EGOHSI3R: ONLINE 4D HAND-SCENE INTERACTION RECONSTRUCTION FROM EGOCENTRIC VIDEO

Anonymous authors
Paper under double-blind review



Figure 1: EgoHSI3R reconstructs camera motion, dense scene geometry, world-space bilateral MANO (Romero et al., 2017) hands, and dense hand contact from egocentric video in a streaming manner. *Top:* input egocentric frames. *Middle:* reconstructed scene, camera, and hands. *Bottom:* reconstructed hands with predicted dense contacts highlighted in red.

ABSTRACT

Egocentric videos offer an abundant source of hand-scene interactions for embodied learning, which requires recovering hands, camera, and scene consistently in a shared 3D world. Existing methods focus on either camera-frame hands or world-space hand trajectories, without jointly recovering the scene. Building a complete reconstruction from such specialized predictions therefore typically requires per-sequence optimization. We propose **EgoHSI3R**, a unified feed-forward framework for online 4D hand-scene interaction reconstruction from egocentric monocular video. To jointly recover these components, it maintains a recurrent scene representation that accumulates geometric evidence over time and provides a shared world space and scale. We introduce hand-centric prompts that fuse pretrained hand priors and local scene context, while MANO vertex queries attend to scene features to predict dense per-vertex contact. These designs enable a single network to jointly estimate camera motion, dense scene geometry, world-space bilateral MANO hands, and per-vertex hand contact. EgoHSI3R achieves the lowest world-space hand joint error on ARCTIC, while running faster than prior world-space hand pipelines and retaining competitive camera and scene reconstruction quality.

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

Egocentric videos provide abundant observations of hand–scene interactions for embodied learning. Recovering these interactions as structured 4D information requires jointly estimating camera motion, persistent dense scene geometry, world-space articulated hands, and dense hand-scene contact in a shared 3D world. Such a representation captures not only how the hands move, but also where and how they physically interact with the surrounding environment.

Recent progress in geometry foundation models has enabled recovering scene geometry from images or videos (Wang et al., 2024a; 2025a;b). For hand reconstruction, Figure 2 summarizes two dominant paradigms in egocentric hand reconstruction. Camera-centric approaches recover detailed hand pose and shape from individual images or videos, but leave their predictions in the camera coordinate system without temporally consistent positions in the world (Potamias et al., 2025; Ye et al., 2026; Wang et al., 2026; Liu et al., 2026a). World-space approaches additionally estimate camera motion to recover hand trajectories in world space (Zhang et al., 2025a; Yu et al., 2025; Zeng et al., 2026; Xu et al., 2026b; Zhu et al., 2026). Both paradigms, however, provide only part of a complete hand-scene interaction reconstruction: they do not jointly recover the scene. Building a complete reconstruction from such specialized predictions therefore requires combining separate estimates and reconciling their coordinates, scales, and relative placement, typically through offline optimization.

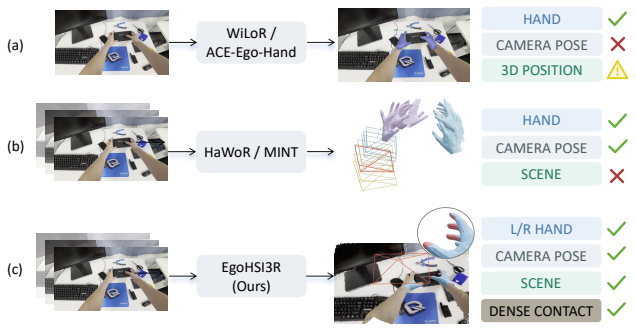

Figure 2: **Comparison of egocentric hand reconstruction paradigms.** (a) Camera-centric methods recover articulated hands in the camera frame, without camera motion and reliable 3D placement. (b) World-space methods additionally estimate camera motion to recover hand trajectories in world space, but do not reconstruct the scene. (c) **EgoHSI3R** jointly recovers bilateral hands, camera pose, scene geometry, and dense hand contact in a shared world frame within a single online, feed-forward model.

We therefore study online 4D hand-scene interaction reconstruction from egocentric video. Given only the frames up to the current time, our task is to incrementally recover camera motion, persistent scene geometry, world-space trajectories of both MANO hands, and dense per-vertex contact in a shared world coordinate system. Here, *online* denotes causal inference without future frames.

This formulation introduces a challenge that does not arise in hand-only reconstruction: the hands and scene must be recovered in a consistent 3D configuration. Independently reconstructed hand trajectories and scene geometry may each appear plausible, yet disagree in their relative position or depth, resulting in floating hands or hand-scene interpenetration. The task therefore requires modeling the spatial relationship between the hands and scene as part of the prediction itself, rather than simply resolving this inconsistency after reconstruction by sequence-specific alignment.

To address this challenge, we present **EgoHSI3R**, an online feed-forward framework that jointly models the hands, scene, and their contact within a shared recurrent representation. The recurrent scene representation accumulates geometric evidence and camera motion over time, providing a persistent world space and scale shared by the hands and scene. We introduce hand-centric prompts that fuse pretrained hand priors and local scene context, and inject them into the recurrent representation to recover world-space trajectories of both MANO hands conditioned on the evolving scene. To explicitly model physical interaction, MANO vertex queries attend to local scene features to predict dense hand contact under contact-aware supervision. As illustrated in Figure 1, EgoHSI3R processes each incoming frame with one recurrent update and jointly reconstructs camera motion, persistent scene geometry, world-space trajectories of both MANO hands, and dense per-vertex hand–scene contact without test-time optimization.

We train EgoHSI3R on multiple egocentric datasets and evaluate it across scene reconstruction, world-space hand reconstruction, and contact prediction. EgoHSI3R achieves state-of-the-art world-

space hand reconstruction on ARCTIC (Fan et al., 2023), reducing W-MPJPE to 108.80 mm, a 15.79 mm improvement over the strongest baseline. Ablation studies further verify that both hand priors and interaction-aware contact feedback contribute to improved world-space hand estimation.

In summary, our main contributions are:

- We propose **EgoHSI3R**, an online feed-forward framework for 4D hand-scene interaction reconstruction from egocentric monocular video, jointly recovering camera motion, persistent scene geometry, world-space bilateral MANO hands, and dense hand-scene contact.
- We design a unified recurrent scene representation that enables joint hand-scene estimation over time, together with hand-centric prompts and vertex-level contact queries for scene-conditioned bilateral hand reconstruction and dense contact modeling.
- EgoHSI3R achieves state-of-the-art performance in world-space hand reconstruction, while retaining competitive camera and scene reconstruction accuracy in an online feed-forward setting.

2 Related Work

2.1 Monocular 3D and 4D Hand Reconstruction

Single-image methods recover detailed 3D hands from unconstrained RGB images using parametric hand models and strong learned priors (Pavlakos et al., 2024; Dong et al., 2024; Chen et al., 2025; Potamias et al., 2025). Egocentric approaches further address severe truncation, distortion, and self-occlusion (Prakash et al., 2024; Millerdurai et al., 2026). However, these methods mainly operate in the camera frame and process frames independently.

