STEREOHOI: RECOVER METRIC HAND-OBJECT INTERACTION FROM EGOCENTRIC STEREO VISION

Anonymous authors
Paper under double-blind review

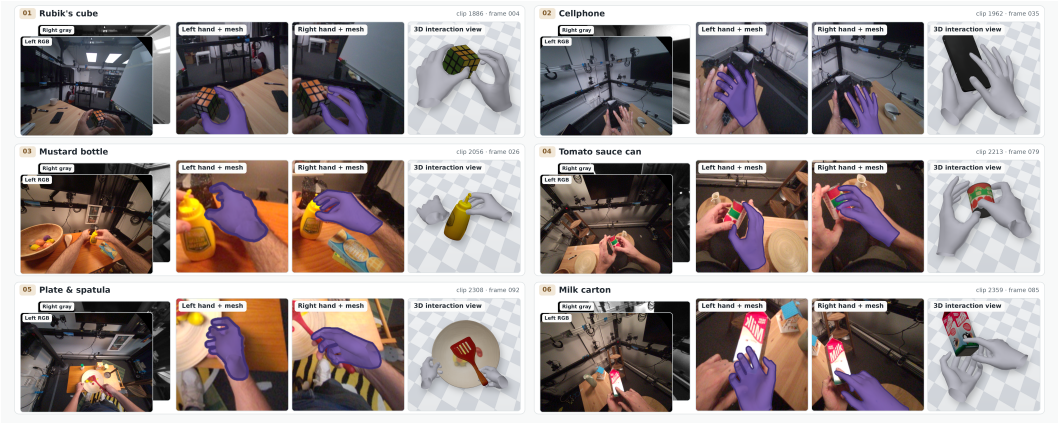

Figure 1: **3D Hand-Object Interaction Estimation from Stereo Egocentric Videos.** Given an input video sequence (left), our method tracks and recovers 3D hand meshes, enabling coherent metric 3D reconstruction of hand-object interaction in a consistent 3D workspace (right).

ABSTRACT

Reconstructing 3D hand-object interaction (HOI) in physical metric space is important for embodied AI and robotic manipulation. Existing HOI approaches mostly rely on RGB observations from monocular images or video streams, where the lack of binocular metric scale makes physical hand-object alignment challenging, often necessitating expensive test-time optimization. While stereo cameras provide binocular disparity cues to resolve metric scale, existing stereo networks focus primarily on dense surface depth estimation, which cannot directly predict hand keypoints or infer their corresponding metric depths. In this work, we explore stereo vision for recovering 3D hand-object interaction and present **StereoHOI**. Rather than recovering metric hand scale through additional fitting to stereo depth estimates, StereoHOI directly recovers metric hands by fusing monocular hand priors extracted from a frozen hand foundation model within an alternating cross-view attention framework. By encoding the monocular hand priors and camera parameters into conditioned special tokens, StereoHOI implicitly captures binocular disparity cues to predict metric wrist translations and pose residuals. This enables placing hands and metric object meshes into a shared metric coordinate frame, simplifying hand-object alignment and supporting video-based interaction analysis. Code will be released to benefit the research community.

1 INTRODUCTION

Reconstructing 3D hand-object interaction (HOI) from egocentric observations is an important problem in computer vision, with broad applications across embodied artificial intelligence, robotic manipulation, and augmented reality. In particular, first-person demonstrations of human manipulation have emerged as an invaluable data source for robot imitation learning (Wang et al., 2024; Kareer et al., 2025; Banerjee et al., 2025), complementing the teleoperated demonstrations used to train visuomotor policies (Chi et al., 2023). Equipping robots with dexterous manipulation skills requires more than isolated hand pose estimation; it crucially demands understanding the physical interaction between the hand and the manipulated object in a shared, metric 3D workspace.

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.

Recent 3D hand-object interaction methods have progressed from template-based tracking with known CAD models (Fan et al., 2023) to category-agnostic, template-free reconstruction of unseen objects (Fan et al., 2024; Liu et al., 2025; Chen et al., 2026b). To handle everyday interactions, modern frameworks (Chen et al., 2026b; Aytekin et al., 2026; Paliwal et al., 2026; Xu et al., 2026a) commonly adopt a modular strategy, pairing monocular hand estimators (Pavlakos et al., 2024; Potamias et al., 2025) with independent object reconstruction pipelines. However, because monocular observations lack direct metric scale, separately estimated hand and object meshes inevitably suffer from depth errors along the line of sight, leading to physical inconsistencies such as unnatural interpenetration or hovering. To resolve this ambiguity, stereo vision offers a natural physical foundation, as stereo cameras are widely deployed on VR headsets (e.g., Quest 3), smart glasses (e.g., Project Aria), and humanoid robots. While 3D foundation models such as SAM 3D (Chen et al., 2026a) and stereo matching networks (Wen et al., 2025) can recover object meshes with reasonable metric scale, directly applying off-the-shelf stereo depth to articulated hand recovery remains difficult. As noted in recent robot perception studies (Han et al., 2026), explicit depth maps and point clouds are often noisy and fragile in real-world manipulation. More importantly, because hands usually lack distinct texture features, stereo depth estimation methods struggle to accurately predict the depth of the wrist and articulated joints, and frequently blend hand regions into the background depth, making subsequent hand model fitting unreliable and computationally expensive.

To address the aforementioned challenges, we present **StereoHOI**, a framework for metric 3D hand and interaction reconstruction from egocentric stereo vision, built around a feed-forward stereo hand network. Instead of relying on fragile depth maps or heavy cost volumes, StereoHOI directly bridges 2D monocular hand priors with 3D metric reasoning (Han et al., 2026). Specifically, we adapt the articulated hand priors from a frozen monocular foundation model (HaMeR (Pavlakos et al., 2024)) to calibrated stereo pairs via an alternating cross-view transformer (Wang et al., 2025b). We introduce conditioned tokens that encode monocular pose and shape priors alongside camera intrinsics and the stereo baseline. By interleaving intra-view and cross-view attention, these tokens resolve binocular disparities to predict metric wrist translation and $\mathrm{SO}(3)$ pose residuals directly. This metric grounding enables direct alignment with stereo-derived object meshes (e.g., from SAM 3D (Chen et al., 2026a)) and consistent sequence-level HOI estimation across continuous video.

Our main contributions are summarized as follows: (1) We formulate metric 3D hand reconstruction from egocentric stereo vision by adapting monocular foundation priors, and leverage an alternating cross-view transformer to fuse binocular features. (2) We show that stereo inputs effectively resolve depth-scale ambiguity, recovering accurate absolute 3D translation while noticeably improving articulated hand mesh quality. (3) We build an egocentric stereo video pipeline that unifies dynamic hands and objects into a shared metric world frame, achieving high-quality interaction tracking at a much lower runtime than optimization-heavy HOI methods.

2 Related Work

2.1 3D Hand Pose and Mesh Recovery

Estimating 3D hand pose and mesh has progressed from direct regression (Lin et al., 2021a;b) to foundation models (Pavlakos et al., 2024; Potamias et al., 2025; Moon, 2023) for parametric recovery (Romero et al., 2017), with recent video frameworks incorporating SLAM smoothing (Yu et al., 2025; Zhang et al., 2025), diffusion priors (Wang et al., 2026; Liu et al., 2026), flow matching (Zeng et al., 2026; Xu et al., 2026b), or joint camera and hand motion prediction (Zhu et al., 2026). However, single-view methods lack absolute depth cues, while multi-view headset trackers such as UmeTrack (Han et al., 2022) recover metric hand pose with dedicated tracking networks rather than monocular foundation models. In contrast, StereoHOI adapts foundation priors to wearable stereo pairs, recovering metric wrist translation.

2.2 3D Hand-Object Interaction Reconstruction

Reconstructing 3D hand-object interaction (HOI) has advanced from template-based tracking with known CAD models (Fan et al., 2023; Wen et al., 2024) to template-free reconstruction of unseen objects using 2D vision models and contact priors (Liu et al., 2025; Pang et al., 2024; 2025). For video sequences, early methods optimize neural implicit geometry per video from scratch (Fan et al.,

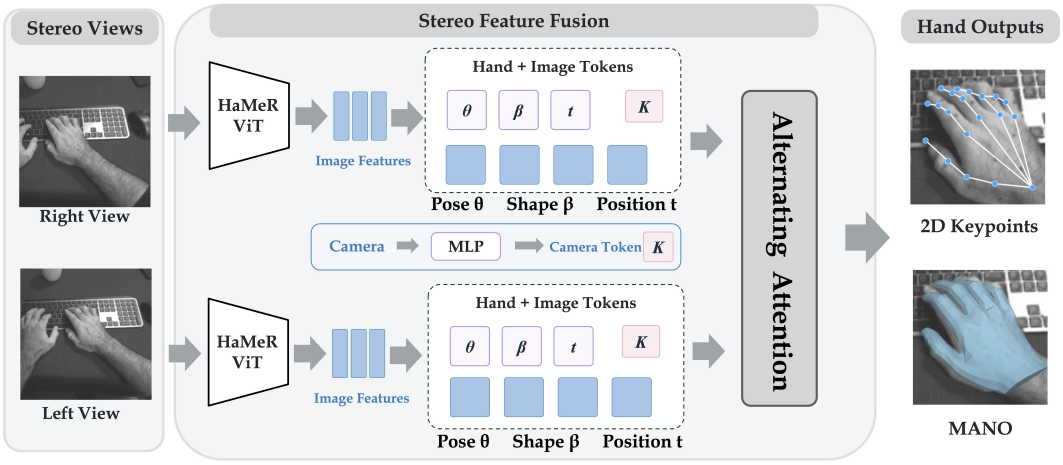

Figure 2: **Overview of StereoHOI.** Given a stereo image pair, visual patch tokens and initial hand parameters from a frozen monocular HaMeR (Pavlakos et al., 2024) backbone are concatenated with camera parameters into paired tokens, which are further processed by an alternating cross-view transformer. The updated tokens directly predict 21 2D hand keypoints alongside refined metric MANO parameters $(\boldsymbol{\theta}', \boldsymbol{\beta}', \mathbf{p}')$. Applied to each frame, this network provides the metric hands for the video pipeline in Figure 3.

