{\beta}}^w \in \mathbb{R}^{10}$ and

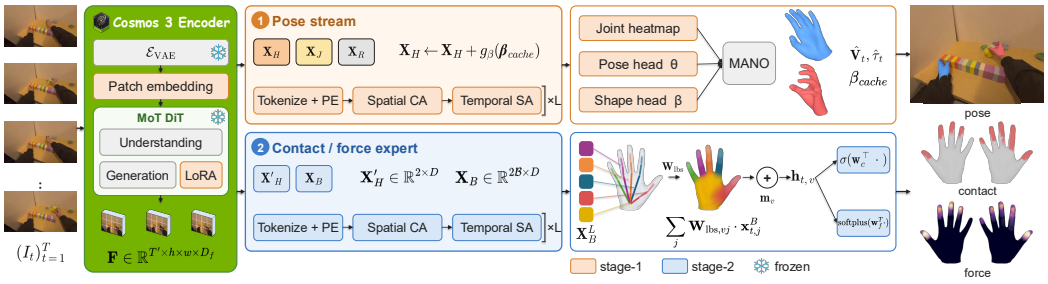



Figure 2: **RLHND overview.** A pre-trained Cosmos 3 backbone receives the clip through its conditioning-frame interface and is adapted with a trainable patch embedding and LoRA adapters. The pose expert stream, trained in stage-1, yields MANO parameters and a metric camera-space translation while the tactile expert stream is trained in stage-2 with the pose stream frozen.

anatomically constrained articulation $\hat{\boldsymbol{\theta}}_t$ with the global wrist rotation, along with metric camera-space translation $\hat{\boldsymbol{\tau}}_t \in \mathbb{R}^3$ and per-frame hand presence and visibility. The tactile output consists of per-vertex tactile signals: contact probability $\hat{\mathbf{c}}_t \in [0,1]^{|V|}$ and force $\hat{\mathbf{f}}_t \in \mathbb{R}^{|V|}_{\geq 0}$, where $|V| = 778$ is the number of MANO vertices. The RLHND architecture consists of three parts: the *clean-latent encoder* converts a pre-trained video-diffusion backbone into a deterministic video feature extractor, the *pose expert stream* predicts pose, shape, and translation in MANO parameter form from the video features, and the *tactile expert stream* estimates contact and force from the same video features with its own weights, trained with the pose stream frozen, thereby decoupling tactile prediction from pose prediction. Figure 2 illustrates the overall process of the proposed method.

### 3.1 Clean-Latent Video Diffusion Encoder

Since video diffusion models are known to internalize strong priors over hand motion and object interactions (Wang et al., 2026), we use the pre-trained encoder of Cosmos 3 (NVIDIA, 2026). Its video tokenizer, *i.e.*, Wan2.2 VAE (Wan Team, 2025), encodes each clip into a clean latent $\mathbf{z} = \mathcal{E}_{\text{VAE}}(\mathcal{C}_w)$. We fine-tune the patch embedding and generation pathway with LoRA (Hu et al., 2022), yielding a deterministic spatiotemporal feature grid $\mathbf{F} \in \mathbb{R}^{T' \times h \times w \times D_f}$.

Our use of the pre-trained VAE differs from prior work (Wang et al., 2026; Liu et al., 2026), which feeds clean videos to Wan2.2 through its denoising interface by setting the out-of-distribution noise level $\sigma = 0$. Instead, we use the conditioning interface of Cosmos 3 for clean context frames, keeping our input scheme aligned with pre-training. We also feed an empty prompt, since the video stream is already conditioned on the clean frames, whereas Liu et al. (2026) uses a fixed caption.

### 3.2 Pose Expert Stream

Following ACE-Ego-Hand (Liu et al., 2026), we tokenize the $\mathbf{F}$ with spatial positional and Fourier-encoded ray embeddings, and decode the resulting tokens with alternating spatial cross-attention and bidirectional temporal self-attention. We depart from this design in two aspects: how the two hand tokens $\mathbf{X}_H$ are formed with the shape conditioning, and how hand articulation is parameterized.

**Shape conditioning.** We construct the two hand tokens $\mathbf{X}_H$ from the cached MANO shape parameter $\boldsymbol{\beta}_{\text{cache}}$, which enforces consistent hand geometry throughout the video and can also be obtained from a one-time calibration of the actor's hand. Unlike ACE-Ego-Hand (Liu et al., 2026), which initializes its hand tokens from fixed learnable parameters, our model can directly incorporate this pre-measured shape. Let $v^w$ denote the fraction of frames in $\mathcal{C}_w$ in which the hand is predicted present. The hand tokens are built from a learnable query $\mathbf{q}_H$ and a zero-initialized MLP $g_\beta$:

$$\mathbf{X}_H^w = \mathbf{q}_H + g_\beta(\boldsymbol{\beta}_{\text{cache}}), \qquad \boldsymbol{\beta}_{\text{cache}} = \begin{cases} \hat{\boldsymbol{\beta}}^{w^*}, & w^* = \min\{\, w' : v^{w'} > T_v \,\} \text{ exists}, \\ \hat{\boldsymbol{\beta}}^{w}, & \text{otherwise}, \end{cases} \tag{1}$$

The shape estimated from the first sufficiently visible clip is shared across the video, and earlier clips are decoded again once it becomes available. The zero initialization of $g_\beta$ gradually introduces

the cached shape during training. We use the ground-truth $\boldsymbol{\beta}$ with probability $0.5$ and disable the shape-head loss for these samples. Since the public benchmarks do not provide pre-calibrated hand shapes, we estimate $\boldsymbol{\beta}$ from the first sufficiently visible clip and reuse it throughout each video.