Video-based methods improve temporal consistency through motion priors, tracking, or generative modeling (Duran et al., 2024; Ye et al., 2026; Lin et al., 2026; Wang et al., 2026; Liu et al., 2026a). Recent systems further reconstruct hands in world coordinates by combining hand observations with estimated camera motion (Zhang et al., 2025a; Yu et al., 2025; Zeng et al., 2026; Xu et al., 2026b). MINT (Zhu et al., 2026) jointly predicts camera and hand trajectories from egocentric video, but does not maintain dense scene geometry or explicitly model hand–scene contact. In contrast, our method jointly reconstructs world-space hands, camera motion, scene geometry, and contact within a causal framework.

2.2 Feed-Forward 3D/4D Reconstruction

Feed-forward geometry models directly infer pointmaps and camera parameters without per-scene optimization. DUSt3R and VGGT establish strong multi-view reconstruction paradigms (Wang et al., 2024a; 2025a), while Spann3R, SLAM3R, and CUT3R introduce spatial memory or recurrent state for sequential reconstruction (Wang & Agapito, 2025; Liu et al., 2025; Wang et al., 2025b). Recent methods further improve long-sequence scalability through bounded memory or relative coordinate representations (Yuan et al., 2026; Chen et al., 2026a; Xu et al., 2026a).

Dynamic extensions reconstruct both scene geometry and motion using time-dependent pointmaps, tracking, or conditional queries (Zhang et al., 2025b; Feng et al., 2025; Zhang et al., 2026; Karhade et al., 2026; Luo et al., 2026). Nevertheless, these models represent humans and hands only as generic dynamic geometry and do not recover articulated hand parameters or interaction states.

2.3 Joint Human-Scene Reconstruction

Joint human–scene reconstruction has been explored through optimization-based systems such as JOSH and TRAM (Liu et al., 2026b; Wang et al., 2024b), as well as feed-forward models including UniSH and Human3R (Li et al., 2026; Chen et al., 2026c). These approaches primarily focus on offline, third-view full-body SMPL reconstruction, leaving detailed hand articulation and hand–scene interaction underexplored. Our work addresses this gap.

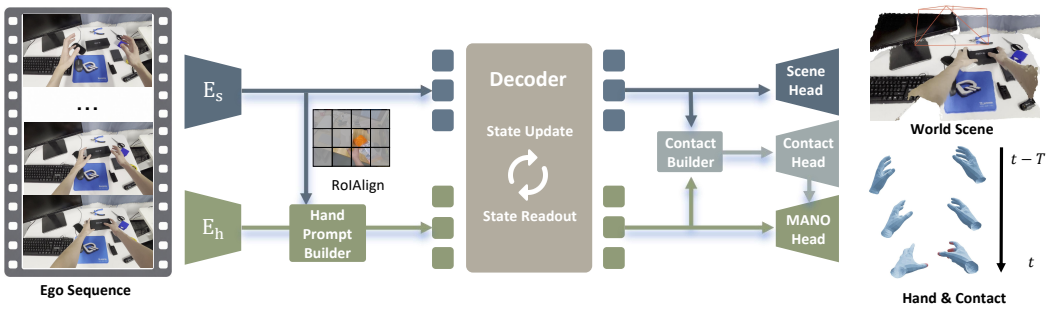

Figure 3: **Overview of EgoHSI3R.** Each incoming egocentric frame is encoded by the scene encoder E_s into image tokens, and each detected hand crop by the pretrained hand encoder E_h. The *Hand Prompt Builder* fuses the hand prior with local scene context obtained via RoIAlign, producing scene-conditioned hand prompts. The decoder jointly processes scene tokens and hand prompts while reading from and updating a persistent scene state. The Scene Head predicts camera pose and world-space pointmaps, and the MANO Head reconstructs bilateral MANO parameters. The interaction-aware *Contact Builder* combines decoded hand tokens with local scene features, which are then used by the Contact Head to predict per-vertex contact and refine hand reconstruction. All outputs are produced online in a shared world space without test-time optimization.

2.4 Dense Hand Contact and Physical Interaction Modeling

ContactPose provides paired hand pose and surface-contact annotations (Brahmbhatt et al., 2020), while CPF and ContactOpt use contact constraints to refine hand–object configurations (Yang et al., 2021; Grady et al., 2021). HACO predicts dense hand contact directly from visual observations (Jung & Lee, 2025), and ContactPrompt explores zero-shot contact estimation using multimodal models (Jung & Lee, 2026). Recent methods jointly estimate hand–object pose (Fu et al., 2026; Bansal et al., 2026) or extend contact prediction toward physical force estimation (Ilic et al., 2026). Most existing approaches estimate contact from individual frames or assume an object mesh is already available. In our formulation, we treat contact as the connection between hand and scene.

3 Method

We address online 4D hand-scene interaction reconstruction by maintaining a recurrent scene representation that is updated as each egocentric frame arrives. The key mechanism is to integrate hand and dense contact estimation into recurrent scene reconstruction, allowing hands, scene geometry, and their interaction to be recovered in a shared world space without future observations or per-sequence optimization. We first formulate the streaming reconstruction task in Sec. 3.1 and summarize the overall framework in Sec. 3.2. We then describe the *Hand Prompt Builder* in Sec. 3.3, the *Interaction-Aware Contact Builder* in Sec. 3.4, and the training objectives in Sec. 3.5.

3.1 Problem Formulation

Given a monocular egocentric RGB stream $\{\mathbf{I}_t\}_{t=1}^{T}$, where $\mathbf{I}_t \in \mathbb{R}^{H\times W\times 3}$, the problem is to use only the observed prefix $\mathbf{I}_{\le t}$ at each time t to recover camera poses $\mathbf{T}_t \in \mathrm{SE}(3)$, dense world-coordinate pointmaps $\mathbf{X}_t \in \mathbb{R}^{H\times W\times 3}$, and bilateral hand states $(\boldsymbol{\Theta}_t^h, \boldsymbol{\beta}_t^h, \boldsymbol{\tau}_t^h, \mathbf{c}_t^h)$ for $h \in \{\mathrm{R}, \mathrm{L}\}$. For each hand, $\boldsymbol{\Theta}_t^h \in \mathrm{SO}(3)^{16}$ contains the rotations of the 16 MANO joints, $\boldsymbol{\beta}_t^h \in \mathbb{R}^{10}$ represents hand shape, $\boldsymbol{\tau}_t^h \in \mathbb{R}^3$ denotes camera-coordinate translation, and $\mathbf{c}_t^h \in [0,1]^{778}$ contains the contact probability of each MANO vertex.

3.2 Framework Overview

To solve the streaming reconstruction task, we propose EgoHSI3R, an online feed-forward framework that jointly reconstructs camera motion, scene geometry, bilateral hands, and dense contact using a recurrent representation, as illustrated in Fig. 3. We instantiate this representation with CUT3R (Wang et al., 2025b), whose persistent state accumulates scene geometry and camera motion over time, and use WiLoR (Potamias et al., 2025) to extract crop-based hand priors. Two components connect these representations. The *Hand Prompt Builder* injects scene-conditioned

hand priors into the recurrent reconstruction process, while the *Interaction-Aware Contact Builder* associates decoded hand features with local scene evidence to predict dense contact.

3.3 HAND PROMPT BUILDER

Visual prompting provides an efficient way to integrate parametric human priors into recurrent scene reconstruction (Chen et al., 2026c). We adapt this principle to bilateral hands and condition each hand prompt on its local scene context.

