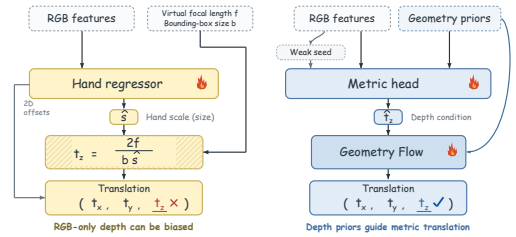
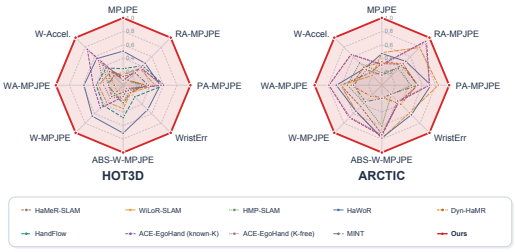
METRICHAND: GEOMETRY-AWARE FLOW MATCHING FOR 4D HAND RECONSTRUCTION IN METRIC WORLD SPACE

Anonymous authors
Paper under double-blind review



(a) Metric Translation Prediction



(b) Quantitative Comparisons on HOT3D and ARCTIC



(c) HOT3D and ARCTIC hand motion

Figure 1: **MetricHand: geometry-guided 4D hand reconstruction.** (a) Weak-perspective projection can bias translation (left), whereas our method uses depth priors to guide metric translation estimation (right). (b) Quantitative comparisons on HOT3D and ARCTIC under a shared evaluation protocol; higher is better. (c) Qualitative world-space comparisons, with GT shown in green, ours in blue, and HaWoR in orange. Closer alignment with the green GT meshes indicates more accurate hand-pose reconstruction in the world coordinate system

ABSTRACT

Egocentric hand motion provides physical motion supervision for dexterous manipulation and vision-language-action pretraining. Reliable use of this supervision requires accurate and temporally coherent 4D hand reconstruction in metric world space. Existing methods typically lift framewise estimates produced under weak perspective into world coordinates. By ignoring joint-dependent depth variation at close range, this approximation couples errors in translation, orientation, and articulation that persist after world lifting and worsen under occlusion. We present METRICHAND, a geometry-aware flow matching framework for 4D hand reconstruction in metric world space. METRICHAND encodes depth priors from hand–object interaction regions into joint-associated representations that preserve local spatial structure and metric scale. Fused with visual features, this geometry guides coarse metric estimation and conditions every step of temporal generation. The model jointly recovers world-space translation, global orientation, MANO articulation, and shape, while structural priors and temporal context support estimation under partial occlusion. METRICHAND reduces camera-space MPJPE by 49.3% and 52.9%, absolute world-space MPJPE by 28.7% and 21.7%, and wrist localization error by 48.5% and 37.9% relative to HaWoR, respectively.

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.

World-space acceleration error decreases by 44.2% and 64.3%, demonstrating improved temporal fidelity alongside more accurate absolute localization. Our project code is available at https://anonymous.4open.science/api/repo/metrichand-review-C230/file/index.html.

1 INTRODUCTION

Hand motion lies at the center of how humans interact with the physical world. In egocentric video, wrist position and orientation specify end-effector targets, while local hand pose defines the articulated configuration that can be transferred to a robot hand through retargeting. These states turn human manipulation videos into scalable physical motion supervision for dexterous manipulation learning and vision-language-action (VLA) pretraining (Kareer et al., 2025; Hoque et al., 2026; Yang et al., 2025). Localization errors can shift hand–object interaction targets, and temporal jitter can propagate into generated trajectories. These applications therefore require accurate and temporally coherent 4D hand motion in a common world coordinate system with metric scale.

Existing egocentric 4D hand reconstruction commonly combines two stages. Image-based methods such as HaMeR and WiLoR recover detailed hand observations and visual features from individual frames (Pavlakos et al., 2024; Potamias et al., 2025). Temporal methods such as HaWoR and HandFlow combine these estimates with camera motion to reconstruct world-space trajectories (Zhang et al., 2025; Xu et al., 2026). These approaches improve temporal consistency under moving cameras, but they still depend on the image frontend to establish the geometric state of the hand in each frame. When that estimate is biased or the hand is occluded, temporal processing must recover motion from incomplete and potentially distorted evidence.

Metric hand reconstruction requires resolving the depth ambiguity left by weak-perspective estimates and camera-normalized rays. HandFlow (Xu et al., 2026) conditions flow matching on visual features and skeletal rays, with detection confidence controlling observation strength. StableHand (Zeng et al., 2026) uses predicted observation quality to adapt the flow path and initialization, alongside hand, camera, and scene-geometry conditions. Our focus is the spatial organization of geometric evidence: associating metric depth and nearby hand–object surfaces with individual joints so that the generator can use these cues for both placement and articulation.

Our key insight is that a topology-aware representation of hand–object interaction (HOI) geometry can serve as a shared condition throughout 4D hand reconstruction. Local surface structure provides cues for recovering hand articulation under occlusion, while metric depth guides hand localization in world coordinates. Integrating this geometric evidence with temporal memory further supports motion consistency across frames and trajectory continuity across windows. Based on this insight, METRICHAND fuses RGB and estimated depth for coarse metric hand estimation, then samples appearance and surface geometry around 21 projected joint anchors. Quality weighting and temporal memory organize these features across windows, and the resulting conditions guide every flow update to jointly recover world-space translation, global orientation, articulation, and shape. Our geometry-conditioning ablations show that maintaining this evidence throughout generation improves reconstruction over using depth only for coarse estimation or supplying geometry only at the first flow step (Table 3).

Under a shared protocol on HOT3D and ARCTIC, the evaluated configuration achieves the lowest wrist localization error and absolute camera- and world-space joint errors among the reevaluated methods. Relative to HaWoR, absolute world-space MPJPE decreases by 28.7% and 21.7%, and world-space acceleration error by 44.2% and 64.3%, respectively. These results demonstrate improvements in absolute localization and motion fidelity (Fig. 1). Component and condition ablations assess geometric representation, flow generation, and temporal memory.

Our contributions are:

- We propose METRICHAND, which associates the spatial structure and metric scale of HOI depth priors with hand topology, providing shared geometric conditions for metric localization and occlusion-robust pose estimation.

- We formulate geometry-aware temporal flow matching that combines quality-aware conditions with overlapping memory to jointly generate translation, global orientation, articulation, and shape.
- We achieve state-of-the-art hand reconstruction on HOT3D, ARCTIC, DexYCB, and HO3D, as well as on unseen real-world HOI4D videos under zero-shot evaluation (Appendix B).

2 Related Work

Image-based Hand Reconstruction. MANO represents hand shape and articulation with a differentiable parametric mesh (Romero et al., 2017). Transformer-based models such as HaMeR and WiLoR recover detailed hands from images (Pavlakos et al., 2024; Potamias et al., 2025); Hamba and HandOS explore alternative spatial reasoning and localization architectures (Dong et al., 2024; Chen et al., 2025). These models provide visual features and articulation priors. Recovering metric motion additionally requires geometric evidence that connects the reconstructed pose to its physical placement.

Hand Motion Estimation and Generation. Temporal methods combine image evidence with motion priors to reconstruct hand trajectories. HMP learns a hand-motion prior (Duran et al., 2024), while Dyn-HaMR and HaWoR recover motion under moving cameras (Yu et al., 2025; Zhang et al., 2025). ViDiHand adapts video-diffusion representations for hand reconstruction (Wang et al., 2026). ACE-Ego-Hand uses a deterministic clean-latent encoder and a bidirectional spatiotemporal decoder to recover bimanual trajectories under occlusion and out-of-sight gaps (Liu et al., 2026). MINT jointly predicts camera motion, field of view, camera-frame hand states, and observability from a shared video representation, using large-scale pipeline supervision (Zhu et al., 2026). HandFlow formulates hand recovery with flow matching and confidence-aware observation conditioning (Xu et al., 2026; Lipman et al., 2023). Our focus is the explicit geometric evidence supplied to temporal generation: topology-aware sampling associates HOI appearance, metric depth, and local surface structure with hand joints, and the resulting conditions remain available throughout the flow process.