**Constrained pose parameterization.** For more faithful tracking of the ground-truth hand articulation, we anatomically constrain the DoF of MANO (Yang et al., 2021). While MANO allows 3-axis rotation at each of its 15 joints, resulting in 45 DoF, the human hand has only 21 DoF. We therefore align the rotation axes with the anatomical structure and disable infeasible axes. Following the anatomy-aware kinematics of Yang et al. (2021), the rotation $\mathrm{R}_j \in SO(3)$ of joint $j$ is factorized about its anatomically defined twist, spread, and bend axes $(\mathbf{a}^{\text{twist}}j, \mathbf{a}_j^{\text{spread}}, \mathbf{a}_j^{\text{bend}})$:

$$\mathrm{R}_j \;=\; \exp\!\big(\varphi_j^{\text{twist}}\,[\mathbf{a}_j^{\text{twist}}]_\times\big)\,\exp\!\big(\varphi_j^{\text{spread}}\,[\mathbf{a}_j^{\text{spread}}]_\times\big)\,\exp\!\big(\varphi_j^{\text{bend}}\,[\mathbf{a}_j^{\text{bend}}]_\times\big), \tag{2}$$

where $[\cdot]_\times$ is the skew-symmetric operator. Then, the feasible set $\mathcal{C}$ is defined as below:

$$\mathcal{C} \;=\; \big\{\boldsymbol{\Theta} \;:\; \varphi_j^{\text{twist}} = \varphi_j^{\text{spread}} = 0 \;\; \forall j \in \text{PIP} \cup \text{DIP}\big\}. \tag{3}$$

We leave the thumb unconstrained because its carpometacarpal joint is a saddle joint with non-orthogonal, non-fixed axes, for which the twist-spread-bend decomposition does not isolate an infeasible axis. This leaves 29 DoF after constraining the remaining joints. To train the model with this representation, we obtain constrained 29-DoF labels by fitting the existing 45-DoF MANO labels to the feasible set $\mathcal{C}$ via inverse kinematics (Zhou et al., 2020; Li et al., 2021):

$$\widehat{\boldsymbol{\Theta}}_c \;=\; \underset{\boldsymbol{\Theta}_c \in \mathcal{C}}{\arg\min}\; \big\|\mathrm{J}(\boldsymbol{\Theta}_c;\boldsymbol{\beta}) - \mathrm{J}(\boldsymbol{\Theta};\boldsymbol{\beta})\big\|_2^2, \tag{4}$$

where $\mathrm{J}(\cdot\,;\boldsymbol{\beta})$ returns the forward-kinematics joint positions under the shape $\boldsymbol{\beta}$ carried by the label. We solve Eq. (4) in the anatomy-aligned Euler coordinates using damped Gauss-Newton steps:

$$\boldsymbol{\varphi} \leftarrow \boldsymbol{\varphi} + \big(\mathbf{G}^\top\mathbf{G} + \lambda\mathbf{I}\big)^{-1}\mathbf{G}^\top\boldsymbol{\epsilon}, \tag{5}$$

where $\boldsymbol{\varphi}$ contains the optimized finger DoF, $\mathbf{G}$ is the Jacobian of the joint positions with respect to $\boldsymbol{\varphi}$, and $\boldsymbol{\epsilon}$ stacks the joint and fingertip residuals. Further details are provided in Appendix A.2.

**Output Heads.** The readout heads decode the hand tokens into pose, shape, depth, and presence predictions. Metric camera-space translation is then recovered using the ray-based mixed-PnP solver of ACE-Ego-Hand (Liu et al., 2026). We refer to Appendix A.3 for details.

### 3.3 Tactile Expert Stream

Tactile supervision is far scarcer than pose supervision, as pressure-labeled datasets often provide constrained or no pose labels (Grady et al., 2022; 2024; Song et al., 2025). We therefore decouple tactile estimation from pose estimation, preventing such training from affecting pose prediction.

**Bone tokens.** Attending over all $|V|$ vertices per frame is computationally prohibitive, so we perform tactile estimation at the hand-skeleton level. Let $\mathcal{B}$ denote the $|\mathcal{B}| = 16$ joints of the MANO kinematic tree. Each joint drives one bone of the skeleton through the linear-blend-skinning weights, so we attach one *bone token* per joint and hand, yielding $\mathbf{X}_B \in \mathbb{R}^{2|\mathcal{B}|\times D}$ with $|\mathcal{B}| \ll |V|$. Each bone token represents the features of the vertices skinned to its bone.

**LBS-based vertex readout.** We keep temporal reasoning at the bone level and expand the features to vertices only at the output using the MANO Linear-Blend-Skinning (LBS) weight $\mathbf{W}_{\mathbf{lbs}} \in \mathbb{R}^{|V|\times|\mathcal{B}|}$, the fixed matrix from MANO that maps features to vertices. Combined with a learned vertex embedding $\mathbf{m}_v$, the per-vertex tactile feature is computed as follows:

$$\mathbf{h}_{t,v} \;=\; \sum_{j\in\mathcal{B}} \mathbf{W}_{\mathbf{lbs},vj}\,\big[\mathbf{X}_B^L\big]_{t,j} \;+\; \mathbf{m}_v, \tag{6}$$

where $[\mathbf{X}_B^L]_{t,j}$ is the token of node $j$ at frame $t$ in $\mathbf{X}_B^L$, obtained by interpolation from $T'$ to $T$.

Two small MLP heads read the per-vertex contact logit and the force-distribution logit from $\mathbf{h}_{t,v}$, and the per-hand total force $\hat{F}_t$ is read from the expert's hand token through a softplus; the per-vertex force is the total distributed over the vertices, $\hat{f}_{t,v} = \hat{F}_t\,\mathrm{softmax}_v(\cdot)$.