For frame $\mathbf{I}_t$, the scene encoder produces patch tokens $\mathbf{F}_t \in \mathbb{R}^{N\times d_s}$. An online hand detector localizes each visible hand with a bounding box $\mathbf{b}_t^h$. We crop the detected region and use the hand encoder to extract a structured prior $\mathbf{A}_t^h \in \mathbb{R}^{M\times d_h}$, where $M = 18$. Its first 16 tokens encode articulated pose cues, while the remaining two summarize shape and global translation.

A hand crop provides detailed articulation cues but removes much of the surrounding geometry needed to place the hand in the scene. We therefore use RoIAlign to extract local scene features $\mathbf{R}_t^h = \mathrm{RoIAlign}(\mathbf{F}_t, \mathbf{b}_t^h)$ and pool them into a context vector $\mathbf{s}_t^h$. The projected hand prior is combined with a learned handedness embedding, and a feature-wise gate controls the injection of local scene context:

$$\mathbf{P}_t^h = \phi_{\mathrm{mix}}\left(\mathbf{q}_t^h + \mathbf{g}_t^h \odot \mathbf{s}_t^h\right), \quad \mathbf{q}_t^h = \phi_{\mathrm{h}}(\mathbf{A}_t^h) + \mathbf{e}_{\mathrm{side}}^h, \quad \mathbf{g}_t^h = \sigma\big(\phi_{\mathrm{g}}([\mathbf{q}_t^h; \mathbf{s}_t^h])\big). \tag{1}$$

This construction preserves the structured hand prior while adapting it to the hand's local scene context. We assign the resulting prompt tokens the spatial position of the bounding-box center. An absent hand is represented by learned mask tokens, maintaining a fixed bilateral token layout.

We append the two hand prompts to CUT3R's camera and scene tokens, where self-attention and bidirectional cross-attention enable the exchange of hand, scene, and historical information. The decoder consequently produces scene-aware hand tokens $\widetilde{\mathbf{P}}_t^h$ for contact and MANO prediction, together with hand-conditioned scene tokens $\widetilde{\mathbf{F}}_t$ for dense reconstruction.

3.4 INTERACTION-AWARE CONTACT BUILDER

The decoded hand tokens provide a compact joint-level representation, whereas contact is defined densely over the MANO surface. To bridge this difference, we first fuse $\widetilde{\mathbf{P}}_t^h$ with the original hand prior $\mathbf{A}_t^h$. We then propagate the first $J = 16$ pose-token features to the $V = 778$ MANO vertices using the fixed MANO skinning matrix $\mathbf{W} \in \mathbb{R}^{V\times J}$. For vertex i, its query feature is

$$\mathbf{v}_{t,i}^h = \mathrm{LN}\left(\sum_{j=1}^{J} W_{ij}\mathbf{u}_{t,j}^h + \overline{\mathbf{u}}_{t,\mathrm{g}}^h + \mathbf{e}_i^{\mathrm{vert}} + \mathbf{e}_{\mathrm{side}}^h\right), \tag{2}$$

where $\mathbf{u}_{t,j}^h$ is the fused feature of joint token j, $\overline{\mathbf{u}}_{t,\mathrm{g}}^h$ summarizes the two non-pose tokens, and $\mathbf{e}_i^{\mathrm{vert}}$ is a learned vertex embedding. The MANO skinning weights preserve the correspondence between articulated joints and surface vertices without introducing an additional learned topology mapping.

To obtain scene evidence, we apply RoIAlign to the decoded scene tokens using the same hand box, yielding local context $\mathbf{C}_t^h$. Each vertex query attends to this context through multi-head cross-attention, followed by a feed-forward block and a binary classifier. The contact probability is therefore computed as $\widehat{c}_{t,i}^h = \sigma(\phi_{\mathrm{c}}(\mathrm{FFN}(\mathbf{v}_{t,i}^h + \mathrm{MHA}(\mathbf{v}_{t,i}^h, \mathbf{C}_t^h, \mathbf{C}_t^h))))$. Restricting the context to the hand neighborhood suppresses irrelevant background features while retaining nearby object surfaces.

Beyond contact classification, the interaction-aware vertex features are averaged and projected to a residual update $\Delta\mathbf{P}_t^h$ for the decoded hand tokens. With a learnable gate α, we update them as $\widetilde{\mathbf{P}}_t^{h\prime} = \widetilde{\mathbf{P}}_t^h + \sigma(\alpha)\Delta\mathbf{P}_t^h$. The final decoded hand tokens are concatenated with $\mathbf{A}_t^h$ and mapped to 16 MANO rotations, 10 shape coefficients, and a 3D translation. We initialize the feedback projection to zero and its gate near zero so that optimization starts from the pretrained reconstruction solution and gradually incorporates interaction cues. In this way, the Contact Builder not only

Table 1: **Evaluation of Camera-space Hand Motion Estimation**. *p-free* denotes penalty-free evaluation, where failed hand detections are not penalized.

Method	3D Pose		Orient. & Position			Temporal
	MPJPE-p↓	PA-p↓	$EPE_{2D\text{-}p}$ ↓	GO-p↓	CT-p↓	Jitter↓
ARCTIC						
HaMeR (Pavlakos et al., 2024)	29.197	14.596	65.289	24.907	0.095	18.279
WildHands (Prakash et al., 2024)	25.704	13.941	50.517	22.320	0.058	12.972
WiLoR (Potamias et al., 2025)	22.012	11.873	71.527	17.358	0.075	24.091
EgoForce‡ (Millerdurai et al., 2026)	22.388	14.322	61.761	21.073	0.069	24.158
OmniHands (Lin et al., 2026)	29.674	14.203	51.505	24.580	0.087	45.312
HaWoR (Zhang et al., 2025a)	45.357	26.375	158.062	43.325	0.149	19.789
ViDiHand (Wang et al., 2026)	21.668	9.821	12.407	14.642	0.047	3.183
ACE-Ego-Hand (Liu et al., 2026a)	16.615	8.756	11.014	**12.208**	**0.028**	**2.770**
Ours (Online)	19.706	7.289	54.131	16.295	0.047	13.938
Ours (Online, p-free)	**16.413**	**7.025**	**10.844**	12.621	0.030	13.938
HOT3D						
HaMeR (Pavlakos et al., 2024)	68.314	21.455	59.077	49.636	0.102	23.632
WildHands (Prakash et al., 2024)	52.791	28.946	111.438	53.933	0.157	22.885
WiLoR (Potamias et al., 2025)	30.966	19.980	72.978	25.746	0.098	17.976
EgoForce‡ (Millerdurai et al., 2026)	43.960	25.709	83.521	37.144	0.130	38.342
OmniHands (Lin et al., 2026)	63.281	22.682	68.437	49.120	0.133	69.510
HaWoR (Zhang et al., 2025a)	71.396	66.031	327.294	79.350	0.262	23.872
ViDiHand (Wang et al., 2026)	21.514	11.383	14.953	15.829	0.040	3.741
ACE-Ego-Hand (Liu et al., 2026a)	**13.535**	6.768	**6.067**	**8.362**	**0.026**	**3.426**
Ours (Online)	23.409	7.041	55.885	18.172	0.059	14.154
Ours (Online, p-free)	16.160	**6.312**	15.470	10.334	0.041	14.154

Table 2: **Evaluation of World-space Hand Motion Estimation.**

Method	W-MPJPE ↓[mm]	WA-MPJPE ↓[mm]	AccEr ↓[m/s^2]	MRRPE ↓[mm]
HaMeR-SLAM (Pavlakos et al., 2024)	232.48	59.35	10.82	163.46
WiLoR-SLAM (Potamias et al., 2025)	183.72	50.03	8.80	122.42
HMP-SLAM (Duran et al., 2024)	166.21	41.62	7.13	131.33
Dyn-HaMR (Yu et al., 2025)	734.31	83.63	6.05	189.09
HaWoR (Zhang et al., 2025a)	171.61	51.98	11.35	112.36
StableHand (Zeng et al., 2026)	124.59	32.66	**4.67**	**93.71**
Ours (Online)	**108.80**	**25.82**	16.91	120.77

predicts surface contact but also closes the hand–scene reasoning loop by refining the reconstructed hand.

