arXiv:2609.35743v1 [cs.CV] 28 Sep 2026

INFINIHAND: STREAMING WORLD-SPACE HAND MOTION ESTIMATION FROM EGOCENTRIC VIDEO

Kerui Ren[1,2] **Kaiwen Song**[1,3] **Weiguang Zhao**[1,4] **Yuxi Wang**[5] **Yufei Liu**[2]
Bo Dai[6] **Haoyu Guo**[1] **Chunhua Shen**[7,1] **Mulin Yu**[1] **Tao Lu**[1] **Junting Dong**[1]
[1]Shanghai Artificial Intelligence Laboratory, [2]Shanghai Jiao Tong University,
[3]University of Science and Technology of China, [4]University of Liverpool,
[5]Nanyang Technological University, [6]The University of Hong Kong, [7]Zhejiang University

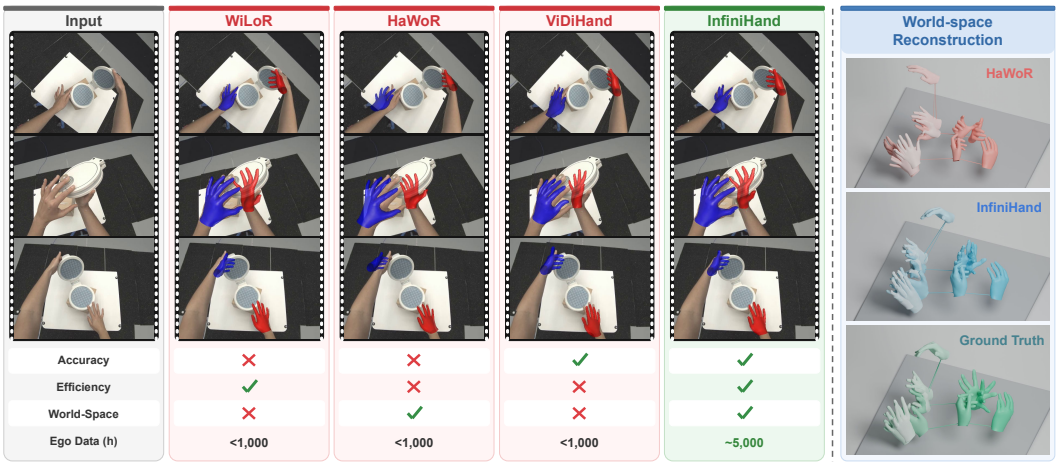

Figure 1: InfiniHand is a streaming feed-forward framework for accurate and efficient world-space hand estimation, pretrained on approximately 5,000 hours of egocentric video. Project page: **https://infinihand.github.io/**.

ABSTRACT

World-space hand motion estimation from egocentric video requires recovering 3D articulated hand geometry while tracking camera egomotion. Existing approaches heavily rely on cascading independent hand pose estimators and SLAM systems, resulting in error accumulation, complex pipelines, and severe computational overhead. To address these limitations, we present InfiniHand, an end-to-end streaming feed-forward framework that jointly estimates MANO parameters, camera trajectories, and hand locations directly from uncalibrated egocentric video. InfiniHand integrates persistent spatiotemporal memory with hand-centered visual features, explicitly coupling camera motion with local hand geometry within a unified architecture. We train InfiniHand in two progressive stages by first learning robust camera-space hand priors and then extending to streaming world-space reconstruction. To support this process, we aggregate a pretraining corpus of approximately 5,000 hours of egocentric data across multiple public datasets. Extensive evaluations demonstrate that InfiniHand outperforms state-of-the-art baselines on in-domain benchmarks, achieving a 21.4% reduction in ARCTIC PA-p compared to ViDiHand while substantially mitigating world-space drift. Furthermore, InfiniHand generalizes robustly to in-the-wild videos and operates at 11.19 FPS, delivering more than twice the throughput of HaWoR.

1 INTRODUCTION

Egocentric video captures diverse human movements and complex hand–object interactions from the first-person perspective, where recovering hand motion in a shared world coordinate system transforms raw visual observations into actionable geometric demonstrations for embodied learning (Hoque et al., 2025). These demonstrations enable training world action models via human-to-robot motion transfer (Li et al., 2026a) and facilitate in-context robot imitation conditioned on retrieved human examples (Papagiannis et al., 2025). However, unconstrained in-the-wild videos rarely come paired with the ground-truth 3D hand poses and camera trajectories necessary to anchor movement within physical environments. Bridging this gap demands a scalable framework for rapid, high-fidelity world-space hand reconstruction, which is a critical prerequisite for downstream embodied learning.

Conventional world-space hand motion estimation pipelines rely on separate models for hand localization, MANO (Romero et al., 2017) parameter prediction, and camera trajectory estimation. HaWoR (Zhang et al., 2025), for example, combines hand detection and tracking, a dedicated camera-space hand reconstruction network, and DROID-SLAM (Teed & Deng, 2021) with Metric3D (Yin et al., 2023) for metric camera motion estimation. Coordinating these components introduces substantial computational and engineering overhead, while detection jitter, camera tracking drift, and scale inconsistencies can propagate through the pipeline and compromise world-space reconstruction. Beyond these architectural constraints, established estimators such as WiLoR (Potamias et al., 2024) and HaWoR (Zhang et al., 2025) lack extensive pretraining on diverse, unconstrained egocentric videos, rendering generalization to complex interactions and rapid camera motion a persistent bottleneck. Collectively, these drawbacks hinder accurate and efficient hand motion reconstruction from in-the-wild egocentric videos.

To address these limitations, we present **InfiniHand**, a streaming feed-forward framework that jointly predicts hand locations, MANO (Romero et al., 2017) parameters, and camera trajectories within a unified architecture. Built upon a streaming 3D foundation model (Chen et al., 2026), InfiniHand establishes a shared spatiotemporal representation to couple global camera motion with local hand articulation. Specifically, as each frame arrives, the model leverages full-image geometric context to first predict hand masks for precise localization. Guided by these masks, it crops local geometric features and fuses them with hand-centered appearance features to regress MANO parameters. Concurrently, the model tracks camera poses in a streaming fashion, supplemented by a lightweight, sparse bundle adjustment (BA) for rapid post-optimization. To power this framework, we aggregate existing public egocentric datasets with MANO or 3D keypoint annotations, designing a dedicated data processing pipeline to filter out noise and convert diverse sources into a unified, high-quality format. Driven by a two-stage training strategy on this curated dataset, InfiniHand achieves rapid, robust world-space hand motion reconstruction, generalizing seamlessly even to complex, in-the-wild video sequences.

We summarize our primary contributions as follows:

- We propose a streaming feed-forward framework that unifies hand localization, MANO parameter prediction, and camera trajectory estimation, enabling fast and efficient world-space hand motion reconstruction.
- We aggregate and clean existing public egocentric datasets into a standardized, high-quality corpus, paired with a dedicated two-stage training scheme to effectively optimize the model.
- Extensive experiments demonstrate that InfiniHand achieves SOTA accuracy in world-space hand motion estimation with remarkable efficiency. Evaluations on in-the-wild videos further confirm its strong generalization in complex real-world scenarios.

2 RELATED WORK

Hand Motion Reconstruction. Hand motion reconstruction has evolved from isolated hand mesh recovery to modeling temporal interactions and trajectories. HaMeR (Pavlakos et al., 2024) leverages large transformers for single-image estimation, whereas WiLoR (Potamias et al., 2024) integrates localization with detailed mesh recovery in unconstrained images. Meanwhile, Hamba (Dong

et al., 2024) introduces graph-guided state-space modeling for joint spatial relations, and WildHands (Prakash et al., 2023) targets egocentric reconstruction. While these methods strengthen local hand estimation, they fail to jointly recover camera motion and world-space trajectories.

Beyond single-hand recovery, InterWild (Moon, 2023) decouples per-hand reconstruction from relative translation estimation to bridge domain gaps. OmniHands (Lin et al., 2024) leverages spatiotemporal reasoning to reconstruct interacting hands, while ViDiHand (Wang et al., 2026) adapts video diffusion priors with hand-overlay supervision for temporally coherent egocentric geometry. However, world-space motion recovery additionally requires disentangling hand movement from camera egomotion. To address this, current approaches like HaWoR (Zhang et al., 2025) rely on egocentric SLAM with motion infilling, whereas Dyn-HaMR (Yu et al., 2025) employs multi-stage optimization combining camera tracking and interacting-hand priors.