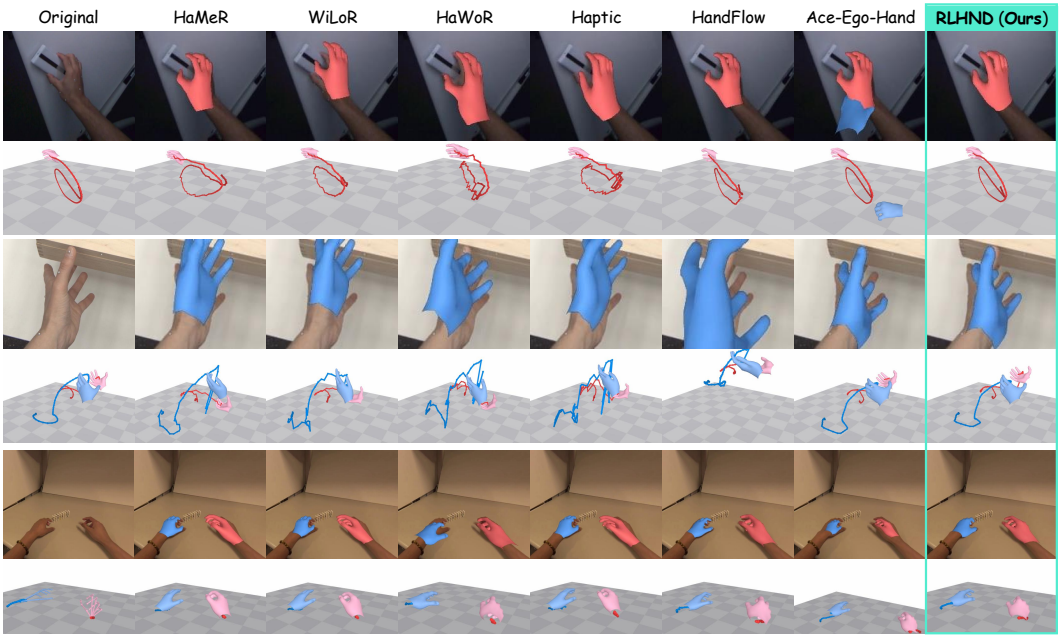



Figure 3: **Qualitative comparison of hand motion reconstruction.** We visualize the reconstruction results from HOT3D (top), ARCTIC (middle), and EgoDex (bottom), with mesh overlays and trajectory visualizations. Extensive visualization results can be found on the project page.

### 3.4 Training

At stage-1 training, we only train the pose expert stream with the following:

$$\mathcal{L}_{\text{pose}} = \lambda_{\text{rot}}\mathcal{L}_{\text{rot}} + \lambda_{\beta}\mathcal{L}_{\beta} + \lambda_{\text{3D}}\mathcal{L}_{\text{3D}} + \lambda_{\text{2D}}\mathcal{L}_{\text{2D}} + \lambda_{\tau}\mathcal{L}_{\tau} + \lambda_{\text{pres}}\mathcal{L}_{\text{pres}} + \lambda_{\text{tmp}}\mathcal{L}_{\text{tmp}} + \lambda_{\text{ray}}\mathcal{L}_{\text{ray}}. \quad (7)$$

At stage-2 training, the pose expert stream is frozen, and only the tactile expert stream is trained:

$$\mathcal{L}_{\text{tactile}} = \lambda_c \,\text{BCE}_w(\hat{c}_{t,v},\, c_{t,v}) + \lambda_F \,\tfrac{1}{|V|}\Big|\textstyle\sum_v \hat{f}_{t,v} - \sum_v f_{t,v}\Big| + \lambda_\pi \,\text{CE}\Big(\hat{\mathbf{f}}_t / \textstyle\sum_v \hat{f}_{t,v},\; \mathbf{f}_t / \sum_v f_{t,v}\Big). \quad (8)$$

Since contact vertices are far rarer than non-contact vertices, we use a positively re-weighted BCE, *i.e.*, $\text{BCE}_w$. The force is supervised through its per-hand total, normalized by the number of vertices, and through its distribution over the vertices, weighted by $\lambda_F$ and $\lambda_\pi$, respectively. We elucidate the details of each loss term in Appendix A.4.

## 4 Experiments

**Setup.** Stage-1 is trained on a weighted mixture of egocentric and exocentric hand-video datasets, and stage-2 on force-labeled glove and pressure-pad recordings together with contact-only labels derived from hand–object meshes, which supervise only the contact head. The datasets and label generation are described in Appendix A.2, and the schedules and hyperparameters in Appendix A.5. Motion reconstruction is evaluated on the held-out HOT3D and ARCTIC egocentric splits and, zero-shot for our method and every baseline, on EgoDex. Tactile estimation is evaluated on the OpenTouch and PressureVisionDB (PVDB) test splits of HOPE (Jeon et al., 2026) and, for contact only, on HOT3D, DexYCB, and ARCTIC with contact labels derived from the hand–object meshes. All baselines use their official weights and inference code. Metric definitions are given in Appendix A.4.

### 4.1 Hand Pose Estimation

In Figure 3, crop-based models exhibit noisy hand trajectories and incorrect mesh overlays, while ACE-Ego-Hand performs better but still shows erroneous hand detection, *e.g.*, the left hand in the first