3.5 Training Objectives

The complete objective combines scene reconstruction, MANO reconstruction, and contact supervision:

$$\mathcal{L} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{mano}}\mathcal{L}_{\text{mano}} + \lambda_{\text{contact}}\mathcal{L}_{\text{contact}}. \tag{3}$$

We retain the original objectives of the scene reconstruction backbone in $\mathcal{L}_{\text{rec}}$, including dense geometry, camera pose, confidence, and RGB supervision. The MANO objective supervises joint rotations, shape, and camera-coordinate translation, together with 3D joints and vertices in both root-relative and camera coordinate systems.

Because contact vertices are sparse, $\mathcal{L}_{\text{contact}}$ uses a focal loss over valid MANO vertices. We further partition the hand surface into six anatomical regions according to the dominant MANO skinning joints. A region is positive if any of its vertices is in contact, leading to $\mathcal{L}_{\text{contact}} = \mathcal{L}_{\text{focal}} + \lambda_{\text{part}}\mathcal{L}_{\text{part}}$.

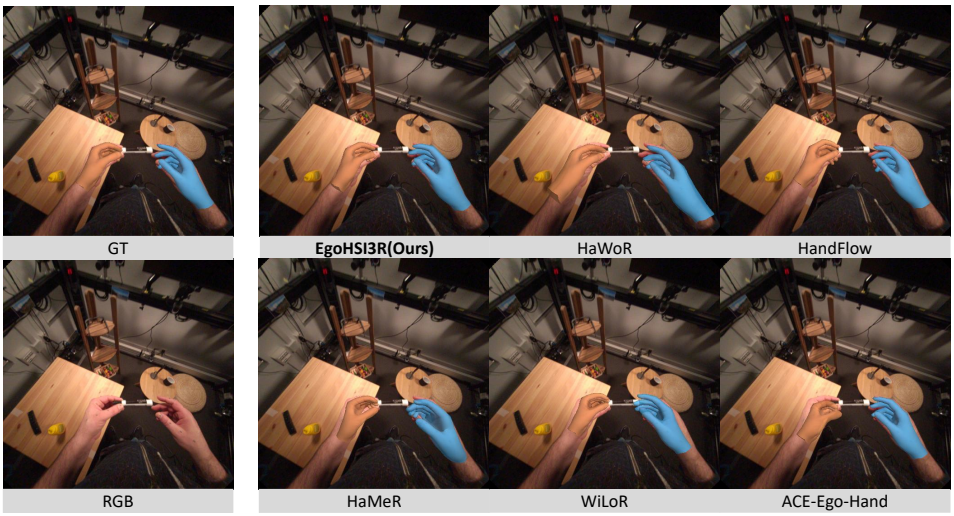

Figure 4: **Qualitative comparison of camera-space hand reconstruction.** The predicted hand meshes are overlaid on the input egocentric frames. Our method produces accurate articulated hand poses and consistent alignment with the observed images.

4 EXPERIMENTS

We use six egocentric hand–object datasets with complementary supervision. The model is trained with the available scene geometry, camera, MANO, and contact annotations from each dataset. We evaluate camera-space hand reconstruction on ARCTIC (Fan et al., 2023) and HOT3D (Banerjee et al., 2025), world-space hand reconstruction on ARCTIC, contact estimation on ARCTIC and DexYCB (Chao et al., 2021), and egocentric scene reconstruction on H2O (Kwon et al., 2021). The dataset composition and runtime comparison are provided in Tab. 6 and Tab. 8 in the Appendix.

4.1 CAMERA-SPACE HAND RECONSTRUCTION

Baselines. We compare with representative monocular and egocentric hand reconstruction methods, including HaMeR (Pavlakos et al., 2024), WildHands (Prakash et al., 2024), WiLoR (Potamias et al., 2025), EgoForce (Millerdurai et al., 2026), OmniHands (Lin et al., 2026), HaWoR (Zhang et al., 2025a), ViDiHand (Wang et al., 2026), and ACE-Ego-Hand (Liu et al., 2026a).

Evaluation Protocols. We evaluate camera-space hand reconstruction on ARCTIC (Fan et al., 2023) and HOT3D (Banerjee et al., 2025) following the evaluation protocols used by the compared methods. We report MPJPE-p, PA-p, 2D projection error $\mathrm{EPE}_{\mathrm{2D\text{-}p}}$, global orientation error (GO-p), camera translation error (CT-p), and temporal jitter. Lower values are better for all metrics. A GT hand is evaluated if at least one projected joint lies inside the image; missing or 2D-unmatched predictions are penalized even with GT boxes. Since causal online methods cannot use future observations to complete missing motion, we also report *p-free* results on successfully matched predictions to measure conditional reconstruction accuracy. Note that this is a matched-only diagnostic, not an evaluation over GT-visible hands.

Results. As shown in Table 1, our *p*-free setting achieves the best PA-p on both datasets, with 7.025 and 6.312 mm, respectively. It also achieves the best MPJPE-p on ARCTIC, the second-best MPJPE-p on HOT3D, and the second-best GO-p on both datasets. This demonstrates high conditional reconstruction accuracy, while the standard results additionally reflect missed or unmatched predictions under partial visibility. Figure 4 provides qualitative comparisons on HOT3D, where our predictions show accurate alignment with the observed hand configurations. Our temporal jitter remains higher than methods designed specifically for motion smoothing, indicating room for improving temporal stability without sacrificing online operation.

4.2 WORLD-SPACE HAND RECONSTRUCTION

Baselines. We compare against single-frame hand estimators coupled with SLAM (Teed & Deng, 2021), including HaMeR-SLAM, WiLoR-SLAM, and HMP-SLAM, as well as world-space hand

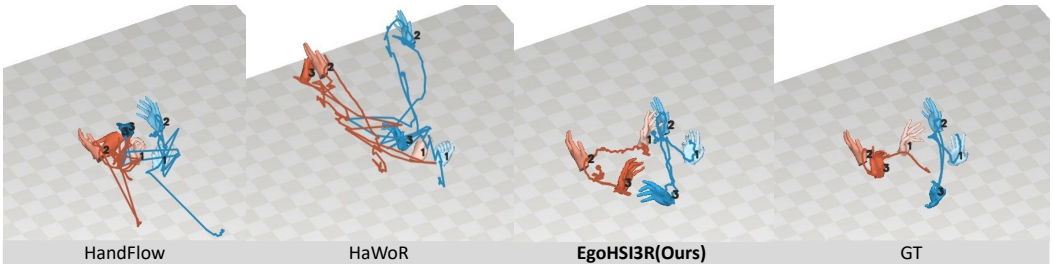

Figure 5: **Qualitative comparison of world-space hand reconstruction.** Existing methods (Xu et al., 2026b; Zhang et al., 2025a) exhibit inaccurate hand trajectories and spatial inconsistencies, while EgoHSI3R produces world-space hand motions closer to the ground truth. The numbers indicate timestamps at 0.83s, 2.50s, and 4.17s, respectively.

