This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# Dyn-HaMR: Recovering 4D Interacting Hand Motion from a Dynamic Camera

Zhengdi Yu Stefanos Zafeiriou Tolga Birdal
Imperial College London

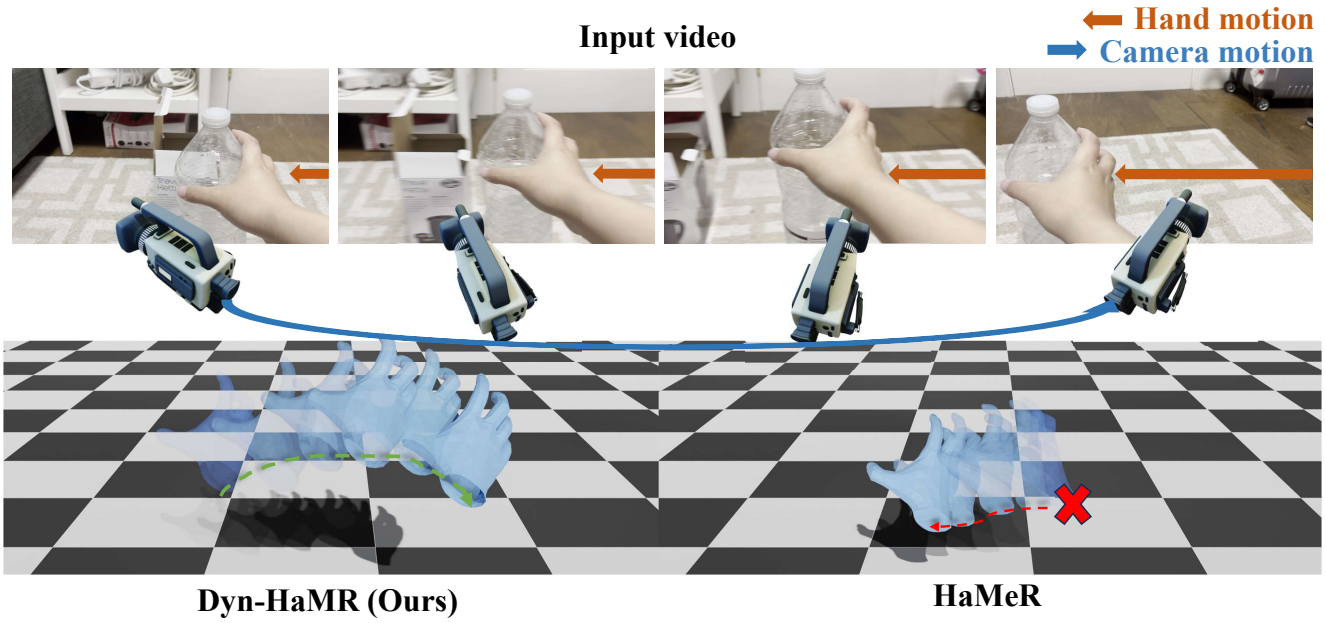



Figure 1. **Dyn-HaMR** as a remedy for the motion entanglement in the wild. The **green** and **red** arrows represent the direction of the hand motion. Dyn-HaMR (Ours) can disentangle the camera and object poses to recover the 4D global hand motion in the real world whilst state-of-the-art 3D hand reconstruction methods like HaMeR [37], IntagHand [25] and ACR [51] fail to do so since they cannot disentangle the sources of motion.

## Abstract

*We propose **Dyn-HaMR**, to the best of our knowledge, the first approach to reconstruct 4D global hand motion from monocular videos recorded by dynamic cameras in the wild. Reconstructing accurate 3D hand meshes from monocular videos is a crucial task for understanding human behaviour, with significant applications in augmented and virtual reality (AR/VR). However, existing methods for monocular hand reconstruction typically rely on a weak perspective camera model, which simulates hand motion within a limited camera frustum. As a result, these approaches struggle to recover the full 3D global trajectory and often produce noisy or incorrect depth estimations, particularly when the video is captured by dynamic or moving cameras, which is common in egocentric scenarios. Our Dyn-HaMR consists of a multi-stage, multi-objective optimization pipeline, that factors in (i) simultaneous localization and mapping (SLAM) to robustly estimate relative camera motion, (ii) an interacting-hand prior for generative infilling and to refine the interaction dynamics, ensuring plausible recovery under (self-)occlusions, and (iii) hierarchical initialization through a combination of state-of-the-art hand tracking methods. Through extensive evaluations on both in-the-wild and indoor datasets, we show that our approach significantly outperforms state-of-the-art methods in terms of 4D global mesh recovery. This establishes a new benchmark for hand motion reconstruction from monocular video with moving cameras. Our project page is at* https://dyn-hamr.github.io/.

## 1. Introduction

In our increasingly digitalized world, capturing and interpreting human movement has become essential for advancing our interaction with computers (HCI) and immersive experiences of augmented and virtual reality (AR/VR). Many of these applications rely on a single, off-the-shelf body-

mounted camera to capture hand motion, usually with complex interactions between two hands. However, as the body moves, the camera follows (*e.g.* egocentric), creating an intricate blend of hand and camera motions. This dynamic setup introduces a fundamental challenge: disentangling the hand motions of interest from the motion of the camera itself–a task often made difficult without auxiliary tracking.

Most current methods for monocular hand reconstruction [2, 4, 19, 23, 25, 32, 33, 51, 54, 55, 57] assume a weak perspective camera model, capturing hand motion in either the camera coordinate frame or in a root-relative coordinate system [32, 33]. These methods, however, overlook camera motion and depend solely on 2D cues, struggling with depth ambiguity, ultimately failing to recover accurate global hand trajectories.

Furthermore, hand interactions present additional obstacles, including frequent occlusions, truncation, and missed detections. Previous work either considered the simpler scenario of single-hand motion [2, 4, 55, 57] or focused on interacting hands without specifically recovering global trajectories [23, 25, 32, 51, 54]. Even with static cameras, the lack of strong interaction priors prevents existing methods from reconstructing two hands realistically under occlusions or truncations. Despite considerable progress, no approach to date has addressed the complete 4D hand reconstruction problem under the challenging conditions posed by dynamic cameras and complex hand interactions. Nor are there publicly available datasets with sufficient temporal information to enable learning of 4D global interactions.

In this work, we present **Dyn-HaMR**: a novel, multi-stage optimization-based framework for reconstructing 4D hand motion trajectories in complex, real-world settings captured by dynamic cameras. Starting with an input RGB video, Dyn-HaMR leverages a robust two-hand tracking system built upon off-the-shelf methods such as MediaPipe [30], ViTPose [48], ACR [51], and HaMeR [37] to initialize the motion state, hierarchically. We then estimate the relative camera motion using a SLAM system [46, 47]. Our multi-objective optimization ensures 3D shape projections align with 2D observations, while handling occlusions, missing detections, and ensuring plausible trajectories through a learned generative hand motion prior inspired by [8], and augmented with biomechanical constraints. We also account for the scale factors in the hands and camera displacement, improving depth reasoning.

Remarkably, as shown in Fig. 1, our approach bypasses the need for precise 3D scene reconstructions, making it adaptable to in-the-wild video data. We demonstrate our method's effectiveness through extensive experiments on dynamic, in-the-wild hand interaction videos and established benchmarks, including H2O [21], EgoDexter [35], FPHA [10], HOI4D [29], and InterHand2.6M [33]. Our main contributions include:

- Introducing the first optimization-based approach capable of disentangling and reconstructing global 4D pose and shape of two hands, and camera trajectory.
- Leveraging a data-driven hand motion prior with biomechanical constraints to guide the optimization, enabling the recovery of realistic and complex hand interactions
- Conducting comprehensive experiments on challenging in-the-wild videos and benchmarks, demonstrating substantial performance improvements over state-of-the-art methods in 4D global motion recovery.

