DEWHAND: DIRECT METRIC WORLD-SPACE HAND MOTION RECONSTRUCTION FROM EGOCENTRIC VIDEOS

Anonymous authors
Paper under double-blind review

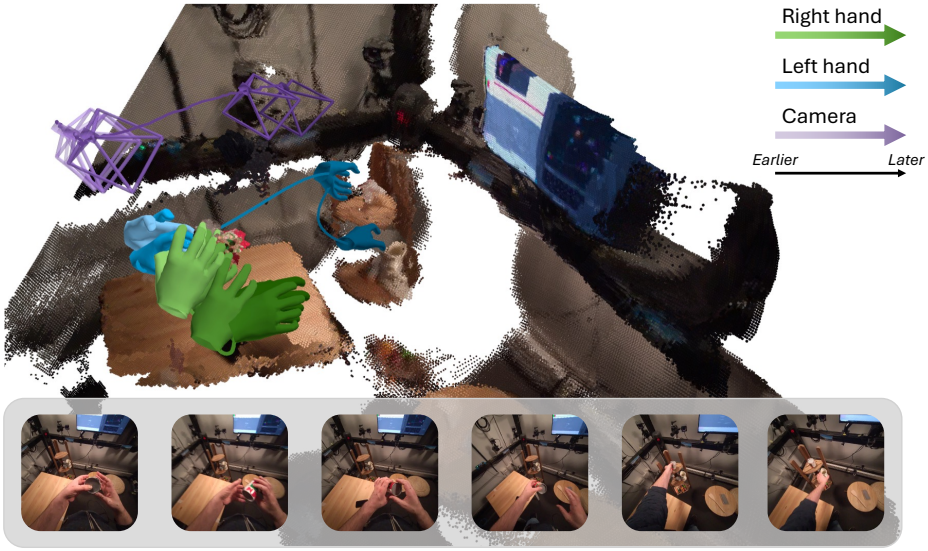


Figure 1: **DewHand** takes monocular egocentric video as input and directly reconstructs the motion of both hands and the camera in a shared metric world coordinate system.

ABSTRACT

Recovering metric world-space hand motion from egocentric video supports understanding human interactions and transferring them to embodied agents. Three challenges complicate this task: multi-stage pipelines propagate errors from independently estimated hand and camera motion; monocular observations leave physical scale ambiguous; and occlusion and out-of-view hands admit multiple plausible motion trajectories. We present **DewHand**, a unified model for direct joint reconstruction of world-space hand and camera motion. To avoid compounding errors from multi-stage estimation, DewHand jointly predicts hand and camera motion in a shared world coordinate system within a single-stage pipeline. To resolve metric-scale ambiguity, it combines scene-level geometric features from VGGT-Ω with metric-space supervision, while hand-specific features preserve fine-grained articulation. To handle occlusion and out-of-view hands, we formulate reconstruction as conditional motion generation with flow matching. Under the same architecture, this generative formulation outperforms direct regression in both accuracy and temporal smoothness with only two sampling steps. Experiments on HOT3D and ARCTIC demonstrate improved world-space accuracy and temporal consistency over state-of-the-art methods. DewHand reduces rigid-aligned world-space error relative to StableHand from 32.9 to 28.4 mm on HOT3D and from 76.2 to 36.2 mm on ARCTIC. It also recovers camera scale within 10% on two thirds of HOT3D clips, versus about a quarter for SLAM. Moreover, in zero-shot evaluations on HoloAssist, H2O, and TACO, captured with different devices, DewHand achieves the lowest rigid-aligned world-space error (WA-MPJPE) on each dataset.

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.

1 INTRODUCTION

Hands are our primary means of interacting with the physical world. Recovering their 3D motion from wearable cameras can support the analysis of human activity, the capture of manipulation skills, and the transfer of human demonstrations to embodied agents. These applications require more than an accurate hand pose in each video frame: they require a metrically scaled trajectory in a stable world coordinate system, so that a hand remains correctly placed as both the wearer and the camera move.

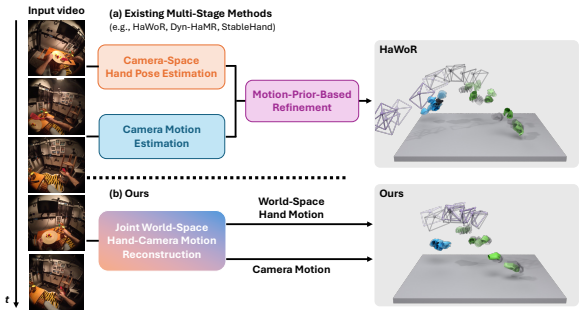


Figure 2: **Direct world-space hand-camera generation.** (a) Prior pipelines (Zhang et al., 2025a; Yu et al., 2025; Zeng et al., 2026) combine separate hand and camera estimates. (b) DewHand jointly generates both motions in metric world space. GT is overlaid in translucent dark gray on each reconstruction; predictions use green for the right hand, blue for the left hand, and purple for the camera.

This work addresses world-space 3D hand motion reconstruction from monocular egocentric video. Given a video, the goal is to recover the time-varying pose of both hands in a common world frame. Consider a wearer reaching for an object while turning their head. In camera coordinates, even a stationary hand appears to move with the camera; conversely, camera motion can hide or amplify the hand's motion in the image. A useful reconstruction must separate these effects and preserve the hand's position relative to the surrounding scene throughout the interaction.

Three challenges motivate our formulation. First, accurate local hand reconstruction does not by itself ensure a consistent world-space trajectory. Despite strong per-frame estimators (Pavlakos et al., 2024; Potamias et al., 2025), existing world-space pipelines typically estimate hands and camera motion separately, compose their outputs, and optionally apply temporal optimization or motion-prior refinement (Zhang et al., 2025a; Yu et al., 2025; Xu et al., 2026; Zeng et al., 2026). Camera drift and mismatches between these estimates can therefore propagate into the reconstructed hand motion. Second, monocular video provides projective observations rather than direct metric measurements, leaving physical scale ambiguous even when local articulation is accurately recovered. Third, self-occlusion, hand-object interactions, motion blur, and frequent exits from the field of view make hand observations incomplete. The visible evidence may be consistent with multiple motion trajectories, motivating a conditional distribution over hand and camera motion.