2024), while recent approaches adopt feed-forward generative object priors and optimize motion or contact, spanning about one minute (Chen et al., 2026b) to tens of minutes or even hours (Wang et al., 2025a; Aytekin et al., 2026; Xu et al., 2026a). Other pipelines attempt to anchor scale using monocular hand priors (Paliwal et al., 2026) or external SLAM trajectories (Ye et al., 2026). However, lacking metric depth, monocular methods suffer from depth-scale ambiguity along the line of sight, often necessitating costly optimization to resolve physical penetration or hovering. In contrast, StereoHOI leverages binocular disparity to ground hands and stereo-derived object meshes (e.g., from SAM 3D (Chen et al., 2026a)) into a shared metric coordinate frame from the outset, enabling direct and scale-consistent hand-object interaction.

3 Method

As illustrated in Figure 2, StereoHOI reconstructs metric 3D hands from a calibrated egocentric stereo pair $(\mathbf{I}_\ell, \mathbf{I}_r)$, with the left view as the reference. It extracts hand priors from each view with a frozen monocular foundation model and fuses them across views with an alternating cross-view transformer (§3.1). A dual-branch decoder then predicts 2D hand keypoints directly, in parallel with camera-space MANO parameters and metric translation (§3.2), and the whole network is trained end-to-end with multi-task geometry supervision (§3.3). Building on these frame-wise predictions, we further develop a sequence-level HOI tracking framework (Figure 3), whose spatiotemporal optimization (§3.4) refines hand trajectories and 6D object poses in a shared world coordinate frame.

3.1 Monocular Hand Priors and Cross-View Encoding

Feature extraction and prior decoding. Compared with large-scale in-the-wild monocular datasets, existing stereo hand datasets remain limited in size and diversity, and training a stereo network from scratch can easily lose the strong priors learned from extensive monocular data. Therefore, we adopt a pre-trained monocular HaMeR (Pavlakos et al., 2024) model as a frozen feature extractor and prior generator. Following HaMeR, each view is cropped based on hand detections and preprocessed into a normalized strip of 256×192 pixels, with left hands flipped to right-hand format. The frozen ViT-H/16 backbone turns each strip into a 16×12 grid of 192 patch features with 1280 channels, and a linear layer projects them to the hidden dimension $d = 1024$ of our stereo transformer, giving the patch tokens $\mathbf{H}_\ell, \mathbf{H}_r \in \mathbb{R}^{192 \times d}$ of the left and right views.

From the same backbone features, HaMeR's pre-trained heads decode a monocular hand estimate $(\boldsymbol{\theta}_v, \boldsymbol{\beta}_v, \mathbf{p}_v)$ for each view $v \in \{\ell, r\}$, where $\boldsymbol{\theta}_v \in \mathrm{SO}(3)^{16}$ denotes the joint rotations, $\boldsymbol{\beta}_v \in \mathbb{R}^{10}$ the

hand shape, and $\mathbf{p}_v \in \mathbb{R}^2$ the 2D wrist position in the image. Inspired by the decoupled prediction strategy of WiLoR (Potamias et al., 2025), we use these estimates as hand priors: the left-view estimate is the initial value that our network refines, and the estimates of both views are given to the transformer as tokens. Unlike WiLoR, which refines its estimate by sampling features along the reprojected mesh, we refine it with binocular disparity cues to recover metric wrist translation.

Prior and camera tokens. To let the transformer reason over these priors, we turn each of them into a d-dimensional special token with a linear layer:

$$\mathbf{q}_{\theta,v} = \mathrm{Linear}_\theta(\boldsymbol{\theta}_v), \quad \mathbf{q}_{\beta,v} = \mathrm{Linear}_\beta(\boldsymbol{\beta}_v), \quad \mathbf{q}_{p,v} = \mathrm{Linear}_p(\mathbf{p}_v), \quad v \in \{\ell, r\}. \tag{1}$$

Since metric depth can only be recovered from disparity when the camera geometry is known, we also encode the camera parameters $\mathbf{k} = (f_x, f_y, c_x, c_y, b)$, including the stereo baseline b, into a camera token $\mathbf{q}_k = \mathrm{Linear}_{\mathrm{cam}}(\mathbf{k})$. For each view, we then place the four special tokens $\mathbf{Q}_v = [\mathbf{q}_{\theta,v}, \mathbf{q}_{\beta,v}, \mathbf{q}_{p,v}, \mathbf{q}_k] \in \mathbb{R}^{4\times d}$ in front of its patch tokens, forming the input sequence $\mathbf{X}_v = [\mathbf{Q}_v, \mathbf{H}_v] \in \mathbb{R}^{(4+192)\times d}$.

Alternating cross-view attention. The two sequences $\mathbf{X}_\ell$ and $\mathbf{X}_r$ are processed jointly by a 24-layer stereo transformer inspired by π^3 (Wang et al., 2025b), which relates the two views without building heavy cost volumes. Its layers alternate between two kinds of attention: even layers apply *intra-view self-attention* within each sequence, so that the special tokens gather spatial context from their own view, while odd layers concatenate the two sequences and apply *cross-view joint attention*, so that tokens of one view can attend to the other and find stereo correspondences. Through this alternation, the right view supplies binocular disparity cues that enrich the left-view tokens. After the last layer, we take the updated left-view sequence $\mathbf{X}'_\ell = [\mathbf{Q}'_\ell, \mathbf{H}'_\ell] \in \mathbb{R}^{(4+192)\times d}$, where $\mathbf{Q}'_\ell$ and $\mathbf{H}'_\ell$ are the updated special tokens and patch tokens. Since the 3D hand is reconstructed in the left (reference) camera frame, only $\mathbf{X}'_\ell$ is used for decoding.

3.2 Dual-Branch Decoding and Hand Mesh Recovery

The enriched sequence $\mathbf{X}'_\ell = [\mathbf{Q}'_\ell, \mathbf{H}'_\ell]$ holds two kinds of tokens, which we decode with two parallel branches: the patch tokens $\mathbf{H}'_\ell$ keep the spatial layout of the image and suit locating 2D keypoints, while the special tokens $\mathbf{Q}'_\ell$ summarize the hand state and suit regressing MANO parameters and the wrist depth.

2D keypoint branch. Existing methods such as HaMeR and WiLoR obtain 2D keypoints indirectly by projecting the reconstructed 3D mesh onto the image, so errors in the 3D estimate carry over to the 2D keypoints. Our 2D keypoint branch instead regresses the 21 keypoint coordinates directly from the patch tokens. A linear layer first reduces $\mathbf{H}'_\ell \in \mathbb{R}^{192\times d}$ to 256 channels, and a Gated Attention Unit (GAU) (Hua et al., 2022) models the relations among the 192 grid positions, giving $\mathbf{F} = \mathrm{GAU}(\mathrm{Linear}(\mathbf{H}'_\ell)) \in \mathbb{R}^{192\times 256}$. Global average pooling turns $\mathbf{F}$ into a hand-level descriptor $\bar{\mathbf{f}} = \mathrm{AvgPool}(\mathbf{F}) \in \mathbb{R}^{256}$, from which a linear head predicts the keypoints $\mathbf{U} = \mathrm{Linear}(\bar{\mathbf{f}}) \in \mathbb{R}^{21\times 2}$ in pixels.