Our supplementary materials provide further qualitative insights in dynamic settings using web-sourced images and videos, emphasizing the robustness of our method. We will release our implementation publicly upon publication.

## 2. Related Work

**Monocular reconstruction of two hands**. Pose and shape estimation of two hands (*bimanual*) has rapidly progressed recently [23, 25, 32, 37, 40, 51]. Among them, [37] scaled up the single pose estimation. Li *et al*. [25] developed a mesh regression network based on GCN that employs pyramid features and learned implicit attention between two hands. Moon el at. [32] proposed to improve in-the-wild hand reconstruction accuracy with a strategy to bridge the domain gap between multi-camera datasets and in-the-wild datasets. Ren *et al*. [40] used a variational autoencoder (VAE) as a prior for interacting hand reconstruction. Yu *et al*. [51] introduced a one-stage hand reconstruction pipeline for two hands by an attention aggregation mechanism with a 2D Gaussian heatmap and cross-hand attention. SignAvatars [52] used a multi-stage optimization pipeline to reconstruct 3D sign actors. All these methods suffer from depth ambiguity and increased jitter in reconstruction due to the coupling of camera and hand poses. Moreover, reliance on weak-perspective camera models inherently limits their ability to perform effectively in dynamic scenarios. In contrast, our optimization-based approach can reconstruct 4D global motion of two hands from complex scenes captured by a moving camera, while more accurately modelling their interaction using a learned hand motion prior.

**4D motion priors**. Many motion models have been learned for computer animation, mostly in the context of human motion [5, 18, 20, 26, 28, 42, 45], including recent recurrent and autoregressive models [12, 14, 17, 27, 49]. These often focus on visual fidelity for a small set of characters and periodic locomotions. Some have explored generating more general motion and body shapes [1, 6, 38, 56], but in the context of short-term future prediction. HuMoR [39] introduced an autoregressive shape prior amenable for test-time-optimization. Other approaches tracked the full hand motion from 3D points on the surface of the hand [7]. InterHandGen [24] proposed a generative model as two-hand prior in close interaction with or without an object. Fourier-

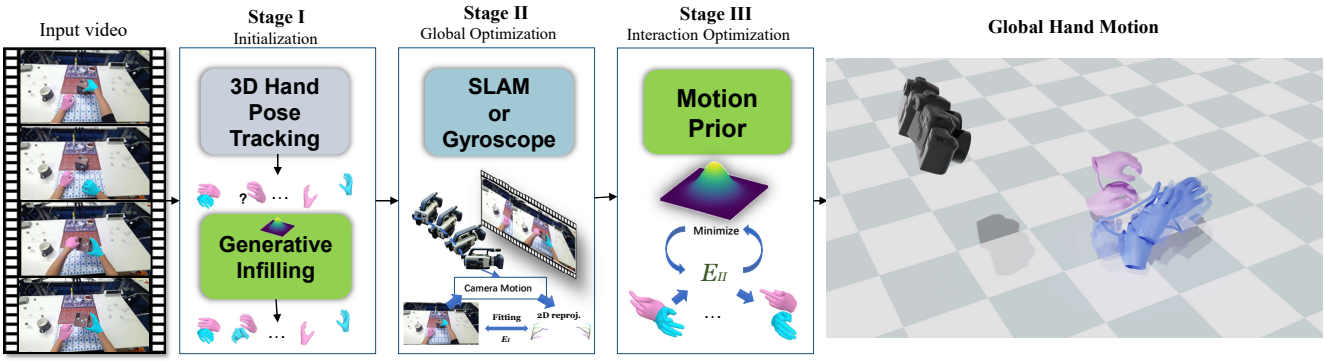



Figure 2. **Overview of our method.** We design a three-stage optimization pipeline to recover the 4D global hand motion from in-the-wild videos even with dynamic cameras. Our method can disentangle hand and camera motion as well as modelling complex hand interactions.

HandFlow [22] devised a spatiotemporal model for 3D hand shape, using a 3D hand occupancy field with articulation-aware query flows along the temporal axis. HMP [8] recently leveraged neural motion fields [15] to capture the plausible space of hand movements by training on the recent Arctic [9] dataset. We leverage their pre-trained models to impose plausibility on the motions we recover.

## 3. Dyn-HAMR

We consider an input video $\mathcal{V} = \{\mathbf{I}_1, \cdots, \mathbf{I}_T\}$ with $T$ frames containing two, possibly interacting hands undergoing arbitrary 6D camera motion. Our goal is to recover the global trajectory of both hands in the *world coordinate system*. As shown in Fig. 2, we design a three stage optimization pipeline, inspired by the recent works in dynamic human motion perception [43, 50]. The first stage (Sec. 3.1) leverages the state-of-the-art interacting hand pose estimation methods [37, 48, 51] to initialize the per-frame hand state for each of the hands in the camera coordinate system. Unlike human bodies, motion sequences extracted from hand images are frequently incomplete due to self-occlusions and rapid movements causing blur. As a remedy, we leverage the recent hand motion priors [8] to perform a generative motion infilling accounting for the missed detections. In the second stage (Sec. 3.2), our goal is to estimate the transformation from the world coordinate system to the camera coordinate system, while optimizing the global motion in the world coordinate system. To this end, we leverage a state-of-the-art SLAM system to compute the relative camera motions. To disambiguate the individual contributions of the camera and hand motion to the global hand motion, we also optimize for the global (world) scale factor. The third stage (Sec. 3.3) once again leverages the learned hand motion prior, this time to further constrain the displacements of hands and refine the complex interactions together with penetration and biomechanical constraints.

**Representing hand motion**. We represent the global motion trajectory as a sequence of hand states, $\mathbf{Q}^h = \{\boldsymbol{q}_t^h\}_{t=1}^T$. At a given time $t$, we parameterize the hand pose and shape via MANO [41] model as $\boldsymbol{q}_t^h = \{\boldsymbol{\theta}_t^h, \boldsymbol{\beta}_t^h, \boldsymbol{\phi}_t^h, \boldsymbol{\tau}_t^h\}$, where $\boldsymbol{\theta}_t^h \in \mathbb{R}^{3\times 15}$ denotes the local hand pose in the form of 15 hand joints, $\boldsymbol{\beta}_t^h \in \mathbb{R}^{10}$ are the shape coefficients, and $(\boldsymbol{\phi}_t^h, \boldsymbol{\tau}_t^h)$ depict the global wrist pose in the form of root orientation, parameterized as the axis-angle $\boldsymbol{\phi}_t^h \in \mathbb{R}^3$, and translation $\boldsymbol{\tau}_t^h \in \mathbb{R}^3$. $h \in [l, r]$ determines the handedness. We assume the identity and the hand shape to remain unchanged through the whole sequence, *i.e.*, $\boldsymbol{\beta}^h := \boldsymbol{\beta}_t^h \, \forall \, t$. We also use ${}^{\mathrm{c}}\boldsymbol{q}_t^h$ and ${}^{\mathrm{w}}\boldsymbol{q}_t^h$ to refer to the pose in camera and world coordinate frames, respectively. Similarly, ${}^{\mathrm{c}}\mathbf{Q}^h$ and ${}^{\mathrm{w}}\mathbf{Q}^h$ denote the trajectories in camera and world frames. At a given time $t$, MANO parameters $\boldsymbol{q}_t^h$ can be used to recover the hand mesh vertices $\mathbf{V} \in \mathbb{R}^{3\times 778}$ and joints $\mathbf{J} \in \mathbb{R}^{3\times 21}$ through the differentiable functions:

$$\mathbf{V}_t^h = W(H(\boldsymbol{J}_t^h, \boldsymbol{\beta}^h), P(\boldsymbol{\beta}^h), \mathbf{S}) + \boldsymbol{\tau}_t^h \mathbf{1}_{778} \quad (1)$$

$$\mathbf{J}_t^h = \mathbf{L}\mathbf{V}_t^h, \quad (2)$$

where $W(\cdot)$ is the skinning function, $H$ is the parametric hand template shape, and $\mathbf{1}_{778} \in \mathbb{R}^{1\times 778}$ is a row vector of ones. $P$ returns the hand joint position at the rest pose, $\mathbf{S}$ depicts the skinning weights and $\mathbf{L}$ is a pre-trained linear regressor.