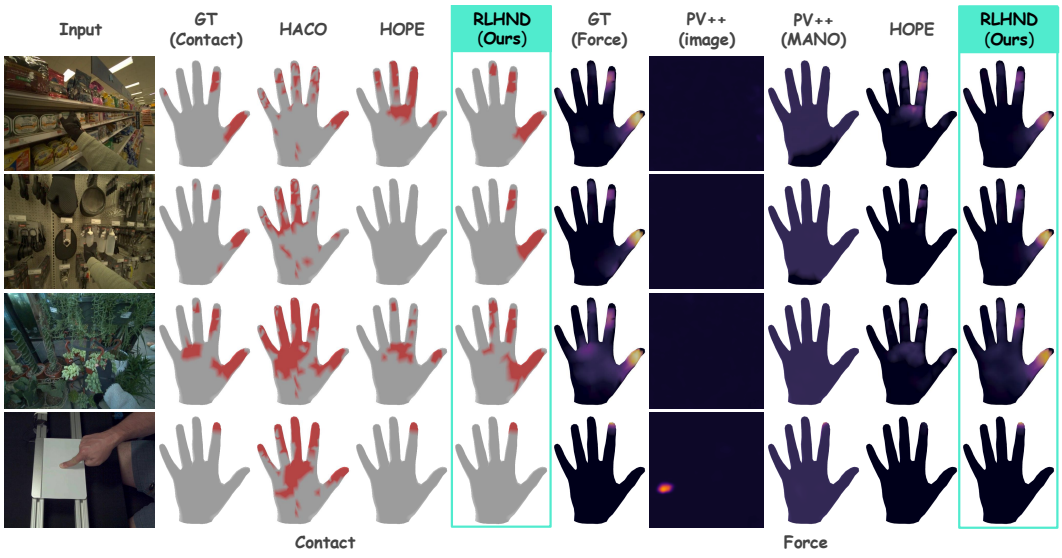



Figure 4: **Qualitative comparison of contact and force estimation.** We visualize the *contact* (left) and *force* (right) estimation from OpenTouch (first to third row) and PVDB (fourth row).

| Test set | Method | $F_{Acc}$ ↑ | MPJPE-p ↓ | PA-p ↓ | MPJPE$^{+OOS}$ ↓ | EPE2D-p ↓ | Jitter ↓ | $\sigma_{shape}$ ↓ |
|---|---|---|---|---|---|---|---|---|
| HOT3D | HaMeR (Pavlakos et al., 2024) | 0.743 | 65.53 | 11.42 | 66.85 | 130.33 | 25.40 | 1.17 |
| | WiLoR (Potamias et al., 2024) | 0.743 | 44.89 | 9.71 | 46.23 | 134.89 | 22.17 | 1.90 |
| | HaWoR (Zhang et al., 2025) | 0.743 | 41.53 | 10.15 | 42.99 | 126.10 | 26.13 | 7.37 |
| | HaPTIC (Ye et al., 2025b) | 0.743 | 62.42 | 11.26 | 63.70 | 130.40 | 38.27 | 3.20 |
| | HandFlow (Xu et al., 2026) | 0.729 | 39.90 | 9.50 | 24.87 | 127.85 | 16.99 | 3.68 |
| | ACE-Ego-Hand (Liu et al., 2026) | 0.928 | 24.41 | 8.60 | 20.30 | 38.09 | 6.48 | 1.22 |
| | **RLHND (Ours)** | **0.996** | 13.01 | 6.72 | 13.08 | **6.84** | **4.38** | 1.30 |
| | **RLHND (Ours)** ($\beta$-cache) | **0.996** | **13.00** | **6.68** | **13.07** | 6.89 | 4.39 | **0.00** |
| ARCTIC ego | HaMeR (Pavlakos et al., 2024) | 0.887 | 31.76 | 9.72 | 52.99 | 86.30 | 21.73 | 0.52 |
| | WiLoR (Potamias et al., 2024) | 0.887 | 24.05 | 7.49 | 46.33 | 87.66 | 25.92 | 6.15 |
| | HaWoR (Zhang et al., 2025) | 0.886 | 31.72 | 9.55 | 52.88 | 82.72 | 29.34 | 5.44 |
| | HaPTIC (Ye et al., 2025b) | 0.886 | 24.72 | 7.92 | 46.88 | 84.05 | 51.09 | 6.19 |
| | HandFlow (Xu et al., 2026) | 0.918 | 40.03 | 9.91 | 47.47 | 74.09 | 13.47 | 5.12 |
| | ACE-Ego-Hand (Liu et al., 2026) | **0.994** | 14.86 | 7.50 | 16.86 | **11.76** | **3.52** | 1.91 |
| | **RLHND (Ours)** | 0.993 | **13.36** | **6.42** | **15.40** | 12.19 | 4.19 | 0.96 |
| | **RLHND (Ours)** ($\beta$-cache) | 0.993 | 13.44 | 6.44 | 15.46 | 12.28 | 4.17 | **0.00** |
| EgoDex | HaMeR (Pavlakos et al., 2024) | 0.970 | 45.97 | 11.79 | 45.90 | 66.31 | 15.48 | 0.43 |
| | WiLoR (Potamias et al., 2024) | 0.970 | 42.57 | 10.87 | 42.50 | 67.68 | 13.05 | 0.22 |
| | HaWoR (Zhang et al., 2025) | 0.970 | 27.76 | 11.22 | 27.73 | 54.20 | 16.94 | 4.58 |
| | HaPTIC (Ye et al., 2025b) | 0.970 | 40.92 | 11.29 | 40.85 | 65.38 | 25.81 | 2.33 |
| | HandFlow (Xu et al., 2026) | 0.978 | 24.60 | 11.45 | 22.34 | 41.44 | 10.89 | 3.57 |
| | ACE-Ego-Hand (Liu et al., 2026) | 0.999 | 20.95 | 10.99 | 20.85 | 22.26 | 3.01 | 0.59 |
| | **RLHND (Ours)** | **1.000** | **19.90** | **10.12** | **19.89** | **20.72** | **2.93** | 0.71 |
| | **RLHND (Ours)** ($\beta$-cache) | **1.000** | 19.93 | 10.14 | 19.91 | **20.72** | 2.94 | **0.00** |

Table 1: **Quantitative comparison of motion reconstruction.** RLHND outperforms baselines on nearly every metric on every benchmark consistently. We denote **best** and second best values.

row, unrealistic motion, *e.g.*, twisted fingers in the second row, and inaccurate hand scale, *e.g.*, the third-row overlay. In contrast, ours produces robust mesh overlays and smooth hand trajectories.