Metric MANO parameter branch. This branch recovers the metric 3D hand from the special tokens $\mathbf{Q}'_\ell$. Three lightweight linear heads decode the pose, shape, and wrist tokens into corrections of the left-view priors and the wrist depth:

$$\Delta\boldsymbol{\theta} = \mathrm{Linear}_\theta(\mathbf{q}'_{\theta,\ell}), \qquad \Delta\boldsymbol{\beta} = \mathrm{Linear}_\beta(\mathbf{q}'_{\beta,\ell}), \qquad (\Delta\mathbf{p},\, \log z) = \mathrm{Linear}_p(\mathbf{q}'_{p,\ell}), \tag{2}$$

where $\Delta\boldsymbol{\theta}$ holds a residual rotation for each of the 16 joints, projected onto $\mathrm{SO}(3)$ by SVD orthogonalization, $\Delta\boldsymbol{\beta}$ is a shape offset, $\Delta\mathbf{p}$ is a 2D wrist offset bounded to a fraction of the crop size, and z is the metric wrist depth, predicted in log space to keep it positive. Because the wrist token has gathered disparity cues from the right view through cross-view attention, z recovers the metric depth that a single view cannot provide. We use these outputs to update the priors,

$$\boldsymbol{\theta}' \;=\; \Delta\boldsymbol{\theta} \cdot \boldsymbol{\theta}_\ell, \qquad \boldsymbol{\beta}' \;=\; \boldsymbol{\beta}_\ell + \Delta\boldsymbol{\beta}, \qquad \mathbf{p}' \;=\; \mathbf{p}_\ell + \Delta\mathbf{p}, \tag{3}$$

and back-project the refined wrist $\mathbf{p}' = (u', v')$ at depth z into the left camera frame with the camera parameters $\mathbf{k}$, which gives the metric wrist translation

$$\mathbf{t} \;=\; \Big(\frac{(u' - c_x)\, z}{f_x},\; \frac{(v' - c_y)\, z}{f_y},\; z\Big). \tag{4}$$

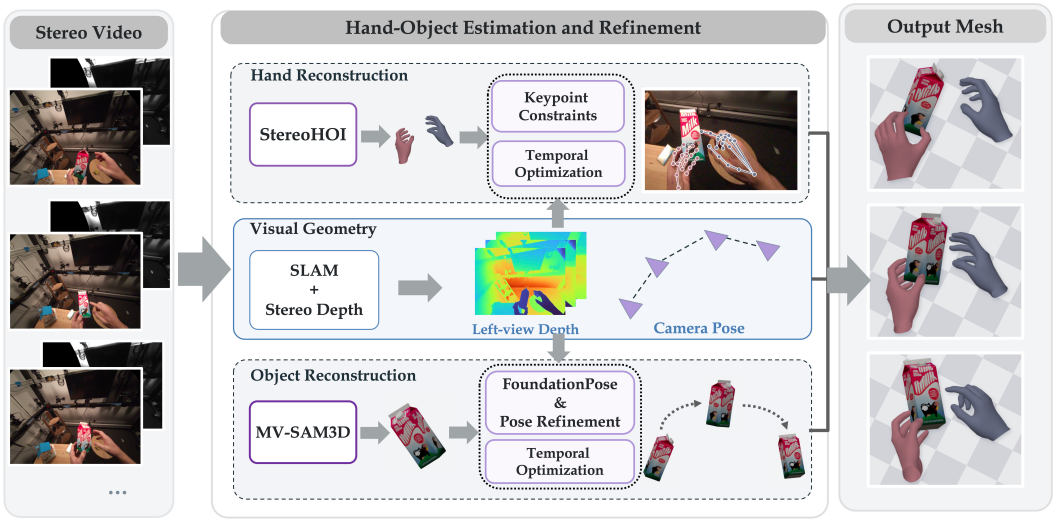

Figure 3: **System Pipeline for Dynamic 3D Hand-Object Interaction Tracking in Videos.** Given egocentric stereo video frames (left), we run the frame-wise StereoHOI network of Figure 2 together with visual SLAM, stereo depth, and 3D object generation (middle), and then optimize hand and object trajectories in parallel, reconstructing metric hand-object interactions in a shared space (right).

Finally, the differentiable MANO layer maps $(\boldsymbol{\theta}', \boldsymbol{\beta}')$ to the hand mesh $\mathbf{M} \in \mathbb{R}^{778\times 3}$ and 21 joints $\mathbf{J} \in \mathbb{R}^{21\times 3}$ (the 16 MANO skeleton joints plus 5 fingertip vertices), which are translated so that the wrist lies at $\mathbf{t}$: $\mathbf{M}, \mathbf{J} = \mathrm{MANO}(\boldsymbol{\theta}', \boldsymbol{\beta}', \mathbf{t})$. During training, a reprojection term (§3.3) keeps the projected joints $\mathbf{J}$ consistent with the directly predicted keypoints $\mathbf{U}$.

3.3 Multi-Task Training and Geometry Supervision

We train StereoHOI end-to-end. Each training sample provides the annotated MANO pose $\boldsymbol{\theta}^*$ and shape $\boldsymbol{\beta}^*$ of the hand, its wrist translation $\mathbf{t}^*$ in the left camera frame, and its 2D keypoints $\mathbf{U}^*$ in the input crop. To compare 3D joints under the same definition, we pass $(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*, \mathbf{t}^*)$ through the same MANO layer to obtain the ground-truth joints $\mathbf{J}^*$, and write $\mathbf{j}_k$ and $\mathbf{j}_k^*$ for the 3D position of joint k in $\mathbf{J}$ and $\mathbf{J}^*$. The objective is a weighted sum of four geometric terms:

$$\mathcal{L} \;=\; \lambda_{\mathrm{2D}}\mathcal{L}_{\mathrm{2D}} + \lambda_{\mathrm{3D}}\mathcal{L}_{\mathrm{3D}} + \lambda_{\mathrm{param}}\mathcal{L}_{\mathrm{param}} + \lambda_{\mathrm{metric}}\mathcal{L}_{\mathrm{metric}}, \tag{5}$$

2D alignment loss. This term supervises the hand in the image plane, $\mathcal{L}_{\mathrm{2D}} = \|\mathbf{U} - \mathbf{U}^*\|_1 + \|\pi(\mathbf{J}, \mathbf{k}) - \mathbf{U}^*\|_1$, computed over validly annotated keypoints. The first term trains the 2D keypoint branch; the second projects the predicted 3D joints with the camera parameters $\mathbf{k}$ onto the same targets, keeping the MANO hand consistent with the image and the 2D branch.

3D articulation loss. This term supervises the finger configuration regardless of where the hand is, $\mathcal{L}_{\mathrm{3D}} = \frac{1}{21}\sum_{k=1}^{21} \|(\mathbf{j}_k - \mathbf{t}) - (\mathbf{j}_k^* - \mathbf{t}^*)\|_1$. Subtracting the wrist from both predicted and ground-truth joints removes the global translation, so depth errors do not affect this term.

Parametric MANO loss. Joint positions do not fully determine the MANO parameters: twisting a finger segment around its bone barely moves the joints, and the shape $\boldsymbol{\beta}$ only mildly changes them. We therefore supervise them directly: writing the refined pose as rotation matrices, $\boldsymbol{\theta}' = (\mathbf{R}_{\mathrm{global}}, \mathbf{R}_1, \ldots, \mathbf{R}_{15})$ for the wrist and the 15 finger joints, and the ground truth $\boldsymbol{\theta}^*$ in the same way, we use $\mathcal{L}_{\mathrm{param}} = \|\mathbf{R}_{\mathrm{global}} - \mathbf{R}_{\mathrm{global}}^*\|_F^2 + \frac{1}{15}\sum_{j=1}^{15} \|\mathbf{R}_j - \mathbf{R}_j^*\|_F^2 + \|\boldsymbol{\beta}' - \boldsymbol{\beta}^*\|_2^2$. Frobenius distances between rotation matrices give smooth gradients without angle singularities.

Metric localization loss. The terms above do not pin down how far the hand is from the camera: $\mathcal{L}_{\mathrm{3D}}$ and $\mathcal{L}_{\mathrm{param}}$ are relative to the wrist, and the reprojection term cannot tell a small nearby hand from a larger distant one. We therefore supervise the absolute position directly, $\mathcal{L}_{\mathrm{metric}} = \|\mathbf{t} - \mathbf{t}^*\|_2 + \frac{1}{21}\sum_{k=1}^{21} \|\mathbf{j}_k - \mathbf{j}_k^*\|_2$, which compares the wrist translation and the absolute joint positions in the left camera frame, in meters. Since this metric depth can be inferred reliably only from the disparity between the two views, the loss drives the wrist token to use the cross-view cues.

3.4 TEMPORAL SEQUENCE OPTIMIZATION

As illustrated in Figure 3, our framework extends frame-wise StereoHOI predictions to continuous video via sequence-level trajectory optimization (Paliwal et al., 2026). To decouple motion refinement, binocular DROID-SLAM (Teed & Deng, 2021) provides continuous 6-DoF camera poses that place hands and objects in a shared world frame. Meanwhile, dense metric depth from FoundationStereo (Wen et al., 2025) is used solely for object mesh reconstruction and 6D pose tracking, avoiding noisy depth artifacts on articulated hands.

Hand sequence optimization. Given per-frame StereoHOI outputs (3D MANO meshes and 2D keypoints), we refine hand trajectories across the video by jointly optimizing temporal motion, 2D projection alignment, and fidelity to network predictions:

$$\min_{\{\mathbf{x}_t, \boldsymbol{\theta}_t\}} \mathcal{L}_{\text{hand}} = \mathcal{L}_{\text{anchor}} + \lambda_T \mathcal{L}_T + \lambda_{\text{KP}} \mathcal{L}_{\text{KP}}. \tag{6}$$

Here, the smoothness term $\mathcal{L}_T$ penalizes abrupt wrist and finger accelerations computed with real timestamps to ensure kinematic continuity. The projection term $\mathcal{L}_{\text{KP}}$ aligns reprojected 3D joints with StereoHOI's 2D keypoint predictions via a robust Charbonnier loss. Meanwhile, the anchor loss $\mathcal{L}_{\text{anchor}}$ keeps the optimized trajectory within a bounded trust region around the initial network estimates. We optimize only the wrist positions $\mathbf{x}_t$ and finger poses $\boldsymbol{\theta}_t$ (Appendix A.1).