**Hand motion priors**. Our framework makes use of priors on hand motion both for motion infilling and for ensuring the plausibility of reconstructed sequences. Unfortunately, due to the difficulties associated with rapid hand movements and global translations, modeling a prior over hands is challenging. We make use of the only available hand prior, HMP [8], which is based upon neural motion fields (NeMF) [15]. NeMF represents the motion as a continuous vector field of kinematic poses in the temporal domain by factoring out the root orientation from local motion: $D : (t, \mathbf{z}^h) \rightarrow \left(\boldsymbol{J}_t^h, (\boldsymbol{\phi}_t^h, \boldsymbol{\tau}_t^h)\right)$, where $\mathbf{z}^h$ is a latent code. To sample novel motions, a variational framework is used where the non-autoregressive decoder $D$ maps the sampled latent vector to a spatiotemporal sequence controlled by the timestep $t$.

### 3.1. Hand Tracking and Hierarchical Initialization

We initialize the per-frame motion state $\boldsymbol{q}_t^h$ by the efficient two-hand tracking system. To this end, we adopt a hierarchical initialization scheme by fusing the state-of-the-art interacting hand reconstruction methods [30, 37, 48, 51]. In particular, we first fine-tune a 2D hand pose estimation model based on ViTPose [48] and utilize it to obtain the bounding box sequence of each hand. Subsequently, we apply [30, 37, 51] on each cropped patch to extract a per frame motion state ${}^{\mathrm{c}}\boldsymbol{q}_t^h$ of each hand in the camera coordinates.

**Motion infilling and temporal consistency**. The aforementioned single-frame interacting hand reconstruction methods naturally lack temporal coherence. Moreover, due to the frequent occlusion during hand interactions, there could be missed detection making the trajectory ${}^{c}Q^h$ *incomplete*. We address both problems by employing the hand motion prior [8] as a generative, smooth motion hallucinator. To do so, for both hands, we optimize for the latent code $\mathbf{z}^h$ in HMP so as to fit the frames where detections are present. We initialize this optimizer from using a canonical *slerp* interpolation in the pose space. We compute the mean shape parameters $\boldsymbol{\beta}^h$ based on this motion sequence, finally leading to the initial 4D hand trajectory in the camera coordinate system ${}^{\mathrm{c}}\mathbf{Q}^h$.

We also initialize the 2D observations by incorporating ViTPose [48] and MediaPipe [30] with the reprojection of [37] and subsequently feed through a confidence-guided filter. To fill in the missed 2D keypoint detections, we reproject the 3D keypoints ${}^{\mathrm{c}}\mathbf{J}_t^h$ from ${}^{\mathrm{c}}\boldsymbol{q}_t^h$ onto the corresponding 2D image plane by weak-perspective camera parameters as $\hat{\mathbf{J}}_{\mathbf{t}}^{\mathbf{h}} \in \mathbb{R}^{3\times 21}$. This way, we get the interpolated 2D keypoints $\{\hat{\mathbf{J}^h}_t\}_{t=0}^T$. For the $i$-th joint at timestep $t$, $\hat{\mathbf{J}}_t^h \in \mathbb{R}^{2\times 21}$ and $\mathbf{J}_t^h \in \mathbb{R}^{3\times 21}$ have $\mathbf{x} = (x_i, y_i, z_i)$ and $\mathbf{p} = (\hat{x_i}, \hat{y_i})$, where the reprojection is defined as:

$$\hat{x_i} = sx_i + \tau_x \qquad \text{and} \qquad \hat{y_i} = sy_i + \tau_y, \tag{3}$$

where $s, \tau_x, \tau_y$ are the scale and the translation for $x, y$ axes separately (*i.e.*, the weak-perspective camera parameters estimated by the hand pose estimation model).

### 3.2. 4D Global Motion Optimization

We now explain the recovery of the original world trajectory of interacting hands. Given the trajectory in the camera coordinate system ${}^{\mathrm{c}}\mathbf{Q}$ (by Sec. 3.1), our key idea here is to compute the relative camera motion with a state-of-the-art data-driven SLAM system, DPVO [47], and to estimate the transform $\mathbf{C}_t = \{\mathbf{R}_t, \boldsymbol{\tau}^{\mathbf{c}}{}_t\}$ at each timestep $t$ from the camera coordinate system to the world coordinate system. Then, the composition of hand motion ${}^{\mathrm{c}}\mathbf{Q}^h$ and camera motion, *i.e.*, ${}^{\mathrm{w}}\boldsymbol{q}_t^h = \mathbf{C}_t \odot {}^{\mathrm{c}}\boldsymbol{q}_t^h$ reveals the global motion. However, the scale of camera motion within the world is inherently uncertain while the hand motions are naturally constrained to be plausible. Hence, we optimize a world scale factor $\omega$ to explicitly model the relative scale between the displacements of the camera and hand motion inspired by [50].

**Optimization variables**. During optimization, we take as input the initialized 2D keypoints sequence $\{\hat{\boldsymbol{J}}_t^h\}$, the 3D motion state sequence ${}^{\mathrm{c}}\mathbf{Q}^h$ in the camera coordinate system, and the world-to-camera transformation $\{\mathbf{R}_t, \boldsymbol{\tau}^{\mathbf{c}}{}_t\}_{t=0}^T$ estimated by the SLAM system, and subsequently propose a global optimization process that jointly optimizes the global trajectories, orientation and local poses of both hands and the camera extrinsics $\mathbf{C}_t$ to match the 2D observations. Specifically, we first initialize the global trajectory ${}^{\mathrm{w}}\mathbf{Q}$ in the world coordinate system as follows:

$$ {}^{\mathrm{w}}\boldsymbol{\phi}_t^h = \mathbf{R}_t^{-1} \cdot {}^{\mathrm{c}}\boldsymbol{\phi}_t^h, \tag{4}$$

$$ {}^{\mathbf{w}}\boldsymbol{\tau}_t^h = \mathbf{R}_t^{-1} \cdot {}^{\mathbf{c}}\boldsymbol{\tau}_t^h - \omega \mathbf{R}_t^{-1} \cdot \boldsymbol{\tau}^{\mathbf{c}}{}_t, \tag{5}$$

while ${}^{\boldsymbol{w}}\boldsymbol{\theta}_t^h = {}^{\boldsymbol{c}}\boldsymbol{\theta}_t^h = \boldsymbol{\theta}_t^h$ and ${}^{\mathrm{w}}\boldsymbol{\beta}_t^h = {}^{\mathrm{c}}\boldsymbol{\beta}_t^h = \boldsymbol{\beta}^h$ remain the same. The initial world-camera scale factor is set as $\omega = 1$. With the initialized motion state ${}^{\mathrm{w}}\boldsymbol{q}_t^h$, the 3D mesh joints at each timestep, ${}^{\mathrm{w}}\mathbf{J}_t^h$, can be extracted as below, where $\mathbf{1}_{21} \in \mathbb{R}^{1\times 21}$ is a row vector of ones:

$$ {}^{\mathrm{w}}\mathbf{J}_t^h = \mathbf{L} \cdot W(H(\boldsymbol{J}_t^h, \boldsymbol{\beta}^h), P(\boldsymbol{\beta}^h), \mathbf{S}) + {}^{\mathbf{w}}\boldsymbol{\tau}_t^h \mathbf{1}_{21}. \tag{6}$$