We introduce **DewHand**, a unified model that directly generates metric world-space motion of both hands and the camera from egocentric video. To avoid error propagation, we jointly generate wrist poses, local articulation, and camera motion in a shared world frame instead of composing independent estimates. To support metric-scale recovery, we condition on geometric features from the frozen VGGT-Ω backbone (Wang et al., 2026) and supervise in metric space; hand-specific features add cues for local articulation. To handle ambiguous and missing observations, we learn a conditional distribution over joint hand-camera motion using flow matching. Compared with regressing a single trajectory, this generative formulation produces more accurate and temporally smoother motion, while exhibiting greater sample diversity where the hand is unobserved.

Our contributions are threefold:

- We formulate egocentric world-space hand reconstruction as direct joint generation of metric dual-hand and camera motion instead of composing separately estimated hands and camera; supervising the generated camera reduces rigid-aligned hand error by 6.4 mm.
- We develop a flow-matching architecture that combines geometry-aware and hand-specific video features to jointly model wrist placement, articulation, and camera motion in a shared world frame; replacing the geometric features with DINOv2 increases hand error by 13 mm and camera error by 57 mm.

- We establish state-of-the-art results on HOT3D and ARCTIC, reducing rigid-aligned world error against StableHand from 32.9 to 28.4 mm and from 76.2 to 36.2 mm, and recovering camera scale within 10% on two thirds of HOT3D clips, versus about a quarter for SLAM.

2 Related Work

3D Hand Pose and Mesh Reconstruction: Large-scale models such as HaMeR (Pavlakos et al., 2024) recover detailed articulated pose in the wild, while WiLoR (Potamias et al., 2025) additionally targets robust 3D hand localization. Interaction and physical priors (Zuo et al., 2023; Suzuki et al., 2026; Zhang et al., 2025b) and spatio-temporal context (Zhao et al., 2024) further address occlusion and motion ambiguity. Recent work has explored complementary strategies for monocular hand reconstruction, including contrastive pre-training in SiMHand (Lin et al., 2025), controllable synthetic data generation in SesaHand (Zhao et al., 2026), exemplar-based in-context reconstruction in EgoHandICL (Xie et al., 2026), and kinematics-aware architectures in Rodrigues Network (Zhang et al., 2026). Egocentric datasets and benchmarks capture hand-object interactions (Ohkawa et al., 2023; Fan et al., 2024). HOPformer (Bansal et al., 2026) models egocentric hand-object pose, and EgoForce (Millerdurai et al., 2026) estimates camera-space hand pose. These advances improve hand pose and shape, but a shared metric world trajectory also requires camera motion.

World-Space Hand Motion Reconstruction: For full-body motion, WHAC (Yin et al., 2024) jointly recovers metric human and camera trajectories. For articulated hands, HaWoR (Zhang et al., 2025a) and Dyn-HaMR (Yu et al., 2025) construct global motion from camera-space hand estimates and separately recovered camera trajectories. HandFlow (Xu et al., 2026) learns camera-space hand motion distributions and transforms its predictions into world coordinates using SLAM. StableHand (Zeng et al., 2026) instead learns world-space refinement of hand observations constructed from separate hand and camera estimates, using quality-aware initialization to preserve reliable observations. UniHand (Sun et al., 2026) generates hand motion through latent diffusion in a fixed coordinate system defined by the first camera frame, but does not generate the camera trajectory. ReViV (Lyu et al., 2026) jointly models camera trajectories and camera-space hand kinematics, then aligns the reconstructed viewer and scene in a metric frame. DewHand jointly generates metric hand and camera trajectories in a single stage, conditioned on geometric and hand-specific features.

3 Method

For a monocular egocentric video $\mathcal{V} = \{I_t\}_{t=1}^{T}$, t indexes video frames and $\tau \in [0, 1]$ denotes flow time. Fig. 3 summarizes our design. Unlike HandFlow (Xu et al., 2026), which models camera-space hand motion before SLAM-based lifting, and StableHand (Zeng et al., 2026), which refines world-space observations assembled from separate hand and camera estimates, DewHand generates both trajectories in a metric world frame without an initial reconstruction. Geometric conditioning resolves scale ambiguity, and conditional flow modeling handles missing observations.

3.1 Problem Formulation and State Representation

Our goal is to recover the metric 3D motion of both hands *and* the camera in a common world frame. We take the camera coordinate system of the first frame as the world frame, so that no external localization, gravity estimate, or scene template is required at inference time. The outputs in Fig. 3 are the camera pose $(\mathbf{R}_t^c, \mathbf{p}_t^c) \in SO(3) \times \mathbb{R}^3$ and, for each hand $h \in \{L, R\}$, global wrist pose $(\mathbf{R}_t^h, \mathbf{p}_t^h) \in SO(3) \times \mathbb{R}^3$, local pose $\boldsymbol{\theta}_t^h \in \mathbb{R}^{15\times 6}$, and clip-level shape $\boldsymbol{\beta}^h \in \mathbb{R}^{10}$ (Romero et al., 2017). The local pose stores 6D rotations of 15 joints, converted to MANO's 45D axis-angle pose for mesh decoding. We group the quantities that live in the world frame into a *global motion stream* $\mathbf{x}_t^{\text{global}} = (\mathbf{R}_t^L, \mathbf{p}_t^L, \mathbf{R}_t^R, \mathbf{p}_t^R, \mathbf{R}_t^c, \mathbf{p}_t^c)$ and the hand-local parameters into a *local hand stream* $\mathbf{x}_t^{\text{local}} = (\boldsymbol{\theta}_t^L, \boldsymbol{\beta}^L, \boldsymbol{\theta}_t^R, \boldsymbol{\beta}^R)$. The clip-level shape target is repeated across frames in this stream. This separation reflects the different geometry and conditioning cues of global rigid motion and local articulation, while the two streams remain coupled throughout the model. Rotations are parameterized by the 6D continuous representation (Zhou et al., 2019).