Object reconstruction and tracking. For each manipulated object, we first reconstruct a static 3D mesh and then track its 6D pose over the video. Given multi-view keyframes with SAM 3.1 (Carion et al., 2025) masks and FoundationStereo (Wen et al., 2025) depth, MV-SAM3D (Li et al., 2026) reconstructs the object mesh, whose absolute metric scale we then set from the back-projected stereo depth. To track the object, FoundationPose (Wen et al., 2024) initializes candidate 6D poses across frames, which are mapped to world coordinates via DROID-SLAM pose estimations to resolve front-back orientation ambiguities. Finally, the 6D object poses are refined jointly over the sequence:

$$\min_{\{(\mathbf{R}_t, \mathbf{t}_t)\}} \mathcal{L}_{\text{obj}} = \mathcal{L}_{\text{align}} + \lambda_{\text{anchor}} \mathcal{L}_{\text{anchor}} + \lambda_{\text{smooth}} \mathcal{L}_{\text{smooth}}, \tag{7}$$

where $\mathcal{L}_{\text{align}}$ uses 2D and 3D cues from tracked object points. In each frame, we detect corner features inside the object mask and track them to the next frame with pyramidal Lucas-Kanade optical flow (Lucas & Kanade, 1981), keeping only points whose forward-backward tracking error is below one pixel, and lift them to 3D with the stereo depth of both frames. Using the relative object pose between the two frames, $\mathcal{L}_{\text{align}}$ carries these 3D points from one frame to the other and requires them to project onto their tracked image positions and to coincide with their stereo-measured 3D positions; the 3D part fixes the object along the viewing direction, where 2D motion alone is ambiguous. $\mathcal{L}_{\text{anchor}}$ keeps the trajectory near the FoundationPose initialization with a soft penalty and bounded updates, preventing drift under finger occlusion, and $\mathcal{L}_{\text{smooth}}$ penalizes the object's linear and angular accelerations in world coordinates. Optimizing hands and objects separately in the same metric frame places them consistently without heuristic collision penalties (Appendix A.2).

4 EXPERIMENTS

We evaluate StereoHOI on single-frame hand reconstruction (§4.2) and video-based hand-object interaction (§4.3), followed by ablation studies (§4.4); more visualizations are given in §A.3.

4.1 EXPERIMENTAL SETUP

Dataset and setup. We conduct our experiments on the HOT3D benchmark (Banerjee et al., 2025), which is recorded with two headsets: the Project Aria recordings provide one central RGB stream and two lateral monochrome streams, while the Quest 3 recordings provide two monochrome streams. Our training set combines 1,272 clips from Project Aria and 1,242 clips from Quest 3. We test exclusively on the Project Aria test set (244 clips, 150 frames per clip), split following StableHand (Zeng et al., 2026) since these clips provide ground-truth 3D poses and meshes for both hands and objects. For a fair comparison with monocular methods, we use the central RGB camera of Project Aria as the left (reference) view and its right monochrome camera as the right view, so that monocular baselines

Table 1: **Quantitative comparison on single-frame 3D hand reconstruction** on the Project Aria test set (53,124 hand instances). Errors in mm except GO (°) and EPE (px). FT: fine-tuned on our data; tri.: stereo-triangulated. POEM-v2 lacks global rotation; UmeTrack uses Aria's SLAM cameras.

Method	CS-J ↓	CS-V ↓	R-J ↓	PA-J ↓	PA-V ↓	CT ↓	GO (°) ↓	EPE ↓
Monocular input								
HaMeR	65.49	64.77	59.32	9.17	9.50	48.99	37.59	10.86
WiLoR	53.47	53.17	20.84	6.48	6.68	47.25	13.13	16.32
Hamba	68.24	67.15	58.55	8.94	9.14	45.60	37.51	11.28
InterWild	87.04	86.08	68.98	10.53	10.61	66.77	45.77	13.41
WildHands	175.57	173.42	64.10	13.18	13.35	145.58	50.98	47.94
OmniHands	138.15	135.67	55.04	10.29	10.25	98.91	35.99	16.69
HaWoR	33.82	34.21	14.28	5.28	5.44	39.65	8.77	7.34
Dyn-HaMR	90.64	90.83	30.40	10.06	10.49	91.55	20.12	7.34
HaMeR-FT	66.23	65.72	11.37	5.65	5.21	64.53	7.46	17.66
Stereo input								
HaMeR-FT (tri.)	20.53	20.48	11.36	5.65	5.21	19.58	7.45	3.80
POEM-v2-FT	14.55	14.56	13.82	6.57	6.76	12.71	–	3.45
UmeTrack	50.61	–	25.91	10.95	–	44.63	14.92	12.29
StereoHOI	**11.17**	**10.49**	**8.83**	**4.95**	**3.99**	**8.72**	**4.87**	**2.60**

Table 2: **Quantitative comparison on video-based 3D hand and object interaction reconstruction.** Hand errors are in mm; units of the other metrics follow Section 4.1.

Method	MPJPE ↓	W-MP ↓	WA-MP ↓	Jitter ↓	RTE ↓	CD ↓	F-5 ↑	F-10 ↑	CD_h ↓	ATE ↓	RPE_t ↓	RPE_r ↓
HaWoR	36.27	42.47	14.62	7.31	4.05	–	–	–	–	–	–	–
Dyn-HaMR	58.98	61.30	28.95	6.34	5.53	–	–	–	–	–	–	–
StableHand	47.69	64.53	25.07	**2.30**	5.17	–	–	–	–	–	–	–
HOLD	152.63	100.81	62.06	61.13	37.83	22.41	0.07	0.16	11.90	0.339	5.89	5.91
BIGS	143.25	124.12	62.79	57.62	39.09	22.76	0.05	0.14	13.45	0.344	6.04	5.89
MagicHOI	70.69	50.86	32.63	24.58	17.86	5.67	0.10	0.20	4.95	0.041	0.69	4.40
Do as I Do	22.18	31.77	13.41	7.28	2.23	4.26	0.15	0.30	3.14	0.038	1.19	8.34
StereoHOI	**10.30**	**18.60**	**9.59**	5.21	**2.02**	**0.76**	**0.48**	**0.79**	**1.08**	**0.015**	**0.29**	**2.99**

are evaluated on the same RGB images. Because the Quest 3 recordings contain only monochrome images, we convert Aria's central RGB images to grayscale to train on both datasets, so our hand model runs on grayscale images throughout.

Evaluation metrics. We evaluate our framework across single-frame hand geometry, video motion tracking, and dynamic hand-object interaction. *(1) For single-frame hand reconstruction* (Pavlakos et al., 2024; Potamias et al., 2025), we report unaligned camera-space joint and vertex errors (CS-MPJPE, CS-MPVPE) retaining translation, orientation, and scale errors, alongside camera translation error (CT in mm) for wrist placement and global geodesic rotation error (GO in °). To isolate intrinsic articulation, we report root-centered joint error (R-MPJPE) by zeroing wrist translation, Procrustes-aligned errors (PA-MPJPE, PA-MPVPE) removing rigid scale and pose ($\mathrm{Sim}(3)$), and 2D keypoint End-Point Error (EPE in px) on 256-crops. *(2) For video-based hand motion* (Zeng et al., 2026; Xu et al., 2026b), we evaluate per-frame tracking via unaligned MPJPE, trajectory drift via warm-up aligned W-MP ($\mathrm{Sim}(3)$ fit on the first two frames) and sequence-aligned WA-MP (global sequence fit), temporal acceleration jitter (mm/frame^2), and Relative Trajectory Error (RTE in %, unscaled trajectory error normalized by GT path length). *(3) For dynamic hand-object interaction* (Fan et al., 2024; Aytekin et al., 2026), we evaluate object surface Chamfer Distance (CD in cm), precision F-scores at 5 mm and 10 mm (F_5, F_{10}), hand-relative Chamfer Distance (CD_h in the local hand frame to decouple relative grasping configuration), centroid Absolute Trajectory Error (ATE in m), and inter-frame Relative Pose Errors (RPE_t in cm, RPE_r in °).