Streaming 3D Reconstruction. Reconstructing scene geometry and camera motion from video has traditionally relied on simultaneous localization and mapping (SLAM), where visual tracking is coupled with bundle adjustment. Classic frameworks like ORB-SLAM2 (Mur-Artal & Tardós, 2017) combine sparse feature tracking with keyframe-based loop closure, whereas DROID-SLAM (Teed & Deng, 2021) replaces handcrafted features with learned recurrent updates and differentiable dense bundle adjustment. Learning-augmented SLAM systems further enhance this pipeline by incorporating feed-forward geometric priors. For instance, MASt3R-SLAM (Murai et al., 2024) builds tracking and global optimization around two-view 3D reconstruction models, VGGT-SLAM (Maggio et al., 2025) constructs submaps via feed-forward predictions and aligns them through projective optimization with loop-closure constraints, and M^3 (Ren et al., 2026) augments multi-view foundation models with dense matching heads for monocular Gaussian splatting SLAM. While these hybrid pipelines benefit from learned priors, they still rely on explicit optimization for cross-view consistency. In contrast, purely feed-forward architectures maintain geometric context natively within the network without post-hoc optimization. For instance, LoGeR (Zhang et al., 2026) processes video chunks by combining sliding-window attention with test-time-training memory to preserve both fine details and long-range consistency. Similarly, LingBot-Map (Chen et al., 2026) deploys a streaming context transformer equipped with anchor context, a pose-reference window, and trajectory memory for incremental reconstruction over extended sequences.

3 METHOD

Fig. 2 presents the overall pipeline of InfiniHand, a streaming framework for joint estimation of articulated hand motion and camera parameters from egocentric video. Given an input RGB sequence $\mathcal{I} = \{\mathbf{I}_t\}_{t=1}^{T}$, the model predicts camera parameters $\hat{\mathcal{P}} = \{\hat{P}_t\}_{t=1}^{T}$, where $\hat{P}_t = (\hat{\mathbf{K}}_t, \hat{\mathbf{R}}_t, \hat{\mathbf{u}}_t)$ denotes camera intrinsics and the camera-to-world pose. We use $s \in \{\mathrm{L}, \mathrm{R}\}$ to index hand side and superscripts h, c, and w to denote hand, camera, and world coordinate systems, respectively. Simultaneously, it estimates the articulated state of each hand s as MANO (Romero et al., 2017) pose $\hat{\mathbf{\Theta}}_t^s \in \mathbb{R}^{15\times 3}$, shape $\hat{\boldsymbol{\beta}}_t^s \in \mathbb{R}^{10}$, global orientation $\hat{\mathbf{\Phi}}_t^{\mathrm{w},s} \in \mathbb{R}^3$, and root translation $\hat{\mathbf{t}}_t^{\mathrm{w},s} \in \mathbb{R}^3$. The model first predicts hand-frame geometry, converts it to camera coordinates, and then obtains world-space motion using the estimated camera trajectory. Specifically, Sec. 3.1 details data preparation and curation, Sec. 3.2 introduces camera-space hand localization and reconstruction and Sec. 3.3 presents joint streaming estimation alongside sparse geometric refinement.

3.1 DATA PREPROCESSING

We construct a comprehensive training corpus of approximately 5,000 hours by aggregating nine public hand-interaction datasets: ARCTIC (Fan et al., 2023), HOT3D (Banerjee et al., 2025), EgoDex (Hoque et al., 2025), DexYCB (Chao et al., 2021), HO3D (Hampali et al., 2020), H2O-3D (Hampali et al., 2022), EgoVerse (Punamiya et al., 2026), EgoLive (Li et al., 2026b), and Xperience-10M (Ropedia, 2026). Video frames are uniformly sampled at 10 FPS. Where explicit MANO annotations are absent, we fit the MANO model to the provided 3D hand keypoints and temporally align the fitted parameters with the sampled frames. To generate spatial supervision, we render the left- and right-hand MANO meshes to produce per-hand segmentation masks, from which tight bounding boxes are subsequently derived. Following standardization of handedness conventions and coordinate systems, each sample comprises an RGB image, camera-space MANO pa-

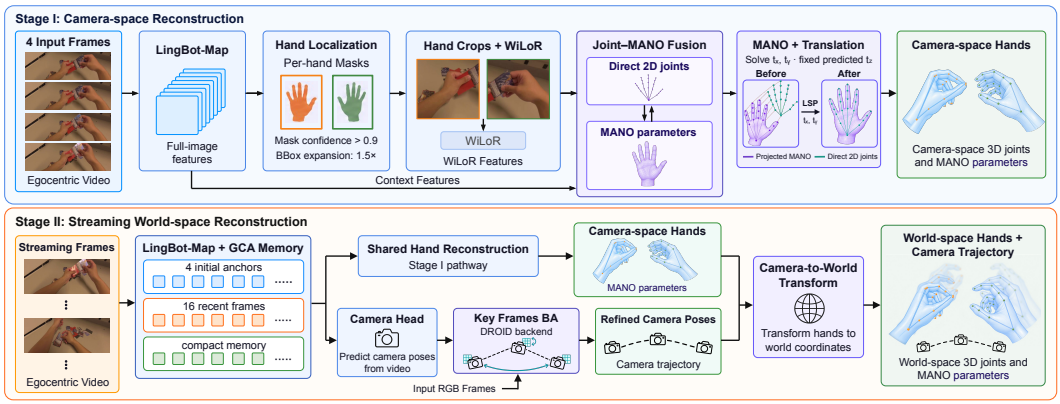

Figure 2: **Overview of InfiniHand.** Stage I learns hand localization and MANO reconstruction from geometric and appearance features, transforming hand-frame predictions into camera coordinates. Stage II jointly estimates hands and camera trajectories with streaming memory, followed by sparse bundle adjustment for camera refinement and world-space reconstruction.

rameters and 3D joints, per-hand masks, and bounding boxes. For sequences featuring ground-truth camera trajectories, we retain camera pose annotations and transform camera-space hand parameters into a shared world coordinate frame, establishing supervision for joint hand and camera trajectory reconstruction.

Upon standardizing the corpus, we observe that certain source annotations remain inconsistent with visual hand states, most prominently in EgoDex. To purge these noisy labels, we utilize Sapiens 2 (Khirodkar et al., 2026) to estimate 2D hand keypoints and align them with the projected 2D locations of annotated MANO joints under standardized topologies. Prior to evaluation, low-confidence detections, invalid hand associations, and anatomically implausible poses are filtered out. For each hand, joint-wise pixel errors are normalized by the bounding-box extent, averaged per frame, and summarized using the 90th percentile error over the sequence. Consequently, sequences exceeding the error threshold for either hand are discarded, whereas uncertain cases are set aside for manual inspection. In total, this cleaning process removes roughly 30–40% of the candidate training data. Lastly, we organize the curated corpus into two stages based on supervision availability: Stage I uses all valid camera-space hand annotations, while Stage II incorporates the subset containing camera trajectories for world-space joint supervision.

3.2 Camera-Space Hand Reconstruction

Hand localization. Our hand localization module leverages full-image geometric features from the pretrained LingBot-Map backbone (Chen et al., 2026) to identify left- and right-hand regions. The module comprises a Dense Prediction Transformer (DPT) (Ranftl et al., 2021) feature decoder and a two-channel convolutional mask head. We initialize the decoder from a pretrained depth head, retaining its feature projections, multi-scale resizing, and coarse-to-fine RefineNet (Lin et al., 2017) fusion, while replacing the final depth prediction layer with a randomly initialized mask head. Given intermediate backbone features $\{\mathbf{F}_t^{(\ell)}\}_{\ell\in\mathcal{S}}$ extracted from four selected layers, the module computes

$$\mathbf{D}_t = \mathcal{D}_{\mathrm{DPT}}(\{\mathbf{F}_t^{(\ell)}\}_{\ell\in\mathcal{S}}), \qquad [\hat{\mathbf{M}}_t^{\mathrm{L}}, \hat{\mathbf{M}}_t^{\mathrm{R}}] = \sigma(\mathcal{H}_{\mathrm{mask}}(\mathbf{D}_t)), \tag{1}$$

where $\mathcal{H}_{\mathrm{mask}}$ upsamples the decoded features $\mathbf{D}_t$ to the input image resolution, and $\sigma(\cdot)$ denotes the element-wise sigmoid function. The localization objective combines binary cross-entropy and Dice losses:

$$\mathcal{L}_{\mathrm{mask}} = \lambda_{\mathrm{BCE}}\mathcal{L}_{\mathrm{BCE}}(\hat{\mathbf{M}}, \mathbf{M}) + \lambda_{\mathrm{Dice}}\mathcal{L}_{\mathrm{Dice}}(\hat{\mathbf{M}}, \mathbf{M}). \tag{2}$$

Validity flags ignore missing annotations to avoid false negative supervision. Subsequently, thresholded masks are converted into expanded bounding boxes $\mathbf{b}_t^s$ for hand reconstruction.