Table 1 reports quantitative comparisons with per-window $\beta$ estimation and $\beta$-caching, which estimates $\beta$ from the first output window and reuses it throughout the video. Ours consistently outperforms across all test sets, including zero-shot EgoDex, with higher $F_{Acc}$ than the external hand detectors used by HaMeR, WiLoR, HaWoR, and HaPTIC, as well as strong 3D and 2D projection metrics. Although $\beta$-caching removes $\sigma_{shape}$, it does not improve jitter, as analyzed in Appendix B.2.

| Method | Contact (vertex-level) | | | | | | | | Force (kPa) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | OpenTouch | | DexYCB | | HOT3D | | ARCTIC | | OpenTouch | | PVDB | |
| | F1 ↑ | AUROC ↑ | F1 ↑ | AUROC ↑ | F1 ↑ | AUROC ↑ | F1 ↑ | AUROC ↑ | MAE ↓ | RMSE ↓ | MAE ↓ | RMSE ↓ |
| PressureVision (Grady et al., 2022) | – | – | – | – | – | – | – | – | 1.930 | 6.190 | 0.471 | 3.726 |
| PressureVision++ (Grady et al., 2024) | – | – | – | – | – | – | – | – | 1.920 | 6.200 | 0.248 | 3.025 |
| HACO (Jung & Lee, 2025) | 0.373 | 0.541 | 0.543 | 0.883 | 0.244 | 0.749 | 0.580 | 0.907 | – | – | – | – |
| HOPE (Jeon et al., 2026) | 0.663 | 0.894 | 0.506 | 0.868 | 0.197 | 0.762 | 0.591 | 0.914 | 1.781 | 5.388 | **0.236** | 2.091 |
| **RLHND (Ours)** | **0.696** | **0.980** | **0.572** | **0.915** | **0.589** | **0.959** | **0.602** | **0.935** | **0.489** | **2.508** | 0.274 | **2.066** |

Table 2: **Quantitative comparison of contact and force estimation.** RLHND outperforms baselines on nearly every benchmark consistently. We denote **best** and second best values.

| Method | Sharpa Wave (22 DoF) | | WUJI v2 (20 DoF) | | Shadow (24 DoF) | | Inspire RH56 (12 DoF) | | ALLEX (15 DoF) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Q-err ↓ | Jerk ↓ | Q-err ↓ | Jerk ↓ | Q-err ↓ | Jerk ↓ | Q-err ↓ | Jerk ↓ | Q-err ↓ | Jerk ↓ |
| Oracle (GT joints) | 0.0 | 0.34 | 0.0 | 0.33 | 0.0 | 0.35 | 0.0 | 0.18 | 0.0 | 0.36 |
| HaMeR (Pavlakos et al., 2024) | 16.6 | 1.45 | 19.4 | 1.31 | 12.6 | 1.31 | 8.1 | 0.90 | 14.1 | 1.19 |
| WiLoR (Potamias et al., 2024) | 14.2 | 1.48 | 17.0 | 1.34 | 10.2 | 1.20 | 6.3 | 0.90 | 11.2 | 1.18 |
| HaWoR (Zhang et al., 2025) | 15.0 | 0.61 | 18.0 | 0.47 | 12.2 | 0.50 | 8.9 | 0.39 | 14.4 | 0.50 |
| HaPTIC (Ye et al., 2025b) | 14.9 | 1.39 | 17.6 | 1.27 | 11.9 | 1.10 | 7.0 | 0.87 | 12.7 | 1.21 |
| HandFlow (Xu et al., 2026) | 18.7 | 0.58 | 21.8 | 0.55 | 14.2 | 0.52 | 9.9 | 0.49 | 17.8 | 0.53 |
| ACE-Ego-Hand (Liu et al., 2026) | 14.0 | 0.20 | 17.4 | 0.21 | 10.8 | 0.21 | 7.4 | 0.11 | 12.0 | 0.20 |
| **RLHND (Ours)** | **12.7** | 0.18 | **16.1** | **0.16** | 9.6 | **0.19** | **6.0** | **0.07** | **10.2** | **0.17** |
| **RLHND (Ours)** ($\beta$-cache) | 12.8 | **0.17** | 16.3 | 0.17 | **9.5** | 0.20 | **6.0** | **0.07** | 10.3 | **0.17** |

Table 3: Retargeting quality on five dexterous right hands (DexPilot, wrist-relative targets), averaged over the HOT3D, ARCTIC ego and EgoDex test sets; per-dataset results are in Appendix B.4. Q-err is the mean absolute deviation (°) of the joint command from the one the same solver produces on the ground-truth joints; Jerk is the median over frames of the joint-averaged second temporal difference of the command (°/frame$^2$), i.e. the frame-to-frame noise on typical frames (Appendix A.4). The grey Oracle row retargets the ground-truth joints (Q-err $= 0$ by construction).

### 4.2 Tactile Estimation

Figure 4 compares predicted contact and pressure maps on OpenTouch and PressureVisionDB. RLHND produces localized contact regions and accurate pressure patterns on the hand surface, whereas HACO over-predicts contact and HOPE produces fragmented contact regions with underestimated pressure. PressureVision++ predicts pressure only in its native planar-sensor representation.

Table 2 quantifies contact on OpenTouch, HOT3D, DexYCB, and ARCTIC and force on OpenTouch and PressureVisionDB. We report both F1 and AUROC for contact. Since the PressureVision family (Grady et al., 2022; 2024) predicts pressure on the image plane rather than the hand surface, its contact metrics are not reported. RLHND achieves the best contact scores across all four datasets in both F1 and AUROC, with competitive force estimation on OpenTouch and PressureVisionDB.

### 4.3 Effectiveness on Robot Learning

**Real-robot experiments.** Dexterous Point Policy (Kim et al., 2026a), *i.e.*, DPP, trains robot policies from human demonstrations using keypoint locations and fingertip contact labels for force application, with the latter *manually* annotated for each demonstration. RLHND provides both signals *automatically* from a single model, with fingertip keypoints from its pose stream and contact labels from its contact head. Table 4 compares rollout results with DPP + RLHND; since the original tracker predicts no contact, both rows use RLHND's contact labels and differ only in the hand keypoints. DPP + RLHND improves performance across the evaluated tasks, with larger gains on the complex bimanual tasks.