Metric Depth Estimation. Monocular geometry models such as Metric3D, Metric3D v2, and MoGe-2 recover depth and metric-scale cues from a single image (Yin et al., 2023; Hu et al., 2024; Wang et al., 2025a). Their spatial predictions complement the appearance features of a normalized hand crop. BLADE uses depth to address close-range perspective distortion in body reconstruction (Wang et al., 2025b). For articulated hand motion, depth provides both absolute placement cues and the local layout of the interaction region. We fuse this evidence with RGB features, sample it around hand joints, and preserve it across time as a shared condition for localization and temporally coherent pose estimation. Camera-motion estimation supplies the reference needed to express this geometry in world coordinates (Teed & Deng, 2021; Zhang et al., 2025).

3 Method

Overview. MetricHand uses interaction geometry to connect metric hand placement with temporal pose recovery (Fig. 2). RGB features and depth priors first produce coarse hand estimates, which are expressed in a metric world frame using the estimated camera trajectory (Sec. 3.2). Topology-aware sampling then gathers appearance and surface geometry around the projected hand joints. Observation quality and cross-window memory organize this evidence into joint-associated conditions (Sec. 3.3). A conditional flow model reads these conditions at every integration step to jointly recover world-space translation, global orientation, articulation, and shape (Sec. 3.4).

3.1 Preliminaries

Problem Definition. Given a monocular video $\{I_t\}_{t=1}^{T}$ and intrinsics K_t, we use estimated depth D_t and camera-to-world poses $\mathcal{T}_t^{WC} = (R_t^{WC}, c_t^{WC})$ to reconstruct both hands $h \in \{L, R\}$ in a common metric world frame.

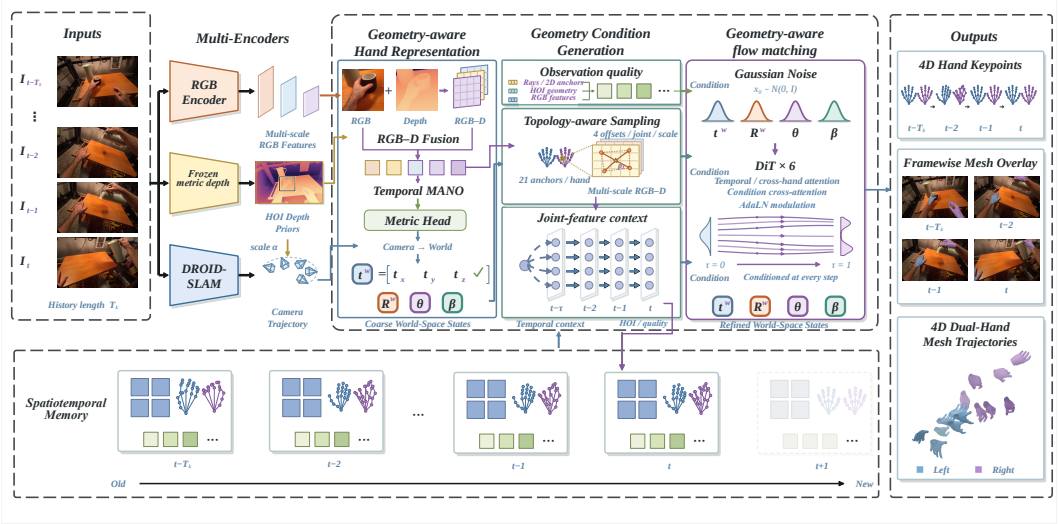


Figure 2: **Overview of MetricHand.** Depth priors supply metric placement cues and local hand–object surface geometry. RGB–D fusion supports coarse hand estimation, while topology-aware sampling associates multi-scale evidence with projected joint anchors. Observation quality and a spatiotemporal memory queue preserve reliable joint-associated context across windows. The resulting conditions are supplied to every flow-matching step to jointly recover world-space wrist translation, global orientation, articulation, and shape. Camera motion defines the metric world reference.

4D Hand Motion Representation. We represent each hand with MANO (Romero et al., 2017). At frame t, $\mathbf{t}_{t,h}^{W}$ is the global translation defined by the physical wrist, $R_{t,h}^{W}$ the global orientation, $\theta_{t,h}$ the 15 local joint rotations, and $\beta_h \in \mathbb{R}^{10}$ the hand shape. Translation, orientation, and articulation vary by frame, while shape is shared within a window. Separate observation and annotation masks control conditioning and supervision.

3.2 Geometry-Aware Hand Representation

Multi-Encoders. Following image-based hand reconstruction (Pavlakos et al., 2024; Potamias et al., 2025), a trainable WiLoR-initialized encoder extracts hand–object appearance features F^I. A frozen monocular geometry model supplies full-frame metric depth (Hu et al., 2024; Wang et al., 2025a), and frozen DROID-SLAM estimates relative camera motion (Teed & Deng, 2021). Depth is aligned to each hand crop with the RGB transform. RGB–D fusion, Temporal MANO, and the Metric Head are trained jointly, while the depth and SLAM networks remain fixed.

RGB–D Fusion. The representation combines dense depth injection with joint-associated geometric aggregation. The depth encoder processes normalized log depth, validity, and spatial gradients to preserve local surface layout, while absolute-depth statistics retain metric context. Depth features are aligned with the RGB feature map and injected through a spatially gated residual:

$$F^{RD} = F^{I} + \text{sigmoid}\left(a + g(\text{GN}(F^{D}))\right) \odot \text{GN}(F^{D}), \tag{1}$$

where g is a learned projection, a a learned bias, and GN denotes group normalization. Temporal MANO uses F^{RD} to predict coarse orientation, articulation, camera translation, and shared shape. The fused map also supplies the multi-scale evidence for topology-aware sampling in Sec. 3.3. Projected coarse joints index this evidence by anatomical identity, allowing the subsequent flow model to use local hand–object geometry together with metric placement cues.

Metric Head. Crop scale and weak perspective leave translation ambiguous at close range (Wang et al., 2025b). Our Metric Head uses fused features, crop geometry, and absolute depth statistics to refine the coarse MANO translation $\tilde{\mathbf{t}}^{C}$ in viewing-ray and log-depth coordinates:

$$\hat{r} = \tilde{r} + 0.25\delta_{xy}, \qquad \hat{\zeta} = \text{clip}(\tilde{\zeta} + \delta_z, \log z_{\min}, \log z_{\max}), \qquad \hat{\mathbf{t}}^{C}_{\text{MANO}} = e^{\hat{\zeta}}[\hat{r}_x, \hat{r}_y, 1]^{\top}, \tag{2}$$

where $\delta = \tanh H_{\text{metric}}(\cdot)$, $\widetilde{r} = \widetilde{\mathbf{t}}^C_{xy} / \max(\widetilde{t}^C_z, \epsilon)$, and $\widetilde{\zeta}$ is clipped log depth. After adding the MANO root offset, the physical wrist and orientation are transformed by the camera-to-world pose. We mask dynamic hand–object regions and recover DROID-SLAM scale by matching its depth to the metric prior on valid static pixels, following metric-scaled world lifting (Zhang et al., 2025). The coarse states and later conditions therefore share one metric world frame.

3.3 Geometry Condition Generation

The condition generator turns crop features into structured evidence. Joint anchors align RGB–D features with hand topology, quality prediction suppresses unreliable content, and a spatiotemporal queue links joint identities across windows. The output preserves local hand–object geometry, global context, and temporal correspondence for the flow model.

Topology-Aware Sampling. We construct an HOI representation indexed by anatomical joints. The 21 projected coarse MANO joints provide stable anchors for gathering appearance and geometry from nearby hand and object surfaces. Each joint query predicts offsets on multi-scale RGB–D features, following deformable sampling (Zhu et al., 2021). Content-based attention is augmented with image-distance, log-depth, and backprojected surface-distance biases, so each joint token retains the spatial organization of its supporting evidence. Object surfaces provide context for pose recovery under occlusion rather than joint-position targets. The query predicts $u_{j,k} = u_j + \Delta u_{j,k}$. Backprojecting valid depth as $P^C(u) = D(u)K^{-1}\bar{u}$, we define a geometric bias:

$$\begin{aligned} b_{j,k} = & -a_u \|\chi(u_{j,k}) - \chi(u_j)\|_2 - a_z |\log D(u_{j,k}) - \log D(u_j)| \\ & - a_P \frac{\|P^C(u_{j,k}) - P^C(u_j)\|_2}{\max(D(u_0), \epsilon)}. \end{aligned} \tag{3}$$

Here $\bar{u}$ is the homogeneous pixel coordinate, χ maps pixels to normalized image coordinates, u_0 is the wrist anchor, and a_u, a_z, a_P are nonnegative learned coefficients. Depth-dependent terms use valid samples only. Combining this bias with the content logit $a_{j,k}$ gives

$$w_{j,k} = \text{softmax}_k(a_{j,k} + b_{j,k}), \qquad f_j = z_j + \nu_j W_O \sum_k w_{j,k} W_V F^{RD}(u_{j,k}), \tag{4}$$

where ν_j gates the update by valid geometric support. Each f_j associates appearance and surface geometry with one joint. Mask pooling, box encoding, and HOI pooling provide complementary hand-level and interaction context; Appendix E specifies the sampling rules.