Hand reconstruction. For each hand box $\mathbf{b}_t^s$, we extract aligned geometric and appearance features. A geometric adapter $\mathcal{A}$ crops and re-embeds the backbone patch features $\mathbf{F}_t$ into a patch grid,

while a WiLoR encoder (Potamias et al., 2024) concurrently processes the corresponding RGB crop to capture fine-grained visual details. Both branches share the same cropping conventions, including horizontal flipping for left-hand canonicalization. The concatenated dual-stream features are then fused via a 1×1 convolutional layer:

$$\mathbf{G}_t^s = \mathcal{A}(\mathbf{F}_t, \mathbf{b}_t^s), \quad \mathbf{A}_t^s = \mathcal{E}_{\mathrm{WiLoR}}(\mathrm{Crop}(\mathbf{I}_t, \mathbf{b}_t^s)), \quad \mathbf{Z}_t^s = \mathrm{Conv}_{1\times1}([\mathbf{G}_t^s; \mathbf{A}_t^s]). \tag{3}$$

Driven by the fused representation $\mathbf{Z}_t^s$, the MANO head regresses hand parameters in a local coordinate frame h, defined by the virtual crop camera. In parallel, an auxiliary head estimates 2D landmarks from $\mathbf{Z}_t^s$ to provide spatial grounding constraints. Omitting the frame index t and hand side s for conciseness, the predictions are expressed as:

$$(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Phi}}^{\mathrm{h}}, \hat{\ell}_{\mathrm{z}}) = \mathcal{H}_{\mathrm{MANO}}(\mathbf{Z}), \qquad \hat{\mathbf{p}} = \mathcal{H}_{\mathrm{2D}}(\mathbf{Z}), \qquad \hat{t}_{\mathrm{z}}^{\mathrm{h}} = \exp(\hat{\ell}_{\mathrm{z}}). \tag{4}$$

After reversing left-hand canonicalization, the MANO decoder (Romero et al., 2017) maps the predicted pose, shape, and orientation into translation-free 3D joints $\bar{\mathbf{J}}^{\mathrm{h}}$. To recover the remaining lateral translation $(\hat{t}_{\mathrm{x}}^{\mathrm{h}}, \hat{t}_{\mathrm{y}}^{\mathrm{h}})$ within the hand frame, we follow ViDiHand (Wang et al., 2026) by aligning the projected 3D joints with the predicted 2D landmarks $\hat{\mathbf{p}}^{\mathrm{pix}}$ and root depth $\hat{t}_{\mathrm{z}}^{\mathrm{h}}$ via differentiable least-squares projection (LSP), which solves for lateral translation while keeping the predicted depth fixed:

$$(\hat{t}_{\mathrm{x}}^{\mathrm{h}}, \hat{t}_{\mathrm{y}}^{\mathrm{h}}) = \arg\min_{a,b} \sum_{j\in\mathcal{V}} \left\| \pi\big(\bar{\mathbf{J}}_j^{\mathrm{h}} + [a, b, \hat{t}_{\mathrm{z}}^{\mathrm{h}}]^\top; \mathbf{K}^{\mathrm{h}}\big) - \hat{\mathbf{p}}_j^{\mathrm{pix}} \right\|_2^2, \tag{5}$$

where $\mathbf{K}^{\mathrm{h}}$ is the virtual crop-camera intrinsic matrix, $\pi(\cdot; \mathbf{K}^{\mathrm{h}})$ is the perspective projection, $\hat{\mathbf{p}}_j^{\mathrm{pix}}$ are the predicted 2D landmarks in crop pixels, and $\mathcal{V}$ represents valid joints with positive depth.

With the recovered translation, we obtain the complete hand-frame MANO representation and transform its global orientation and translation into the original camera coordinate system:

$$\mathrm{Rot}(\hat{\boldsymbol{\Phi}}^{\mathrm{c}}) = \mathbf{R}_{\mathrm{h}\leftarrow\mathrm{c}}^\top \mathrm{Rot}(\hat{\boldsymbol{\Phi}}^{\mathrm{h}}), \qquad \hat{\mathbf{t}}^{\mathrm{c}} = \mathbf{R}_{\mathrm{h}\leftarrow\mathrm{c}}^\top \hat{\mathbf{t}}^{\mathrm{h}}. \tag{6}$$

Here $\mathbf{R}_{\mathrm{h}\leftarrow\mathrm{c}}$ rotates from the original camera to the virtual hand camera, and $\mathrm{Rot}(\cdot)$ converts an axis-angle vector to its rotation matrix via the Rodrigues formula.

For training losses, the 2D landmark head is first supervised via a mean L_1 loss $\mathcal{L}_{\mathrm{2D}}$ computed over valid joints in normalized crop coordinates. Concurrently, the MANO head combines pose, shape, orientation, translation, 3D joint, and reprojection supervision. Global orientation, translation, and 3D joints are compared in the original camera frame c, while local pose and shape are frame-invariant and reprojection uses the corresponding image coordinates:

$$\mathcal{L}_{\mathrm{MANO}} = \lambda_{\mathrm{g}}\mathcal{L}_{\mathrm{orient}} + \lambda_{\theta}\mathcal{L}_{\mathrm{pose}} + \lambda_{\beta}\mathcal{L}_{\mathrm{shape}} + \lambda_{\mathrm{t}}\mathcal{L}_{\mathrm{trans}} + \lambda_{\mathrm{J}}\mathcal{L}_{\mathrm{joints}}^{\mathrm{c}} + \lambda_{\mathrm{r}}\mathcal{L}_{\mathrm{reproj}}. \tag{7}$$

In Stage I, each head is optimized separately on available camera-space ground truth. Detailed formulations of all loss terms are elaborated in Appendix B.

3.3 World-Space Hand Reconstruction

Joint streaming training. In Stage II, we integrate the pretrained hand modules with the camera pose head to jointly optimize the localization, 2D landmark, MANO, and camera heads. Training is performed on video clips consisting of four anchor frames followed by two consecutive 16-frame windows, forming a $4 + 16 \times 2$ temporal layout. The streaming state persists across adjacent windows within a clip and is reset between independent clips. Ground-truth and predicted bounding boxes are sampled in a $1 : 1$ ratio, exposing hand reconstruction to realistic localization noise while retaining direct supervision from clean crops. Specifically, this streaming memory state $\mathcal{M}_k = (\mathcal{M}_{\mathrm{anchor}}, \mathcal{W}_k, \mathcal{T}_k)$ is managed via the Geometric Context Attention (GCA) mechanism (Chen et al., 2026), which integrates anchor features $\mathcal{M}_{\mathrm{anchor}}$, a recent dense-feature window $\mathcal{W}_k$, and compressed trajectory memory $\mathcal{T}_k$. Here, anchor features establish a shared reference frame, the dense window captures fine-grained local correspondences, and trajectory tokens preserve historical context after their corresponding dense features are evicted. Formally, for window k, the backbone updates its geometric features and memory state as:

$$(\mathbf{F}_k, \mathcal{M}_k) = \mathcal{B}_{\mathrm{GCA}}(\mathbf{I}_k, \mathcal{M}_{k-1}). \tag{8}$$

From this updated feature representation $\mathbf{F}_k$, the camera and hand modules decode predictions for each frame t in this window:

$$\hat{P}_t = \mathcal{H}_{\text{cam}}(\mathbf{F}_k)_t, \quad \hat{\mathbf{M}}_t = \mathcal{H}_{\text{loc}}(\mathbf{F}_k)_t, \quad \hat{\mathcal{Y}}_t^s = \mathcal{H}_{\text{rec}}(\mathbf{I}_t, \mathbf{F}_t, \mathbf{b}_t^s), \tag{9}$$

where $\hat{P}_t = (\hat{\mathbf{K}}_t, \hat{\mathbf{R}}_t, \hat{\mathbf{u}}_t)$ and $\hat{\mathcal{Y}}_t^s = (\hat{\mathbf{p}}_t^s, \hat{\mathbf{\Theta}}_t^s, \hat{\boldsymbol{\beta}}_t^s, \hat{\mathbf{\Phi}}_t^{\text{c},s}, \hat{\mathbf{t}}_t^{\text{c},s})$ collect the camera and hand predictions. The localization and reconstruction modules follow Sec. 3.2, including hand-to-camera conversion. Subsequently, the global orientation and translation are transformed from camera coordinates into the first-anchor world frame:

$$\text{Rot}(\hat{\mathbf{\Phi}}_t^{\text{w},s}) = \hat{\mathbf{R}}_t \, \text{Rot}(\hat{\mathbf{\Phi}}_t^{\text{c},s}), \qquad \hat{\mathbf{t}}_t^{\text{w},s} = \hat{\mathbf{R}}_t \hat{\mathbf{t}}_t^{\text{c},s} + \hat{\mathbf{u}}_t. \tag{10}$$

Meanwhile, local articulated pose and shape remain unchanged under this transformation, completing the world-space MANO representation. Finally, for supervision, predictions and targets use the same anchor transformation and sample-level spatial normalization $\tilde{\mathbf{x}} = \mathbf{x}/\kappa$, where $\kappa > 0$ is the sample-level normalization scale.

For training losses, we combine 2D landmark, mask, camera-space MANO, world-space joint, temporal, and camera supervision:

$$\mathcal{L}_{\text{joint}} = \lambda_{\text{cam}} \mathcal{L}_{\text{camera}} + \lambda_{\text{w}} \mathcal{L}_{\text{joints}}^{\text{w}} + \lambda_{\text{temp}} \mathcal{L}_{\text{temp}} + \mathcal{L}_{\text{MANO}} + \lambda_{\text{2D}} \mathcal{L}_{\text{2D}} + \mathcal{L}_{\text{mask}}. \tag{11}$$

Here, $\mathcal{L}_{\text{MANO}}$ retains the camera-space supervision defined in Sec. 3.2, while $\mathcal{L}_{\text{joints}}^{\text{w}}$ constrains the reconstructed joints in the shared world frame. Specifically, the camera loss $\mathcal{L}_{\text{camera}}$ combines absolute pose and field-of-view supervision with relative-motion supervision between valid frame pairs. Meanwhile, $\mathcal{L}_{\text{temp}}$ enforces temporal consistency across camera predictions, MANO parameters, hand masks, and 2D landmarks. Each loss term is evaluated conditionally based on annotation availability. Further loss details are detailed in Appendix B.