**Retargeting to robot hands.** We evaluate how well tracker outputs serve as robot action spaces in Table 3. We retarget per-frame predictions to five dexterous right hands using the same DexPilot (Handa et al., 2020) solver with wrist-relative targets, and report Q-err and joint jerk averaged over HOT3D, ARCTIC ego, and EgoDex. RLHND, with or without the $\beta$-cache, achieves the lowest mean Q-err across all five embodiments and produces among the smoothest commands.

| Method | Human : Robot episodes | Pick and Place | | | | | Bimanual | | | All ↑ |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Bottle | Box | Ball | Bird | Avg. ↑ | Plastic bag | Assemble tissue | Avg. ↑ | |
| DPP (Kim et al., 2026a) | 100 : 100 | **93.8** | **100.0** | 81.3 | 81.3 | 89.1 | 53.1 | 75.0 | 64.1 | 80.8 |
| **DPP + RLHND (Ours)** | 100 : 100 | **93.8** | **100.0** | **87.5** | **93.8** | **93.8** | **59.4** | **90.6** | **75.0** | **87.5** |

Table 4: **Real-robot success rate of Dexterous Point Policy.** DPP is trained on human demonstrations labeled by its original tracker, and DPP + RLHND on the same demonstrations labeled by RLHND.

| | ARCTIC ego | | | | EgoDex | | | |
|---|---|---|---|---|---|---|---|---|
| Pose variant | MPJPE-p ↓ | PA-p ↓ | EPE2D-p ↓ | Q-err ↓ | MPJPE-p ↓ | PA-p ↓ | EPE2D-p ↓ | Q-err ↓ |
| A0: w/o Cosmos 3 | 16.62 | 7.55 | 13.40 | 12.48 | 21.57 | 10.63 | 22.45 | 15.64 |
| A1: w/o constrained pose | 14.26 | 6.52 | 12.33 | 10.85 | 19.91 | 10.38 | 20.76 | 15.31 |
| A2: w/o $\beta$-conditioning | 14.87 | 6.72 | 12.27 | 10.48 | 20.06 | 10.33 | 20.95 | 15.29 |
| A3: Full model | 13.36 | **6.42** | **12.19** | 10.12 | **19.90** | **10.12** | **20.72** | **15.22** |
| A4: Full model + GT $\beta$ | **13.15** | 6.45 | 12.21 | **9.69** | – | – | – | – |

| | DexYCB | | HOT3D | | ARCTIC | | OpenTouch force (kPa) | |
|---|---|---|---|---|---|---|---|---|
| Tactile variant | F1 ↑ | AUROC ↑ | F1 ↑ | AUROC ↑ | F1 ↑ | AUROC ↑ | MAE ↓ | RMSE ↓ |
| B0: w/o Cosmos 3 | 0.549 | 0.900 | 0.548 | 0.940 | 0.557 | 0.919 | 0.490 | 2.686 |
| B1: w/o LBS spread | 0.532 | 0.892 | 0.569 | 0.948 | 0.557 | 0.913 | 0.537 | 2.623 |
| B2: w/o contact-only data | 0.047 | 0.683 | 0.145 | 0.616 | 0.117 | 0.755 | 0.500 | 2.524 |
| B3: Full tactile expert | **0.572** | **0.915** | **0.589** | **0.959** | **0.602** | **0.935** | **0.489** | **2.508** |

Table 5: **Leave-one-out ablations of RLHND.** *Top (pose):* variants A0-A4, evaluated on ARCTIC ego and EgoDex. *Bottom (tactile):* variants B0–B3, evaluated on contact benchmarks and OpenTouch.

### 4.4 Ablation Studies

In the motion reconstruction, we ablate the three components that distinguish RLHND from its base tracker by removing each from the full model and retraining under the same 20k-step schedule and data mixture (Table 5). Variant (A0) replaces the Cosmos 3 Nano encoder with the Wan2.2 backbone following ACE-Ego-Hand, (A1) replaces the anatomically constrained parameterization of Section 3.2 with 16 free 6D rotations supervised on the raw 45-DoF MANO labels, and (A2) disables external-$\boldsymbol{\beta}$ injection and teacher forcing, so shape is predicted by the shape head alone, without the one-time hand calibration of Section 3.2, as in the full model (A3).

The encoder is the dominant factor, with A0 degrading performance across all metrics and test sets, confirming that leveraging the internal representations of a strong video backbone provides substantial gains. The other two components also contribute substantially to pose accuracy: removing either the anatomically constrained pose representation (A1) or shape caching (A2) degrades MPJPE in ARCTIC, while their combination yields a larger improvement in the full model. We also evaluate the full model with ground-truth $\boldsymbol{\beta}$ (A4), which yields no noticeable gains in joint-level metrics but improves Q-err, suggesting that fixed hand shape may stabilize downstream retargeting.

Each tactile variant B0–B2 retrains the stage-2 expert on the frozen full pose model with the stage-2 data mixture of Section 4. B0 uses the Wan2.2 stage-1 model of A0, B1 replaces the fixed MANO skinning weights $\mathbf{W}_{\mathbf{lbs}}$ of Eq. (6) with a learnable $778\times16$ spread matrix, and B2 removes the contact-only sources from the stage-2 mixture and trains on OpenTouch and PVDB alone.

B0 and B1 both degrade performance relative to the full model, showing that each ablated component contributes to tactile estimation. Removing the contact-only data (B2) causes a larger drop in contact performance on DexYCB, HOT3D, and ARCTIC, with no notable change in force estimation.