4.2 Single-Frame 3D Hand Reconstruction

Comparison with state-of-the-art methods. We benchmark StereoHOI on the Project Aria test set (53,124 hand instances) against four groups of methods: monocular foundation models (HaMeR (Pavlakos et al., 2024), WiLoR (Potamias et al., 2025), Hamba (Dong et al., 2024)), interacting and egocentric hand models (InterWild (Moon, 2023), WildHands (Prakash et al., 2024),

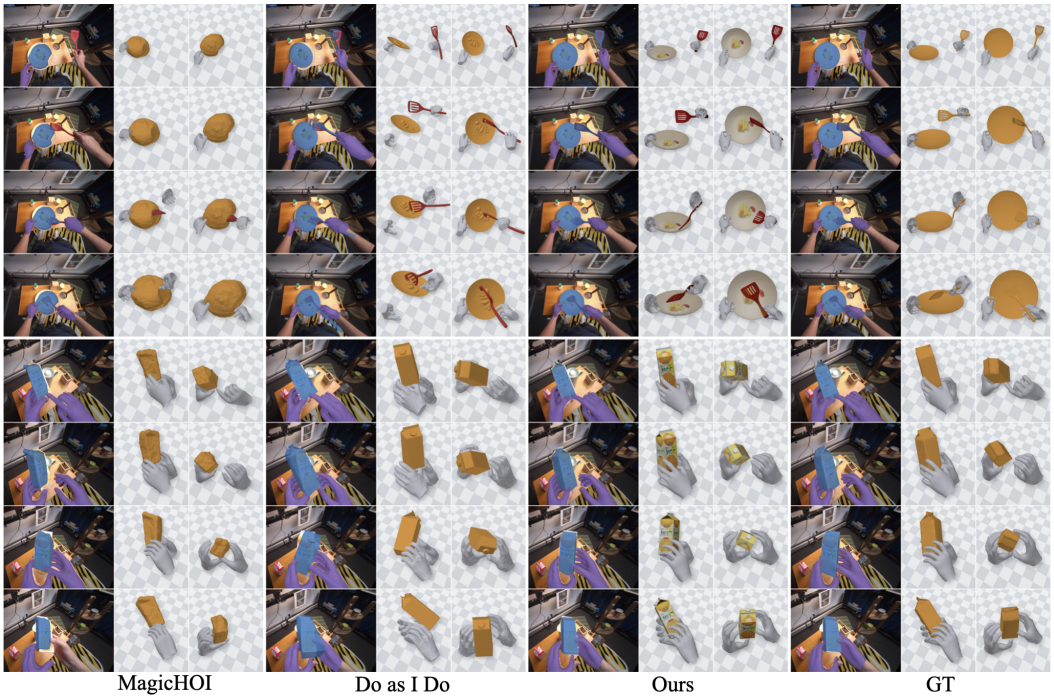

Figure 4: **Qualitative comparison of dynamic 3D hand-object interaction reconstruction** across representative egocentric video sequences on the HOT3D benchmark. For each sequence, we compare our reconstructed 3D hands and metric objects against state-of-the-art methods MagicHOI (Wang et al., 2025a) and Do as I Do (Paliwal et al., 2026), alongside the Ground Truth (GT).

OmniHands (Lin et al., 2024)), video-based egocentric motion models (HaWoR (Zhang et al., 2025), Dyn-HaMR (Yu et al., 2025)), and in-domain fine-tuned and stereo baselines (HaMeR-FT, HaMeR-FT (tri.), POEM-v2-FT (Yang et al., 2025), and UmeTrack (Han et al., 2022)).

Analysis. As shown in Table 1, StereoHOI consistently and substantially outperforms the baselines of all four paradigms. While monocular foundation models (HaMeR, WiLoR, Hamba) struggle with severe depth-scale ambiguities along the line of sight (yielding CT errors exceeding $45\,\mathrm{mm}$ and CS-J above $53\,\mathrm{mm}$), our framework leverages calibrated binocular parallax to recover physically grounded 3D hands in camera space. Crucially, under the identical calibrated stereo camera setup, StereoHOI still markedly surpasses the stereo baselines HaMeR-FT (tri.), POEM-v2-FT, and UmeTrack, suggesting that triangulating per-view predictions or running dedicated multi-view networks is less effective than jointly reasoning over stereo features built on strong monocular priors. Furthermore, StereoHOI achieves state-of-the-art accuracy on root-relative articulation (R-J of **8.83 mm**) and Procrustes-aligned pose (PA-J of **4.95 mm**), showing that stereo reasoning improves not only absolute hand placement but also fine-grained finger articulation.

4.3 Video-Based Hand-Object Interaction Reconstruction

Comparison with state-of-the-art methods. We compare StereoHOI with video-based hand motion methods (HaWoR (Zhang et al., 2025), Dyn-HaMR (Yu et al., 2025), and StableHand (Zeng et al., 2026)) and dynamic hand-object reconstruction methods (HOLD (Fan et al., 2024), BIGS (On et al., 2025), MagicHOI (Wang et al., 2025a), and Do as I Do (Paliwal et al., 2026)), and report hand tracking, motion smoothness, object geometry, and trajectory metrics in Table 2.

Analysis. As shown in Table 2, StereoHOI outperforms both monocular video hand motion baselines and dynamic HOI frameworks on all metrics except jitter, where StableHand produces smoother motion. Specifically, our method achieves an RTE of only 2.02%, more than halving the trajectory drift of video baselines (HaWoR at 4.05% and StableHand at 5.17%). Furthermore, compared with dynamic HOI methods (HOLD, BIGS, MagicHOI, and Do as I Do), StereoHOI achieves better

Table 3: **Ablation studies and efficiency comparisons.** (a) Single-frame stereo hand reconstruction ablations on Project Aria (errors in mm). (b) Hand sequence trajectory optimization ablations (MPJPE and WA-MP in mm). (c) Object dynamic tracking ablations. (d) Computational efficiency comparison across video methods on a 150-frame clip (peak GPU memory in MiB). ↓ denotes lower is better; best results are in bold. HaMeR† lifts monocular predictions to 3D with FoundationStereo depth.

(a) Single-Frame Hand Model Ablations

Configuration	CS-J ↓	R-J ↓	PA-J ↓	CT ↓
StereoHOI (Ours)	**11.17**	**8.83**	**4.95**	8.72
Stereo Disparity and Input Views				
w/o Right View	14.13	9.95	5.34	11.66
Duplicated Left	13.76	9.83	5.23	11.30
Transformer Depth Scalability				
18 Layers	11.58	9.27	5.08	8.98
36 Layers	11.27	9.14	5.00	**8.65**
Priors and Architecture Design				
w/o Prior Token Init	11.88	9.30	5.15	9.29
Direct Parameter Dec.	11.45	9.24	5.10	8.89
w/o 2D Keypoint Head	11.52	9.30	5.18	8.77
HaMeR† (Stereo Lift)	108.33	11.51	5.72	108.70

(b) Hand Sequence Refinement Ablation

Method	MPJPE ↓	WA-MP ↓	Jitter ↓	RTE (%) ↓
Full Model	**10.302**	**9.588**	**5.207**	**2.017**
w/o $\mathcal{L}_{\mathrm{KP}}$	10.819	10.056	5.214	**2.017**
w/o $\mathcal{L}_{T}$	10.511	9.821	7.495	2.089

(c) Object Sequence Tracking Ablation

Method	CD (cm) ↓	ATE (m) ↓	RPE_t (cm) ↓	RPE_r (°) ↓
Full Model	**0.760**	**0.0149**	**0.293**	**2.990**
w/o $\mathcal{L}_{\mathrm{align}}$	2.097	0.0432	0.479	4.872
w/o $\mathcal{L}_{\mathrm{smooth}}$	1.068	0.0243	0.788	79.942

(d) Video Efficiency Comparison (150-frame clip)

Metric	StereoHOI	Do as I Do	MagicHOI	HOLD	BIGS
Time (min) ↓	**14.6**	103.3	185.9	890.2	1,210
Peak Mem. ↓	20,812	23,936	19,032	18,384	**3,600**

object geometry (CD of 0.76 cm and higher F-scores), more accurate hand-relative object placement (CD_h of 1.08 cm), and more accurate object trajectories (ATE of 0.015 m and lower RPE). As shown in Figure 4, the shared metric space visibly reduces hovering and interpenetration compared with monocular baselines.

4.4 Ablation Studies and Baseline Comparisons

Single-frame hand ablations (Table 3a). Dropping the right view or duplicating the left view degrades camera translation and joint accuracy (CT rises to 11.66 and 11.30 mm), indicating that the gains come from true binocular parallax rather than dual-stream capacity. Using 18 layers increases all errors, while 36 layers bring no clear gain (only CT improves slightly, from 8.72 to 8.65 mm) with 50% more layers, so we use 24 layers by default. Furthermore, monocular token initialization and residual parameterization ease stereo learning (removing the prior token initialization raises CS-J to 11.88 mm), while direct 2D keypoint supervision aids 3D MANO mesh fitting. To benchmark modular pipelines, HaMeR† lifts monocular predictions with FoundationStereo depth but suffers from large translation errors, since dense stereo matching is ambiguous on textureless hands and general stereo networks recover the visible skin surface rather than the hand joints.

Sequence trajectory optimization (Table 3b, c). For hands (Table 3b), the smoothness term $\mathcal{L}_T$ suppresses high-frequency jitter while the keypoint term $\mathcal{L}_{\mathrm{KP}}$ preserves tracking accuracy; for objects (Table 3c), $\mathcal{L}_{\mathrm{align}}$ and $\mathcal{L}_{\mathrm{smooth}}$ prevent trajectory drift and orientation flips.

Video computational efficiency (Table 3d). By recovering metric hand-object alignment directly, StereoHOI needs far less optimization than HOLD, BIGS, and MagicHOI, and runs about seven times faster than the fastest baseline, Do as I Do (14.6 vs. 103.3 min), with lower peak memory.

5 Conclusion

We presented StereoHOI, a framework for metric 3D hand and hand-object interaction reconstruction from egocentric stereo videos. By lifting hand priors from a frozen monocular hand foundation model to the stereo setting through alternating cross-view attention, StereoHOI accurately recovers metric hand pose and shape and places the hands with the reconstructed objects in a shared metric space. Experiments on HOT3D demonstrate clear improvements over monocular and multi-view baselines in hand reconstruction accuracy, temporal tracking, and object reconstruction. Our model is still limited by the scale of calibrated stereo training data, and the sequence optimization takes minutes per clip; larger datasets and feed-forward multi-frame models are promising next steps.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing and language polishing of the manuscript. All AI-assisted text was reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses the publicly released HOT3D dataset (Banerjee et al., 2025) under its license terms, and we did not collect any new data involving human subjects. Our method reconstructs 3D hand motion and hand-object interaction from egocentric stereo video, with intended applications in robot learning, augmented reality, and human activity understanding. Because egocentric recordings may capture bystanders or private environments, applying such methods to new recordings should follow informed consent and applicable privacy regulations.