Observation Quality. Inspired by confidence-aware conditioning in HandFlow (Xu et al., 2026), we separate availability from reliability. Hand- and joint-level heads combine RGB–D descriptors with detection confidence, depth support, camera features, and the coarse state into z_i. For condition slot i, a learned quality score weights its content c_i against a placeholder:

$$q_i = Q_\phi(z_i) \in [0, 1], \quad \widetilde{c}_i = o_i q_i c_i + (1 - o_i q_i) e_{\text{placeholder}} + e^{\text{id}}_i. \tag{5}$$

Here o_i indicates availability, while e^{id}_i preserves source, time, hand, and joint identity. Hand-level scores weight regional tokens and joint-level scores weight joint tokens; modality masks retain source validity. Quality scores are learned jointly through the reconstruction objectives, without explicit quality targets or an auxiliary loss.

Spatiotemporal Memory Queue. The queue carries topology-aware HOI features across windows. Each joint token $g_{t,h,j}$ is obtained from the sampled descriptor f_j in Eq. 4, with observation quality and identity encoding applied as in Eq. 5. It therefore retains local appearance and surface geometry alongside its timestamp, hand identity, and joint identity. Spatial attention exchanges evidence between joints within a frame, and temporal attention follows the same anatomical joint through the current window and queued features. This correspondence allows a partially occluded joint to retrieve contextual evidence from visible joints and preceding observations. We maintain the FIFO queue $\mathcal{Q}_n$ indexed by timestamp, hand, and joint identity (Zhang et al., 2025; Xu et al., 2026):

$$\begin{aligned} g^S_{t,h,j} &= g_{t,h,j} + \text{Attn}\left(g_{t,h,j}, \{g_{t,h',k}\}_{h',k}\right), \\ g^{ST}_{t,h,j} &= g^S_{t,h,j} + \text{Attn}\left(g^S_{t,h,j}, \{g^S_{s,h,j}\}_{s \in \mathcal{W}_n \cup \mathcal{T}(\mathcal{Q}_n)}\right). \end{aligned} \tag{6}$$

Here $\mathcal{W}_n$ is the current window; $\mathcal{T}(\mathcal{Q}_n)$ contains only earlier timestamps, so the two sets are disjoint. Invalid observations are masked as keys, while placeholder queries can retrieve valid context. As the window advances, departed timestamps enter the history with detached features and predicted states; duplicate entries are merged and the oldest are evicted. Predicted states retain their observation-validity labels. Appendix E specifies the window and history lengths.

Condition Encoding. The encoder combines camera-ray cues, 2D skeletal anchors, topology-aware HOI features, and observation quality into C_n, projecting current and queued evidence to a shared width while retaining source, time, hand, and joint identities. C_n supplies cross-attention, while its pooled representation $\bar{C}_n$ modulates the flow blocks.

3.4 Geometry-Aware Flow Matching

Direct framewise regression lacks a unified temporal update when local evidence is biased or partially occluded. Building on flow matching and its use for generative hand recovery (Lipman et al., 2023; Xu et al., 2026), we map a two-hand window to motion tokens with component, time, and hand identities. Rotations use continuous 6D coordinates (Zhou et al., 2019), translation is normalized, and shape is shared within the window. The velocity field $v_\phi(x_\tau, \tau, C_n)$ transports Gaussian noise toward the joint state distribution, with the coarse estimate supplied through C_n rather than used as the initial sample.

Condition-Guided Generation. DiT-style blocks (Peebles & Xie, 2023; Xu et al., 2026) couple motion tokens across state components, time, and hands through self-attention, and retrieve C_n through cross-attention. Flow time τ and pooled context $\bar{C}_n$ modulate normalization and residual gates. Every velocity evaluation reads C_n: depth guides placement, while joint-associated HOI and memory features guide articulation and temporal consistency. Appendix E gives the block equations.

Flow Integration. Following the rectified-flow formulation used by HandFlow (Lipman et al., 2023; Xu et al., 2026), we connect Gaussian motion x_0 and normalized ground truth x_1 with a linear path:

$$x_\tau = (1-\tau)x_0 + \tau x_1, \qquad u_\tau = x_1 - x_0,$$
$$\mathcal{L}_{\mathrm{FM}} = \sum_k \mathbb{E}\big[\|v_\phi^k(x_\tau, \tau, C_n) - u_\tau^k\|^2\big]_{\mathrm{valid},k}. \tag{7}$$

Here k indexes shape, global orientation, articulation, and translation. Each component is averaged over its valid annotations before summation, preventing high-dimensional articulation from dominating the other states. Following HandFlow's few-step setting, inference integrates $dx_\tau/d\tau = v_\phi(x_\tau, \tau, C_n)$ with three Euler steps (Xu et al., 2026). For long videos, repeated global timestamps share their initial noise as in HandFlow (Xu et al., 2026); our implementation then blends window endpoints with triangular weights. Appendix E gives the network and blending details.

Training Objectives. We supervise the MANO-decoded endpoint $\widehat{x}_1 = x_\tau + (1-\tau)v_\phi(x_\tau, \tau, C_n)$. For the physical wrist residual $\mathbf{e}^W_{t,h} = \widehat{\mathbf{t}}^W_{t,h} - \mathbf{t}^W_{t,h}$ in metres, temporal and world-space translation supervision are

$$\mathcal{L}_{\mathrm{temp}} = \frac{1}{N_{\mathrm{temp}}} \sum_{t=2}^{L-1} \sum_{h\in\{L,R\}} m^{\mathrm{temp}}_{t,h}\, \rho_{\delta_{\mathrm{temp}}}\big(\mathbf{e}^W_{t+1,h} - 2\mathbf{e}^W_{t,h} + \mathbf{e}^W_{t-1,h}\big),$$
$$\mathcal{L}_{\mathrm{trans}} = \frac{1}{N_{\mathrm{trans}}} \sum_{t=1}^{L} \sum_{h\in\{L,R\}} m_{t,h}\, \rho_{\delta_{\mathrm{trans}}}\Big(\widehat{\mathbf{t}}^W_{t,h} - \mathbf{t}^W_{t,h}\Big), \tag{8}$$

Here ρ_δ is coordinate-averaged Smooth L1; m masks valid wrists or consecutive triplets, and N counts them. The temporal term matches target accelerations. The total objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{FM}} + \mathcal{L}_{\mathrm{metric}} + \mathcal{L}_{\mathrm{pose}} + \lambda_{\mathrm{temp}}\mathcal{L}_{\mathrm{temp}}, \tag{9}$$

The metric group includes $\lambda_{\mathrm{trans}}\mathcal{L}_{\mathrm{trans}}$. Appendix E specifies the remaining metric and pose terms, weights, thresholds, masks, and normalizers.

Table 1: **Comparison on HOT3D and ARCTIC.** Position errors are in mm and world acceleration error in m/s^2 (↓). 1st , 2nd , and 3rd rank available values within each dataset. Gain compares Ours with the actual second-ranked error.