Table 3: **Contact estimation** on ARCTIC (Fan et al., 2023) and DexYCB (Chao et al., 2021). We report per-vertex contact precision, recall, and F1-score.

	ARCTIC			DexYCB		
Method	Prec.↑	Rec.↑	F1↑	Prec.↑	Rec.↑	F1↑
HACO (Jung & Lee, 2025)	43.609	**65.181**	**48.917**	43.993	**61.304**	**49.115**
Ours	**52.152**	45.015	44.298	**51.619**	50.376	48.197

motion methods Dyn-HaMR (Yu et al., 2025), HaWoR (Zhang et al., 2025a), and StableHand (Zeng et al., 2026).

Evaluation Protocols. We evaluate world-space hand motion on ARCTIC (Fan et al., 2023). W-MPJPE measures absolute world-coordinate joint error, WA-MPJPE evaluates the trajectory after global alignment, AccEr measures temporal acceleration error, and MRRPE evaluates the relative placement of the two hands. Lower values indicate better performance.

Results. Table 2 shows that our online model achieves the best W-MPJPE and WA-MPJPE, reducing them to 108.80 mm and 25.82 mm. Compared with the strongest baseline, StableHand, this corresponds to improvements of 15.79 mm and 6.84 mm, respectively. Figure 5 provides qualitative comparisons on ARCTIC, where EgoHSI3R maintains temporally consistent world-space hand poses and recovers more accurate world-space hand motions. The higher AccEr and MRRPE are partly attributable to causal inference, which precludes future-frame smoothing and sequence-level correction of bilateral trajectories under intermittent visibility.

4.3 Contact Evaluation

Baselines. We compare with HACO (Jung & Lee, 2025), a representative method for dense hand–object contact estimation on the MANO surface.

Evaluation Protocols. We evaluate binary contact prediction on ARCTIC (Fan et al., 2023) and DexYCB (Chao et al., 2021) using the standard 778-vertex MANO topology. Precision, recall, and F1-score are computed over valid annotated hand vertices with a threshold of 0.5; higher values indicate better performance.

Results. As reported in Table 3, our method improves contact precision over HACO on both ARCTIC and DexYCB, reaching 52.152 and 51.619, respectively. Its lower recall leads to lower F1-scores, suggesting that the current model produces more conservative contact predictions. These results show improved reliability for predicted contact vertices while leaving recall as the main direction for further improvement.

4.4 Egocentric 3D Reconstruction

Baselines. We compare with EgoMono4D (Yuan et al., 2025) and ReViV (Lyu et al., 2026) as offline reconstruction baselines, and with TTT3R (Chen et al., 2026b) as an online recurrent reconstruction method.

Table 4: **Evaluation of Egocentric 3D Reconstruction** on the H2O (Kwon et al., 2021) dataset.

Method	**Optim.**	**Onl.**	Abs Rel ↓	$\delta < 1.25$ ↑	ATE ↓[mm]	RPEt ↓[mm]	RPEr ↓[°]
EgoMono4D (Yuan et al., 2025)	✓	×	0.049	<u>0.976</u>	7.149	2.606	0.781
ReViV (Lyu et al., 2026)	×	×	0.260	0.608	**2.383**	**0.865**	<u>0.075</u>
TTT3R (Chen et al., 2026b)	×	✓	<u>0.045</u>	0.965	3.578	1.271	**0.074**
Ours	×	✓	**0.027**	**0.993**	<u>2.573</u>	<u>0.923</u>	0.079

Table 5: **Ablation study.** The two variants with a contact head use the same contact-head warm-up and joint training procedure.

Variant	WiLoR FT	Contact head	Contact feedback	ARCTIC MPJPE [mm]↓	HOT3D MPJPE [mm] ↓	ARCTIC W-MPJPE [mm] ↓	ARCTIC F1 ↑	DexYCB F1 ↑
Base model	–	–	–	22.916	28.974	127.855	–	–
Hand-only baseline	✓	–	–	19.952	23.653	114.067	–	–
Without contact feedback	✓	✓	–	20.276	24.114	113.296	43.349	47.657
EgoHSI3R (full)	✓	✓	✓	**19.706**	**23.409**	**108.801**	**44.298**	**48.197**

Evaluation Protocols. We evaluate scene depth and camera tracking on H2O (Kwon et al., 2021). Depth quality is measured by absolute relative error (Abs Rel) and threshold accuracy $\delta < 1.25$. Camera trajectories are evaluated using absolute trajectory error (ATE) and relative translation and rotation errors (RPEt and RPEr). Abs Rel, ATE, RPEt, and RPEr are minimized, whereas $\delta < 1.25$ is maximized.

Results. As shown in Table 4, our method achieves the best depth accuracy, with an Abs Rel of 0.027 and $\delta < 1.25$ of 0.993. It also ranks second in ATE and RPEt, closely approaching the offline ReViV method, while maintaining causal online inference without test-time optimization. Its RPEr of $0.079°$ is comparable to the best online result.

4.5 Ablation Study

Table 5 examines the contributions of hand fine-tuning, contact prediction, and contact feedback.

Fine-tuning WiLoR substantially improves hand reconstruction over the base model, reducing ARCTIC MPJPE from 22.916 mm to 19.952 mm and W-MPJPE from 127.855 mm to 114.067 mm. Because the WiLoR prior was originally trained on right hands, fine-tuning substantially improves bilateral hand reconstruction.

Without contact feedback, the jointly trained contact head achieves F1 scores of 43.349 on ARCTIC and 47.657 on DexYCB. Its ARCTIC and HOT3D MPJPE are slightly higher than those of the hand-only baseline, whereas ARCTIC W-MPJPE improves modestly from 114.067 to 113.296 mm. This mixed outcome may reflect optimization instability introduced by jointly training the hand reconstruction and contact prediction objectives.

With contact feedback enabled, ARCTIC and HOT3D MPJPE decrease by 0.570 and 0.705 relative to the no-feedback variant, while W-MPJPE decreases by 4.495 mm. Contact F1 also improves on both datasets. Since the final two variants share the same warm-up and joint training procedure, these gains support the value of incorporating contact cues into the reconstruction process.

5 Conclusion

We presented **EgoHSI3R**, an online feed-forward framework for unified 4D hand–scene interaction reconstruction from egocentric video. It maintains a recurrent scene representation that provides a shared world space for hand and scene reconstruction, while scene-conditioned hand prompts and interaction-aware contact modeling couple hand reconstruction with the evolving scene. Together, these designs enable joint reconstruction of camera motion, persistent scene geometry, world-space bilateral MANO hands, and dense hand-scene contact without future observations or test-time optimization. Experiments across multiple egocentric benchmarks demonstrate state-of-the-art world-space hand reconstruction while retaining competitive camera and scene reconstruction accuracy. These results highlight the benefit of jointly modeling hands, scene geometry, and physical interaction within a unified recurrent representation for egocentric interaction reconstruction.

AI USE STATEMENT

We used AI tools only to improve the readability of the writing and to assist with identifying potential bugs in the code. All research ideas, methods, experimental results, and conclusions were developed and verified by the authors. The authors reviewed all AI-assisted content and take full responsibility for the final manuscript and code.