**Optimization scheme**. We recover the trajectory in the world coordinate frame by minimizing the following loss:

$$\begin{aligned} E_I({}^{\mathrm{w}}\boldsymbol{q}^h, \omega, \mathbf{R}_t, \boldsymbol{\tau}^{\mathbf{c}}{}_t) = \lambda_{2\mathrm{d}}\mathcal{L}_{2\mathrm{d}} + \lambda_{\mathrm{s}}\mathcal{L}_{\mathrm{smooth}} + \lambda_{\mathrm{cam}}\mathcal{L}_{\mathrm{cam}} \\ + \lambda_{\boldsymbol{J}}\mathcal{L}_{\boldsymbol{J}} + \lambda_{\boldsymbol{\beta}}\mathcal{L}_{\boldsymbol{\beta}}. \end{aligned} \tag{7}$$

The first term aligns the reprojection of ${}^{\mathrm{w}}\mathbf{Q}^h$ with the initial 2D observations $\{\hat{\mathbf{j}}_t^h\}$:

$$\mathcal{L}_{2\mathrm{d}} = \sum_{t=0}^{T} \sum_{h\in\{l,r\}} \rho\left(\mathbf{C}_t^h \left(\tilde{\boldsymbol{J}}_t^h - \hat{\boldsymbol{J}}_t^h\right)\right). \tag{8}$$

Here, $\tilde{\boldsymbol{J}}_t^h = \Pi({}^{\mathrm{w}}\mathbf{J}_t^h, \mathbf{R}_t, \omega, \boldsymbol{\tau}^{\mathbf{c}}{}_t, \mathbf{K})$ and $\Pi$ is the perspective camera projection with camera intrinsics $\mathbf{K} \in \mathbb{R}^{3\times 3}$. $\mathbf{C}_t^h$ is a mask obtained from the joint visibility and $\tilde{\mathbf{j}}_t^h$ is the reprojected 2D keypoints from the current 3D keypoints ${}^{\mathrm{w}}\mathbf{j}_t^h$. $\rho(\cdot)$ is the Geman-McClure robust function [11].

In practice, we minimize Eq. (7) by first optimizing the root orientation $\{\boldsymbol{\phi}_t^h\}_{t=0}^T$ and the translation $\{\boldsymbol{\tau}_t^h\}_{t=0}^T$ in world coordinate system by minimizing Eq. (8) for 20 steps followed by 60 steps for updating the local hand pose, shape, the scale factor $\omega$, and the camera extrinsics $\{\mathbf{R}_t, \boldsymbol{\tau}^{\mathbf{c}}{}_t\}_{t=0}^T$. Essentially, the motions of the two hands will further constrain the camera scale factor $\omega$ and improve the performance of complex hand interaction videos. However, the reprojection loss lacks constraints and could introduce implausible poses. To this end, we leverage the natural

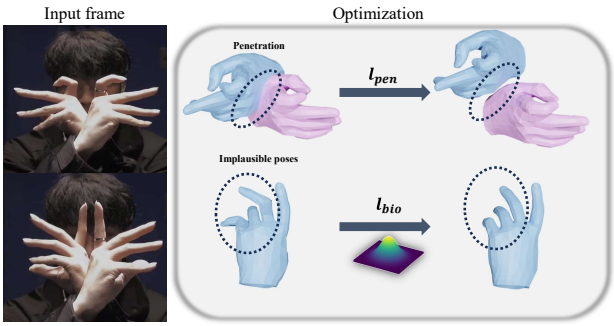



Figure 3. **Effectiveness of interaction optimization module**. Our penetration loss helps with separating hand meshes while the biomechanical constraints prevent implausible configurations that are not explained away by HMP.

temporal information and introduce regularization terms for both hand and camera motion to prevent implausible poses and jittering trajectories caused by depth ambiguity:

$$\mathcal{L}_{\text{smooth}} = \sum_{t=0}^{T} \sum_{h\in\{l,r\}} \|{}^{\text{w}}\mathbf{J}^h_{t+1} - {}^{\text{w}}\mathbf{J}^h_t\|^2 + d_{\boldsymbol{\theta}}(\boldsymbol{\theta}^h_{t+1}, \boldsymbol{\theta}^h_t)^2$$

$$\mathcal{L}_{\text{cam}} = \sum_{t=0}^{T} d_{\mathbf{R}}(\mathbf{R}_{t+1}, \mathbf{R}_t)^2 + \sum_{t=0}^{T} \|{\boldsymbol{\tau}^{\mathbf{c}}}_{t+1} - {\boldsymbol{\tau}^{\mathbf{c}}}_t\|^2,$$

where rotational terms $d$ are computed as geodesic distances. After that, in terms of reasonable regularization terms, to reduce the jittery poses, we employ standard pose $\mathcal{L}_{\boldsymbol{J}} = \sum_{t=0}^{T}\sum_{h\in\{l,r\}} \|\boldsymbol{J}^h_t\|^2$ and shape prior $\mathcal{L}_{\boldsymbol{\beta}}(\boldsymbol{\beta}^h) = \sum_{h\in\{l,r\}} \|\boldsymbol{\beta}^h\|^2$ term [41].

### 3.3. Interacting Motion Prior Optimization

After the optimization of Stages I and II, we obtained global hand motion in the world coordinate system. In Stage III, we introduce an interacting hand motion prior optimization module to better model the interactions. Moreover, it constrains the displacement of the hands to be plausible, which helps to determine the contribution of the camera from the hand motion with a well-learned scale factor $\omega$.

**Optimization variables**. For the latent optimization with motion prior [8], we omit the decoded global orientation as it is inherently less constrained and less correlated to the local pose compared to the body pose. Specifically, we initialize the latent code $\mathbf{z}^h$ from the pre-trained encoder. Our objective is to perform the optimization over the latent code $\mathbf{z}^h$, global motion state ${}^{\text{w}}\mathbf{Q}^h$ and the scale factor $\omega$ during the optimization. Similarly, we only optimize the root orientation $\{\boldsymbol{\phi}^h_t\}^T_{t=0}$ and the translation $\{\boldsymbol{\tau}^h_t\}^T_{t=0}$ in the first 200 steps and add $\mathbf{z}^h$ with hand local pose and camera parameters into the optimization variables in the following 200 steps. We recover the final interacting plausible hand motion by minimizing the following combined objective:

$$\begin{aligned} E_{II}({}^{\text{w}}\boldsymbol{q}^h, \omega, \mathbf{R}_t, {\boldsymbol{\tau}^{\mathbf{c}}}_t) = &\mathcal{L}_{\text{prior}} + \mathcal{L}_{\text{pen}} + \mathcal{L}_{\text{bio}} \\ &+ \lambda_{\text{2d}}\mathcal{L}_{\text{2d}} + \lambda_s\mathcal{L}_{\text{smooth}} \\ &+ \lambda_{\text{cam}}\mathcal{L}_{\text{cam}} + \lambda_{\boldsymbol{J}}\mathcal{L}_{\boldsymbol{J}} + \lambda_{\boldsymbol{\beta}}\mathcal{L}_{\boldsymbol{\beta}}. \end{aligned} \tag{9}$$

We now define each of the sub-objectives.