	Method	Camera				World			
		MPJPE ↓	RA-MPJPE ↓	PA-MPJPE ↓	WristErr ↓	ABS-W-MPJPE ↓	W-MPJPE ↓	WA-MPJPE ↓	Accel. ↓
HOT3D	HaMeR-SLAM	67.87	58.23	9.57	52.29	73.28	148.23	38.43	15.91
	WiLoR-SLAM	53.78	16.58	6.70	48.89	59.21	75.90	24.65	11.39
	HMP-SLAM	73.44	62.89	11.70	59.54	78.77	152.00	40.31	7.39
	Dyn-HaMR	66.06	55.46	9.19	49.98	274.08	207.97	67.98	8.85
	HaWoR	16.97	12.65	5.77	14.99	28.88	41.09	15.47	5.57
	HandFlow	34.89	14.26	5.23	31.45	42.59	63.27	22.56	8.63
	ACE-EgoHand (known-K)	113.92	16.19	5.31	102.06	118.43	55.95	20.24	4.13
	ACE-EgoHand (K-free)	118.66	15.35	5.66	110.66	122.61	98.49	31.12	6.03
	MINT	66.47	22.68	8.50	70.49	197.46	111.93	41.83	10.98
	Ours	**8.60**	**5.84**	**3.18**	**7.72**	**20.58**	**26.71**	**9.00**	**3.11**
	Gain vs. 2nd	49.3%	53.8%	39.2%	48.5%	28.7%	35.0%	41.8%	24.7%
ARCTIC	HaMeR-SLAM	54.56	25.67	9.68	56.05	166.19	83.51	31.21	17.14
	WiLoR-SLAM	36.23	14.23	6.18	34.56	152.89	68.80	31.39	21.62
	HMP-SLAM	95.82	32.63	11.89	100.61	199.30	99.51	37.31	9.38
	Dyn-HaMR	50.56	25.20	9.87	50.12	671.54	160.91	41.91	11.59
	HaWoR	37.26	22.45	7.10	32.23	154.36	68.52	28.17	12.28
	ACE-EgoHand (known-K)	56.57	12.04	7.30	57.56	160.71	54.66	24.25	6.76
	ACE-EgoHand (K-free)	58.70	12.28	7.32	60.39	163.84	55.78	24.23	6.84
	MINT	116.45	27.23	9.68	106.91	621.82	107.94	45.37	13.49
	Ours	**17.54**	**11.28**	**5.18**	**20.02**	**120.86**	**37.40**	**19.08**	**4.39**
	Gain vs. 2nd	51.6%	6.3%	16.2%	37.9%	20.9%	31.6%	21.3%	35.1%

4 Experiments

4.1 Experimental Setup

Datasets. We follow the observed-track evaluation convention of HaWoR (Zhang et al., 2025) and compare methods on the same valid observed frames of HOT3D and ARCTIC (Banerjee et al., 2025; Fan et al., 2023), both captured with moving egocentric cameras. Dataset splits and frame-selection rules are provided in Appendix D.

Metrics. We report camera MPJPE and absolute world MPJPE (ABS-W-MPJPE) for joint localization. WristErr is the Euclidean error of the single physical wrist point in camera coordinates. RA-MPJPE subtracts the wrist, whereas PA-MPJPE fits a per-frame similarity transform. World W-MPJPE and WA-MPJPE use blockwise similarity alignment fitted to the first two frames or the full block, respectively (Zhang et al., 2025). Position errors are in millimetres. World acceleration error, in m/s^2, measures disagreement with ground-truth joint dynamics on the original timeline. The world reference is established by rigid camera alignment, with no further hand alignment for ABS-W-MPJPE. Appendices C and D provide metric definitions and evaluation details.

4.2 Comparison with State of the Art

Baselines. We compare against SLAM-lifted HaMeR, WiLoR, and HMP (Pavlakos et al., 2024; Potamias et al., 2025; Duran et al., 2024), and video-based methods including Dyn-HaMR, HaWoR, HandFlow, ACE-Ego-Hand, and MINT (Yu et al., 2025; Zhang et al., 2025; Xu et al., 2026; Liu et al., 2026; Zhu et al., 2026). Baseline configurations are detailed in Appendix D. Dyn-HaMR retains its native optimized world trajectory.

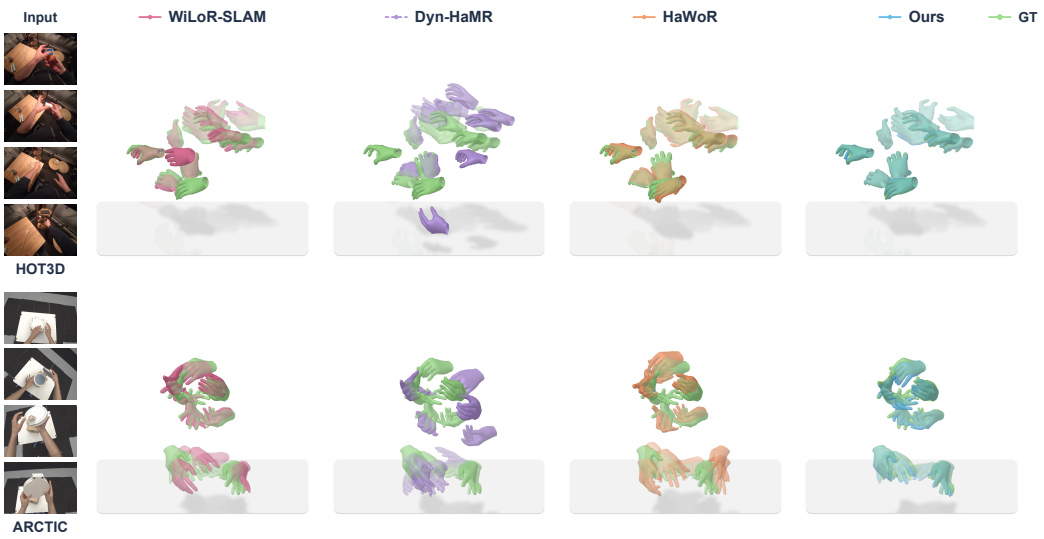


Figure 3: **Qualitative comparison of world-space hand trajectories on HOT3D and ARCTIC.** Input frames run from top to bottom within each row. Green meshes show ground truth; pink, purple, orange, and blue show WiLoR-SLAM, Dyn-HaMR, HaWoR, and MetricHand, respectively. Closer overlap with the green GT meshes indicates higher hand-pose accuracy along the displayed world-space trajectories.

Quantitative Results. Table 1 shows that METRICHAND achieves the lowest absolute joint and wrist localization errors among the reevaluated methods on both datasets. We analyze the results in terms of metric localization, pose accuracy, and temporal consistency.

For *metric localization*, ABS-W-MPJPE reaches 20.58 and 120.86 mm on HOT3D and ARCTIC, reducing HaWoR's errors by 28.7% and 21.7%, respectively. Wrist localization error also decreases by 48.5% and 37.9%, indicating more accurate physical placement of the reconstructed hands.

For *pose accuracy*, PA-MPJPE is 3.18 and 5.18 mm on the two datasets, the lowest values in the tables. Camera-space MPJPE reaches 8.60 and 17.54 mm, showing improvements in both similarity-aligned pose and absolute hand geometry.

For *temporal consistency*, world-space acceleration error is 3.11 and 4.39 m/s^2, improving over HaWoR by 44.2% and 64.3%, respectively. These results show closer agreement with ground-truth motion dynamics alongside improved localization.

Qualitative Comparison. Figure 3 compares the reconstructed motion during object manipulation. Read the RGB frames from top to bottom, then follow the overlaid hand states from translucent early poses to opaque late poses. Each method is shown from the same view with green GT meshes overlaid. MetricHand follows the changing palm orientation and finger configuration in both examples. In HOT3D, its palm and finger contours remain close to GT as the hands move around the object; in ARCTIC, it preserves the transition between open and flexed fingers. These examples illustrate the articulated geometry recovered by the RGB-D-conditioned flow. Further examples appear in Fig. 4 (Appendix F).

4.3 ABLATION STUDY

We emphasize camera-space MPJPE and evaluate every ablation with the same SLAM trajectory. Table 2 compares depth representations; Table 3 reports component and geometry-conditioning ablations on HOT3D and ARCTIC.

Depth Representation. Our RGB-D baselines concatenate image and depth features or fuse them through cross-attention before direct regression. On ARCTIC, these reduce RGB-only MPJPE from 36.34 to 31.65 and 29.64 mm. Topology-aware sampling associates HOI geometry with hand joints

Table 2: **Depth representation ablation on ARCTIC.** The first four controls share a metric regression head. Position errors are in mm (↓).

	Camera			World
Configuration	MPJPE	PA-MPJPE	WristErr	ABS-W
RGB-only regression	36.34	6.78	35.27	151.74
RGB-D regression (concatenation)	31.65	6.26	30.98	147.36
RGB-D regression (cross-attention)	29.64	6.01	28.56	142.14
RGB-D regression (topology-aware)	19.98	5.79	23.89	127.49
Topology-aware conditions + flow	17.54	5.18	20.02	120.86

and reaches 19.98 mm, improving on the two fusion baselines by 36.9% and 32.6%. In the full-model ablation, removing topology-aware sampling raises MPJPE from 17.54 to 38.91 mm. Both comparisons support the value of structured geometric conditions.