Hands are frequently outside the field of view, occluded, or missed by per-frame detectors, so the observations can admit multiple plausible trajectories. A deterministic regressor can collapse these

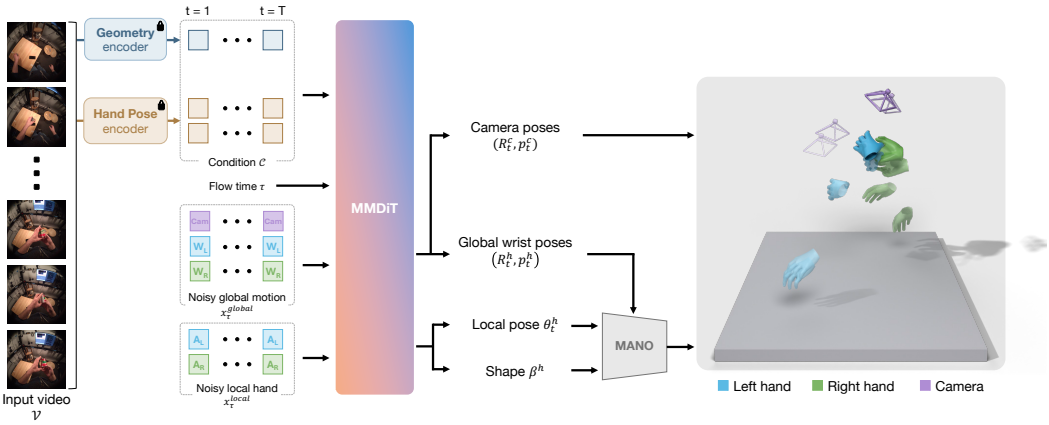


Figure 3: **DewHand architecture.** Frozen encoders provide $\mathcal{C} = (\mathcal{C}_{\text{geo}}, \mathcal{C}_{\text{hand}})$. At flow time τ, the model first applies spatio-temporal attention within each stream, then joint attention across the global motion, local hand, and condition streams at each frame. The Multimodal Diffusion Transformer (MMDiT) denoises $\mathbf{x}_\tau^{\text{global}}$ and $\mathbf{x}_\tau^{\text{local}}$. The outputs are camera pose $(\mathbf{R}_t^c, \mathbf{p}_t^c)$ and, for each hand $h \in \{L, R\}$, global wrist pose $(\mathbf{R}_t^h, \mathbf{p}_t^h)$, local pose $\boldsymbol{\theta}_t^h$, and shape $\boldsymbol{\beta}^h$. Cam, W_h, and A_h are the corresponding tokens; MANO converts the hand outputs to meshes.

possibilities into an averaged motion. Let $\mathcal{C} = (\mathcal{C}_{\text{geo}}, \mathcal{C}_{\text{hand}})$ denote the geometric and hand-specific features extracted from $\mathcal{V}$, as described in Section 3.2. Rather than directly regressing a single trajectory, we therefore learn the conditional motion distribution $p_\theta(\mathbf{x}_{1:T}^{\text{global}}, \mathbf{x}_{1:T}^{\text{local}} \mid \mathcal{C})$. Moreover, apparent hand motion in egocentric video is induced jointly by hand movement and camera ego-motion. We therefore generate hands and camera together, allowing the model to separate their contributions while maintaining their consistency in the shared world frame.

3.2 Visual Conditioning and Missing Observations

Metric world-space reconstruction requires scene-level geometric cues to constrain scale and global placement, alongside fine-grained hand cues for local articulation. We therefore encode the video with two complementary frozen backbones: VGGT-Ω (Wang et al., 2026) provides clip-level scene geometry and camera cues, while HaMeR (Pavlakos et al., 2024) provides detailed per-hand appearance features. Together with metric-space supervision (Section 3.4), the geometric features support physical-scale recovery without fixing the output to an externally estimated camera trajectory. VGGT-Ω jointly processes all T frames, so each per-frame output already contains temporal context. A learned query attention-pools its patch features. We add a learned projection of its camera token to the pooled feature and project the sum into one geometry token $\mathbf{c}_t^G$ per frame. HaMeR patch features are spatially pooled and projected into separate left- and right-hand tokens $\mathbf{c}_t^h$, $h \in \{L, R\}$; separate tokens keep a missing or corrupted observation of one hand from affecting the other.

Hand-specific features are unavailable when a hand is out of view or its detection fails. Let $m_t^h \in \{0, 1\}$ indicate whether a feature is used for hand h at frame t (1 if observed, 0 if missing or masked during training). We replace unused features with a learned, side-specific mask token $\mathbf{e}_{\text{mask}}^h$:

$$\tilde{\mathbf{c}}_t^h = m_t^h \mathbf{c}_t^h + (1 - m_t^h)\mathbf{e}_{\text{mask}}^h. \tag{1}$$

The resulting $\tilde{\mathbf{c}}_t^h$ is the hand-conditioning token passed to the flow model. During training, we additionally mask individual observations, contiguous spans, and complete hand tracks. The model must therefore recover missing evidence from the remaining video, scene, and temporal context, improving robustness to intermittent detector failures. The condition stream has three tokens per frame: geometry, left hand, and right hand.

3.3 JOINT HAND-CAMERA FLOW MODEL

Flow matching (Lipman et al., 2023; Hu et al., 2023) models the trajectories compatible with incomplete observations by learning transport from noise to data. We apply it to the joint hand-camera state, conditioned on visual features. The model uses two integration steps at inference.

At flow time τ, $\mathbf{x}_\tau^{\text{global}}$ and $\mathbf{x}_\tau^{\text{local}}$ denote the clip-wide states of the two generated streams. Here the subscript τ indexes flow time, while t continues to index video frames within each state; the subscript 1 below denotes the clean endpoint of the flow. We use one token per entity and frame. The global motion stream contains three 9D tokens for the left wrist, right wrist, and camera. Each token represents 3D translation and 6D rotation in world coordinates. The local hand stream contains two 100D tokens, one per hand, containing the 6D rotations of the 15 articulated MANO joints and the 10D MANO shape. The condition stream contains the three tokens defined above. Each input is projected to width D, and entity- and frame-specific embeddings retain its identity and temporal position. The three streams contain eight tokens in total at every frame. In Fig. 3, Cam denotes the camera token, W_h the global wrist token, and A_h the local hand token for side $h \in \{L, R\}$.