**Prior loss** ($\mathcal{L}_{\text{prior}}$). We define $\mathcal{L}_{\text{prior}} = \lambda_{\mathbf{z}}\mathcal{L}_{\mathbf{z}} + \lambda_{\boldsymbol{\phi}}\mathcal{L}_{\boldsymbol{\phi}} + \lambda_{\boldsymbol{\tau}}\mathcal{L}_{\boldsymbol{\tau}}$, where $\mathcal{L}_{\mathbf{z}}$ ensures that the motion is likely under the hand-motion prior by penalizing the negative log-likelihood for each of the hands:

$$\mathcal{L}_{\mathbf{z}} = \sum_{h\in\{l,r\}} \sum_{t=0}^{T} -\log \mathcal{N}(\mathbf{z}^h; \mu^h(\{\boldsymbol{J}^h_t\}), \sigma^h(\{\boldsymbol{J}^h_t\})).$$

The two other terms ensure an as jitter-free as possible trajectory by encouraging global consistency over the global orientation $\Phi^h$ and translation ${}^{\text{w}}\boldsymbol{\tau}$:

$$\mathcal{L}_{\boldsymbol{\phi}} = \sum_{t=0}^{T} d_{\boldsymbol{\phi}}({}^{\text{w}}\boldsymbol{\phi}_t, {}^{\text{w}}\hat{\boldsymbol{\phi}}_t) \text{ and } \mathcal{L}_{\boldsymbol{\tau}} = \sum_{t=0}^{T} \|{}^{\mathbf{w}}\boldsymbol{\tau}^h_t - {}^{\mathbf{w}}\hat{\boldsymbol{\tau}}^h_t\|^2.$$

**Biomechanical loss** ($\mathcal{L}_{\text{bio}}$). While the hand prior helps correcting certain implausible configurations, it is still necessary to explicitly constrain the hand pose for improved motion quality. Hence, we further add biomechanical constraints [44] to our objective function, which consists of three terms: $\mathcal{L}_{\text{bio}} = \lambda_{ja}\mathcal{L}_{\text{ja}} + \lambda_{bl}\mathcal{L}_{\text{bl}} + \lambda_{palm}\mathcal{L}_{\text{palm}}$. For simplicity, we omit the handedness here. For $i^{\text{th}}$ finger bone, each of the terms is defined as:

$$\mathcal{L}_{\text{ja}} = \sum_i d_{\alpha,H}(\boldsymbol{\alpha}^i_{1:T}, \mathbf{H}^i), \tag{10}$$

$$\mathcal{L}_{\text{bl}} = \sum_i \mathcal{I}(\|\mathbf{b}^i_{1:T}\|_2; b^i_{\text{min}}, b^i_{\text{max}}), \tag{11}$$

$$\begin{aligned} \mathcal{L}_{\text{palm}} = &\sum_i \mathcal{I}(\|\mathbf{c}^i_{1:T}\|_2; c^i_{\text{min}}, c^i_{\text{max}}) \\ &+ \sum_i \mathcal{I}(\|\mathbf{d}^i_{1:T}\|_2; d^i_{\text{min}}, d^i_{\text{max}}), \end{aligned} \tag{12}$$

where $\mathcal{L}_{\text{bl}}$ is for bone length, $\mathcal{L}_{\text{palm}}$ is for palmar region optimization, and $\mathcal{L}_{\text{ja}}$ is for joint angle priors. $\mathcal{L}_{\text{ja}}$ constrains the sequence of joint angles for the $i$-th finger bone $\boldsymbol{\alpha}^i_{1:T} = (\boldsymbol{\alpha}^f_{1:T}, \boldsymbol{\alpha}^a_{1:T})$ by approximating the convex hull on $(\boldsymbol{\alpha}^f_{1:T}, \boldsymbol{\alpha}^a_{1:T})$ plane with the point set $\mathbf{H}^i$, and the objective is to minimize the distance $d_{\boldsymbol{\alpha},\mathbf{H}}$ between them. $\mathcal{I}$ is the interval loss penalizing the outliers, and $b_i$ is the bone length of $i$-th bone. Finally, $\mathcal{L}_{\text{palm}}$ penalizes the outliers of curvature range $(c^i_{\text{min}}, c^i_{\text{max}})$ and angular distance range $(d^i_{\text{min}}, d^i_{\text{max}})$ to constraint for the 4 root bones of palm.

**Penetration loss** ($\mathcal{L}_{\text{pen}}$).The final loss in this section enhances the reconstruction quality under challenging hand

interactions by incorporating an interpenetration penalty to explicitly constrain the contact and the undesired penetrations between hand meshes:

$$\mathcal{L}_{\text{pen}} = \sum_{t=0}^{T} \left( \sum_{\mathbf{v}_t^r \in \mathbf{V}_t^r} \min_{\mathbf{v}_t^l \in \mathbf{V}_t^l} \|\mathbf{v}_t^l - \mathbf{v}_t^r\|^2 + \sum_{\mathbf{v}_t^l \in \mathbf{V}_t^l} \min_{\mathbf{v}_t^r \in \mathbf{V}_t^r} \|\mathbf{v}_t^l - \mathbf{v}_t^r\|^2 \right). \tag{13}$$

where $\mathbf{V}_t^l$ and $\mathbf{V}_t^r$ are the intersected vertices of the predicted left hand and right hand, respectively.

## 4. Experimental Evaluation

We evaluate our method qualitatively and quantitatively across a variety of datasets through various metrics. Our suppl. material further provides additional tables, visualizations and a video demonstration. We start by explaining our implementation, used metrics, datasets and baselines before moving onto presenting results.

**Implementation details**. We implement our method in PyTorch [36]. We use the L-BFGS algorithm for our three-stage optimization with a learning rate $lr = 1$. The following hyperparameters were used for each stage:

- Stage II: $\lambda_{2d} = 0.001$, $\lambda_{smooth} = 10$, $\lambda_{cam} = 100$ (camera displacement) and $\lambda_\theta = 0.04$, $\lambda_\beta = 0.05$.
- Stage III: $\lambda_z = 200$, $\lambda_\phi = 2$, $\lambda_\gamma = 10$, $\lambda_{pen} = 10$, $\lambda_\beta = 0.05$, $\lambda_{ja} = 1$, $\lambda_{palm} = 1$, $\lambda_{bl} = 1$.

For the baselines of [8, 25, 37, 51], we adhere to their original implementation.

**Evaluation metrics**. To thoroughly evaluate the performance of our pipeline, we conduct our evaluation on two fronts: (**i**) local pose and shape evaluation, and (**ii**) global motion reconstruction evaluation.

- **Local pose and shape evaluation**: We report the Mean Per Joint Error (MPJPE), Mean Per Vertex Position Error (MPVPE) and Acceleration Error (Acc Err) measured in $mm/s^2$ after root alignment to evaluate the smoothness of reconstructed hand motion.
- **Global motion evaluation**: We quantify the errors that accumulate over time due to camera motion following [43, 50, 53]. We split sequences into 128-frame segments to optimize and align them separately with ground truth using the first two frames (G-MPJPE) or the whole segment (GA-MPJPE) in $mm$.

We provide more details on implementation and evaluation in our supplementary material.