Sparse bundle adjustment. Although trajectory memory preserves long-range context, dense historical features are inevitably compressed beyond the anchor and recent windows. Consequently, fine-grained geometric constraints from earlier observations are not explicitly revisited during optimization, leading to residual drift over extended sequences. To address this, we integrate a DROID bundle-adjustment backend (Teed & Deng, 2021) into streaming prediction, enforcing explicit cross-frame constraints to correct long-term drift efficiently. Instead of optimizing over the entire frame history, we maintain a binary keyframe pool that balances local temporal continuity with long-range geometric constraints. By selectively retaining keyframes for sparse refinement, BA can revisit past observations and reduce cumulative camera drift. The refined camera poses are subsequently used to transform local hand predictions into a unified world coordinate system. Implementation details and pool update rules are elaborated in Appendix C.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets & Metrics. For camera-space evaluation, we follow ViDiHand (Wang et al., 2026) and use 34 test scenes from ARCTIC (Fan et al., 2023), 10 from HOT3D (Banerjee et al., 2025), and 166 from HOI4D (Liu et al., 2022). We additionally evaluate on 100 test scenes from EgoDex (Hoque et al., 2025) to cover more diverse scenarios. ARCTIC, HOT3D, and EgoDex also support our world-space evaluation. We further assess in-the-wild reconstruction qualitatively on Ego4D (Grauman et al., 2022) test videos. For hand detection, we report Frame Accuracy (FAcc), Recall, and F1 score based on hand presence. For camera-space reconstruction, we evaluate MP-p, PA-p, EPE-p, GO-p, and CT-p, which measure root-relative 3D joint error, Procrustes-aligned 3D joint error, 2D projection error, global orientation error, and camera-space wrist position error, respectively. The suffix $-\text{p}$ denotes that penalties are incorporated for missed hands. For world-space reconstruction, we report PA-MPJPE, W-MPJPE, and WA-MPJPE to measure 3D joint error after per-hand, per-frame alignment, without additional alignment, and after a single sequence-level alignment shared by both hands, respectively. Detailed metric definitions are provided in Appendix A.

Baselines. For camera-space reconstruction, we compare with InterWild (Moon, 2023), HaMeR (Pavlakos et al., 2024), Hamba (Dong et al., 2024), WildHands (Prakash et al., 2023),

Table 1: **Camera-space quantitative comparison.** Evaluating detection and camera-space hand motion metrics across four benchmarks, InfiniHand consistently yields the lowest PA-p. **Bold** and underlined denote best and second-best results.

Method	ARCTIC								HOT3D							
	FAcc↑	Recall↑	F1↑	MP-p↓	PA-p↓	EPE-p↓	GO-p↓	CT-p↓	FAcc↑	Recall↑	F1↑	MP-p↓	PA-p↓	EPE-p↓	GO-p↓	CT-p↓
InterWild	0.878	0.943	0.959	30.82	15.95	53.89	25.39	0.097	0.669	0.881	0.868	77.17	24.81	71.48	58.50	0.213
HaMeR	0.875	0.943	0.957	29.20	14.60	65.29	24.91	0.095	0.692	0.904	0.883	68.31	21.46	59.08	49.64	0.102
Hamba	0.833	0.912	0.941	31.23	17.17	87.05	27.82	0.110	0.632	0.828	0.853	71.73	29.62	107.63	56.53	0.128
WildHands	0.879	0.946	0.960	25.70	13.94	50.52	22.32	0.058	0.655	0.863	0.844	52.79	28.95	111.44	53.93	0.157
OmniHands	0.866	0.949	0.954	29.67	14.20	51.51	24.58	0.087	0.649	0.895	0.868	63.28	22.68	68.44	49.12	0.133
WiLoR	0.919	0.951	0.974	22.01	11.87	71.53	17.36	0.075	0.827	0.897	0.937	30.97	19.98	72.98	25.75	0.098
Dyn-HaMR	0.842	0.918	0.951	27.90	17.02	85.72	25.95	0.121	0.614	0.811	0.802	74.21	38.20	171.62	43.85	0.571
HaWoR	0.700	0.817	0.895	45.36	26.38	158.06	43.33	0.149	0.348	0.499	0.654	71.40	66.03	327.29	79.35	0.262
ViDiHand	**0.997**	**0.999**	**0.999**	21.67	9.82	**12.41**	14.64	0.047	**0.948**	0.974	**0.983**	**21.51**	11.38	**14.95**	15.83	**0.040**
Ours	0.993	0.996	0.998	**17.09**	**7.72**	23.13	**12.80**	**0.044**	0.916	**0.981**	0.971	25.52	**11.10**	17.02	**15.49**	0.059

Method	EgoDex								HOI4D							
	FAcc↑	Recall↑	F1↑	MP-p↓	PA-p↓	EPE-p↓	GO-p↓	CT-p↓	FAcc↑	Recall↑	F1↑	MP-p↓	PA-p↓	EPE-p↓	GO-p↓	CT-p↓
InterWild	0.937	**1.000**	0.983	69.34	30.00	153.42	65.83	0.204	0.731	0.922	0.864	53.07	22.91	80.55	41.74	0.228
HaMeR	0.949	0.984	0.985	57.67	16.49	53.62	41.42	0.129	0.731	0.923	0.864	44.48	21.58	79.49	33.56	0.187
Hamba	0.710	0.824	0.896	65.68	31.66	188.48	64.86	0.170	0.710	0.885	0.849	47.16	25.92	115.79	37.39	0.204
WildHands	0.932	0.966	0.979	48.64	17.32	66.15	42.60	0.107	0.730	0.924	0.864	45.62	23.60	82.25	45.65	0.159
OmniHands	0.937	**1.000**	0.983	49.49	17.56	64.49	43.77	0.125	0.655	0.937	0.834	44.26	18.69	70.66	34.39	**0.108**
WiLoR	0.887	0.940	0.968	48.29	19.59	91.55	40.02	0.137	0.962	0.966	0.972	33.71	14.90	41.58	25.53	0.115
Dyn-HaMR	0.942	0.971	0.984	41.67	15.12	39.55	30.79	0.083	0.750	0.863	0.845	45.10	29.26	144.64	40.18	0.258
HaWoR	0.838	0.913	0.954	33.39	18.99	96.86	28.90	0.127	0.869	0.864	0.919	47.33	28.85	135.75	43.09	0.139
ViDiHand	0.911	0.950	0.974	42.33	17.22	65.05	33.28	0.303	0.984	0.991	0.990	**30.09**	13.96	24.46	23.42	0.117
Ours	**0.961**	0.997	**0.995**	**16.43**	**7.29**	**24.08**	**15.51**	**0.031**	**0.989**	**0.993**	**0.991**	33.64	**12.11**	**22.75**	**23.00**	0.136

Table 2: **World-space quantitative comparison.** We evaluate world-space hand motion and trajectory metrics across three benchmarks, with InfiniHand achieving the lowest W-MPJPE.

Method	ARCTIC			HOT3D			EgoDex		
	PA-MPJPE↓	W-MPJPE↓	WA-MPJPE↓	PA-MPJPE↓	W-MPJPE↓	WA-MPJPE↓	PA-MPJPE↓	W-MPJPE↓	WA-MPJPE↓
WiLoR-SLAM	7.23	65.86	46.31	6.46	106.77	45.20	10.25	96.02	40.95
HaWoR	9.03	95.39	45.45	5.86	93.70	**35.02**	10.36	103.85	35.31
Dyn-HaMR	10.85	114.10	63.60	10.17	296.72	132.08	11.58	78.41	37.11
Ours	**7.07**	**59.21**	**43.59**	**5.71**	**87.37**	35.76	**4.91**	**28.77**	**16.47**

OmniHands (Lin et al., 2024), WiLoR (Potamias et al., 2024), Dyn-HaMR (Yu et al., 2025), HaWoR (Zhang et al., 2025), and ViDiHand (Wang et al., 2026). For world-space reconstruction, we compare with HaWoR, Dyn-HaMR, and WiLoR-SLAM. WiLoR-SLAM transforms WiLoR hand predictions into world coordinates using DROID-SLAM (Teed & Deng, 2021) camera poses scaled by Metric3D v2 (Hu et al., 2024).

Implementation details. Stage I and Stage II are trained for 300k and 100k optimization steps, respectively, using a global batch size of 64 and the AdamW (Loshchilov & Hutter, 2019) optimizer with a weight decay of 0.01. Stage I processes 4-frame clips, setting learning rates to 10^{-4} for the mask and joint heads and 5×10^{-5} for the MANO head. Stage II employs the 36-frame layout $(4 + 16 \times 2)$, using a learning rate of 3×10^{-5} for the camera head and 10^{-5} for all other trainable modules. All learning rates follow linear warmup with cosine annealing, and global gradient norms are clipped at 1.0. Full-resolution images are processed at 378×518 $(H \times W)$, while hand bounding boxes are expanded by $1.5\times$ and resized to 256×256 for the crop branch.