REPRODUCIBILITY STATEMENT

Section 3.1 and Section 3.2 describe the input preprocessing, network architecture, and prediction branches, and Section 3.3 gives the training objectives. Section 4.1 specifies the HOT3D training and test splits, the camera selection, and the evaluation metrics. The sequence-level hand and object optimization is described in Section 3.4, with its loss terms and main hyper-parameters detailed in Appendix A.

REFERENCES

Ayce Idil Aytekin, Xu Chen, Zhengyang Shen, Thabo Beeler, Helge Rhodin, Rishabh Dabral, and Christian Theobalt. Grasp in Gaussians: Fast monocular reconstruction of dynamic hand-object interactions. *arXiv preprint arXiv:2604.12929*, 2026.

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, et al. 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, 2025.

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

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. SAM 3D: 3Dfy anything in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7220–7232, 2026a.

Yuantao Chen, Jiahao Chang, Chongjie Ye, Chaoran Zhang, Zhaojie Fang, Chenghong Li, and Xiaoguang Han. ForeHOI: Feed-forward 3D object reconstruction from daily hand-object interaction videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8868–8879, 2026b.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion Policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems (RSS)*, 2023.

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 (NeurIPS)*, volume 37, pp. 2127–2160, 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, 2023.

Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Muhammed Kocabas, Xu Chen, Michael J Black, and Otmar Hilliges. HOLD: Category-agnostic 3D reconstruction of interacting hands and objects from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 494–504, 2024.

Evans Han, Yunfan Jiang, Yingke Wang, Haoyue Xiao, Huang Huang, Jianwen Xie, Jiajun Wu, Li Fei-Fei, and Ruohan Zhang. StereoPolicy: Improving robotic manipulation policies via stereo perception. *arXiv preprint arXiv:2605.09989*, 2026.

Shangchen Han, Po-Chen Wu, Yubo Zhang, Beibei Liu, Linguang Zhang, Zheng Wang, Weiguang Si, Peizhao Zhang, Yujun Cai, Tomas Hodan, et al. UmeTrack: Unified multi-view end-to-end hand tracking for VR. In *SIGGRAPH Asia 2022 Conference Papers*, 2022.

Weizhe Hua, Zihang Dai, Hanxiao Liu, and Quoc Le. Transformer quality in linear time. In *International Conference on Machine Learning (ICML)*, pp. 9099–9117, 2022.

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 *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13226–13233, 2025.

Baicheng Li, Dong Wu, Jun Li, Shunkai Zhou, Zecui Zeng, Lusong Li, and Hongbin Zha. MV-SAM3D: Adaptive multi-view fusion for layout-aware 3D generation. *arXiv preprint arXiv:2603.11633*, 2026.

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.

Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1954–1963, 2021a.

Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh Graphormer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12939–12948, 2021b.

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.

Yumeng Liu, Xiaoxiao Long, Zemin Yang, Yuan Liu, Marc Habermann, Christian Theobalt, Yuexin Ma, and Wenping Wang. EasyHOI: Unleashing the power of large models for reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7037–7047, 2025.

Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 674–679, 1981.

Gyeongsik Moon. Bringing inputs to shared domains for 3D interacting hands recovery in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17028–17037, 2023.

Jeongwan On, Kyeonghwan Gwak, Gunyoung Kang, Junuk Cha, Soohyun Hwang, Hyein Hwang, and Seungryul Baek. BIGS: Bimanual category-agnostic interaction reconstruction from monocular videos via 3D Gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17437–17447, 2025.

Bhawna Paliwal, Haritheja Etukuru, William Liang, Pieter Abbeel, Nur Muhammad Mahi Shafiullah, and Jitendra Malik. Do as I do: Dexterous manipulation data from everyday human videos. *arXiv preprint arXiv:2606.19333*, 2026.

Yik Lung Pang, Changjae Oh, and Andrea Cavallaro. Sparse multi-view hand-object reconstruction for unseen environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 803–810, 2024.

Yik Lung Pang, Alessio Xompero, Changjae Oh, and Andrea Cavallaro. Stereo hand-object reconstruction for human-to-robot handover. *IEEE Robotics and Automation Letters*, 10(6):5761–5768, 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*, pp. 9826–9836, 2024.

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)*, 2025.

Aditya Prakash, Ruisen Tu, Matthew Chang, and Saurabh Gupta. 3D hand pose estimation in everyday egocentric images. In *European Conference on Computer Vision (ECCV)*, pp. 183–202, 2024.

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

Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 16558–16569, 2021.

Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C Karen Liu. DexCap: Scalable and portable mocap data collection system for dexterous manipulation. In *Robotics: Science and Systems (RSS)*, 2024.

Shibo Wang, Haonan He, Maria Parelli, Christoph Gebhardt, Zicong Fan, and Jie Song. MagicHOI: Leveraging 3D priors for accurate hand-object reconstruction from short monocular video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5957–5968, 2025a.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 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.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: Unified 6D pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17868–17879, 2024.

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. FoundationStereo: Zero-shot stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5249–5260, 2025.

Hao Xu, Yilin Liu, Yinqiao Wang, Chi-Wing Fu, and Niloy J Mitra. CHOIR: Contact-aware 4D hand-object interaction reconstruction. *arXiv preprint arXiv:2605.20992*, 2026a.

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*, 2026b.

Lixin Yang, Licheng Zhong, Pengxiang Zhu, Xinyu Zhan, Junxiao Kong, Jian Xu, and Cewu Lu. Multi-view hand reconstruction with a point-embedded transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 47(11):10680–10695, 2025.

Yufei Ye, Jiaman Li, Ryan Rong, and C Karen Liu. WHOLE: World-grounded hand-object lifted from egocentric videos. *arXiv preprint arXiv:2602.22209*, 2026.

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)*, 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.

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, 2025.

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.

A SEQUENCE-LEVEL TEMPORAL OPTIMIZATION DETAILS

This section provides comprehensive mathematical formulations, implementation details, and hyperparameter configurations for the sequence-level temporal optimization introduced in §3.4. The entire pipeline is designed to run on continuous egocentric video sequences without requiring manual keyframing, external marker tracking, or predefined object CAD models.

A.1 HAND SEQUENCE TRAJECTORY OPTIMIZATION

In continuous egocentric video, frame-by-frame hand pose estimation inevitably suffers from sensor noise, momentary finger self-occlusions, and temporal jitter. To recover smooth and physically compliant hand trajectories across continuous video, our sequence-level optimization framework balances three complementary objectives: temporal motion smoothness, 2D visual grounding, and 3D trust-region anchoring.

Temporal motion smoothness ($\mathcal{L}_T$). Because human hands possess physical inertia, natural hand motions are continuous and do not exhibit abrupt acceleration spikes. To ensure motion continuity and temporal stability across the sequence, our smoothness objective $\mathcal{L}_T$ evaluates discrete second-order accelerations using actual capture timestamps:

$$\mathcal{L}_T \;=\; \frac{1}{T-2}\sum_{t=2}^{T-1}\left(\left\|\frac{\mathbf{a}_{\text{root},t}}{\alpha_{\text{root}}}\right\|^2 + \frac{1}{15}\sum_{j=1}^{15}\left\|\frac{\mathbf{a}_{\text{finger},t,j}}{\alpha_{\text{finger}}}\right\|^2\right), \tag{8}$$

where $\mathbf{a}_{\text{root},t} \in \mathbb{R}^3$ and $\mathbf{a}_{\text{finger},t,j} \in \mathbb{R}^3$ denote the second-order linear acceleration of the wrist and angular acceleration of the 15 articulated finger joints, computed via central finite differences over non-uniform intervals Δt. The constants $\alpha_{\text{root}} = 3.0\,\text{m/s}^2$ and $\alpha_{\text{finger}} = 30.0\,\text{rad/s}^2$ serve as normalization scales for wrist and finger accelerations, so that the penalty suppresses high-frequency jitter without strongly damping intentional rapid gestures.

2D visual grounding ($\mathcal{L}_{\text{KP}}$). Kinematic smoothing alone can cause rapid hand movements to "cut corners" away from visual evidence. To anchor the 3D hand to the image plane, the reprojection loss $\mathcal{L}_{\text{KP}}$ tethers 3D MANO joints to the 2D keypoints predicted by StereoHOI:

$$\mathcal{L}_{\text{KP}} \;=\; \frac{1}{21T}\sum_{t=1}^{T}\sum_{j=1}^{21}\rho\left(\frac{\pi(\mathbf{j}_{t,j}) - \mathbf{u}_{t,j}}{\boldsymbol{\sigma}_{\text{crop}}}\right), \tag{9}$$

where $\pi(\cdot)$ denotes pinhole projection with the camera parameters $\mathbf{k}$, and $\mathbf{u}_{t,j} \in \mathbb{R}^2$ is the predicted 2D coordinate of joint j across all 21 landmarks (1 wrist $+20$ finger joints). To handle frequent finger occlusions during manipulation, we apply a smooth Charbonnier loss $\rho(\mathbf{r}) = \sqrt{1 + \|\mathbf{r}\|^2} - 1$, which behaves quadratically for small errors of a few pixels and grows linearly for larger deviations, reducing the influence of occluded outliers on the hand. Normalizing residuals by $\boldsymbol{\sigma}_{\text{crop}} = 0.01 \times \mathbf{s}_{\text{crop}}$, where $\mathbf{s}_{\text{crop}} = (256, 192)$ pixels is the input crop size, ensures scale invariance across varying hand distances.