Each layer performs attention in two successive stages: within-stream attention, then joint attention across streams. *Within-stream attention* first relates the entities at each frame and then aggregates temporal context per entity. They run independently on the global motion, local hand, and condition streams, with separate parameters, to build per-stream spatio-temporal representations.

Joint attention across streams then lets global motion, local hand, and condition tokens attend to one another at each frame in a Multimodal Diffusion Transformer (MMDiT) block (Esser et al., 2024). As in HandFlow (Xu et al., 2026), condition tokens are updated rather than kept fixed.

The flow time τ and the clip-averaged VGGT camera token provide adaptive modulation throughout the network. Separate output heads map the final global motion and local hand tokens to clean-trajectory estimates $\hat{\mathbf{x}}_1^{\text{global}}$ and $\hat{\mathbf{x}}_1^{\text{local}}$. The predicted wrist poses and articulation are composed through MANO with the clip-averaged shape $\bar{\boldsymbol{\beta}}^h = T^{-1}\sum_t \hat{\boldsymbol{\beta}}_t^h$, where $\hat{\boldsymbol{\beta}}_t^h$ is the predicted shape component for hand h at video frame t. The predicted camera entity directly defines the camera trajectory; no separately estimated camera trajectory is composed with the hands after generation.

3.4 FLOW MATCHING AND TRAINING OBJECTIVES

For training, $\mathbf{x}_\tau = (\mathbf{x}_\tau^{\text{global}}, \mathbf{x}_\tau^{\text{local}})$ denotes the full clip state at flow time τ; subscripts 0 and 1 denote flow endpoints, not video frames. Let $\mathbf{x}_1 = (\mathbf{x}_{1:T}^{\text{global}}, \mathbf{x}_{1:T}^{\text{local}})$ denote the ground-truth trajectory after per-dimension standardization, and $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ a noise sample of the same shape. We draw a single flow time τ per clip from a logit-normal distribution, share it across both streams so that they are always denoised at a common noise level, and form $\mathbf{x}_\tau = (1-\tau)\,\mathbf{x}_0 + \tau\,\mathbf{x}_1$. The network predicts the clean sample $\hat{\mathbf{x}}_1$ rather than the velocity, which gives the flow-matching term

$$\mathcal{L}_{\text{fm}} = \left\|\hat{\mathbf{x}}_1^{\text{global}} - \mathbf{x}_1^{\text{global}}\right\|^2 + \left\|\hat{\mathbf{x}}_1^{\text{local}} - \mathbf{x}_1^{\text{local}}\right\|^2. \tag{2}$$

This term is computed in the standardized parameter space. Let $\mathcal{J}(\mathbf{x})$ denote the world-space 3D hand joints obtained by decoding state $\mathbf{x}$ through MANO, and let $\mathcal{J}^{\text{gt}}$ denote the ground-truth joints. As in HandFlow (Xu et al., 2026), we supervise the decoded joints, here in metric world space, $\mathcal{L}_{\text{joint}} = \mathrm{SmoothL1}(\mathcal{J}(\hat{\mathbf{x}}_1), \mathcal{J}^{\text{gt}})$, over frames with valid hand ground truth. This converts errors in wrist pose and local articulation into their physical effect on the complete hand. We supervise the camera directly in metric space with $\mathcal{L}_{\text{cam}}$, a squared error on its position and 6D rotation. To supervise temporal changes in wrist translation and rotation, let $\mathbf{q}$ denote world-space reference points fixed to each wrist. We match their first and second differences along video time t to the ground truth using $\mathcal{L}_{\text{vel}} = \mathrm{SmoothL1}(\Delta\hat{\mathbf{q}}, \Delta\mathbf{q}^{\text{gt}})$ and $\mathcal{L}_{\text{acc}} = \mathrm{SmoothL1}(\Delta^2\hat{\mathbf{q}}, \Delta^2\mathbf{q}^{\text{gt}})$, respectively. With loss weights $\lambda_j = \lambda_c = 1$ and $\lambda_v = \lambda_a = 800$, the total objective is

$$\mathcal{L} = \mathcal{L}_{\text{fm}} + \lambda_j\mathcal{L}_{\text{joint}} + \lambda_c\mathcal{L}_{\text{cam}} + \lambda_v\mathcal{L}_{\text{vel}} + \lambda_a\mathcal{L}_{\text{acc}}. \tag{3}$$

The predicted clean endpoint defines a flow field for each stream $S \in \{\text{global}, \text{local}\}$ as $\mathbf{v}_\theta^S(\mathbf{x}_\tau, \tau, \mathcal{C}) = (\hat{\mathbf{x}}_1^S(\mathbf{x}_\tau, \tau, \mathcal{C}) - \mathbf{x}_\tau^S)/(1-\tau)$. We integrate this field with two Euler steps; sample aggregation is detailed in Appendix C.1.

Table 1: **World-space hand motion on HOT3D and ARCTIC.** Errors in mm, AccEr in $\mathrm{m/s^2}$; wWA is wrist-level WA-MPJPE. Pred. shape: estimated MANO shape; GT shape: privileged ground-truth shape. $*$: zero-shot on ARCTIC. †: training data include HaWoR-split evaluation sequences. **Best**/second best: per split, among rows with the same shape input.