Flow Matching. Replacing flow with direct regression retains all conditions, the queue, and the window length. Flow reduces ARCTIC MPJPE from 19.98 to 17.54 mm (12.2%) and acceleration error from 7.93 to 4.39 m/s^2 (44.6%). HOT3D shows the same trend. Flow therefore improves pose accuracy and agreement with ground-truth dynamics under the same geometric conditions.

Memory Queue. Removing history raises MPJPE from 8.60 to 15.42 mm on HOT3D and from 17.54 to 25.67 mm on ARCTIC. ARCTIC acceleration error also rises from 4.39 to 6.31 m/s^2, showing that cross-window context improves both reconstruction and temporal consistency.

HOI Context and Quality. On ARCTIC, removing regional HOI context or quality weighting raises MPJPE to 32.42 or 25.35 mm, while PA-MPJPE changes from 5.18 to 5.49 or 5.40 mm. Their larger effects on unaligned error indicate a stronger contribution to metric placement than to similarity-aligned pose. Every component removal worsens all reported metrics on both datasets.

5 Conclusions and Limitations

Conclusion. We presented MetricHand, a geometry-aware flow matching framework for reconstructing 4D hand motion in metric world space. Our key insight is to use hand–object interaction geometry as a shared condition for metric localization and occlusion-robust pose estimation. HOI depth and visual features guide coarse metric estimation and form joint-associated conditions that retain spatial structure and metric scale. Quality-aware conditioning and temporal memory maintain this evidence throughout flow generation, jointly recovering world-space translation, global orientation, articulation, and shape. Experiments on HOT3D and ARCTIC demonstrate improved absolute localization and temporal fidelity. These results highlight the value of persistent geometric conditioning for accurate and temporally coherent hand reconstruction from egocentric video.

Limitations. World-space accuracy remains dependent on the estimated camera trajectory and monocular depth prior. Errors in DROID-SLAM rotation, translation, or recovered metric scale propagate to the reconstructed hands and can produce substantial world-space offsets even when camera-space pose is accurate. On ARCTIC, camera MPJPE is 17.54 mm, whereas absolute world MPJPE remains 120.86 mm. MetricHand achieves lower absolute world error than the compared baselines, including those using the shared metric-scaled DROID trajectory. Baselines with separate camera estimates follow the configurations in Appendix D. Improving camera motion and depth estimation remains necessary for more accurate global hand placement.

AI USE STATEMENT

In this work, we used generative AI tools solely to improve the language and clarity of the manuscript. We did not use generatively AI tools for any other purpose. We reviewed all AI-assisted revisions to verify that they preserved the intended meaning and accurately reflected our work. We take responsibility for the final content of this work, including all text, claims, and artifacts.

REPRODUCIBILITY STATEMENT

We provide details of our method and experimental protocols in the main paper and appendices to enable reproduction of our method and experiments.

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 (CVPR)*, pp. 7061–7071, June 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 (CVPR)*, pp. 9044–9053, June 2021.

Xingyu Chen, Zhuheng Song, Xiaoke Jiang, Yaoqing Hu, Junzhi Yu, and Lei Zhang. HandOS: 3D Hand Reconstruction in One Stage. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17304–17314, June 2025.

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 *Advances in Neural Information Processing Systems*, volume 37, pp. 2127–2160. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-0069.

Enes Duran, Muhammed Kocabas, Vasileios Choutas, Zicong Fan, and Michael J. Black. HMP: Hand Motion Priors for Pose and Shape Estimation From Video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6353–6363, January 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 (CVPR)*, pp. 12943–12954, June 2023.

Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HONnotate: A Method for 3D Annotation of Hand and Object Poses. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3193–3203, 2020. URL https://arxiv.org/abs/1907.01481.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. EgoDex: Learning Dexterous Manipulation from Large-Scale Egocentric Video. In *International Conference on Learning Representations*, volume 2026, pp. 4218–4237, 2026.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3D v2: A Versatile Monocular Geometric Foundation Model for Zero-Shot Metric Depth and Surface Normal Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10579–10596, Dec 2024. doi: 10.1109/tpami.2024.3444912.

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. EgoMimic: Scaling Imitation Learning via Egocentric Video. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13226–13233, 2025. URL https://arxiv.org/abs/2410.24221.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

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. *arXiv preprint arXiv:2608.20308*, 2026. URL https://arxiv.org/abs/2608.20308.

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 (CVPR)*, pp. 9826–9836, June 2024.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, October 2023.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. WiLoR: End-to-end 3D Hand Localization and Reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12242–12254, June 2025.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6):1–17, Nov 2017. doi: 10.1145/3130800.3130883.

Zhihao Sun, Tong Wu, Ruirui Tu, Daoguo Dong, and Zuxuan Wu. UniHand: A Unified Model for Diverse Controlled 4D Hand Motion Modeling. *arXiv preprint arXiv:2602.21631*, 2026. URL https://arxiv.org/abs/2602.21631.

Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16558–16569. Curran Associates, Inc., 2021.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. MoGe-2: Accurate Monocular Geometry with Metric Scale and Sharp Details. In *Advances in Neural Information Processing Systems*, volume 38, pp. 35928–35959. Curran Associates, Inc., 2025a. doi: 10.52202/085713-1207.

Shengze Wang, Jiefeng Li, Tianye Li, Ye Yuan, Henry Fuchs, Koki Nagano, Shalini De Mello, and Michael Stengel. BLADE: Single-view Body Mesh Estimation through Accurate Depth Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21991–22000, June 2025b.

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. URL https://arxiv.org/abs/2606.30308.

Mingxi Xu, Bowen Duan, Yi Gu, Zhengyang Shen, Renjing Xu, and Yutao Yue. HandFlow: Fully Generative 4D Hand Recovery with Flow Matching. *arXiv preprint arXiv:2607.11221*, 2026. URL https://arxiv.org/abs/2607.11221.

Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, Hongxu Yin, Sifei Liu, Song Han, Yao Lu, and Xiaolong Wang. EgoVLA: Learning Vision-Language-Action Models from Egocentric Human Videos. *arXiv preprint arXiv:2507.12440*, 2025. URL https://arxiv.org/abs/2507.12440.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9043–9053, October 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 (CVPR)*, pp. 27716–27726, June 2025.

Huajian Zeng, Chaohua Yao, Yuantai Zhang, Jiaqi Yang, Rolandos Alexandros Potamias, and Xingxing Zuo. StableHand: Quality-Aware Flow Matching for World-Space Dual-Hand Motion Estimation from Egocentric Video. *arXiv preprint arXiv:2605.18553*, 2026. URL https://arxiv.org/abs/2605.18553.

Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. HaWoR: World-Space Hand Motion Reconstruction from Egocentric Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1805–1815, June 2025.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the Continuity of Rotation Representations in Neural Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5738–5746, 2019. URL https://arxiv.org/abs/1812.07035.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=gZ9hCDWe6ke.

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. *arXiv preprint arXiv:2609.04958*, 2026. URL https://arxiv.org/abs/2609.04958.

A Ablation Studies

Table 3 extends the component controls in Sec. 4.3; constant quality retains validity masks. The full model leads all reported component variants. The two additional controls retain the same depth-based coarse estimate, generator, three flow steps, and training budget. *Depth for coarse estimation only* withholds direct depth-derived geometry from all flow updates; *Geometry at first step only* supplies it at the first update. Both restrict cross-attention, pooled modulation, and cached geometry consistently during training and inference. All variants share scored frames and the SLAM trajectory.

Table 3: **Component and geometry-conditioning ablations on HOT3D and ARCTIC.** Position errors are in mm and world acceleration error in m/s^2 (↓). Controls follow Sec. 4.3 and Appendix A; Table 2 compares depth representations.