3D trust-region anchoring ($\mathcal{L}_{\text{anchor}}$). Because StereoHOI already resolves metric 3D depth and global hand orientation from binocular disparity, our optimizer acts as a local temporal refiner rather than re-estimating ungrounded geometry. We construct a trust region centered at initial network predictions:

$$\mathcal{L}_{\text{anchor}} \;=\; \frac{1}{T}\sum_{t=1}^{T}\left(\left\|\frac{\mathbf{x}_t - \mathbf{x}_t^{(0)}}{\sigma_x}\right\|^2 + \frac{1}{15}\sum_{j=1}^{15}\left\|\frac{\boldsymbol{\theta}_{t,j} - \boldsymbol{\theta}_{t,j}^{(0)}}{\sigma_\theta}\right\|^2\right), \tag{10}$$

where $\mathbf{x}_t^{(0)}$ and $\boldsymbol{\theta}_{t,j}^{(0)}$ denote the initial wrist position and finger pose parameters. The factor 15 reflects the kinematic tree of the MANO hand model (Romero et al., 2017): each of the 5 digits comprises 3 articulated joints (metacarpophalangeal, proximal interphalangeal, and distal interphalangeal joints for the four fingers, and carpometacarpal, metacarpophalangeal, and interphalangeal joints for the thumb), totaling $5 \times 3 = 15$ finger joints, while the 16th root joint (wrist) is parameterized separately

by $\mathbf{x}_t$. Tolerances $\sigma_x = 5.0\,\mathrm{mm}$ and $\sigma_\theta = 10.0^\circ$ bound expected feed-forward noise. We also enforce hard box constraints ($\|\mathbf{x}_t - \mathbf{x}_t^{(0)}\|_\infty \le 5.0\,\mathrm{mm}$, $\|\boldsymbol{\theta}_{t,j} - \boldsymbol{\theta}_{t,j}^{(0)}\|_\infty \le 10.0^\circ$) at each step, and fix hand shape $\boldsymbol{\beta}$, bone lengths, and wrist orientation to preserve biometric identity.

Joint sequence optimization. The overall objective unifies these physical forces:

$$\min_{\{(\mathbf{x}_t, \{\boldsymbol{\theta}_{t,j}\})\}} \mathcal{L}_{\text{hand}} \;=\; \mathcal{L}_{\text{anchor}} + \lambda_T \mathcal{L}_T + \lambda_{\text{KP}} \mathcal{L}_{\text{KP}}, \tag{11}$$

where we set $\lambda_T = \lambda_{\text{KP}} = 0.1$ to weight temporal smoothness and 2D landmark fidelity against the trust-region anchor. We optimize all sequence frames simultaneously using L-BFGS for 50 iterations, backpropagating gradients directly through the differentiable MANO layer.

A.2 Object Reconstruction and Dynamic 6D Tracking

In everyday egocentric interactions, manipulated objects vary widely, are heavily occluded by interacting fingers, and lack pre-existing 3D CAD models. To track dynamic objects across continuous video, our framework proceeds in three stages: object mesh creation, object pose initialization, and continuous 6D trajectory refinement.

Object mesh creation. Rather than relying on pre-existing CAD models, we reconstruct the object directly from the input video. We first segment the manipulated object across frames in the reference view using SAM 3.1 (Carion et al., 2025). To capture complete 3D geometry without redundancy, we compute the Structural Similarity Index (SSIM) (Wang et al., 2004) between masked crops to select keyframes with maximal viewpoint diversity. These keyframes are fed into MV-SAM3D (Li et al., 2026) to generate a canonical 3D surface mesh $\mathcal{M}_c$. To ground the mesh into true metric scale, we align its dimensions with the stereo point cloud back-projected from FoundationStereo (Wen et al., 2025). Crucially, this scale factor is computed and locked once across the sequence, yielding a strictly rigid 3D object mesh $\mathcal{M}$ immune to temporal scale breathing.

Object mesh pose initialization. Given the reconstructed mesh $\mathcal{M}$, we initialize candidate 6D poses using FoundationPose (Wen et al., 2024). However, symmetric or planar objects under finger occlusion often suffer from sudden 180° flips, while wearer head rotation conflates camera movement with manipulation. To decouple egomotion and resolve orientation ambiguities, we initialize N discrete pose hypotheses on keyframes, propagate them across the video, and transform them into the static world coordinate system using egocentric SLAM poses $\mathbf{T}_{w\leftarrow c,t}$ (Teed & Deng, 2021): $\mathbf{T}_{w\leftarrow o,t}^{(k)} = \mathbf{T}_{w\leftarrow c,t}\,\mathbf{T}_{c\leftarrow o,t}^{(k)}$. We then select the globally consistent trajectory k^* by minimizing a sequence-level energy:

$$k^* \;=\; \arg\min_{k\in\{1,\dots,N\}} \sum_{t=1}^{T} \Big(E_{\text{obs}}(\mathbf{T}_{c\leftarrow o,t}^{(k)}) + \lambda_{\text{cont}} E_{\text{cont}}(\mathbf{T}_{w\leftarrow o,t}^{(k)}, \mathbf{T}_{w\leftarrow o,t-1}^{(k)}) \Big), \tag{12}$$

where E_{obs} rewards silhouette overlap and stereo depth alignment (discounting hand-occluded pixels), while E_{cont} penalizes sudden angular flips and velocity jumps in world space.

Continuous 6D trajectory refinement. Starting from the selected hypothesis k^*, we refine the entire 6D trajectory $\{(\mathbf{R}_t, \mathbf{t}_t)\}_{t=1}^T$ by minimizing

$$\min_{\{(\mathbf{R}_t, \mathbf{t}_t)\}} \mathcal{L}_{\text{obj}} \;=\; \mathcal{L}_{\text{align}} + \lambda_{\text{anchor}} \mathcal{L}_{\text{anchor}} + \lambda_{\text{smooth}} \mathcal{L}_{\text{smooth}}, \tag{13}$$

with $\lambda_{\text{anchor}} = \lambda_{\text{smooth}} = 1$ and the weights inside each term given below. We optimize small rotation and translation updates of all frames jointly with Adam for 100 iterations, with learning rates of 6×10^{-3} for rotation and 8×10^{-4} for translation, and keep the iterate with the lowest objective. All residuals use the Charbonnier penalty $\rho(s) = \sqrt{1 + s^2} - 1$.

Tracked-point alignment ($\mathcal{L}_{\text{align}}$). This term ties the poses to point tracks on the object. In each frame, up to 180 corner features are detected inside the eroded object mask and tracked to the next frame with pyramidal Lucas-Kanade optical flow (Lucas & Kanade, 1981). A track is kept only if tracking it back returns within one pixel of its start, and both endpoints lie on the object with valid

stereo depth; additional tracks connect selected keyframe pairs over longer spans. Each kept track provides a 3D point $\mathbf{x}$ in the source frame a, lifted with stereo depth, its tracked pixel $\mathbf{u}$ in the target frame b, and its stereo 3D point $\mathbf{y}$ in frame b. Let $\mathbf{T}_{b\leftarrow a}$ denote the relative object motion implied by the current poses, which maps object points seen in frame a into frame b. We then define

$$\mathcal{L}_{\text{align}} \;=\; \mathcal{L}_{\text{fwd}} + 0.25\,\mathcal{L}_{\text{bwd}} + 0.12\,\mathcal{L}_{\text{depth}}, \tag{14}$$

where $\mathcal{L}_{\text{fwd}}$ penalizes the pixel distance between the projection $\pi(\mathbf{T}_{b\leftarrow a}\mathbf{x})$ and the tracked pixel $\mathbf{u}$, $\mathcal{L}_{\text{bwd}}$ applies the same check from frame b back to frame a using $\mathbf{y}$, and $\mathcal{L}_{\text{depth}}$ penalizes the 3D distance between $\mathbf{T}_{b\leftarrow a}\mathbf{x}$ and $\mathbf{y}$, normalized by an expected stereo error that grows with depth. Each track is weighted by the number of tracks in its frame pair and by the stability of the stereo depth around its endpoints. The two image terms make the poses follow the observed motion of the object, while the depth term fixes the object along the optical axis, where image motion alone is ambiguous.