	Method		PA ↓	W-MPJPE ↓ Sim(3)	W-MPJPE ↓ SE(3)	WA-MPJPE ↓ Sim(3)	WA-MPJPE ↓ SE(3)	wWA ↓ Sim(3)	wWA ↓ SE(3)	MRRPE-A ↓	AccEr ↓
HOT3D	*HaWoR split*										
	WiLoR + ViPE†		8.21	171.78	176.26	64.52	92.06	53.00	86.33	166.29	16.98
	HaWoR		7.24	151.62	162.04	55.43	84.80	46.10	79.20	138.53	14.77
	HandFlow†		7.49	162.23	170.94	64.61	95.24	53.46	90.00	134.36	17.71
	EgoForce + ViPE†		7.11	158.31	167.30	60.13	79.31	49.24	73.08	114.14	15.69
	HOPformer + ViPE		13.54	223.86	231.17	85.55	106.01	60.56	85.19	199.68	20.87
	Ours	(Pred. shape)	**6.76**	**78.60**	**77.59**	**32.27**	**35.94**	**24.71**	**29.49**	**56.91**	**1.89**
	StableHand split										
	StableHand	(Pred. shape)	11.56	107.43	117.69	37.19	43.74	23.32	32.86	55.78	3.14
	Ours	(Pred. shape)	**6.25**	**64.21**	**63.46**	**26.45**	**28.56**	**20.47**	**24.07**	**52.00**	**2.65**
	StableHand	(GT shape)	**4.94**	81.02	87.31	28.71	32.90	**20.13**	26.92	**45.35**	2.92
	Ours	(GT shape)	5.71	**63.58**	**63.33**	**26.26**	**28.40**	20.47	**24.08**	51.97	**2.66**
ARCTIC	WiLoR + ViPE		8.12	231.97	233.98	72.38	77.64	59.36	66.76	112.45	20.70
	HaWoR		12.40	311.77	339.56	94.84	140.57	74.34	124.68	191.89	24.51
	HandFlow*		11.01	272.98	275.82	77.62	94.70	62.08	81.96	144.62	15.14
	EgoForce + ViPE		8.88	242.60	243.37	76.83	89.12	62.39	77.33	126.72	15.79
	HOPformer + ViPE		8.36	176.74	178.65	56.13	60.65	44.34	49.43	73.94	15.96
	StableHand	(Pred. shape)	11.37	282.79	285.59	74.68	78.63	54.05	59.40	101.48	3.47
	Ours	(Pred. shape)	**8.10**	**100.05**	**97.11**	**33.17**	**36.35**	**26.05**	**29.68**	**49.34**	**2.22**
	StableHand	(GT shape)	10.44	282.45	282.69	72.89	76.22	52.52	56.99	97.90	3.55
	Ours	(GT shape)	**8.14**	**99.20**	**97.00**	**33.03**	**36.18**	**26.02**	**29.65**	**49.33**	**2.22**

4 Experiments

4.1 World-Space Hand Motion

Datasets: We evaluate on HOT3D (Banerjee et al., 2025) and ARCTIC (Fan et al., 2023), following prior work on world-space hand motion reconstruction, including HaWoR (Zhang et al., 2025a) and HandFlow (Xu et al., 2026) on HOT3D and StableHand (Zeng et al., 2026) on both datasets. We train on HOT3D (for the StableHand split, a separate model that excludes its test windows; Appendix B.3) and fine-tune the resulting checkpoint on ARCTIC. For both training and inference, we use sequences of $T = 150$ frames at 30 fps, corresponding to 5-second clips.

Evaluation Metrics: We report PA-MPJPE, W-MPJPE, WA-MPJPE, MRRPE-A, and acceleration error (AccEr) on the 21-joint MANO skeleton. MRRPE-A measures inter-wrist displacement after a shared rigid alignment. We report WA-MPJPE with similarity ($\mathrm{Sim}(3)$) and rigid ($\mathrm{SE}(3)$) alignment. Rigid alignment removes differences in world-frame origin and orientation while preserving scale, so scale errors remain visible; similarity alignment additionally corrects scale. We also report wrist-level WA-MPJPE (wWA), which scores only the wrist trajectory and thus separates world placement from finger articulation. Appendix B details the scoring protocol. Unless stated otherwise, inference uses two Euler steps and five samples.

Baselines: We compare per-frame hand estimators (WiLoR, EgoForce (Millerdurai et al., 2026), HOPformer (Bansal et al., 2026)) lifted with ViPE, and the video-based methods HaWoR (Zhang et al., 2025a), HandFlow (Xu et al., 2026), and StableHand (Zeng et al., 2026); all but HandFlow are trained on ARCTIC. Baseline implementation details are in Appendix B.

Results: DewHand achieves the best world-space accuracy on both datasets (Table 1). Against StableHand with predicted shape, it reduces $\mathrm{SE}(3)$-aligned WA-MPJPE from 43.74 to 28.56 mm on the

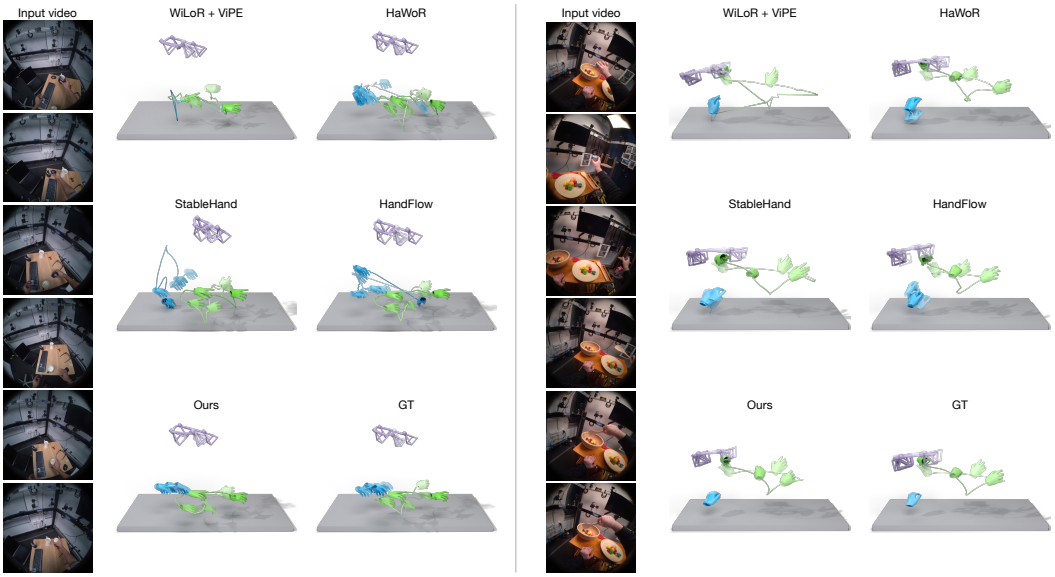


Figure 4: **Qualitative comparison on HOT3D.** Two selected sequences show six input frames and reconstructions from representative methods and ground truth. For each method, one $\mathrm{SE}(3)$ transform fitted jointly to both hands is also applied to its camera poses; no scale correction is used. Methods share the same view and crop within each sequence. Green and blue denote the right and left hands; purple frustums indicate camera poses.