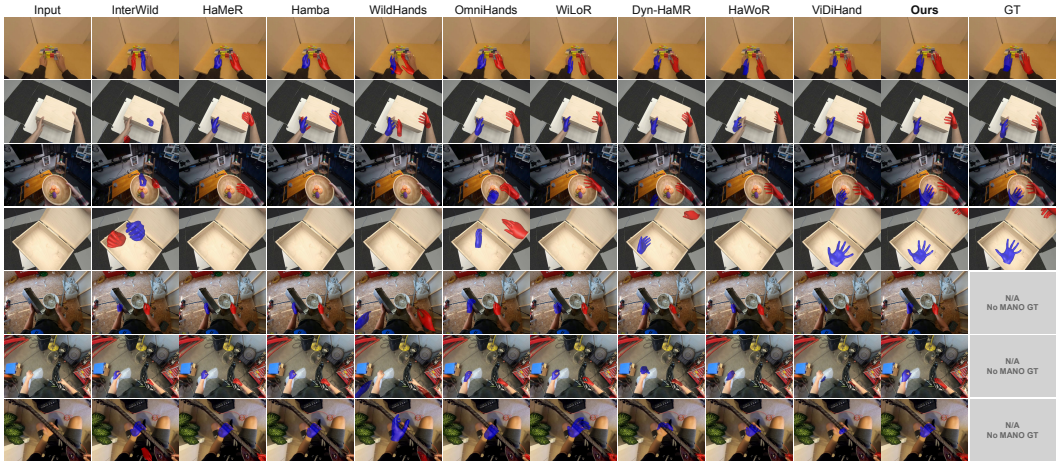

Figure 3: **Camera-space qualitative comparison.** InfiniHand recovers both hands without missed detections while producing precise hand motion under severe occlusion and in-the-wild scenes.

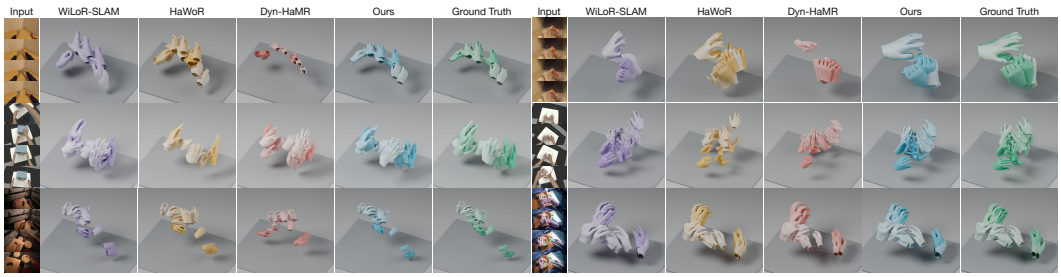

Figure 4: **World-space qualitative comparison.** Visual results on diverse datasets show that our reconstructed articulation and trajectories achieve the highest fidelity to ground truth.

4.2 Camera-space results analysis

Table 1 and Fig. 3 demonstrate robust hand reconstruction capabilities across both in-domain and out-of-domain scenarios. Notably, InfiniHand achieves the lowest PA-p across all four datasets, reducing the error relative to ViDiHand (Wang et al., 2026) from 9.82 to 7.72 mm on ARCTIC (a 21.4% reduction) and from 17.22 to 7.29 mm on EgoDex (a 57.7% reduction). Qualitatively, in-domain comparisons highlight precise finger articulation and reliable recovery under severe hand occlusions. For in-the-wild sequences lacking ground-truth MANO annotations, InfiniHand preserves plausible hand geometry and accurate image alignment under heavy clutter and unusual viewpoints, whereas ViDiHand exhibits visible degradation, highlighting superior cross-dataset generalization.

4.3 World-space results analysis

Table 2 and Fig. 4 demonstrate superior world-space placement and relative motion tracking. InfiniHand achieves the lowest W-MPJPE across all three datasets, reducing errors from 65.86 to 59.21 mm (10.1%) on ARCTIC compared to WiLoR-SLAM (Potamias et al., 2024; Teed & Deng, 2021) and from 78.41 to 28.77 mm (63.3%) on EgoDex compared to Dyn-HaMR (Yu et al., 2025). Lower WA-MPJPE scores further confirm that trajectory fidelity persists after sequence-level alignment. Qualitatively, InfiniHand preserves hand scale, inter-hand spacing, and curved motion trajectories with minimal drift. Specifically, our reconstruction avoids spatial displacement (Fig. 4, top-left) and aligns two-hand orientations more faithfully with the reference trajectory (bottom-right). These visualizations confirm faithful relative motion alongside accurate absolute placement.

Table 3: **Quantitative ablation study on ARCTIC.** Results show the contributions of appearance features, translation recovery, sparse BA, joint training, and hand localization.

Variant	Detection			Camera-space					World-space		
	FAcc↑	Recall↑	F1↑	MP-p↓	PA-p↓	EPE-p↓	GO-p↓	CT-p↓	PA-MPJPE↓	W-MPJPE↓	WA-MPJPE↓
w/o WiLoR Features	0.993	0.996	0.998	33.31	10.91	32.89	22.08	0.087	7.46	60.36	45.12
w/o LSP	0.993	0.996	0.998	18.63	8.02	26.30	12.97	0.064	7.21	59.57	44.35
w/o BA	0.993	0.996	0.998	17.09	7.72	23.13	12.80	0.044	7.07	80.48	51.20
w/o Stage II	0.985	0.989	0.990	17.29	7.81	23.56	13.01	0.045	7.40	185.76	111.05
w/o Mask Head	0.700	0.817	0.895	33.31	25.50	167.66	36.97	0.143	7.43	60.39	45.26
Full Model	0.993	0.996	0.998	17.09	7.72	23.13	12.80	0.044	7.07	59.21	43.59

4.4 Efficiency analysis

InfiniHand achieves the highest prediction throughput among the compared methods, reaching 11.19 FPS. This significantly outperforms existing baselines, including WiLoR-SLAM (Potamias et al., 2024; Teed & Deng, 2021) (8.53 FPS), HaWoR (Zhang et al., 2025) (5.48 FPS), and Dyn-HaMR (Yu et al., 2025) (0.81 FPS), yielding a $2.04\times$ speedup over HaWoR. This efficiency is achieved by reusing streaming features and overlapping model execution across temporal windows. Specifically, two primary workers run in a pipeline, where the first worker processes the upcoming 16-frame window using four-anchor context, while the second worker simultaneously decodes hand and camera predictions for the prior window. This parallel workflow reduces idle wait time between feature extraction and decoding. In full-system execution, a third worker runs sparse bundle adjustment in the background whenever the binary keyframe pool exceeds its capacity threshold, decoupling periodic refinement from window prediction while still incurring additional computation.

4.5 Ablation Study

We conduct ablation studies on all 34 ARCTIC (Fan et al., 2023) test scenes, as reported in Table 3. (1) Removing WiLoR features nearly doubles MP-p, from 17.09 to 33.31 mm, highlighting the importance of hand-centered appearance for detailed reconstruction. (2) Replacing LSP with direct translation regression degrades translation accuracy and image-space alignment, supporting the use of projection constraints for hand placement. (3) Disabling sparse BA increases W-MPJPE from 59.21 to 80.48 mm while leaving camera-space metrics unchanged, demonstrating its contribution to global reconstruction accuracy. (4) Omitting Stage II increases W-MPJPE from 59.21 to 185.76 mm and WA-MPJPE from 43.59 to 111.05 mm despite only minor changes in camera-space errors, emphasizing the importance of joint hand–camera training for world-space motion estimation. (5) Replacing the learned mask head with HaWoR masks substantially degrades detection and increases PA-p from 7.72 to 25.50 mm, underscoring the role of reliable localization in hand motion reconstruction.

5 Limitations

While InfiniHand achieves accurate world-space 3D hand reconstruction, there remains clear room for further improvement across several key technical aspects. In terms of scale recovery, the underlying LingBot-Map framework lacks inherent metric scale, requiring an auxiliary post-processing alignment model whose downstream estimation errors can inevitably propagate to predicted hand positions and global motion trajectories. Regarding data coverage, high-quality egocentric datasets featuring complex, large-amplitude two-hand interactions and camera dynamics remain scarce, restricting generalization to unconstrained in-the-wild sequences with rapid viewpoint changes, severe occlusions, and intermittent hand visibility. On the supervision front, monocular reconstruction pipelines often inherit time-varying scale drift from derived monocular annotations, creating temporally inconsistent training targets that compromise long-sequence trajectory accuracy and temporal smoothness despite global scale alignment.