**Datasets**. Data used for evaluation of 3D interacting hands is typically captured with static cameras, such as InterHand2.6M [33], Re:InterHand [34]. To assess performance under dynamic camera movement, we employ the following egocentric hand-object interaction motion datasets:

- **H2O** [21] is a multiview dataset of two hands manipulating objects, which has 4 subjects performing 36 actions in 8 scenes. It provides the MANO and 3D pose annotations for both hands along with the object poses. In this work, we only focus on the egocentric view.
- **HOI4D** [29] is a large-scale egocentric dataset that contains 4000 video sequences of 9 subjects interacting with 800 different objects from 16 categories. HOI4D provides rich annotations of panoptic segmentation, motion segmentation, 3D hand pose, object pose, and hand action.
- **FPHA** [10] is an egocentric, dynamic, RGB-D hand dataset capturing hand in motion interacting with 3D objects, labeled with poses captured by magnetic sensors.
- **EgoDexter** [35] is an RGB-D dataset for hand tracking evaluation, which contains 4 challenging occlusions and clutter scenarios with 3D joints annotation lifted from manually annotated 2D joints.
- **InterHand2.6M** [33] is the first publicly available interacting hands dataset with 3D mesh annotations from multiview static cameras. For our video-based experiments, we use the interacting hand (IH) subset with both human and machine annotation (H+M) of the 30 fps version.
- **Ego-Exo4D** [13] is large-scale multi-view dataset with multi-modal annotations capturing simultaneous ego and multiple exo videos. We evaluated on the subset of hand ego-pose benchmark.
- **HOT3D** [3] is an egocentric dataset with multiview synchronized image streams recorded with two head-mounted devices, which provides accurate 3D poses and shapes of hands and objects.

Among them, we use the validation set for HOT3D and Ego-Ego4D datasets since their test set annotations (required to obtain GT global motion) are not publicly available this point. For other datasets used, we follow the official split to conduct our experiments.

**Baselines**. We compare our method against state-of-the-art approaches of ACR [51], IntagHand [25] and HaMeR [37] in terms of monocular hand reconstruction and global hand motion estimation. We initialize **Ours (Dyn-HaMR)** from [37] to get the results in the tables.

### 4.1. Results

We now present our experimental results for the tasks of bimanual motion estimation from static cameras, and local and global motion estimation in dynamic settings. We also provide ablations on different stages and components.

**Motion estimation from static cameras**. We start by assessing our approach in the classical regime of static cameras using the 30 FPS InterHand2.6M [33]. Tab. 1 reports our results across all compared methods evaluated using their official configurations for a fair comparison. Despite being designed to excel in dynamic camera scenarios, our method achieves state-of-the-art performance even in the

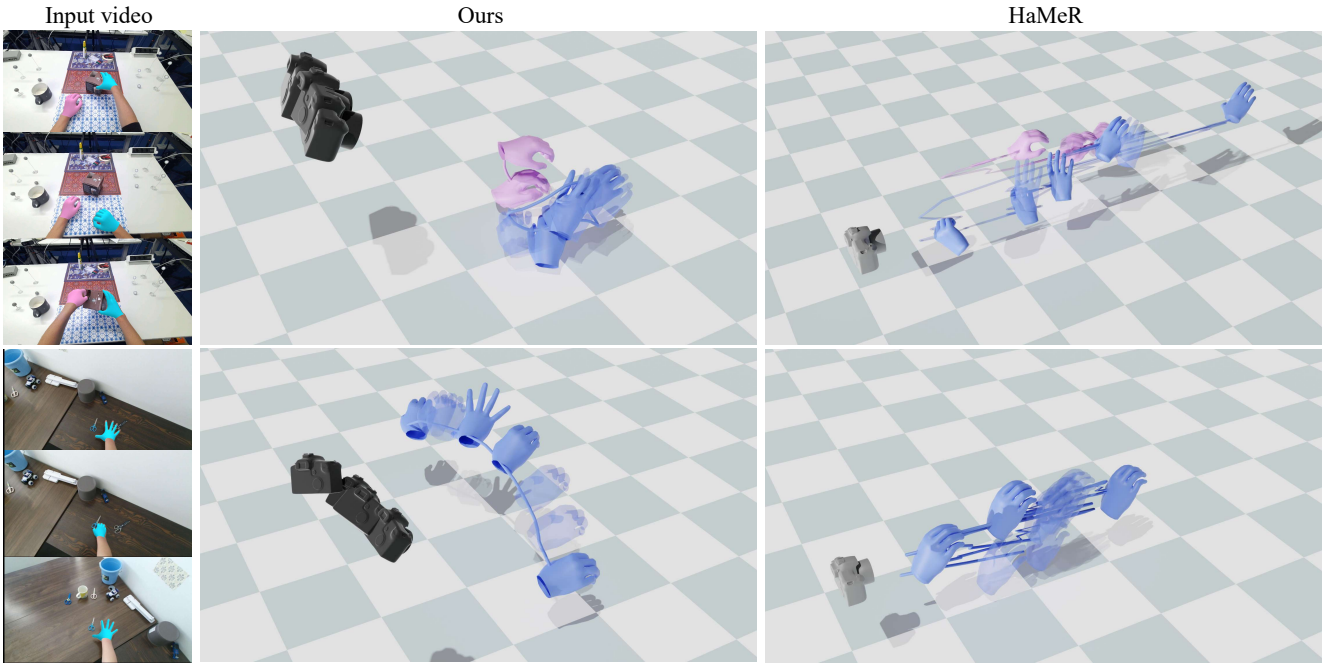



Figure 4. **Qualitative comparison with state-of-the-art method HaMeR [37].** It can be seen that our method recovers significantly more plausible global hand motion. **first row** is from H2O dataset [21], while the second row is from HOI4D dataset [29].

Table 1. **Quantitative evaluation results for InterHand2.6M [34] 30 fps dataset.** We compare our method with the state-of-the-art hand reconstruction methods on local hand poses.

| Method | MPJPE ↓ | MPVPE ↓ | Acc Err ↓ |
|---|---|---|---|
| InterWild [31] | 12.35 | 13.45 | 6.68 |
| DIR [40] | 9.09 | 9.43 | 8.92 |
| ACR [51] | 8.75 | 9.01 | 3.99 |
| IntagHand [25] | 9.26 | 9.71 | 4.41 |
| HaMeR [37] | 9.84 | 10.13 | 5.13 |
| Ours (w/o III) | 8.98 | 9.25 | 4.72 |
| Ours (Dyn-HaMR) | **7.94** | **8.15** | **2.76** |

static settings, consistently surpasses existing approaches across all reported metrics. This particularly showcases the effectiveness of our stage III, which significantly improves the plausibility of hand interactions by incorporating constraints from motion prior, biomechanical and penetration constraints, effectively mitigating depth ambiguities in the re-projection phase of earlier optimization stages. We present multiple views of our high-quality bimanual hand mesh reconstructions and their plausible interactions in Fig. 5. More qualitative results including results of in-the-wild videos can be found in our suppl. material.

**Local motion estimation**. We gauge the quality of reconstructing intricate details of the hands by measuring the joint reconstruction error, in relative to the root, a.k.a. *local motion*. Results reported in Tabs. 2 and 3 show the superiority of Dyn-HaMR on bimanual reconstruction on H2O and HOI4D datasets, respectively, in comparison to the state-of-the-art approaches of [25, 37, 51], where we consistently achieves the lowest MPJPE. We also assess the inter-frame smoothness and realism of motion by examining the acceleration error (Acc Err). Our approach achieves significantly lower Acc Err compared to existing methods, reflecting improved temporal consistency and smoother transitions in hand pose trajectories. Further evaluations including the FPHA dataset [10] are included in our suppl. material.