Table 2: **Camera motion on HOT3D.** $\delta{<}\kappa$ is the percentage of clips whose trajectory scale is within a factor κ of the ground truth. ATE is reported without alignment (raw) and after $\mathrm{SE}(3)$ or $\mathrm{Sim}(3)$ alignment; ATE/RPE in mm, rot in degrees. **Best** in bold, second best underlined.

Method	$\delta{<}1.10\uparrow$	$\delta{<}1.25\uparrow$	$\mathrm{ATE}_{\mathrm{raw}}\downarrow$	$\mathrm{ATE}_{\mathrm{SE3}}\downarrow$	$\mathrm{ATE}_{\mathrm{Sim3}}\downarrow$	rot $\downarrow$	$\mathrm{RPE}_{\mathrm{t}}$@1 $\downarrow$
HaWoR	26.1%	42.2%	142.6	71.2	36.2	8.27	7.61
ViPE	24.6%	53.6%	96.4	45.6	24.9	4.61	10.21
VGGT-Ω	20.6%	48.8%	77.0	38.8	27.7	6.00	12.78
Ours	**66.2%**	**91.9%**	**31.5**	**16.7**	**13.1**	**1.73**	**3.79**

HOT3D StableHand split and from 78.63 to 36.35 mm on ARCTIC, and also improves PA-MPJPE. With the same privileged ground-truth shape, ours leads in every metric on ARCTIC and trails StableHand on HOT3D only in PA-MPJPE, MRRPE-A, and (by 0.3 mm) $\mathrm{Sim}(3)$-aligned wWA, so the world-space gains do not come from hand shape. The small gap between our $\mathrm{Sim}(3)$- and $\mathrm{SE}(3)$-aligned errors indicates that metric scale is retained; ours also has the lowest AccEr, and the lowest MRRPE-A among predicted-shape methods. Fig. 4 illustrates this: DewHand's hand trajectories remain close to ground truth, whereas several baselines show greater dispersion in the left sequence or displace the nearly stationary left hand in the right. Its margin persists when more than half of the frames lack a hand detection (71.0 vs. 134.3 mm for the best baseline; Appendix C.3), and it runs 2.6–3.7$\times$ faster than the SLAM-based pipelines (Appendix C.4).

4.2 Camera Motion

Evaluation: On HOT3D, we report camera scale accuracy, raw and aligned ATE, mean rotation error, and 1-frame RPE (Table 2); all trajectories are relative to frame 0.

Baselines: We compare against VGGT-Ω, the frozen geometry backbone we condition on, together with the two SLAM systems used to lift the hand baselines: HaWoR's (Zhang et al., 2025a) internal DROID-SLAM (Teed & Deng, 2021) and ViPE (Huang et al., 2025).

Table 3: **Zero-shot generalization to HoloAssist, H2O, and TACO.** All methods are scored on common frames. Camera-space methods are lifted to world coordinates with ViPE run on each 150-frame evaluation window with ground-truth intrinsics; DewHand uses no camera intrinsics. † marks target-domain training; **best** and second best are marked within each dataset.

	Method	PA ↓	W-MPJPE ↓		WA-MPJPE ↓		wWA ↓		AccEr ↓	MRRPE-A ↓
			Sim(3)	SE(3)	Sim(3)	SE(3)	Sim(3)	SE(3)		
HoloAssist	HaWoR	13.47	128.07	148.96	51.59	77.36	38.90	67.87	24.91	112.25
	HandFlow	12.53	142.57	233.89	55.39	161.42	43.36	153.23	29.91	109.69
	ReViV†	14.17	224.11	253.67	63.02	125.99	46.82	113.63	23.71	136.57
	WiLoR + ViPE	12.69	167.78	263.88	56.24	179.21	43.73	174.46	30.31	139.42
	EgoForce + ViPE	13.78	172.61	279.13	68.35	184.37	52.17	176.22	46.20	156.19
	HOPformer + ViPE	16.54	150.57	232.90	59.40	153.31	40.65	137.55	31.86	91.82
	Ours	**11.96**	**105.00**	**112.86**	**40.89**	**52.67**	**29.42**	**42.25**	**7.33**	**84.55**
H2O	HaWoR	7.22	62.61	62.86	22.81	35.57	14.46	31.63	13.35	34.52
	HandFlow	7.32	32.76	62.67	17.26	25.13	11.51	20.87	8.72	36.59
	ReViV†	9.44	56.76	55.46	23.72	27.45	12.93	18.96	3.05	45.28
	WiLoR + ViPE	**5.68**	35.93	64.65	14.80	22.04	**7.94**	16.26	7.50	23.76
	EgoForce + ViPE†	6.08	**27.68**	58.75	**14.16**	22.50	9.08	18.50	7.58	**22.76**
	HOPformer + ViPE	9.67	51.92	81.68	25.22	32.87	13.59	23.02	12.18	48.49
	Ours	6.80	40.88	**40.77**	19.54	**20.68**	12.10	**15.05**	**2.60**	46.17
TACO	HaWoR	9.77	80.84	84.61	32.34	43.63	23.95	38.01	15.47	57.17
	HandFlow	10.49	**55.51**	**56.76**	27.91	36.05	19.89	29.80	9.82	72.61
	ReViV†	11.41	88.09	90.67	40.07	47.77	27.86	39.72	5.63	**54.28**
	WiLoR + ViPE	**8.51**	56.98	60.94	27.38	33.45	19.71	29.29	12.37	55.82
	EgoForce + ViPE	10.13	66.52	71.55	33.91	43.06	25.04	38.17	17.28	75.41
	HOPformer + ViPE	12.70	74.58	75.48	38.50	41.73	22.97	28.58	17.08	90.59
	Ours	9.26	56.91	58.99	**26.76**	**31.29**	**18.58**	**24.15**	**3.35**	60.95

Results: Despite conditioning on frozen VGGT-Ω features, DewHand improves camera scale and trajectory accuracy over the backbone (Table 2). Gains after similarity alignment indicate improvements beyond scale correction.