## 5 Conclusion

We presented RLHND, a physically grounded hand tracker that leverages a pre-trained video foundation model for robot learning. RLHND achieves state-of-the-art performance on hand motion reconstruction and tactile estimation benchmarks while producing accurate and smooth retargeted commands across five dexterous hands. On a real robot, training Dexterous Point Policy on RLHND-labeled demonstrations outperforms the original tracker while eliminating manual annotation.

## Reproducibility Statement

The architecture and training procedure are specified in Section 3.1–3.4, with implementation details, label construction, losses, evaluation metrics, training schedules, and baseline protocols provided in Appendix A. All training and evaluation data come from publicly available datasets, and the retargeting solver and its settings are described in Section 4.3. After the review, we will release the implementation of RLHND, along with the pre-trained weights.

## Ethics Statement

RLHND uses only publicly released human hand-object datasets under their respective licenses and consent procedures. We collect no new human data and release no raw video. Since the method can recover fine-grained hand activity from ordinary footage, including egocentric recordings containing bystanders, we intend it for robot learning from consenting demonstrators and discourage surveillance use. Contact and force are inferred rather than measured, so we recommend joint-torque or current limits and human supervision when deploying retargeted commands. Finally, the limited diversity of public datasets may lead to performance variation across subjects, hand sizes, and skin tones that is not captured by our evaluation sets. The authors declare no conflicts of interest.

## AI Use Statement

We used generative AI tools to assist with figure and table editing, and routine coding. AI-assisted code was reviewed and tested by the authors, and all reported results were independently produced and verified by the authors. All AI-assisted text and figures were reviewed against the underlying experiments and data. The authors take full responsibility for the content of this paper.

## References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, et al. Introducing HOT3D: An egocentric dataset for 3D hand and object tracking. *arXiv preprint arXiv:2406.09598*, 2024. 1, 17, 18, 21

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025. 1

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. $\pi_0$: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. 1

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. 1

Xiongyi Cai, Ri-Zhao Qiu, Geng Chen, Lai Wei, Isabella Liu, Tianshu Huang, Xuxin Cheng, and Xiaolong Wang. In-N-On: Scaling egocentric manipulation with in-the-wild and on-task data. *arXiv preprint arXiv:2511.15704*, 2025. 1

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. 23, 26

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. DexYCB: A benchmark for capturing

hand grasping of objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9044–9053, 2021. 16, 17, 18, 21

Baijun Chen, Weijie Wan, Tianxing Chen, Xianda Guo, Congsheng Xu, Yuanyang Qi, Haojie Zhang, Longyan Wu, Tianling Xu, Zixuan Li, Yizhe Wu, Rui Li, Xiaokang Yang, Ping Luo, Wei Sui, and Yao Mu. UniVTAC: A unified simulation platform for visuo-tactile manipulation data generation, learning, and benchmarking. *arXiv preprint arXiv:2602.10093*, 2026. 2

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*, pp. 12943–12954, 2023. 16, 17, 18, 21

Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1

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)*, 2022. 2, 3, 5, 8, 19

Patrick Grady, Jeremy A Zhao, Chengcheng Tang, Aditya Kumar, Christopher D Twigg, Yafei Ye, Etienne Vouga, and Charles C Kemp. PressureVision++: Estimating fingertip pressure from diverse RGB images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 8698–8708, 2024. 5, 8

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1

Siddhant Haldar and Lerrel Pinto. Point policy: Unifying observations and actions with key points for robot manipulation. In *Conference on Robot Learning*, 2025. 1

Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3D annotation of hand and object poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3196–3206, 2020. 17, 18

Ankur Handa, Karl Van Wyk, Wei Yang, Jacky Liang, Yu-Wei Chao, Qian Wan, Stan Birchfield, Nathan Ratliff, and Dieter Fox. DexPilot: Vision-based teleoperation of dexterous robotic hand-arm system. In *IEEE International Conference on Robotics and Automation*, 2020. 1, 8, 23

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. EgoDex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations*, 2026. 1, 17

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 4

Subin Jeon, Byungjun Kim, and Hanbyul Joo. HOPE: Hand-object pressure estimation from monocular videos. *arXiv preprint arXiv:2608.06192*, 2026. 2, 3, 6, 8, 18, 22, 23, 33

Daniel Jung and Kyoung Mu Lee. Learning dense hand contact estimation from imbalanced data. *Advances in Neural Information Processing Systems*, 38:120351–120384, 2025. 2, 3, 8, 18, 22, 23

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. EgoMimic: Scaling imitation learning via egocentric video. *arXiv preprint arXiv:2410.24221*, 2024. 1

Beomjun Kim, Seong Hyeon Park, Seunghoon Sim, Seungjun Moon, Sanghyeok Lee, and Jinwoo Shin. Dexterous point policy: Learning point-based dexterous hand policies from human demonstrations. *arXiv preprint arXiv:2606.10614*, 2026a. 1, 2, 8, 9

Dongyoung Kim, Huiwon Jang, Myungkyu Koo, et al. RLDX-1 technical report. *arXiv preprint arXiv:2605.03269*, 2026b. 1

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 1

Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2O: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10138–10148, 2021. 17

Marion Lepert, Jiaying Fang, and Jeannette Bohg. Phantom: Training robots without robots using only human videos. In *Conference on Robot Learning*, 2025. 1

Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. HybrIK: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3383–3393, 2021. 5

Qixiu Li, Yu Deng, Yaobo Liang, Lin Luo, Lei Zhou, Chengtang Yao, Lingqi Zeng, Zhiyuan Feng, Huizhi Liang, Sicheng Xu, Yizhong Zhang, Xi Chen, Hao Chen, Lily Sun, Dong Chen, Jiaolong Yang, and Baining Guo. VITRA: Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. *arXiv preprint arXiv:2510.21571*, 2025. 1

Jongbin Lim, Taeyun Ha, Mingi Choi, Jisoo Kim, Byungjun Kim, Subin Jeon, and Hanbyul Joo. HRDexDB: A paired human-robot dataset for cross-embodiment dexterous grasping. *arXiv preprint arXiv:2604.14944*, 2026. 17, 18