6 CONCLUSION

We present InfiniHand, a streaming feed-forward framework for jointly estimating hand locations, MANO parameters, and camera trajectories from egocentric video. To power this architecture, we aggregate approximately 5,000 hours of public video into a clean, large-scale training corpus, followed by a two-stage training scheme that develops precise MANO recovery and aligns world-space hand and camera motions. Extensive evaluations across in-domain benchmarks and out-of-domain videos demonstrate strong reconstruction quality and generalization, validating the value of scaling up diverse egocentric supervision. Beyond accuracy, the streaming design achieves more than twice the inference throughput of HaWoR under standard timing protocols. Collectively, these advances enable faster and more reliable 3D annotation of egocentric video, unlocking downstream applications in dexterous data augmentation, human-to-robot motion transfer, and manipulation policy learning. By converting abundant human video into structured 3D hand-motion supervision, InfiniHand offers a scalable path toward data generation for embodied intelligence.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with translation. We have not used generative AI tools for designing research methods and experiments, implementing methodologies, interpreting results, proposing or refining hypotheses, cleaning and reformatting datasets, or supporting qualitative and thematic data analysis, and generating synthetic datasets, proposing mathematical claims, providing key elements for proving mathematical claims, and assisting in writing proofs are not applicable to this work. Additionally, we used generative AI tools for summarizing or analyzing existing literature, and editing the manuscript to enhance readability. We have reviewed all AI-assisted work: translated or polished text was manually cross-checked sentence-by-sentence to ensure that the original intent remained uncompromised. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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. *arXiv preprint arXiv:2604.14141*, 2026.

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. *arXiv preprint arXiv:2407.09646*, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano

Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. EgoDex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Kaixuan Wang, Hao Chen, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3Dv2: A Versatile Monocular Geometric Foundation Model for Zero-shot Metric Depth and Surface Normal Estimation. *arXiv preprint arXiv:2404.15506*, 2024.

Rawal Khirodkar, He Wen, Julieta Martinez, Yuan Dong, Zhaoen Su, and Shunsuke Saito. Sapiens2. *arXiv preprint arXiv:2604.21681*, 2026.

Baoyu Li, Xinchen Yin, Mengying Lin, Yixin Zhang, and Danfei Xu. EgoWAM: World action models beyond pixels with in-the-wild egocentric human data. *arXiv preprint arXiv:2607.08436*, 2026a.

Yihang Li, Xuelong Wei, Jingzhou Luo, Yingjing Xiao, Yibo Bai, Guangyuan Zhou, Teng Zou, Chenguang Gui, Jiajun Wen, He Zhang, Kangliang Chen, Xing Pan, Shuaiyan Liu, Daming Wang, Tao An, Jiayi Li, Shibo Jin, Wanwan Zhang, Tianyu Wang, Boren Wei, Zhixuan Huang, Fangsheng Liu, Ruodai Li, Hui Zhang, Anson Li, Yicheng Gong, Peng Cao, Jiaming Liang, and Liang Lin. EgoLive: A large-scale egocentric dataset from real-world human tasks. *arXiv preprint arXiv:2604.23570*, 2026b.

Dixuan Lin, Yuxiang Zhang, Mengcheng Li, Wei Jing, Qi Yan, Qianying Wang, Yebin Liu, and Hongwen Zhang. OmniHands: Towards Robust 4D Hand Mesh Recovery via A Versatile Transformer. *arXiv preprint arXiv:2405.20330*, 2024.

Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Dominic Maggio, Hyungtae Lim, and Luca Carlone. VGGT-SLAM: Dense RGB SLAM Optimized on the SL(4) Manifold. *arXiv preprint arXiv:2505.12549*, 2025.

Gyeongsik Moon. Bringing Inputs to Shared Domains for 3D Interacting Hands Recovery in the Wild. *arXiv preprint arXiv:2303.13652*, 2023.

Raúl Mur-Artal and Juan D. Tardós. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017.

Riku Murai, Eric Dexheimer, and Andrew J. Davison. MASt3R-SLAM: Real-Time Dense SLAM with 3D Reconstruction Priors. *arXiv preprint arXiv:2412.12392*, 2024.

Georgios Papagiannis, Norman Di Palo, Pietro Vitiello, and Edward Johns. R+X: Retrieval and execution from everyday human videos. In *IEEE International Conference on Robotics and Automation*, 2025.

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

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. WiLoR: End-to-end 3d hand localization and reconstruction in the wild. *arXiv preprint arXiv:2409.12259*, 2024.

Aditya Prakash, Ruisen Tu, Matthew Chang, and Saurabh Gupta. 3D Hand Pose Estimation in Everyday Egocentric Images. *arXiv preprint arXiv:2312.06583*, 2023.

Ryan Punamiya, Simar Kareer, Zeyi Liu, Josh Citron, Ri-Zhao Qiu, Xiongyi Cai, Alexey Gavryushin, Jiaqi Chen, Davide Liconti, Lawrence Y. Zhu, Patcharapong Aphiwetsa, Baoyu Li, Aniketh Cheluva, Pranav Kuppili, Yangcen Liu, Dhruv Patel, Aidan Gao, Hye-Young Chung, Ryan Co, Renee Zbizika, Jeff Liu, Xiaomeng Xu, Haoyu Xiong, Geng Chen, Sebastiano Oliani, Wenkai Xuan, Chenyu Yang, Xi Wang, James Fort, Richard Newcombe, Josh Gao, Jason Chong, Garrett Matsuda, Aseem Doriwala, Marc Pollefeys, Robert Katzschmann, Xiaolong Wang, Shuran Song, Judy Hoffman, and Danfei Xu. EgoVerse: An egocentric human dataset for robot learning from around the world. *arXiv preprint arXiv:2604.07607*, 2026.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Kerui Ren, Guanghao Li, Changjian Jiang, Yingxiang Xu, Tao Lu, Linning Xu, Junting Dong, Jiangmiao Pang, Mulin Yu, and Bo Dai. M^3: Dense Matching Meets Multi-View Foundation Models for Monocular Gaussian Splatting SLAM. *arXiv preprint arXiv:2603.16844*, 2026.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6):245:1–245:17, 2017.

Ropedia. Xperience-10M: A large-scale egocentric multimodal dataset with structured 3d/4d annotations. Hugging Face dataset, 2026. URL `https://huggingface.co/datasets/ropedia-ai/xperience-10m`.

Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. *Advances in Neural Information Processing Systems*, 34, 2021.

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

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3D: Towards Zero-shot Metric 3D Prediction from A Single Image. *arXiv preprint arXiv:2307.10984*, 2023.

Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. Dyn-HaMR: Recovering 4d interacting hand motion from a dynamic camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. HaWoR: World-space hand motion reconstruction from egocentric videos. *arXiv preprint arXiv:2501.02973*, 2025.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Chen Sun, Ming-Hsuan Yang, Forrester Cole, Trevor Darrell, and Deqing Sun. LoGeR: Long-context geometric reconstruction with hybrid memory. *arXiv preprint arXiv:2603.03269*, 2026.

This supplementary material provides additional technical details and qualitative results to complement the main paper. Section A defines the detection, camera-space, and world-space metrics. Section B specifies the loss functions. Section C explains the binary keyframe pool for sparse bundle adjustment. Section D presents additional qualitative comparisons.

A Evaluation Metrics

Detection metrics. We follow ViDiHand (Wang et al., 2026) for prediction–target association. Projected MANO mesh boxes are matched greedily in descending IoU order, requiring identical handedness and IoU > 0.1 after enlarging target boxes by 10%. Matching is one-to-one. Predictions with an available presence score must exceed 0.5; unconditional tracker outputs are treated as positive. Unmatched target hands and predictions contribute FN and FP, respectively. A target is off-screen if none of its 21 joints projects inside the image at depth > 0.01 m; such targets and their matched predictions are excluded. Let $\mathcal{F}$ be frames containing at least one target hand, and let $\mathrm{FP}_t^{\mathrm{os}}$ and $\mathrm{FN}_t^{\mathrm{os}}$ be the remaining counts after off-screen exclusion. Then

$$\mathrm{FAcc} = \frac{1}{|\mathcal{F}|} \sum_{t \in \mathcal{F}} \mathbf{1}[\mathrm{FP}_t^{\mathrm{os}} = 0 \ \wedge \ \mathrm{FN}_t^{\mathrm{os}} = 0], \tag{12}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \qquad \mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{13}$$

Counts are pooled across frames and hand sides. Correct side presence alone is insufficient: a spatially unmatched hand contributes a detection error.