4.3 Cross-Device Zero-Shot Generalization

We evaluate the HOT3D-trained model without fine-tuning on HoloAssist (Wang et al., 2023), testing transfer from Aria fisheye video to HoloLens 2 perspective video. HoloAssist provides camera poses and articulated hand tracks, enabling world-space evaluation across capture devices. On randomly sampled validation windows, DewHand achieves the lowest errors across all reported metrics, including PA-MPJPE, world-space placement, acceleration, and inter-hand placement (Table 3). Camera-space baselines are lifted with per-window ViPE (Huang et al., 2025) rather than ground-truth cameras; target-domain-trained references are marked in the table.

We further evaluate zero-shot transfer on the held-out subject of the H2O (Kwon et al., 2021) pose benchmark and on TACO (Liu et al., 2024), lifting the camera-space methods with per-window ViPE trajectories. On both datasets, DewHand achieves the lowest $\mathrm{SE}(3)$-aligned joint and wrist-level WA-MPJPE, as well as AccEr (Table 3). Other metrics, particularly local pose and inter-hand placement, show mixed results.

4.4 Ablation Study

We ablate camera supervision, visual conditioning, and the training objective on HOT3D.

Joint Camera Estimation (Table 4(a))**:** Removing camera supervision increases $\mathrm{SE}(3)$-aligned WA-MPJPE from 35.94 to 42.36 mm while leaving PA-MPJPE nearly unchanged. Camera supervision therefore primarily improves global hand placement rather than local articulation.

Table 4: **Ablations on HOT3D.** (a) Camera supervision, (b) visual conditioning, and (c) training objective. All variants use mask tokens and the same 64-epoch training recipe; each changes only the named component. **Best** in bold and second best underlined within each block. Metrics follow Tables 1 and 2; scale accuracy is reported as a percentage.

	Hands				Camera			
Variant	PA ↓	$\mathrm{WA}_{\mathrm{SE}(3)}$ ↓	$\mathrm{wWA}_{\mathrm{SE}(3)}$ ↓	AccEr ↓	$\delta<1.10$ ↑	$\delta<1.25$ ↑	$\mathrm{ATE}_{\mathrm{raw}}$ ↓	rot ↓
(a) Camera supervision								
Full	**6.76**	**35.94**	**29.49**	1.89	**66.2**	**91.9**	**31.5**	**1.73**
w/o camera supervision	6.86	42.36	34.63	**1.80**	–	–	–	–
(b) Conditioning features								
HaMeR **+ VGGT-Ω**	6.76	**35.94**	**29.49**	1.89	**66.2**	**91.9**	**31.5**	1.73
HaMeR + DA3	6.77	36.40	29.86	**1.79**	64.2	90.6	**31.5**	**1.60**
HaMeR + DINOv2	**6.71**	49.15	43.20	1.97	34.0	63.0	88.9	6.12
∅ + VGGT-Ω	10.72	46.57	35.07	2.04	64.0	90.3	32.8	1.83
(c) Training objective								
L2 regression	7.04	45.13	38.13	3.34	62.0	89.8	**31.4**	1.83
Flow matching	**6.76**	**35.94**	**29.49**	**1.89**	**66.2**	**91.9**	31.5	**1.73**

Conditioning Features (Table 4(b))**:** Replacing VGGT-Ω with DINOv2 (Oquab et al., 2024) raises $\mathrm{SE}(3)$-aligned WA-MPJPE from 35.94 to 49.15 mm and generated-camera ATE from 31.5 to 88.9 mm, supporting the value of 3D geometric features. DA3 (Lin et al., 2026) instead yields comparable hand and camera accuracy, despite its backbone camera ATE being 113.3 versus 77.0 mm for VGGT-Ω. This suggests that the camera trajectory is generated rather than copied from the backbone estimate. Removing hand-specific features substantially worsens local pose while leaving camera accuracy stable, indicating that the two condition streams play complementary roles.

Training Objective (Table 4(c))**:** Compared with L2 regression, flow matching improves local pose (PA-MPJPE: 7.04 to 6.76 mm), $\mathrm{SE}(3)$-aligned WA-MPJPE (45.13 to 35.94 mm), and temporal accuracy (AccEr: 3.34 to $1.89\,\mathrm{m/s^2}$), while camera accuracy remains comparable. The gain does not come from the architecture, which the regressor shares, and inference does not reduce to a single prediction: one Euler step, which returns the average of the network's direct estimates, is 3.4 mm worse and nearly doubles AccEr (Appendix C.1). Flow samples also spread more where the hand is undetected (8.9 vs. 23.8 mm), and the spread tracks the error (Appendix C.2).

5 Conclusion

We addressed metric world-space reconstruction of two hands and the camera from monocular egocentric video. DewHand jointly generates global hand and camera motion, local articulation, and hand shape in a shared world frame, avoiding the composition of separately estimated trajectories. Geometric and hand-specific features provide complementary cues for global placement and articulation, while flow matching models motion under incomplete observations. On HOT3D and ARCTIC, DewHand achieves the lowest rigid-aligned world-space hand error. The small gap between rigid- and similarity-aligned hand errors further supports metric-scale recovery without scale correction at evaluation. On HOT3D, it also improves metric camera scale recovery over the compared SLAM systems. Camera supervision primarily improves global hand placement, and flow matching yields more accurate and temporally consistent motion than direct regression. Zero-shot results show the lowest rigid-aligned WA-MPJPE on HoloAssist, H2O, and TACO, although other metrics do not always lead. Future work could extend the model beyond fixed offline windows and incorporate manipulated objects and hand-object contact.

AI USE STATEMENT

Generative AI tools were used in three ways. (i) *Writing*: drafting and polishing text, translation, and preparing figure layouts. (ii) *Retrieval and discovery*: searching for related work and checking bibliographic entries against the original papers. (iii) *Research ideation and execution*: suggesting experiments and ablations; writing and debugging code for data preprocessing, baseline execution, evaluation, and visualization; running experiments; and summarizing results. The authors made all research decisions, including which experiments to run and report, reviewed all AI-assisted outputs, cross-checked reported numbers against evaluation logs, and take full responsibility for the manuscript, analyses, figures, and reported results.