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023. 1

Vincent Liu, Ademi Adeniji, Haotian Zhan, Siddhant Haldar, Raunaq Bhirangi, Pieter Abbeel, and Lerrel Pinto. EgoZero: Robot learning from smart glasses. *arXiv preprint arXiv:2505.20290*, 2025. 1

Yufei Liu, Xixi Wang, Hao Li, Ganlong Zhao, Kaitong Cai, Chengkai Jin, Chunxiao Liu, Jianbo Liu, Siyuan Huang, Xingang Pan, and Hongsheng Li. ACE-Ego-Hand: Repurposing video diffusion models for occlusion-robust egocentric 3d hand motion recovery, 2026. URL https://arxiv.org/abs/2608.20308. 2, 3, 4, 5, 7, 8, 19, 20, 23, 33

Hao Luo, Yicheng Feng, Wanpeng Zhang, Sipeng Zheng, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Qin Jin, and Zongqing Lu. Being-H0: Vision-language-action pretraining from large-scale human videos. *arXiv preprint arXiv:2507.15597*, 2025. 1

Dantong Niu, Zhuoyang Liu, Zekai Wang, Boning Shao, Zhao-Heng Yin, Anirudh Pai, Yuvan Sharma, Stefano Saravalle, Ruijie Zheng, Jing Wang, Ryan Punamiya, Mengda Xu, Yuqi Xie, Yunfan Jiang, Letian Fu, Konstantinos Kallidromitis, Matteo Gioia, Junyi Zhang, Jiaxin Ge, Haiwen Feng, Fabio Galasso, Wei Zhan, David M. Chan, Yutong Bai, Roei Herzig, Jiahui Lei, Li Fei-Fei, Ken Goldberg, Jitendra Malik, Pieter Abbeel, Yuke Zhu, Danfei Xu, Linxi Fan, and Trevor Darrell. T-Rex: Tactile-reactive dexterous manipulation. *arXiv preprint arXiv:2606.17055*, 2026. 2

NVIDIA. Cosmos 3: A mixture-of-transformers omni foundation model for physical ai. *arXiv preprint arXiv:2606.02800*, 2026. 2, 4

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024. 1

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9826–9836, 2024. 3, 7, 8, 23, 33

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025. 1

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild, 2024. 3, 7, 8, 23, 33

Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. AnyTeleop: A general vision-based dexterous robot arm-hand teleoperation system. In *Robotics: Science and Systems*, 2023. 1

Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6):1–17, 2017. 3

Kenneth Shaw, Shikhar Bahl, and Deepak Pathak. VideoDex: Learning dexterity from internet videos. In *Conference on Robot Learning*, 2022. 1

Yuxin Ray Song, Jinzhou Li, Rao Fu, Devin Murphy, Kaichen Zhou, Rishi Shiv, Yaqi Li, Haoyu Xiong, Crystal Elaine Owens, Yilun Du, et al. OpenTouch: Bringing full-hand touch to real-world interaction. *arXiv preprint arXiv:2512.16842*, 2025. 2, 3, 5, 18, 21

Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 4

Yuxi Wang, Chengkai Jin, Yufei Liu, Wenqi Ouyang, Tianyi Wei, Zhiwei Zeng, Siyuan Huang, Zhiqi Shen, and Xingang Pan. The surprising effectiveness of video diffusion models for hand motion reconstruction. *arXiv preprint arXiv:2606.30308*, 2026. 2, 3, 4

Bowen Wen, Shaurya Dewan, and Stan Birchfield. Fast-foundationstereo: Real-time zero-shot stereo matching. *arXiv preprint arXiv:2512.11130*, 2025a. 23, 26

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025b. 26

Mingxin Xu et al. Handflow: Fully generative 4d hand recovery with flow matching. https://github.com/mxxu00/HandFlow, 2026. V1 open-source release. 3, 7, 8, 23, 33

Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. CPF: Learning a contact potential field to model the hand-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11097–11106, 2021. 5

Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Hongxu Yin, Sifei Liu, et al. EgoVLA: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025. 1

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. In *International Conference on Learning Representations*, 2025a. 1

Yufei Ye, Yao Feng, Omid Taheri, Haiwen Feng, Shubham Tulsiani, and Michael J. Black. Predicting 4D hand trajectory from monocular videos, 2025b. 3, 7, 8, 23, 33

Yuan Zeng, Yujia Shi, Tiao Tan, Xingting Li, Yaqi Qin, Zongqing Lu, Wenming Yang, Jing-Hao Xue, and Qingmin Liao. EgoTactile: Learning grasp pressure for everyday objects from egocentric video. In *International Conference on Machine Learning*, 2026. 3

Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. OakInk2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 445–456, 2024. 17, 18

Chi Zhang, Penglin Cai, Ziheng Xi, Haoqi Yuan, Hao Luo, Wanpeng Zhang, Sipeng Zheng, Chaoyi Xu, and Zongqing Lu. Human-centric transferable tactile pre-training for dexterous robotic manipulation. *arXiv preprint arXiv:2607.01067*, 2026a. 2, 17

Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. HaWoR: World-space hand motion reconstruction from egocentric videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1805–1815, 2025. 3, 7, 8, 23, 33

Xidong Zhang, Yichi Zhang, Jiaxin Shi, Fucai Zhu, Siyu Zhu, Michael Yu Wang, Xiaojun Wu, and Weihao Yuan. UniTacVLA: Unified tactile understanding and prediction in vision language action models. *arXiv preprint arXiv:2606.31723*, 2026b. 2

Yiming Zhao, Taein Kwon, Paul Streli, Marc Pollefeys, and Christian Holz. EgoPressure: A dataset for hand pressure and pose estimation in egocentric vision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025. 3, 34

Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5346–5355, 2020. 5