REPRODUCIBILITY STATEMENT

We provide the experimental settings and implementation details to facilitate reproducibility in Sec. 3 and Sec. 4, including the model architecture, training objectives, datasets, and evaluation protocols. We are also preparing the code for public release.

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 *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (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 *Proc. of European Conference on Computer Vision (ECCV)*, 2026.

Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. ContactPose: A dataset of grasps with object contact and hand pose. In *Proc. of European Conference on Computer Vision (ECCV)*, August 2020.

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

Lin-Zhuo Chen, Jian Gao, Shangzhan Zhang, Yihang Chen, Ka Leong Cheng, Yipengjing Sun, Liangxiao Hu, Nan Xue, Xing Zhu, Yujun Shen, Yao Yao, and Yinghao Xu. LingBot-Map: Geometric context transformer for streaming 3d reconstruction, 2026a.

Xingyu Chen, Zhuheng Song, Xiaoke Jiang, Yaoqing Hu, Junzhi Yu, and Lei Zhang. Handos: 3d hand reconstruction in one stage. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. TTT3r: 3d reconstruction as test-time training. In *Proc. of International Conference on Learning Representations (ICLR)*, 2026b.

Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Gerard Pons-Moll. Human3r: Everyone everywhere all at once. In *Proc. of International Conference on Learning Representations (ICLR)*, 2026c.

Haoye Dong, Aviral Chharia, Wenbo Gou, Francisco Vicente Carrasco, and Fernando De la Torre. Hamba: Single-view 3d hand reconstruction with graph-guided bi-scanning mamba. In *Proc. of Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 2127–2160, 2024.

Enes Duran, Muhammed Kocabas, Vasileios Choutas, Zicong Fan, and Michael J. Black. Hmp: Hand motion priors for pose and shape estimation from video. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

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 *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Haiwen Feng, Junyi Zhang, Qianqian Wang, Yufei Ye, Pengcheng Yu, Michael J. Black, Trevor Darrell, and Angjoo Kanazawa. St4RTrack: Simultaneous 4d reconstruction and tracking in the world. In *Proc. of International Conference on Computer Vision (ICCV)*, 2025.

Hongming Fu, Wenjia Wang, Xiaozhen Qiao, Rolandos Alexandros Potamias, Taku Komura, Shuo Yang, Zheng Liu, and Bo Zhao. EgoGrasp: World-space hand-object interaction estimation from egocentric videos, 2026.

Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. ContactOpt: Optimizing contact to improve grasps. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Andela Ilic, Rachel Schuchert, Yijing Jiang, and Christian Holz. EgoPHI: Estimating 3d hand-object contact and force from egocentric vision. In *Proc. of European Conference on Computer Vision (ECCV)*, 2026.

Daniel Sungho Jung and Kyoung Mu Lee. Learning dense hand contact estimation from imbalanced data. In *Proc. of Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Daniel Sungho Jung and Kyoung Mu Lee. Training-free dense hand contact estimation with multimodal large language models, 2026.

Jay Karhade, Nikhil Keetha, Yuchen Zhang, Tanisha Gupta, Akash Sharma, Sebastian Scherer, and Deva Ramanan. Any4d: Unified feed-forward metric 4d reconstruction. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14578–14589, 2026.

Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proc. of International Conference on Computer Vision (ICCV)*, pp. 10138–10148, October 2021.

Mengfei Li, Peng Li, Zheng Zhang, Jiahao Lu, Chengfeng Zhao, Wei Xue, Qifeng Liu, Sida Peng, Wenxiao Zhang, Wenhan Luo, Yuan Liu, and Yike Guo. Unish: Unifying scene and human reconstruction in a feed-forward pass. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Dixuan Lin, Yuxiang Zhang, Mengcheng Li, Wei Jing, Qi Yan, Qianying Wang, Yebin Liu, and Hongwen Zhang. Omnihands: Robust motion capture of interactive hands via a versatile transformer. *ACM Transactions on Graphics*, 45(4), 2026. doi: 10.1145/3807943.

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, 2026a.

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 *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21740–21751, 2024.

Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21013–21022, June 2022.

Yuzheng Liu, Siyan Dong, Shuzhe Wang, Yingda Yin, Yanchao Yang, Qingnan Fan, and Baoquan Chen. SLAM3R: Real-time dense scene reconstruction from monocular RGB videos. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Zhizheng Liu, Joe Lin, Wayne Wu, and Bolei Zhou. Joint optimization for 4d human-scene reconstruction in the wild. In *Proc. of International Conference on Learning Representations (ICLR)*, 2026b.

Yihang Luo, Shangchen Zhou, Yushi Lan, Xingang Pan, and Chen Change Loy. 4RC: 4d reconstruction via conditional querying anytime and anywhere. In *Proc. of the International Conference on Machine Learning (ICML)*, 2026.

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 *Proc. of European Conference on Computer Vision (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 *Proc. of SIGGRAPH*, 2026.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (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 *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Aditya Prakash, Ruisen Tu, Matthew Chang, and Saurabh Gupta. 3d hand pose estimation in everyday egocentric images. In *Proc. of European Conference on Computer Vision (ECCV)*, 2024.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *Proc. of SIGGRAPH*, 36(6), November 2017.

Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. *Proc. of Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. In *International Conference on 3D Vision (3DV)*, pp. 78–89, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3d vision made easy. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709, 2024a.

Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *Proc. of European Conference on Computer Vision (ECCV)*, pp. 467–487. Springer, 2024b.

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, 2026.

Congrong Xu, Huachen Gao, Xingyu Chen, Yuliang Xiu, Jun Gao, and Anpei Chen. R^3: 3d reconstruction via relative regression, 2026a.

Mingxi Xu, Bowen Duan, Yi Gu, Zhengyang Shen, Renjing Xu, and Yutao Yue. Handflow: Fully generative 4d hand recovery with flow matching, 2026b.

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 *Proc. of International Conference on Computer Vision (ICCV)*, 2021.

Yufei Ye, Yao Feng, Omid Taheri, Haiwen Feng, Shubham Tulsiani, and Michael J. Black. Predicting 4d hand trajectory from monocular videos. In *International Conference on 3D Vision (3DV)*, 2026.

Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. Dyn-hamr: Recovering 4d interacting hand motion from a dynamic camera. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025.

Chengbo Yuan, Geng Chen, Li Yi, and Yang Gao. Self-supervised monocular 4d scene reconstruction for egocentric videos. In *Proc. of International Conference on Computer Vision (ICCV)*, pp. 8863–8874, October 2025.

Shuai Yuan, Yantai Yang, Xiaotian Yang, Xupeng Zhang, Zhonghao Zhao, Lingming Zhang, and Zhipeng Zhang. InfiniteVGGT: Visual geometry grounded transformer for endless streams, 2026.

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.

Chuhan Zhang, Guillaume Le Moing, Skanda Koppula, Ignacio Rocco, Liliane Momeni, Junyu Xie, Shuyang Sun, Rahul Sukthankar, Joëlle K. Barral, Raia Hadsell, Zoubin Ghahramani, Andrew Zisserman, Junlin Zhang, and Mehdi S. M. Sajjadi. Efficiently reconstructing dynamic scenes one d4rt at a time. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3R: A simple approach for estimating geometry in the presence of motion. In *Proc. of International Conference on Learning Representations (ICLR)*, 2025b.

Zijie Zhu, Weiren Cai, Yizhou Wang, Zhenjie Yang, Yide Liu, Jiahao Chen, and Guanqi He. MINT: A unified model for world-space camera and hand motion estimation from scalable egocentric pipeline supervision, 2026.