Configuration	HOT3D					ARCTIC				
	Camera			World		Camera			World	
	MPJPE	PA-MPJPE	WristErr	ABS-W-MPJPE	Accel.	MPJPE	PA-MPJPE	WristErr	ABS-W-MPJPE	Accel.
Full design	**8.60**	**3.18**	**7.72**	**20.58**	**3.11**	**17.54**	**5.18**	**20.02**	**120.86**	**4.39**
w/o Topology-aware sampling	17.85	4.75	12.39	31.39	4.58	38.91	6.78	35.21	154.36	5.47
w/o Flow matching	10.16	3.39	9.11	23.14	4.16	19.98	5.79	23.89	127.49	7.93
w/o Memory queue	15.42	3.98	16.55	27.78	4.05	25.67	5.96	27.37	136.24	6.31
w/o HOI region context	19.73	4.38	18.84	31.12	4.53	32.42	5.49	28.67	139.67	5.08
w/o Quality weighting	11.26	3.78	10.29	21.94	3.78	25.35	5.40	23.86	132.69	4.96
Geometry conditioning										
Depth for coarse estimation only	23.76	4.97	21.33	38.49	5.93	38.95	6.56	36.04	168.93	8.19
Geometry at first step only	15.96	4.14	17.47	29.35	4.59	28.95	5.99	28.12	141.66	6.28

B Additional Results

We train on HOT3D, ARCTIC, DexYCB, and HO3D. The following tables report DexYCB, HO3D, and HOI4D results; HOI4D is held out for zero-shot evaluation of MetricHand.

B.1 Depth Sources on DexYCB

Table 4 compares depth sources on DexYCB. Sensor depth improves all four metrics over MoGe-3, reducing MPJPE from 14.6 to 12.8 mm (12.3%), WristErr from 12.4 to 10.9 mm, and PA-MPJPE from 3.42 to 3.39 mm. MoGe-3 improves all four metrics over Metric3D v2.

Table 4: **Depth sources on DexYCB.** Camera-space errors in mm (↓).

Depth source	MPJPE	RA-MPJPE	PA-MPJPE	WristErr
MoGe-3	14.6	7.95	3.42	12.4
Metric3DV2	16.4	8.98	3.97	14.5
Sensor depth	12.8	6.31	3.39	10.9

B.2 Inference Speed

Table 5 reports model inference throughput on H200.

Table 5: **MetricHand inference speed.** Model throughput counts 16 frames per window without overlap. Timing uses one 16-frame, single-hand window, batch size 1, and three flow steps (10 warmups, 50 repeats). It includes the RGB-D encoder, flow, and MANO, with inputs resident on the GPU; depth estimation, detection, SLAM, and data transfer are excluded.

		Model stage	
Device	Precision	ms/window ↓	FPS ↑
H200	FP32	120.36	132.9
H200	BF16	44.44	360.0

B.3 DexYCB Occlusion Results

Table 6 reports overall and occlusion-stratified DexYCB results. MetricHand's PA-MPJPE is 3.42 mm overall and 4.03, 4.12, and 4.14 mm in the three increasing-occlusion groups. Literature values are reproduced from UniHand and HandFlow (Sun et al., 2026; Xu et al., 2026).

Table 6: **DexYCB reconstruction under occlusion.** Errors are in mm (↓); AUC values are normalized to [0,1] (↑). Literature values follow UniHand and HandFlow; Deformer's 25–50% PA-MPJPE follows the original UniHand table. Dashes denote unreported literature values.

	All				Occl. 25–50%		Occl. 50–75%		Occl. 75–100%	
Method	RA-MPJPE	AUC_{RA}	PA-MPJPE	AUC_{PA}	PA-MPJPE	AUC_{PA}	PA-MPJPE	AUC_{PA}	PA-MPJPE	AUC_{PA}
MeshGraphormer	16.21	0.691	6.41	0.872	6.85	0.863	7.22	0.856	7.76	0.845
SemiHandObj	–	–	6.33	0.874	6.70	0.866	7.17	0.857	8.96	0.821
HandOccNet	–	–	5.80	0.884	6.22	0.876	6.43	0.872	7.37	0.853
WiLoR	–	–	5.01	0.900	–	–	5.42	0.892	5.68	0.887
S2HAND(V)	19.67	0.625	7.27	0.855	7.74	0.845	7.71	0.846	7.87	0.843
VIBE	16.95	0.675	6.43	0.871	6.72	0.865	6.84	0.864	7.06	0.858
TCMR	16.03	0.701	6.28	0.875	6.56	0.869	6.58	0.868	6.95	0.861
Deformer	13.64	0.740	5.22	0.896	5.71	0.886	5.70	0.886	6.34	0.873
HaWoR	14.22	0.779	4.76	0.905	–	–	5.03	0.899	5.07	0.899
UniHand	–	–	4.08	0.918	4.22	0.913	4.25	0.912	4.26	0.912
HandFlow	8.12	0.839	3.88	0.922	4.21	0.916	4.35	0.913	4.85	0.903
MetricHand	7.95	0.864	3.42	0.938	4.03	0.931	4.12	0.925	4.14	0.919

B.4 HO3D Results

Table 7 reports HO3D results. MetricHand obtains 6.5 mm PA-MPJPE and mesh F-scores of 0.697 and 0.993 at 5 and 15 mm.

Table 7: **HO3D reconstruction.** MPJPE and PA-MPJPE are in mm (↓); AUC_{PA}, F@5, and F@15 are normalized to [0,1] (↑). MPJPE reports our unaligned camera-space reevaluation on 20,137 frames;Dashes denote unavailable results.

Method	MPJPE	PA-MPJPE	AUC_{PA}	F@5	F@15
HandOccNet	–	9.1	0.819	0.564	0.963
AMVUR	–	8.3	0.835	0.608	0.965
HaMeR	53.23	7.7	0.846	0.635	0.980
WiLoR	53.63	7.5	0.851	0.646	0.983
Deformer	–	9.4	–	0.546	0.963
Dyn-HaMR	45.01	7.8	0.849	0.656	0.969
HaWoR	52.39	6.9	0.858	0.658	0.974
UniHand	–	6.7	0.866	0.671	0.988
MetricHand	15.32	6.5	0.873	0.697	0.993

B.5 HOI4D Results

Table 8 reports camera-space reconstruction errors on the internal HOI4D held-out split of 76,480 frames. ACE with unknown intrinsics obtains 50.93 mm MPJPE and 10.01 mm PA-MPJPE, compared with 54.88 and 10.09 mm for its known-intrinsics variant.

Table 8: **HOI4D reconstruction on the internal held-out split.** Camera-space errors are in mm (↓). RR-MPJPE subtracts the wrist position before computing joint errors.

Method	MPJPE	RR-MPJPE	PA-MPJPE
HandFlow	135.05	28.10	10.59
ACE (known K)	54.88	21.03	10.09
ACE (K-free)	50.93	20.82	10.01
MetricHand (Ours)	18.73	6.44	2.69

C Evaluation Metrics

We report geometric errors in millimetres and acceleration error in m/s^2. The alignment-based metrics follow the conventions used in HaWoR, UniHand, and StableHand (Zhang et al., 2025; Sun et al., 2026; Zeng et al., 2026), with the fitting frames and coordinate systems specified below.

- **MPJPE and WristErr.** MPJPE is the mean Euclidean error over the 21 hand joints without hand-based alignment. Camera MPJPE uses camera coordinates; ABS-W-MPJPE uses the world reference defined in Appendix D. World errors additionally reflect camera-pose error. WristErr measures the camera-space error of the physical wrist alone.
- **RA-MPJPE.** Root-relative MPJPE subtracts each hand's wrist position before comparing joints. It removes translation while retaining orientation and scale errors.
- **PA-MPJPE.** A separate Procrustes similarity transform (translation, rotation, and uniform scale) aligns each predicted hand to ground truth before joint error is computed.
- **W-MPJPE and WA-MPJPE.** Both apply one similarity transform across a trajectory block. W-MPJPE fits the first two frames and evaluates the full block, retaining subsequent drift. WA-MPJPE fits all frames jointly and measures the residual trajectory and pose error. Our evaluator uses blocks of at most 100 frames. UniHand denotes these metrics G-MPJPE and GA-MPJPE; StableHand uses first-frame alignment for W-MPJPE, so the fitting convention must be retained when quoting its results.
- **Acceleration error.** We average the Euclidean difference between predicted and ground-truth world-space joint accelerations, computed by second-order temporal differences at 30 Hz. Lower error indicates closer agreement with true motion dynamics.
- **AUC.** The normalized area under the percentage-of-correct-keypoints curve summarizes joint accuracy over error thresholds. $\mathrm{AUC}_{\mathrm{RA}}$ and $\mathrm{AUC}_{\mathrm{PA}}$ use root-relative and Procrustes-aligned joint errors, respectively. The HO3D benchmark uses thresholds from 0 to 50 mm; quoted literature entries retain their source protocol.
- **F@5 and F@15.** Following the official HO3D evaluation (Hampali et al., 2020), these are mesh F-scores at 5 and 15 mm: the harmonic mean of precision and recall from bidirectional nearest-neighbour vertex distances after alignment. They are distinct from the fraction of frames below an MPJPE threshold.