Anchoring ($\mathcal{L}_{\text{anchor}}$). Grasping fingers often cover large parts of the object, so some frames have few reliable tracks and an unconstrained optimizer could drift. The anchor term keeps the trajectory close to its initialization:

$$\mathcal{L}_{\text{anchor}} \;=\; 0.2\,\mathcal{L}_{\text{pos}} + \mathcal{L}_{\text{prior}} + 0.03\,\mathcal{L}_{\text{trust}}. \tag{15}$$

$\mathcal{L}_{\text{prior}}$ is a quadratic penalty on the rotation and translation updates of each frame, with scales of 0.15 rad and 15 mm, and is weighted more strongly for frames whose initial pose is reliable and already agrees with the point tracks. $\mathcal{L}_{\text{trust}}$ leaves updates below 0.35 rad and 4 cm free but penalizes larger ones steeply, and $\mathcal{L}_{\text{pos}}$ keeps the transformed points at least 1 cm in front of the camera.

Motion smoothness ($\mathcal{L}_{\text{smooth}}$). Real objects have mass and inertia, so their position and orientation cannot change abruptly between adjacent frames. We therefore penalize the second-order differences of the object translation and rotation in world coordinates over consecutive frames, $\mathcal{L}_{\text{smooth}} = 0.025\,(\mathcal{L}_{\text{acc},\mathbf{t}} + \mathcal{L}_{\text{acc},\mathbf{R}})$, where the translation term uses a scale of 1 cm. Both terms are weighted more strongly where the observations are less reliable, which suppresses high-frequency tracking chatter while still allowing rapid manipulation.

Importantly, because hand and object trajectories are optimized independently within a shared, calibrated 3D metric coordinate system, their relative placement follows directly from the metric estimates, which reduces hovering and interpenetration without explicit penetration penalties or contact heuristics.

A.3 Qualitative Hand Reconstruction and Baseline Comparisons

We provide qualitative evaluations to complement the quantitative benchmarks, examining single-frame stereo baselines, continuous video methods, and detection robustness across diverse sequences.

Single-frame stereo baseline comparisons. Figure 5 compares StereoHOI against stereo baselines, namely HaMeR-FT (tri.), POEM-v2-FT (Yang et al., 2025), and UmeTrack. POEM-v2 is fine-tuned from its official weights on our training set, using joints triangulated from its own 2D predictions as training references; it outputs meshes without global rotation, so GO is not reported in Table 1, and its zero-shot version yields a CS-J above 3.8 m. StereoHOI generally recovers more accurate 3D hand poses, more natural hand shapes, and closer metric root locations, while the baselines show larger depth errors and distorted finger articulations under complex viewpoints and occlusions.

Video-based temporal hand reconstruction comparisons. Figure 6 compares continuous video reconstruction against temporal baselines (HaWoR (Zhang et al., 2025) and StableHand (Zeng et al., 2026)). While existing video methods accumulate tracking drift during dynamic interactions, StereoHOI keeps hand poses and trajectories closer to the ground truth.

Hand detection and reconstruction across diverse video clips. Figure 7 demonstrates the robustness of our framework across diverse real-world egocentric video clips. Across rapid movements, varying illumination, and heavy finger occlusions, our model reliably localizes hand regions and generates high-fidelity metric 3D hand meshes.



Figure 5: Qualitative comparison of single-frame hand pose and shape estimation between StereoHOI and stereo baselines, namely HaMeR-FT (tri.), POEM-v2-FT, and UmeTrack. CS and PA denote the camera-space and Procrustes-aligned joint errors of each example; for UmeTrack they are computed on the 20 joints shared with its hand skeleton (CS20/PA20). StereoHOI generally captures complex finger articulations and natural hand shape, and places the hand closer to the ground truth in camera space.

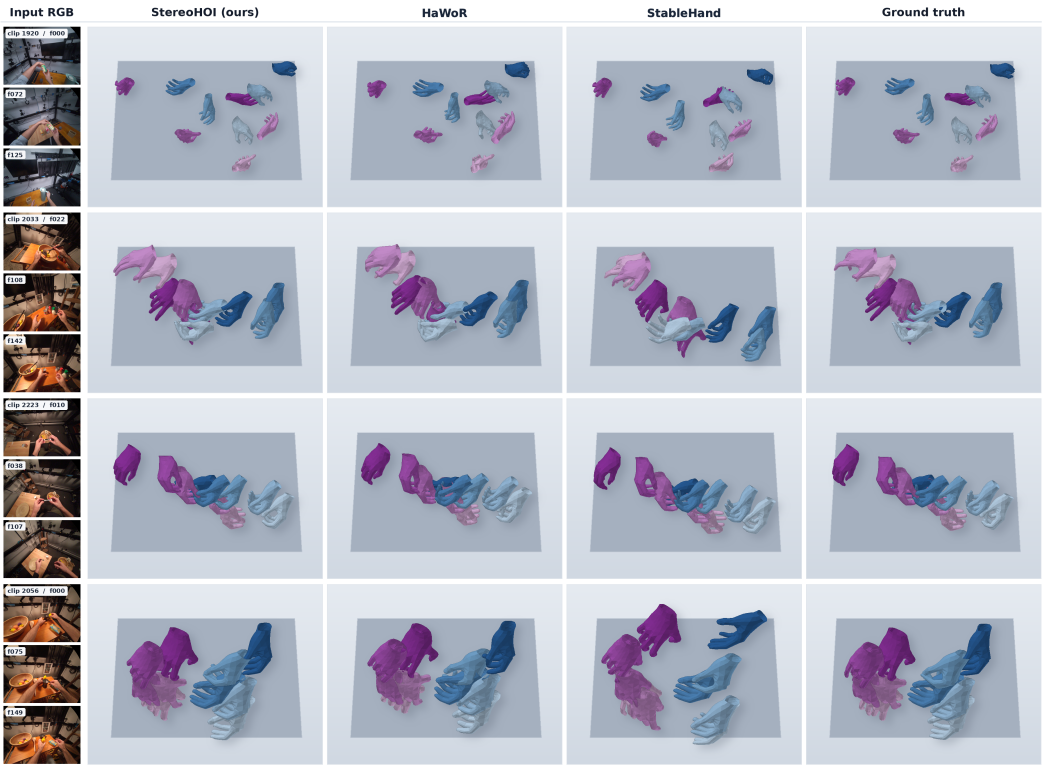

Figure 6: Qualitative comparison of continuous video hand reconstruction against temporal baselines (HaWoR and StableHand). StereoHOI reconstructs more accurate hand poses and trajectories with less drift across dynamic egocentric manipulation sequences.

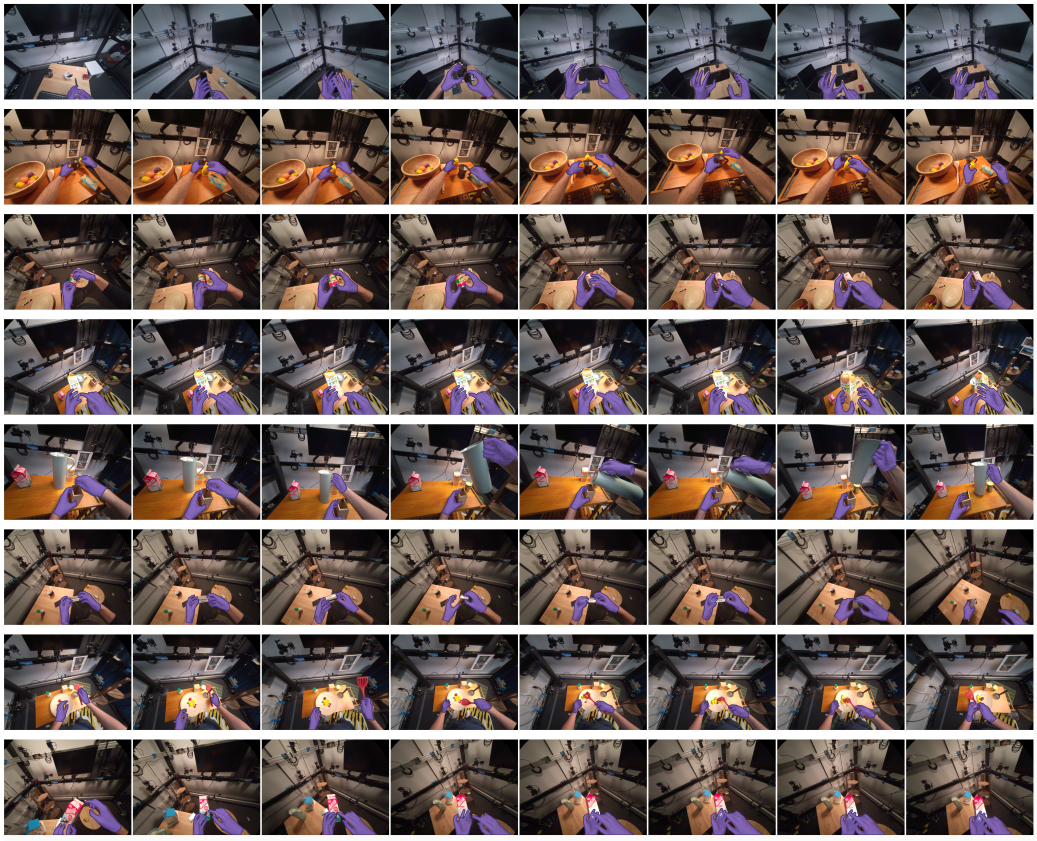

Figure 7: Qualitative visualizations of hand detection and 3D reconstruction across diverse egocentric video clips. Across challenging illumination conditions, varying hand-object scales, and rapid movements, our model consistently localizes hand regions and generates high-fidelity metric 3D hand meshes.