A Appendix

A.1 Datasets and Data Labeling

As shown in Table 6, we train on six sources: HOT3D (Banerjee et al., 2025), ARCTIC (Fan et al., 2023), TACO (Liu et al., 2024), H2O (Kwon et al., 2021), HOI4D (Liu et al., 2022), and DexYCB (Chao et al., 2021). Each processed frame provides RGB input and available scene, camera, hand, and contact supervision, with explicit validity masks. Splits are defined at the recording level: windows, frames, and camera streams from the same recording remain in the same split.

Coordinate and validity conventions. The processed representation uses meters and OpenCV camera axes ($+x$ right, $+y$ down, $+z$ forward), with separate right/left MANO slots. Intrinsics and bounding boxes are transformed consistently with image rectification, resizing, cropping, and padding. Native side-specific MANO layers preserve left/right geometry. Parameter validity, visible-hand validity, and contact validity are distinguished, so unavailable supervision is not silently converted into a negative label. Camera poses are exported as camera-to-world transforms in the dataset's reference frame; the temporal pose is fixed within each DexYCB camera stream.

Table 6: Supervision available in the processed training sources. A star marks generated supervision: rendered/completed depth or geometry-derived contact. Availability does not imply that every frame or hand is valid.

Dataset	RGB	Depth	Pose	Hand	Contact
HOT3D (Aria)	✓	✓*	✓	✓	✓*
ARCTIC	✓	✓*	✓	✓	✓*
TACO	✓	✓	✓	✓	✓*
H2O	✓	✓	✓	✓	✓*
HOI4D	✓	✓	✓	✓	✓*
DexYCB	✓	✓	✓	✓	✓*

Depth labeling. ARCTIC depth is camera-z depth rendered from the annotated hand and articulated-object meshes; background without mesh support is invalid. HOT3D uses hybrid LingBot-Depth completion: the foreground branch receives native mesh depth together with sparse SLAM prompts, while the static background branch receives SLAM-only prompts. Foreground/background regions are selected using the exported segmentation and combined into the final depth map. The stored source-dependent confidence map is not used to reweight the final HOT3D depth supervision in these runs. TACO, H2O, HOI4D, and DexYCB use their processed camera-aligned sensor depth, with dataset specific scale conversion and invalid-depth masks.

Contact labeling. For MANO vertex $\mathbf{v}$ and the relevant posed object surface $\mathcal{S}$, the generated binary label is

$$y(\mathbf{v}) = \mathbb{1}[d(\mathbf{v}, \mathcal{S}) \leq 0.01 \text{ m}]\,. \tag{4}$$

This 1 cm geometric labeling threshold is distinct from the 0.5 probability threshold used to classify predicted contacts. ARCTIC uses the articulated object; TACO uses the tool and target meshes; H2O uses the annotated object; DexYCB uses the grasped object. HOI4D uses unsigned point-to-triangle distance to valid posed object parts. HOT3D approximates the surface with deterministic object-vertex and triangle-centroid samples, capped at 30,000 samples per object. Contact is supervised only when the required hand and object annotations are valid. These labels describe geometric proximity, rather than force measurements or independently annotated physical contact.

A.2 Evaluation Protocol and Data Splits

Split unit and model selection. We use a frozen six-source split defined over complete recordings, rather than splitting frames or temporal windows at random. All windows of a recording and all its camera streams remain in the same partition. Table 7 gives the available processed sequences; these counts are distinct from the evaluation-clip counts below. Validation is used for checkpoint and inference-setting selection; test data are reserved for final scoring.

Table 7: Frozen processed-data split sizes. DexYCB counts camera streams; all other rows count processed sequences.

Dataset	Train	Val	Test
HOT3D	103	11	22
ARCTIC	235	26	33
TACO	806	90	1,313
H2O	124	14	46
HOI4D	1,683	–	–
DexYCB	6,400	320	1,280

Table 8: Efficiency of the measured RGB-to-world pipelines. FPS is median end-to-end throughput; memory is the maximum allocated CUDA memory across measured clips in GiB.

Method	Params (M)	FPS	Memory (GiB)
WiLoR+ViPE	708.954	7.545	6.142
HaWoR	1170.122	7.756	8.099
HandFlow	907.017	7.645	10.063
EgoHSI3R	1470.250	**10.070**	6.650

A.3 Efficiency Analysis

Baselines and measurement scope. We compare the final feedback model against a local WiLoR (Potamias et al., 2025)+ViPE pipeline, HaWoR (Zhang et al., 2025a), and HandFlow (Xu et al., 2026b). WiLoR+ViPE is a local engineering composition, rather than a reproduction of an unspecified SLAM backend. Measurements use one NVIDIA H200, the same 150-frame ARCTIC input clips, and a 512-pixel input long edge. Method-specific internal resizing is retained. One clip is used for warmup and ten clips for measurement. Model loading and final visualization are excluded. Necessary intermediate file I/O in HaWoR is included.

Results. As shown in Table 8, the final model achieves 10.070 FPS end to end and 15.285 FPS for model inference alone. These are different timing scopes: the former includes image loading, preprocessing, detection, crop preparation, inference, and output decoding. Model-only throughput is therefore not substituted into the end-to-end comparison. Peak allocated memory is 6.650 GiB, below HaWoR and HandFlow and above the local WiLoR+ViPE composition. The parameter count includes the detector. For serial multi-process baselines, memory is the maximum phase peak, rather than the sum of phase peaks.

A.4 Training Loss Details

We distinguish scene reconstruction, articulated hand reconstruction, and contact prediction. The final training objective is

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{MANO}} \\ + 0.3\big(\mathcal{L}_{\text{focal}} + 0.1\mathcal{L}_{\text{part}}\big),\end{aligned} \tag{5}$$

where $\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{points}} + \mathcal{L}_{\text{pose}} + \mathcal{L}_{\text{RGB}}$ comes from CUT3R.

MANO parameter supervision. Each valid hand has 16 joint rotation matrices R, shape coefficients $\boldsymbol{\beta}$, and native MANO translation $\mathbf{t}$. The side-specific differentiable MANO layer maps these parameters to camera space joints J and 778 vertices V. Define $\ell(A, B) = |A|^{-1} \sum_k |A_k - B_k|$, a mean elementwise L_1 error. The parameter loss is

$$\mathcal{L}_{\text{param}} = \ell(\hat{R}, R) + 0.1\ell(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) + \ell(\hat{\mathbf{t}}, \mathbf{t}). \tag{6}$$

Rotation supervision compares matrix entries, rather than axis-angle vectors or geodesic distances. Native left and right MANO layers are used without replacing one side with the other's geometry.

MANO joint and surface supervision. Let $\mathbf{r} = J_0$ and $\hat{\mathbf{r}} = \hat{J}_0$ denote the reference and predicted wrists. Camera-space terms supervise absolute hand placement:

$$\mathcal{L}_{\text{cam}} = \ell(\hat{J}, J) + \ell(\hat{V}, V). \tag{7}$$

Wrist-relative terms separately constrain articulation and shape:

$$\begin{aligned}\mathcal{L}_{\text{rel}} = \ell(\hat{J} - \hat{\mathbf{r}}, J - \mathbf{r}) \\ + \ell(\hat{V} - \hat{\mathbf{r}}, V - \mathbf{r}).\end{aligned} \tag{8}$$

The same wrist is subtracted from every vertex or joint of a hand. All four 3D terms have unit weight and use meter-valued coordinates; they are not rescaled by the scene normalization factor s. Parameter, joint, and vertex losses require image validity, a visible-hand mask, and valid MANO parameters. Each term is averaged over its selected entries and then over views. Batches with no valid hands contribute a zero MANO loss.