**Global motion estimation**. Tabs. 2 and 3 (as well as our suppl. material) reports our quantitative results on recovering local and global 4D motion under the presented metrics, using the official test split for the egocentric dynamic camera views. It can be observed that our Dyn-HaMR outperforms all of the state-of-the-art methods [25, 37, 51] by a large margin in terms of **G-MPJPE** and **GA-MPJPE**.

Fig. 4 presents a qualitative comparison against state-of-the-art methods, revealing significant improvements in generating realistic and plausible motion trajectories, particularly with enhanced depth reasoning under inter-hand occlusions. Our method can better recover the actual global trajectories and respect the motion of both the hands and the moving camera, whilst other methods suffer from depth ambiguity, being agnostic to camera pose & displacements. The entanglement of hand and camera motion further degenerates the scale estimation and lead to implausible motions. Moreover, Tab. 4 and Tab. 5 show the results on challenging two-hand object interaction scenarios, where further compare our method with the canonical HaMeR + DPVO demonstrating the vitality of our optimization. We present further qualitative results in motion reconstruction

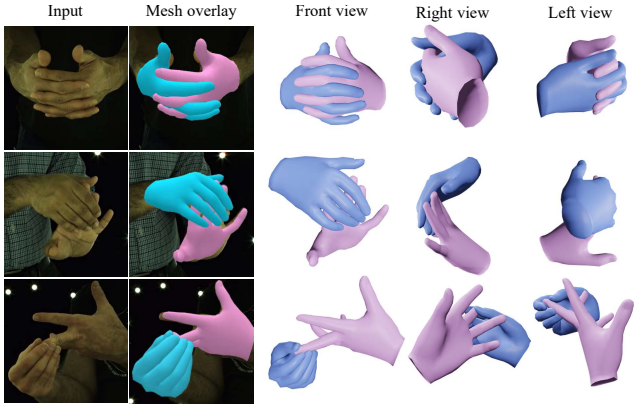



Figure 5. **Qualitative evaluation on InterHand2.6M [33].** In each row, we show the mesh overlay and detailed reconstruction from different views.

Table 2. **Quantitative evaluation results for H2O [21] dataset**. Our method demonstrates significant improvements over state-of-the-art approaches in recovering both local and global 4D hand motion, with additional gains achieved incorporating Stage III.

| Method | G-MPJPE ↓ | GA-MPJPE ↓ | MPJPE ↓ | Acc Err ↓ |
|---|---|---|---|---|
| ACR [51] | 113.6 | 88.5 | 46.8 | 14.3 |
| IntagHand [25] | 105.5 | 81.5 | 45.6 | 13.5 |
| HaMeR [37] | 96.9 | 75.7 | 32.9 | 9.21 |
| Ours (w/o III) | 51.9 | 41.2 | 24.9 | 9.5 |
| Ours (Dyn-HaMR) | **45.6** | **34.2** | **22.5** | **4.2** |

Table 3. **Quantitative evaluation results for HOI4D [29] dataset.** We compare our method with the state-of-the-art hand reconstruction methods [25, 37, 51].

| Method | G-MPJPE ↓ | GA-MPJPE ↓ | MPJPE ↓ | Acc Err ↓ |
|---|---|---|---|---|
| ACR [51] | 251.1 | 153.5 | 36.4 | 12.5 |
| IntagHand [25] | 291.3 | 145.6 | 40.9 | 14.1 |
| HaMeR [37] | 201.6 | 129.7 | 27.6 | 11.6 |
| Ours (w/o III) | 69.2 | 48.5 | 23.7 | 10.9 |
| Ours (Dyn-HaMR) | **58.5** | 45.6 | **19.5** | **4.1** |

on HOI4D, FPHA and EgoDexter datasets in the suppl. material. Our suppl. video provides in-the-wild capture of hand interactions from dynamic cameras.

**Ablation studies**. To fully assess the effectiveness of Dyn-HaMR, we perform further ablation studies on the pipeline design analyzing the contribution of each component. In particular, we investigate the effectiveness of the key components: (i) $\mathcal{L}_{bio}$, (ii) $\mathcal{L}_{pen}$, (iii) the interacting hand motion prior module in Stage III, and (iv) generative infilling, where we replace the generative infilling module with a simple interpolation. It can be seen from Tab. 2 and Sec. 4.1 that incorporating Stage III can boost the performance by a considerable margin as it provides a well-learned motion prior information for the final stage optimization and yields more plausible and smoother 4D trajectory reconstructions. Moreover, as shown in Fig. 3, adding the $\mathcal{L}_{bio}$ and $\mathcal{L}_{pen}$ can significantly improve the motion quality making it more plausible and realistic.

Table 4. **Quantitative evaluation Results on Ego-Exo4D [13] dataset.** We only report the positional error metrics as the GT annotation is too discrete to extract a meaningful acc.

| Method | Jerk ↓ | PA-MPJPE ↓ | G-MPJPE ↓ |
|---|---|---|---|
| ACR | 245.12 | 19.45 | 291.34 |
| IntagHand | 175.65 | 24.32 | 275.89 |
| HaMeR | 195.45 | 15.87 | 267.75 |
| HaMeR + DPVO | 200.45 | 15.87 | 213.75 |
| Ours (Dyn-HaMR) | **5.26** | **14.34** | **53.89** |

Table 5. **Qualitative results on HOT3D [3] dataset**. We compare our method with the canonical HaMeR + DPVO baseline.

| Method | Jerk ↓ | PA-MPJPE ↓ | G-MPJPE ↓ | Acc Err ↓ |
|---|---|---|---|---|
| ACR | 153.45 | 16.41 | 159.34 | 16.45 |
| IntagHand | 171.24 | 21.75 | 165.42 | 15.12 |
| HaMeR | 189.62 | 10.43 | 155.98 | 13.78 |
| HaMeR + DPVO | 195.77 | 10.43 | 129.45 | 12.78 |
| Ours (Dyn-HaMR) | **4.18** | **8.87** | **42.36** | **4.95** |

Table 6. **Ablation of pipeline components on H2O [21] dataset.** It shows the impact of removing different components from the pipeline on various performance metrics.

| Method | G-MPJPE ↓ | GA-MPJPE ↓ | MPJPE ↓ | Acc Err ↓ |
|---|---|---|---|---|
| Stage I | 84.5 | 72.5 | 25.6 | 8.8 |
| Stage I+II | 51.9 | 41.2 | 24.9 | 9.5 |
| w/o bio. const. | 49.6 | 43.1 | 24.5 | 4.3 |
| w/o pen. const. | 46.3 | 34.7 | 23.6 | **4.1** |
| w/o gen. infill. | 48.9 | 37.8 | 24.1 | 5.6 |
| Ours (Dyn-HaMR) | **45.6** | **34.2** | **22.5** | 4.2 |

## 5. Conclusion

We introduced **Dyn-HaMR**, to the best of our knowledge, the only data-driven work, which could reliably recover the 4D global motion of two interacting hands from complex, in-the-wild videos containing complex scenes, acquired by moving, dynamic cameras. Dyn-HaMR achieves this by leveraging a state-of-the-art SLAM system in conjunction with proposed interacting hand priors. Our method consists of a multi-objective optimization pipeline in which we estimate the relative camera motions and address motion entanglement as well as depth ambiguity problems by aligning the reconstruction with 2D observations. We further incorporate a learnable world scale factor to disambiguate the contributions of local hand and camera motions to the global hand motion. Our interaction priors allow for filling in the missing detections while ensuring plausible hand trajectories as demonstrated with extensive evaluations.