D Evaluation Protocols

Datasets. For HOT3D, we follow HaWoR's validation split and evaluation protocol (Zhang et al., 2025), using 27 videos. ARCTIC follows the P2-val protocol (Fan et al., 2023): 34 sequences from subject s05, egocentric camera 0. The common valid right-hand tracks contain 82,850 and 20,127 scored frames, respectively. We score continuous observation- and annotation-valid segments of at least 16 frames, retaining occluded joints within scored hands. DexYCB uses the `s0` protocol (Chao et al., 2021), and HO3D uses the official evaluation set (Hampali et al., 2020).

World reference and temporal metrics. Camera trajectories are rigidly aligned to the ground-truth camera reference in blocks of at most 1,000 frames, without scale fitting. ABS-W-MPJPE adds no hand alignment. Ground-truth camera poses are used for evaluation only. Acceleration is computed within continuous segments; discarding the boundary acceleration samples leaves $N-4$ samples per N-frame segment.

Comparison settings. HaMeR-SLAM, WiLoR-SLAM, HMP-SLAM, HandFlow, ACE-EgoHand, and MetricHand use the shared metric-scaled DROID trajectory (Teed & Deng, 2021). MINT and Dyn-HaMR retain their native camera or optimized world estimates. Dataset-specific evaluation protocols are specified above. ACE-EgoHand uses its known-K or K-free configuration, and MINT predicts intrinsics. Training mixtures and frontend inputs differ across methods; common scored frames define the comparison in Table 1.

E Model Details

Depth Features and Metric Initialization. The depth branch receives normalized log depth, a validity mask, and two spatial gradients. It retains the mean and standard deviation of valid log depths, valid fraction, and mean edge magnitude as absolute context. Nonfinite and out-of-range depth values are masked before encoding. The RGB encoder, RGB-D fusion, temporal hand estimator, and metric head are trained with the hand model; the depth predictor and SLAM network are fixed.

The metric head refines coarse translation $\widetilde{t}^C$ with $\delta = \tanh H_{\text{metric}}(\cdot)$:

$$\widehat{r} = \widetilde{r} + 0.25\delta_{xy}, \qquad \widehat{\zeta} = \text{clip}(\widetilde{\zeta} + \delta_z, \log z_{\min}, \log z_{\max}), \qquad \widehat{t}^C = e^{\widehat{\zeta}}[\widehat{r}_x, \widehat{r}_y, 1]^\top. \tag{10}$$

Here $\widetilde{r} = \widetilde{t}^C_{xy} / \max(\widetilde{t}^C_z, \epsilon)$ and $\widetilde{\zeta} = \log \text{clip}(\widetilde{t}^C_z, z_{\min}, z_{\max})$, with $z_{\min} = 0.05$ and $z_{\max} = 5$ metres. The ray is formed before clipping depth. A zero-initialized output layer starts from this bounded proposal. Camera scale is obtained by robustly matching relative SLAM depth to the metric prior over valid static pixels, and the scale is applied to camera translations.

Coordinate and Fusion Operations. The inverse camera transform used for projection is

$$p^C_{t,h} = (R^{WC}_t)^\top (p^W_{t,h} - c^{WC}_t), \qquad R^C_{t,h} = (R^{WC}_t)^\top R^W_{t,h}. \tag{11}$$

The depth residual in the RGB-D encoder is

$$F^{RD} = F^I + \text{sigmoid}\left(a + g(\text{GN}(F^D))\right) \odot \text{GN}(F^D), \tag{12}$$

Here g is a learned projection, a is a learned bias, and GN denotes group normalization. The depth and image feature maps have matching spatial resolution and channel width.

Geometric Sampling and Quality. Each joint query predicts four offsets per feature scale in Eqs. 3–4. Anchor positions are detached when constructing the sampling grid. Depth-dependent bias terms require valid depth at both sample and anchor; the 3D term additionally requires valid wrist depth. Absolute geometry is encoded separately from normalized sampling distances. Viewing rays are $K^{-1}\bar{u}$, and surface positions are lifted with $P^W(u) = R^{WC}P^C(u) + c^{WC}$. These points describe visible surfaces in the interaction region, including nearby objects. Mask pooling uses the hand mask as spatial weights, and the box token encodes normalized center and extent. Each source retains its own validity indicator.

The quality encoder in Sec. 3.3 uses pooled joint and regional descriptors for the hand-level head, and each joint descriptor with hand context for the joint-level heads. Both receive raw detection and keypoint confidence, depth validity and support, camera/crop features, and the coarse hand state. For a hand or joint descriptor z_i, the scoring head produces

$$q_i = Q_\phi(z_i) \in [0, 1]. \tag{13}$$

The score controls the content–placeholder mixture in Eq. 5. Gradients from the existing reconstruction objectives pass through this weighting operation to train the scoring heads jointly with the model. There are no explicit quality targets or auxiliary quality losses. Unavailable inputs retain their validity masks and placeholder tokens, and identity embeddings remain present regardless of quality. The hand segmentation mask used for pooling is distinct from observation-validity and annotation masks.

Queue and Motion Tokens. We use 16-frame windows with eight-frame overlap and up to 32 historical timestamps strictly before the current window. For a window starting at a_n, the history spans $[\max(0, a_n - 32), a_n - 1]$; for example, window $[32, 47]$ reads history $[0, 31]$. Overlap belongs to the current window and does not consume the history budget. As the window advances by eight frames, its departed timestamps enter the FIFO history. Entries are keyed by timestamp, hand, and joint, retaining the most recent cached version. Each stores spatially aggregated features g^S, generated state embeddings, quality, and validity, with gradients detached across windows. Historical features enter Eq. 6, and state embeddings enter C_n. Invalid observation keys are masked; predicted states keep their original validity indicators. The queue resets between sequences. Within-window attention is bidirectional and requires the complete 16-frame window.

With continuous 6D rotations (Zhou et al., 2019), each hand window contains $10 + 99L$ normalized state dimensions: ten shared shape parameters and, per frame, six global-rotation, ninety articulation, and three translation coordinates. Translation uses training-set normalization and is restored to metric units for decoding. Each component instance maps to three tokens with component, time, subtoken, and hand embeddings. The velocity field has six Transformer blocks of width 512 with eight attention heads, query/key normalization, and SwiGLU feed-forward layers. Cross-attention gates start at a small nonzero value, and training samples flow time from a logit-normal distribution.

Condition Injection. The condition encoder projects the source groups in Sec. 3.3 to a shared width while preserving validity, quality, and source, time, hand, and joint identities. Joint and regional descriptors, geometric coordinates, coarse states, and queued states form C_n; attention pooling produces $\bar{C}_n$. Flow time τ and $\bar{C}_n$ modulate motion tokens H through adaptive layer normalization:

$$\mathrm{AdaLN}(H; \tau, \bar{C}_n) = \big(1 + a(\tau, \bar{C}_n)\big) \odot \mathrm{LN}(H) + b(\tau, \bar{C}_n), \tag{14}$$

where a and b are learned scale and shift functions. With $\widetilde{H} = \mathrm{AdaLN}(H; \tau, \bar{C}_n)$, cross-attention retrieves the conditions through

$$\begin{aligned} Q &= \widetilde{H} W_Q, \qquad K = C_n W_K, \qquad V = C_n W_V, \\ H &\leftarrow H + \gamma_{\mathrm{cross}}(\tau, \bar{C}_n) \odot \left[\mathrm{softmax}\left(\frac{Q K^\top}{\sqrt{d_k}} \right) V W_O \right], \end{aligned} \tag{15}$$

Here γ_{cross} gates the residual update. AdaLN supplies global modulation, while cross-attention retrieves token-specific geometry and context. The same available evidence is supplied at every velocity evaluation within a window, with each condition source retaining its own validity mask.