MANO image-plane supervision. With the reference intrinsics, perspective projection is

$$\pi_K([x, y, z]^\top) = [f_x x/z + c_x,\ f_y y/z + c_y]^\top. \tag{9}$$

The joint and vertex reprojection terms are

$$\begin{aligned}\mathcal{L}_{J2D} &= \ell\big(\pi_K(\hat{J}), \pi_K(J)\big), \\ \mathcal{L}_{V2D} &= \ell\big(\pi_K(\hat{V}), \pi_K(V)\big).\end{aligned} \tag{10}$$

These errors are measured in pixels. Only finite prediction/reference pairs with both depths above 0.01 m are used, and the projected reference point must lie strictly inside the image. Safe clamping before division prevents invalid gradients near the camera plane. To avoid unstable projection supervision before coarse geometry is learned, both terms are gated by $g = \mathbb{1}[\text{stopgrad}(\ell(\hat{V}, V)) < 1.0]$, using the camera-space vertex loss. The full final-stage hand objective is

$$\begin{aligned}\mathcal{L}_{\text{MANO}} = \mathcal{L}_{\text{param}} + \mathcal{L}_{\text{cam}} + \mathcal{L}_{\text{rel}} \\ + 0.005g\big(\mathcal{L}_{J2D} + \mathcal{L}_{V2D}\big).\end{aligned} \tag{11}$$

Dense contact and hand-part supervision. The contact branch predicts one logit for each of the 778 MANO vertices. With binary target y_v and probability $p_v = \sigma(z_v)$, define $p_{t,v} = y_v p_v + (1 - y_v)(1 - p_v)$. The vertex loss is

$$\mathcal{L}_{\text{focal}} = -\frac{1}{778}\sum_{v=1}^{778} 0.5(1 - p_{t,v})^2 \log p_{t,v}. \tag{12}$$

The implementation evaluates this expression through numerically stable binary cross-entropy with logits. Vertices are assigned to six coarse regions (palm and five fingers) using dominant MANO skinning weights. For region $\mathcal{V}_k$, the pooled logit and target are

$$\bar{z}_k = \log \sum_{v \in \mathcal{V}_k} \exp z_v - \log |\mathcal{V}_k|, \qquad \bar{y}_k = \max_{v \in \mathcal{V}_k} y_v. \tag{13}$$

The part loss averages binary cross-entropy over the six regions. Contact loss is averaged over valid visible hand-frames, using contact validity, image validity, and the hand mask.

A.5 Additional Qualitative Examples

Figs. 6–7 show some additional qualitative results of contact prediction and depth.

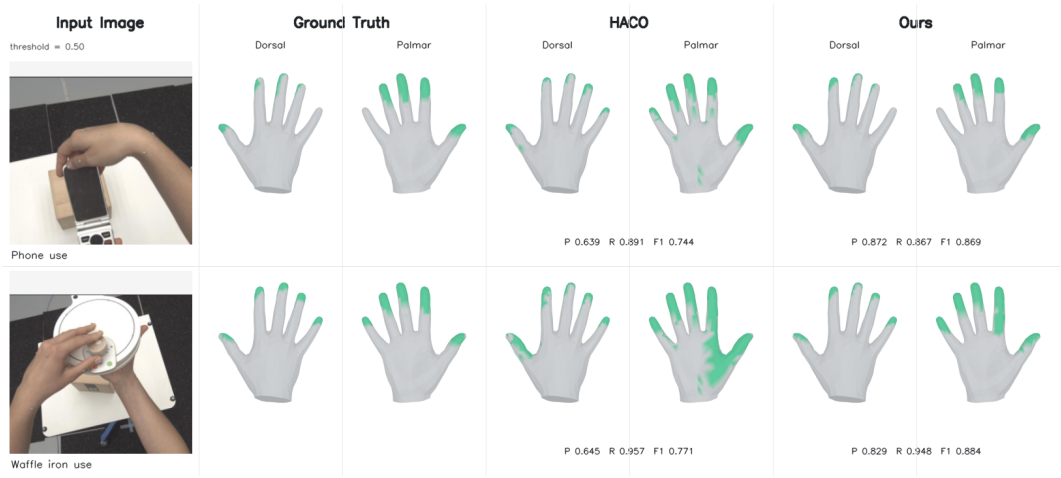

Figure 6: Contact visualization on ARCTIC: phone use (frame 332 of `s05/phone_use_04`) and waffle-iron use (frame 308 of `s05/waffleiron_use_01`). Both methods use exported GT hand boxes and contact threshold 0.5. Green marks predicted/reference contact on canonical hand surfaces, viewed from dorsal and palmar sides; these panels visualize vertex labels, not recovered hand pose. The examples were selected for clear crops and a precision advantage of our predictions, so their per-frame scores are illustrative rather than an unbiased dataset-level comparison. No contact postprocessing is applied.

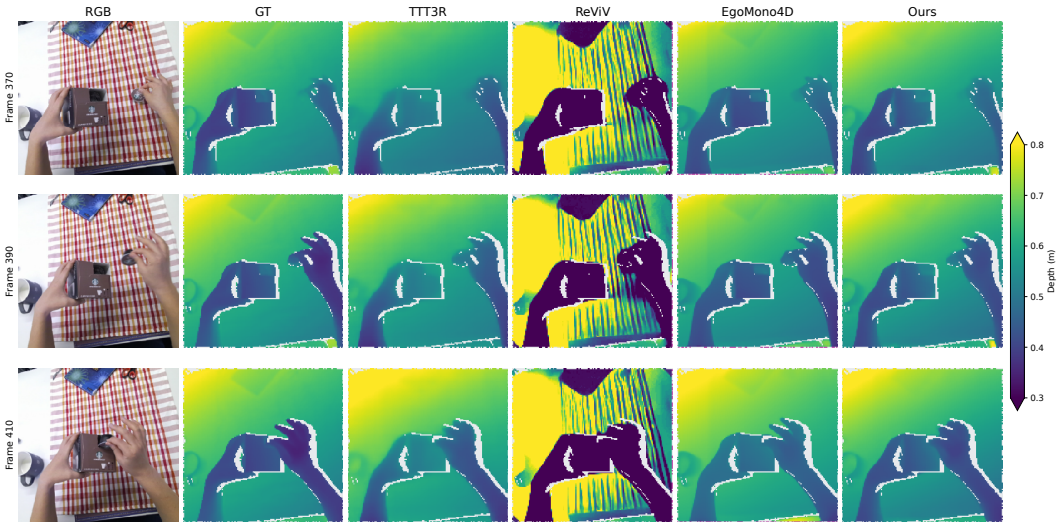

Figure 7: H2O depth examples from `subject4_ego/h1/0`, frames 370, 390, and 410 of a 60-frame clip. Each method uses one median depth scale for the whole clip, with the same reference camera rays and common field of view. Colors share a fixed depth range; unsupported reference regions are masked. The example is the fixed manifest index 0, not a selection by relative model accuracy. The figure shows depth maps, not a fused global scene reconstruction.