Camera-space metrics. We use the detection-penalized geometry metrics of ViDiHand (Wang et al., 2026). Let $\hat{\mathbf{J}}_j^{\mathrm{c}}$ and $\mathbf{J}_j^{\mathrm{c}}$ be predicted and target joints in metres, with wrist index $j = 0$ and $J = 21$. For a matched hand, set $\hat{\mathbf{Q}}_j = \hat{\mathbf{J}}_j^{\mathrm{c}} - \hat{\mathbf{J}}_0^{\mathrm{c}}$ and $\mathbf{Q}_j = \mathbf{J}_j^{\mathrm{c}} - \mathbf{J}_0^{\mathrm{c}}$. The root-relative and Procrustes-aligned errors are

$$e_{\mathrm{MP}} = \frac{1}{J} \sum_j \|\hat{\mathbf{Q}}_j - \mathbf{Q}_j\|_2, \qquad e_{\mathrm{PA}} = \frac{1}{J} \sum_j \|A^*(\hat{\mathbf{Q}}_j) - \mathbf{Q}_j\|_2, \tag{14}$$

where A^* is the proper similarity transform minimizing squared joint distances for that hand. Let $\hat{\mathbf{R}}_{\mathrm{hand}}$ and $\mathbf{R}_{\mathrm{hand}}$ denote its predicted and target global rotations. Orientation and wrist-position errors are

$$e_{\mathrm{GO}} = \frac{180}{\pi} \arccos\left(\mathrm{clamp}\left(\frac{\mathrm{tr}(\hat{\mathbf{R}}_{\mathrm{hand}}^\top \mathbf{R}_{\mathrm{hand}}) - 1}{2}, -1, 1 \right) \right), \tag{15}$$

$$e_{\mathrm{CT}} = \|\hat{\mathbf{J}}_0^{\mathrm{c}} - \mathbf{J}_0^{\mathrm{c}}\|_2. \tag{16}$$

The matching and off-screen exclusion rules are shared with the detection metrics. For metric $m \in \{\mathrm{MP}, \mathrm{PA}, \mathrm{GO}, \mathrm{CT}\}$, the penalized mean is

$$E_{m\text{-p}} = \frac{\sum_{i \in \mathcal{H}_{\mathrm{matched}}} e_m(i) + \sum_{i \in \mathcal{H}_{\mathrm{missed}}} e_m^{\mathrm{miss}}(i)}{|\mathcal{H}_{\mathrm{matched}}| + |\mathcal{H}_{\mathrm{missed}}|}. \tag{17}$$

Missed-hand MP and PA both use the unaligned canonical-hand joint error; GO uses the target's rotation distance from identity, and CT uses the target wrist's distance from the camera origin. MP-p and PA-p are converted to millimetres, GO-p is in degrees, and CT-p is in metres. EPE-p is a joint-weighted pixel error over on-screen target joints $\mathcal{U}$:

$$\text{EPE-p} = \frac{1}{|\mathcal{U}|} \sum_{(i,j) \in \mathcal{U}} \begin{cases} \min\Big(\|\pi_{\mathbf{K}}(\hat{\mathbf{J}}_{i,j}^{\mathrm{c}}) - \pi_{\mathbf{K}}(\mathbf{J}_{i,j}^{\mathrm{c}})\|_2, D_i \Big), & i \in \mathcal{H}_{\mathrm{matched}}, \\ D_i, & i \in \mathcal{H}_{\mathrm{missed}}, \end{cases} \tag{18}$$

where $D_i = \sqrt{W_i^2 + H_i^2}$ is the image diagonal. Target joints must project inside the image with depth above 0.01 m; predicted depth is floored at 0.01 m for projection. Thus detection and geometric errors use the same matched-hand set.

World-space metrics. World-space evaluation uses the same geometric matching and 21-joint ordering, with predictions and targets expressed in the complete scene's first-camera frame. For matched hand-frame pairs $\mathcal{H}$, define

$$E(A) = \frac{1}{J|\mathcal{H}|} \sum_{(t,s)\in\mathcal{H}} \sum_{j} \|A_{t,s}(\hat{\mathbf{J}}^{\mathrm{w},s}_{t,j}) - \mathbf{J}^{\mathrm{w},s}_{t,j}\|_2. \tag{19}$$

W-MPJPE uses the identity transform. PA-MPJPE fits a separate similarity transform for each hand-frame pair. WA-MPJPE uses one transform shared by both hands and all frames of the complete scene, obtained from

$$(\alpha^*, \mathbf{R}^*, \mathbf{v}^*) = \arg \min_{\alpha>0,\, \mathbf{R}\in\mathrm{SO}(3),\, \mathbf{v}} \sum_{(t,s)\in\mathcal{H}} \sum_{j} \|\alpha \mathbf{R}\hat{\mathbf{J}}^{\mathrm{w},s}_{t,j} + \mathbf{v} - \mathbf{J}^{\mathrm{w},s}_{t,j}\|_2^2. \tag{20}$$

All three errors are reported in millimetres and aggregated with matched hand-frame weighting. Alignment is used only for the corresponding metric; it does not modify W-MPJPE or the saved predictions.

B Loss Functions

All losses are averaged over valid annotations; an empty valid set contributes zero. For the mask, landmark, and MANO terms below, we write the per-mask, per-joint, or per-hand error and omit the outer sample average. Coordinate-wise and joint-wise normalization factors are retained explicitly. Predictions carry hats, and unhatted quantities are targets.

B.1 Mask and 2D Landmark Losses

For a valid hand mask containing P pixels, binary cross-entropy and soft Dice are

$$\ell_{\mathrm{BCE}} = -\frac{1}{P} \sum_{p=1}^{P} \left[M_p \log \hat{M}_p + (1 - M_p) \log(1 - \hat{M}_p) \right], \tag{21}$$

$$\ell_{\mathrm{Dice}} = 1 - \frac{2 \sum_p \hat{M}_p M_p + 1}{\sum_p \hat{M}_p + \sum_p M_p + 1}. \tag{22}$$

BCE is evaluated from logits for numerical stability, while Dice uses sigmoid probabilities. Averaging each term over valid masks gives $\mathcal{L}_{\mathrm{mask}} = \lambda_{\mathrm{BCE}}\mathcal{L}_{\mathrm{BCE}} + \lambda_{\mathrm{Dice}}\mathcal{L}_{\mathrm{Dice}}$. The landmark head is supervised directly in normalized crop coordinates:

$$\mathcal{L}_{\mathrm{2D}} = \|\hat{\mathbf{p}} - \mathbf{p}\|_1. \tag{23}$$

The valid set excludes missing, nonfinite, and out-of-range target landmarks.

B.2 MANO Reconstruction Losses

Let $R(\cdot) = \mathrm{Rot}(\cdot)$ convert an axis-angle vector to a rotation matrix, and let $d_{\mathrm{SO}(3)}$ denote the rotation distance defined below. Orientation and local pose use

$$\mathcal{L}_{\mathrm{orient}} = d_{\mathrm{SO}(3)}\Big(R(\hat{\mathbf{\Phi}}^{\mathrm{c}}), R(\mathbf{\Phi}^{\mathrm{c}})\Big), \tag{24}$$

$$\mathcal{L}_{\mathrm{pose}} = \frac{1}{15} \sum_{k=1}^{15} d_{\mathrm{SO}(3)}\Big(R(\hat{\mathbf{\Theta}}_k), R(\mathbf{\Theta}_k)\Big). \tag{25}$$

Shape and camera-frame MANO translation use coordinate-wise squared error:

$$\mathcal{L}_{\mathrm{shape}} = \frac{1}{10}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2, \qquad \mathcal{L}_{\mathrm{trans}} = \frac{1}{3}\|\hat{\mathbf{t}}^{\mathrm{c}} - \mathbf{t}^{\mathrm{c}}\|_2^2. \tag{26}$$

The 3D joint and reprojection terms are

$$\mathcal{L}^{\mathrm{c}}_{\mathrm{joints}} = \frac{1}{3}\|\hat{\mathbf{J}}^{\mathrm{c}} - \mathbf{J}^{\mathrm{c}}\|_1, \qquad \mathcal{L}_{\mathrm{reproj}} = \|\bar{\pi}_{\mathbf{K}}(\hat{\mathbf{J}}^{\mathrm{c}}) - \mathbf{p}^{\mathrm{img}}\|_1. \tag{27}$$

Here $\bar{\pi}_{\mathbf{K}}$ projects into normalized image coordinates and $\mathbf{p}^{\mathrm{img}}$ is the corresponding target. Validity is applied per joint; projections and targets use matching intrinsics and image normalization. Unlike $\mathcal{L}_{\mathrm{2D}}$, reprojection supervises joints decoded from MANO. The six weighted components form $\mathcal{L}_{\mathrm{MANO}}$ as defined in Sec. 3.2.

B.3 World-Space Joint and Camera Losses

The camera loss groups the absolute and relative pose terms as $\mathcal{L}_{\text{camera}} = \alpha_{\text{abs}}\mathcal{L}_{\text{abs-pose}} + \alpha_{\text{rel}}\mathcal{L}_{\text{rel-pose}}$. After applying the anchor transformation in Equation 10 and the same sample-level normalization to predictions and targets, the world-joint term compares the prediction with the dataset-provided joint annotations using a robust element-wise smooth-L_1 penalty ρ_ϵ (summed over coordinates):

$$\mathcal{L}^{\text{w}}_{\text{joints}} = \frac{1}{N_{\text{J}}} \sum_{t,s,j} m^{\text{J}}_{t,s} \rho_\epsilon \left(\hat{\tilde{\mathbf{J}}}^{\text{w},s}_{t,j} - \tilde{\mathbf{J}}^{\text{w},s}_{t,j} \right). \tag{28}$$

Here $\rho_\epsilon(r) = r^2/(2\epsilon)$ for $|r| < \epsilon$ and $|r| - \epsilon/2$ otherwise, and $m^{\text{J}}_{t,s}$ denotes hand validity and N_{J} counts valid joint coordinates. Importantly, the target is derived from the original joint annotation rather than joints decoded from ground-truth MANO parameters.