REFERENCES

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. HOT3D: Hand and object tracking in 3D from egocentric multi-view videos. In *CVPR*, 2025.

Siddhant Bansal, Zhifan Zhu, Shashank Tripathi, Jiahe Zhao, Michael J. Black, and Dima Damen. Towards in-the-wild egocentric 3D hand-object pose estimation. In *ECCV*, 2026.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

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 *CVPR*, 2023.

Zicong Fan, Takehiko Ohkawa, Linlin Yang, Nie Lin, Zhishan Zhou, Shihao Zhou, Jiajun Liang, Zhong Gao, Xuanyang Zhang, Xue Zhang, et al. Benchmarks and challenges in pose estimation for egocentric hand interactions with objects. In *ECCV*, pp. 428–448. Springer, 2024.

Vincent Tao Hu, Wenzhe Yin, Pingchuan Ma, Yunlu Chen, Basura Fernando, Yuki M. Asano, Efstratios Gavves, Pascal Mettes, Bjorn Ommer, and Cees G. M. Snoek. Motion flow matching for human motion synthesis and editing. *arXiv preprint arXiv:2312.08895*, 2023.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixe, and Sanja Fidler. ViPE: Video pose engine for 3D geometric perception. *arXiv preprint arXiv:2508.10934*, 2025.

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 (ICCV)*, pp. 10138–10148, 2021.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. In *ICLR*, 2026.

Nie Lin, Takehiko Ohkawa, Yifei Huang, Mingfang Zhang, Minjie Cai, Ming Li, Ryosuke Furuta, and Yoichi Sato. SiMHand: Mining similar hands for large-scale 3D hand pose pre-training. In *ICLR*, 2025. URL https://openreview.net/forum?id=96jZFqM5E0.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. TACO: Benchmarking generalizable bimanual tool-ACtion-object understanding. In *CVPR*, 2024.

Xiaozhong Lyu, Gen Li, Zhiyin Qian, Xucong Zhang, Marc Pollefeys, and Siyu Tang. ReViV: Reconstructing the viewer and the view in 4D from monocular egocentric video. In *ECCV*, 2026.

Christen Millerdurai, Shaoxiang Wang, Yaxu Xie, Vladislav Golyanik, Didier Stricker, and Alain Pagani. EgoForce: Forearm-guided camera-space 3D hand pose from a monocular egocentric camera. In *ACM SIGGRAPH Conference Papers*, 2026.

Takehiko Ohkawa, Kun He, Fadime Sener, Tomáš Hodan, Luan Tran, and Cem Keskin. AssemblyHands: towards egocentric activity understanding via 3d hand pose estimation. In *CVPR*, pp. 12999–13008, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. WiLoR: End-to-end 3D hand localization and reconstruction in-the-wild. In *CVPR*, 2025.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 2017.

Zhihao Sun, Tong Wu, Ruirui Tu, Daoguo Dong, and Zuxuan Wu. UniHand: A unified model for diverse controlled 4D hand motion modeling. In *ICLR*, 2026. URL https://openreview.net/forum?id=upUl6hMYwy.

Naru Suzuki, Takehiko Ohkawa, Tatsuro Banno, Jihyun Lee, Ryosuke Furuta, and Yoichi Sato. Affordance-guided diffusion prior for 3D hand reconstruction. In *ECCV*, 2026.

Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. In *NeurIPS*, 2021.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *CVPR*, 2026.

Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. HoloAssist: an egocentric human interaction dataset for interactive AI assistants in the real world. In *ICCV*, 2023.

Binzhu Xie, Shi Qiu, Sicheng Zhang, Yinqiao Wang, Hao Xu, Muzammal Naseer, Chi-Wing Fu, and Pheng-Ann Heng. EgoHandICL: Egocentric 3D hand reconstruction with in-context learning. In *ICLR*, 2026.

Mingxi Xu, Bowen Duan, Yi Gu, Zhengyang Shen, Renjing Xu, and Yutao Yue. HandFlow: Fully generative 4D hand recovery with flow matching. *arXiv preprint arXiv:2607.11221*, 2026.

Wanqi Yin, Zhongang Cai, Ruisi Wang, Fanzhou Wang, Chen Wei, Haiyi Mei, Weiye Xiao, Zhitao Yang, Qingping Sun, Atsushi Yamashita, Ziwei Liu, and Lei Yang. WHAC: World-grounded humans and cameras. In *ECCV*, 2024.

Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. Dyn-HaMR: Recovering 4D interacting hand motion from a dynamic camera. In *CVPR*, 2025.

Huajian Zeng, Chaohua Yao, Yuantai Zhang, Jiaqi Yang, Rolandos Alexandros Potamias, and Xingxing Zuo. StableHand: Quality-aware flow matching for world-space dual-hand motion estimation from egocentric video. *arXiv preprint arXiv:2605.18553*, 2026.

Jialiang Zhang, Haoran Geng, Yang You, Congyue Deng, Pieter Abbeel, Jitendra Malik, and Leonidas Guibas. Rodrigues network for learning robot actions. In *ICLR*, 2026. URL https://openreview.net/forum?id=IZHk6BXBST.

Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. HaWoR: World-space hand motion reconstruction from egocentric videos. In *CVPR*, 2025a.

Yufei Zhang, Zijun Cui, Jeffrey O. Kephart, and Qiang Ji. Diffusion-based 3D hand motion recovery with intuitive physics. In *ICCV*, 2025b.

Weichao Zhao, Hezhen Hu, Wengang Zhou, Li Li, and Houqiang Li. Exploiting spatial-temporal context for interacting hand reconstruction on monocular RGB video. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 20(6):171:1–171:18, 2024.

Zhuoran Zhao, Xianghao Kong, Linlin Yang, Zheng Wei, Pan Hui, and Anyi Rao. SesaHand: Enhancing 3D hand reconstruction via controllable generation with semantic and structural alignment. In *ICLR*, 2026. URL https://openreview.net/forum?id=sKMgGQQy7g.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019.

Binghui Zuo, Zimeng Zhao, Wenqian Sun, Wei Xie, Zhou Xue, and Yangang Wang. Reconstructing interacting hands with interaction prior from monocular images. In *ICCV*, 2023.