**Limitation & future work**. While Dyn-HaMR successfully works for limited time-horizons, extending it to long sequences with generative, extrapolation capabilities remains to be explored. Developing a regression-based method is one of the possible directions to deal with long sequences. We will also work on improving the hand priors extending [16] and incorporate object interactions.

## Acknowledgments

T. Birdal and S. Zafeiriou acknowledge support from the Engineering and Physical Sciences Research Council [grant EP/X011364/1]. T. Birdal was supported by a UKRI Future Leaders Fellowship [grant number MR/Y018818/1] as well as a Royal Society Research Grant RG/R1/241402.

## References

[1] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7144–7153, 2019. 2

[2] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1067–1076, 2019. 2

[3] P. Banerjee, S. Shkodrani, P. Moulon, S. Hampali, S. Han, F. Zhang, L. Zhang, J. Fountain, E. Miller, S. Basol, R. Newcombe, R. Wang, J. Engel, and T. Hodan. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. *arXiv preprint arXiv:2411.19167*, 2024. 6, 8

[4] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10843–10852, 2019. 2

[5] Matthew Brand and Aaron Hertzmann. Style machines. In *ACM SIGGRAPH*, pages 183–192, 2000. 2

[6] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[7] Guillaume Dewaele, Frédéric Devernay, and Radu Horaud. Hand motion from 3d point trajectories and a smooth surface model. In *Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part I 8*, pages 495–507. Springer, 2004. 2

[8] 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)*, 2024. 2, 3, 4, 5, 6

[9] 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 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[10] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 6, 7

[11] Stuart Geman. Statistical methods for tomographic image restoration. *Bull. Internat. Statist. Inst.*, 52:5–21, 1987. 4

[12] Saeed Ghorbani, Calden Wloka, Ali Etemad, Marcus A Brubaker, and Nikolaus F Troje. Probabilistic character motion synthesis using a hierarchical deep latent variable model. In *Computer Graphics Forum*. Wiley Online Library, 2020. 2

[13] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, K. Ashutosh, V. Baiyya, S. Bansal, B. Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 6, 8

[14] Ikhsanul Habibie, Daniel Holden, Jonathan Schwarz, Joe Yearsley, and Taku Komura. A recurrent variational autoencoder for human motion synthesis. In *28th British Machine Vision Conference*, 2017. 2

[15] Chengan He, Jun Saito, James Zachary, Holly Rushmeier, and Yi Zhou. Nemf: Neural motion fields for kinematic animation. *Advances in Neural Information Processing Systems*, 35:4244–4256, 2022. 3

[16] Yannan He, Garvita Tiwari, Tolga Birdal, Jan Eric Lenssen, and Gerard Pons-Moll. Nrdf: Neural riemannian distance fields for learning articulated pose priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1661–1671, 2024. 8

[17] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 2

[18] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017. 2

[19] Changlong Jiang, Yang Xiao, Cunlin Wu, Mingyang Zhang, Jinghong Zheng, Zhiguo Cao, and Joey Tianyi Zhou. A2j-transformer: Anchor-to-joint transformer network for 3d interacting hand pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8846–8855, 2023. 2

[20] Lucas Kovar, Michael Gleicher, and Frédéric Pighin. Motion graphs. In *ACM Transactions on Graphics 21(3), Proc. SIGGRAPH*, pages 473–482, 2002. 2

[21] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10138–10148, 2021. 2, 6, 7, 8

[22] Jihyun Lee, Junbong Jang, Donghwan Kim, Minhyuk Sung, and Tae-Kyun Kim. Fourierhandflow: Neural 4d hand representation using fourier query flow. In *NeurIPS*, 2023. 3

[23] Jihyun Lee, Minhyuk Sung, Honggyu Choi, and Tae-Kyun Kim. Im2hands: Learning attentive implicit representation of interacting two-hand shapes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[24] Jihyun Lee, Shunsuke Saito, Giljoo Nam, Minhyuk Sung, and Tae-Kyun Kim. Interhandgen: Two-hand interaction generation via cascaded reverse diffusion. In *CVPR*, 2024. 2

[25] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 6, 7, 8

[26] Yan Li, Tianshu Wang, and Heung-Yeung Shum. Motion texture: A two-level statistical model for character motion synthesis. In *ACM Transactions on Graphics 21(3), Proc. SIGGRAPH*, pages 465–472, 2002. 2

[27] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel van de Panne. Character controllers using motion vaes. In *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*. ACM, 2020. 2

[28] C. Karen Liu, Aaron Hertzmann, and Zoran Popović. Learning physics-based motion style with nonlinear inverse optimization. *ACM Trans. Graph*, 2005. 2

[29] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21013–21022, 2022. 2, 6, 7, 8

[30] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 2, 4

[31] 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*, pages 17028–17037, 2023. 7

[32] Gyeongsik Moon. Bringing inputs to shared domains for 3D interacting hands recovery in the wild. In *CVPR*, 2023. 2

[33] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 6, 8

[34] Gyeongsik Moon, Shunsuke Saito, Weipeng Xu, Rohan Joshi, Julia Buffalini, Harley Bellan, Nicholas Rosen, Jesse Richardson, Mize Mallorie, Philippe Bree, Tomas Simon, Bo Peng, Shubham Garg, Kevyn McPhail, and Takaaki Shiratori. A dataset of relighted 3D interacting hands. In *NeurIPS Track on Datasets and Benchmarks*, 2023. 6, 7

[35] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2017. 2, 6

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 2019. 6

[37] 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*, pages 9826–9836, 2024. 1, 2, 3, 4, 6, 7, 8

[38] Dario Pavllo, Christoph Feichtenhofer, Michael Auli, and David Grangier. Modeling human motion with quaternion-based neural networks. *International Journal of Computer Vision*, pages 1–18, 2019. 2

[39] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[40] Pengfei Ren, Chao Wen, Xiaozheng Zheng, Zhou Xue, Haifeng Sun, Qi Qi, Jingyu Wang, and Jianxin Liao. Decoupled iterative refinement framework for interacting hands reconstruction from a single rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 7

[41] Javier Romero, Dimitris Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6), 2017. 3, 5

[42] Charles Rose, Michael F. Cohen, and Bobby Bodenheimer. Verbs and adverbs: Multidimensional motion interpolation. *IEEE Computer Graphics and Applications*, 18(5):32–40, 1998. 2

[43] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2070–2080, 2024. 3, 6

[44] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *European conference on computer vision*, pages 211–228. Springer, 2020. 5

[45] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Trans. Graph.*, 38(6):209–1, 2019. 2

[46] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34, 2021. 2

[47] Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 4

[48] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 2, 3, 4

[49] Dongseok Yang, Doyeon Kim, and Sung-Hee Lee. LoBSTr: Real-time Lower-body Pose Prediction from Sparse Upper-body Tracking Signals. *Computer Graphics Forum*, 2021. 2

[50] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 4, 6

[51] Zhengdi Yu, Shaoli Huang, Fang Chen, Toby P. Breckon, and Jue Wang. Acr: Attention collaboration-based regressor for arbitrary two-hand reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 3, 4, 6, 7, 8

[52] Zhengdi Yu, Shaoli Huang, Yongkang Cheng, and Tolga Birdal. Signavatars: A large-scale 3d sign language holistic motion dataset and benchmark. In *European Conference on Computer Vision*, pages 1–19. Springer, 2024. 2

[53] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11038–11049, 2022. 6

[54] Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang. Interacting two-hand 3d pose and shape reconstruction from single color image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

[55] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2354–2364, 2019. 2

[56] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3372–3382, 2021. 2

[57] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5346–5355, 2020. 2