The rotation distance is

$$d_{\text{SO}(3)}(\hat{\mathbf{R}}, \mathbf{R}) = \arccos \left[\text{clamp} \left(\frac{\text{tr}(\hat{\mathbf{R}}^\top \mathbf{R}) - 1}{2}, -1 + \epsilon, 1 - \epsilon \right) \right]. \tag{29}$$

The absolute-pose term supervises anchor-normalized camera translation, quaternion, and field of view. Because $\mathbf{q}$ and $-\mathbf{q}$ represent the same rotation, we first define the sign-aligned target

$$\mathbf{q}^\star_t = \begin{cases} \mathbf{q}_t, & \langle \hat{\mathbf{q}}_t, \mathbf{q}_t \rangle \geq 0, \\ -\mathbf{q}_t, & \text{otherwise}, \end{cases} \tag{30}$$

We then compute robust means over valid frames for $\hat{\tilde{\mathbf{u}}}_t - \tilde{\mathbf{u}}_t$, $\hat{\mathbf{q}}_t - \mathbf{q}^\star_t$, and $\hat{\mathbf{f}}_t - \mathbf{f}_t$, where $\mathbf{f}_t$ contains horizontal and vertical field-of-view angles and $\mathbf{q}_t$ is a unit quaternion.

Equivalently, with ρ_ϵ applied coordinate-wise and summed, the nine-coordinate absolute loss is

$$\mathcal{L}_{\text{abs-pose}} = \frac{1}{9|\mathcal{V}_{\text{cam}}|} \sum_{t \in \mathcal{V}_{\text{cam}}} \left[\rho_\epsilon(\hat{\tilde{\mathbf{u}}}_t - \tilde{\mathbf{u}}_t) + \rho_\epsilon(\hat{\mathbf{q}}_t - \mathbf{q}^\star_t) + \rho_\epsilon(\hat{\mathbf{f}}_t - \mathbf{f}_t) \right], \tag{31}$$

where $\mathcal{V}_{\text{cam}}$ contains frames with valid camera annotations.

To constrain local camera motion independently of the anchor choice, the relative-pose term considers ordered pairs of distinct valid frames in the pose-reference window $\mathcal{W}_{\text{pose}}$. Let $\mathbf{C}_t = \begin{bmatrix} \mathbf{R}_t & \mathbf{u}_t \\ \mathbf{0}^\top & 1 \end{bmatrix}$ denote the homogeneous camera-to-world transform associated with P_t. For frames i and j in this window, we define

$$\mathbf{C}_{j \leftarrow i} = \mathbf{C}_j^{-1} \mathbf{C}_i, \tag{32}$$

and optimize

$$\mathcal{L}_{\text{rel-pose}} = \frac{1}{N_{\text{pair}}} \sum_{\substack{i \neq j \\ i,j \in \mathcal{W}_{\text{pose}}}} m^{\text{cam}}_i m^{\text{cam}}_j \left[d_{\text{SO}(3)}(\Delta\hat{\mathbf{R}}_{ji}, \Delta\mathbf{R}_{ji}) + \lambda_{\text{rel,u}} \left\| \Delta\hat{\mathbf{u}}_{ji} - \Delta\mathbf{u}_{ji} \right\|_1 \right]. \tag{33}$$

Here m^{cam}_t indicates camera validity, N_{pair} counts valid ordered pairs, and $\Delta\mathbf{R}_{ji}$ and $\Delta\mathbf{u}_{ji}$ are the rotation and translation of $\mathbf{C}_{j \leftarrow i}$. The camera translations entering this term have already been normalized by κ and are not divided by the scale again.

Each term in $\mathcal{L}_{\text{joint}}$ is evaluated only where its required supervision is available. Camera and world-joint terms are omitted for samples without the corresponding valid annotations. The relative-pose term is active only when at least two valid poses are present.

C Binary Keyframe Pool for Sparse Bundle Adjustment

We maintain a time-ordered pool $\mathcal{P} = \{(f_i, b_i)\}_{i=1}^n$, where f_i stores a frame and its associated geometric state, and $b_i \in \{0, 1\}$ records whether it has been retained after a previous BA selection. Each incoming frame is appended with $b_i = 0$. When $|\mathcal{P}| > N$, we select every K-th pool entry, starting from the oldest:

$$\mathcal{S} = \{ f_{1+mK} \mid m \geq 0,\ 1 + mK \leq |\mathcal{P}| \}. \tag{34}$$

The DROID-SLAM backend (Teed & Deng, 2021) first refines the selected camera poses. We then toggle the selected entries' bits and remove every entry whose updated bit is zero:

$$b_i^+ = b_i \oplus \mathbf{1}[f_i \in \mathcal{S}], \qquad \mathcal{P}^+ = \{(f_i, b_i^+) \mid (f_i, b_i) \in \mathcal{P},\ b_i^+ = 1\}, \tag{35}$$

where $\oplus$ denotes binary exclusive-or. A newly selected frame changes from 0 to 1 and remains in the pool; a retained frame selected again changes from 1 to 0 and is retired after contributing to that BA update. Unselected retained frames remain available, while unselected new frames are discarded. Thus recent observations enter densely, whereas only sampled historical frames survive between refinement calls. Starting selection from the oldest entry also ensures that retained frames are eventually revisited and retired. This policy lets BA reuse historical constraints alongside recent observations without keeping a permanent set of old keyframes. The refined camera poses are used for world-space hand conversion; the hand predictor itself remains feed-forward.

The trigger is evaluated after new frames are inserted, and selection follows chronological pool order rather than a fixed stride in the original video. Refinement uses the selected frames before their retention flags are updated. Thus an old frame selected for retirement still contributes to that BA call. The trigger threshold N controls when refinement runs, while K controls sampling sparsity; this retention policy does not impose a strict exponential distribution over frame ages.

D Additional Qualitative Results

Figure 5 provides additional camera-space comparisons on Xperience-10M (Ropedia, 2026). Figure 6 provides additional world-space comparisons on ARCTIC (Fan et al., 2023), HOT3D (Banerjee et al., 2025), and EgoDex (Hoque et al., 2025). Selected clips illustrate behavior under changing visibility and viewpoint; numerical conclusions use the complete evaluated scenes.

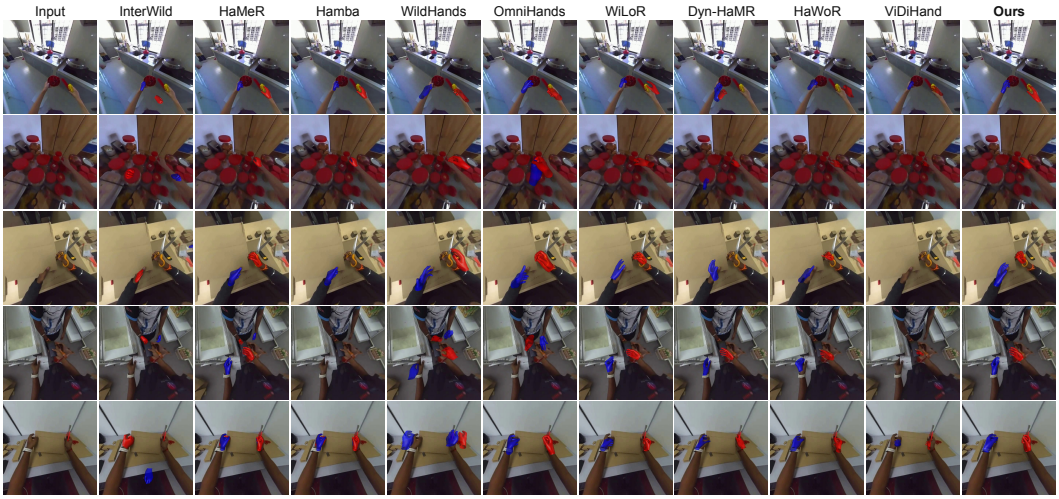

Figure 5: **Additional in-the-wild camera-space results.** Comparisons on Ego4D and Xperience illustrate hand reconstruction under diverse viewpoints, occlusions, and interactions.

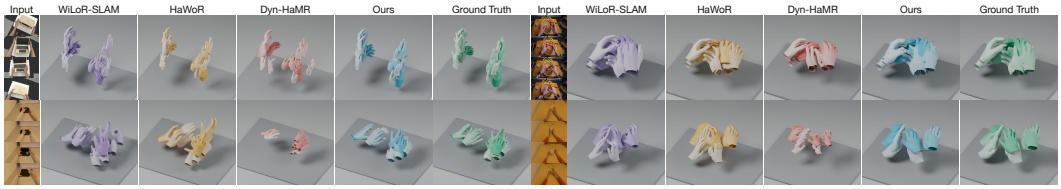

Figure 6: **Additional world-space qualitative results.** Comparisons on ARCTIC, HOT3D, and EgoDex show reconstructed hand configurations and motion across further sequences.