Loss Details. The world-space wrist-translation loss $\mathcal{L}_{\mathrm{trans}}$ in Eq. 8 supervises the physical wrist, rather than the MANO translation parameter before its root offset. For Eq. 8, $m_{t,h}$ marks a valid wrist target, and $m^{\mathrm{temp}}_{t,h} = m_{t-1,h} m_{t,h} m_{t+1,h}$ selects consecutive valid timestamps. The normalizers $N_{\mathrm{trans}} = \sum_{t=1}^{L} \sum_h m_{t,h}$ and $N_{\mathrm{temp}} = \sum_{t=2}^{L-1} \sum_h m^{\mathrm{temp}}_{t,h}$ count contributing hand observations and triplets. The groups in Eq. 9 expand to

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\mathrm{FM}} &+ 0.5\mathcal{L}_{\mathrm{trans}} + 0.3\mathcal{L}_{\log z} + 0.05\mathcal{L}_{\mathrm{abs}} + 0.5\mathcal{L}_{\mathrm{2D}} + 0.05\mathcal{L}_{\mathrm{3D}} \\ &+ 0.05\mathcal{L}_R + 0.05\mathcal{L}_\theta + 0.02\mathcal{L}_\beta + 0.05\mathcal{L}_+ + 0.01\mathcal{L}_{\mathrm{temp}} + 0.8a(s)\mathcal{L}_{\mathrm{localV}}, \end{aligned} \tag{16}$$

For Eq. 8, $\delta_{\mathrm{trans}} = 0.02$ m and $\delta_{\mathrm{temp}} = 0.01$ m; the corresponding weights are $\lambda_{\mathrm{trans}} = 0.5$ and $\lambda_{\mathrm{temp}} = 0.01$. For a three-dimensional residual $\mathbf{r}$, $\rho_\delta(\mathbf{r}) = \frac{1}{3}\sum_{d=1}^{3} \ell_\delta(r_d)$, where $\ell_\delta(r) = r^2/(2\delta)$ for $|r| < \delta$ and $|r| - \delta/2$ otherwise. The masks and normalizers in Eq. 8 average this penalty over valid observations; a term with no valid observations is omitted. The metric group is the weighted sum of $\mathcal{L}_{\mathrm{trans}}, \mathcal{L}_{\log z}, \mathcal{L}_{\mathrm{abs}}, \mathcal{L}_{\mathrm{2D}}, \mathcal{L}_+$. The pose group is the weighted sum of $\mathcal{L}_{\mathrm{3D}}, \mathcal{L}_R, \mathcal{L}_\theta, \mathcal{L}_\beta, \mathcal{L}_{\mathrm{localV}}$, with the coefficients in Eq. 16. Flow matching and the temporal term are added separately, as in Eq. 9. World wrist and absolute-joint losses use Smooth L1 with transitions 0.02 and 0.01 metres. The log-depth term acts on camera-space MANO translation with transition 0.02, accounting for the root offset when converting from physical wrist position. Reprojection uses each joint's own perspective depth and the same crop transform for prediction and target. The positive-depth barrier has threshold 0.03 metres and softness 0.02 metres.

Rotation matrices and shape parameters receive squared-error supervision; $\mathcal{L}_{\mathrm{3D}}$ supervises wrist-relative joints. For $\mathcal{L}_{\mathrm{localV}}$, predicted and target meshes are decoded with identity global orientation, wrist-centered, and divided by their own root-mean-square radii before comparison. This

term measures relative surface geometry. Its weight warms up with $a(s) = \min((s+1)/2000, 1)$. The temporal loss applies Smooth L1 with transition 0.01 to the difference between predicted and ground-truth wrist second differences on consecutive valid timestamps. Training uses frame-index differences; physical acceleration evaluation additionally divides by the squared frame interval.

Long Sequences and Implementation Scope. For overlapping windows, Gaussian noise is shared by global timestamp and hand identity, with one shape-noise sample per track. Endpoint estimates are blended with positive triangular weights, $\omega(r) = 1 - |r - (L-1)/2|/((L+1)/2)$. Rotation coordinates are blended before matrix decoding, and shape is aggregated across windows. This output blending complements the queue's role in supplying conditions.

Table 1 reports the full method, and Table 2 reports the ARCTIC depth representation controls under the protocol in Appendix D.

F QUALITATIVE RESULTS

Figure 4 shows two HOT3D and three ARCTIC examples of world-space hand reconstruction, with ground-truth meshes in green. Opacity increases over time to show the progression of motion.

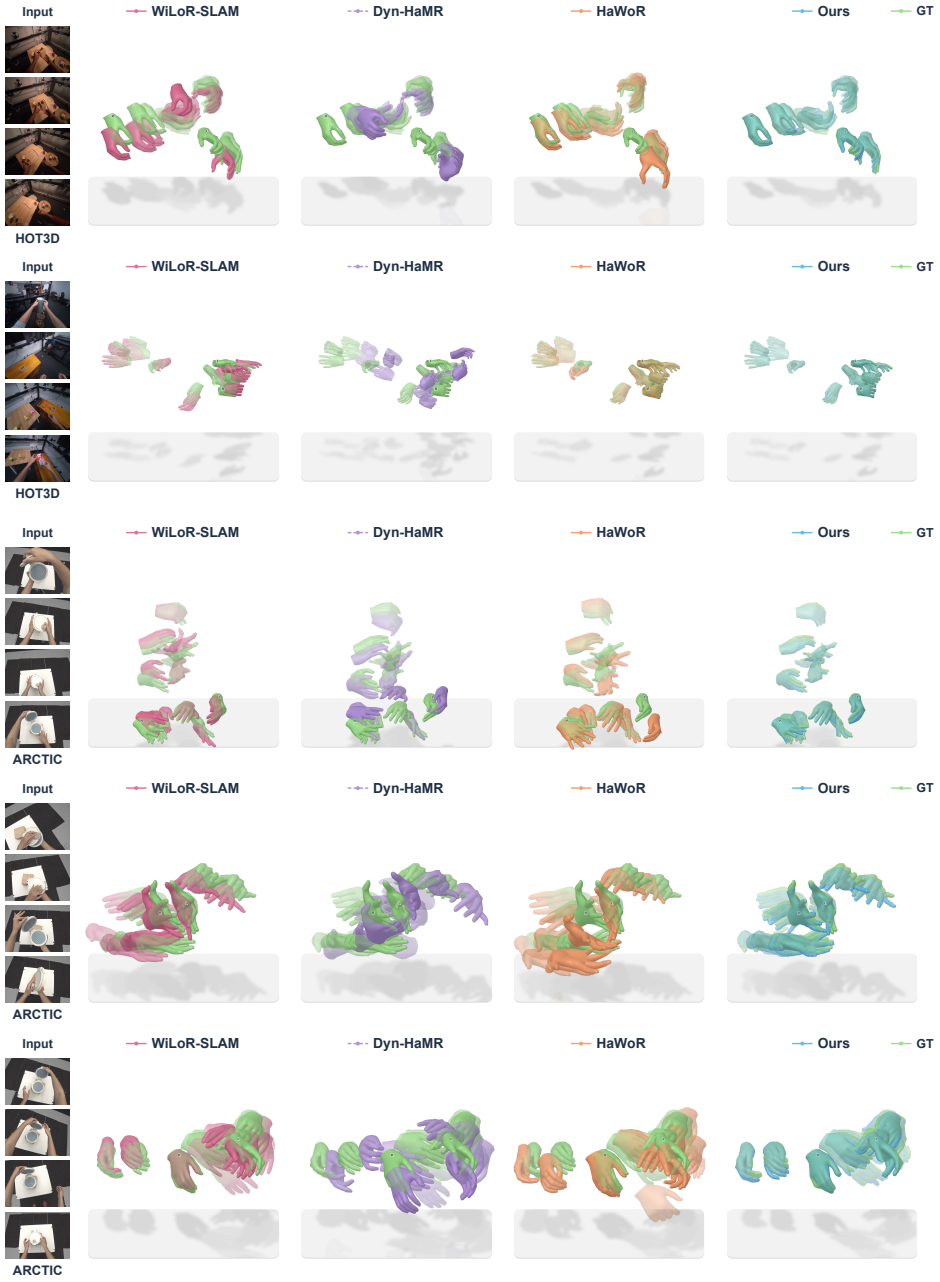


Figure 4: **Qualitative comparisons on HOT3D and ARCTIC Datasets** Columns show input, WiLoR-SLAM, Dyn-HaMR, HaWoR, and MetricHand. GT is green.