Estimating Body and Hand Motion in an Ego-sensed World

Brent Yi[1] Vickie Ye[1] Maya Zheng[1] Lea Müller[1]
Georgios Pavlakos[2] Yi Ma[1] Jitendra Malik[1] Angjoo Kanazawa[1]

[1]UC Berkeley [2]UT Austin

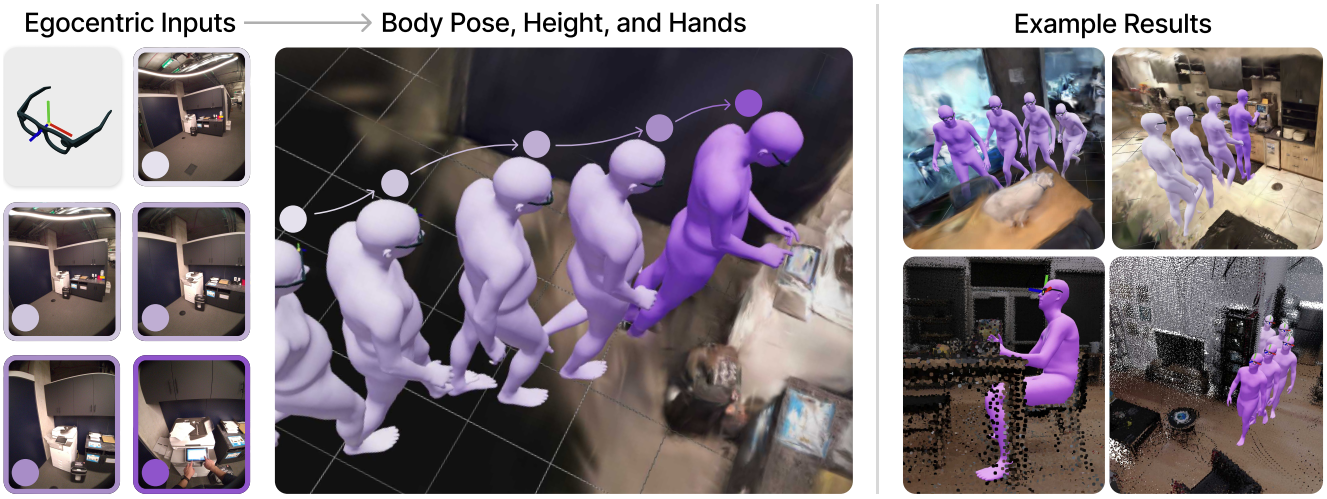


Figure 1. **EgoAllo.** We present a system that estimates human body pose, height, and hand parameters from egocentric SLAM poses and images. Outputs capture the wearer's actions in the allocentric reference frame of the scene, which we visualize here with 3D reconstructions.

Abstract

We present EgoAllo, a system for human motion estimation from a head-mounted device. Using only egocentric SLAM poses and images, EgoAllo guides sampling from a conditional diffusion model to estimate 3D body pose, height, and hand parameters that capture a device wearer's actions in the allocentric coordinate frame of the scene. To achieve this, our key insight is in representation: we propose spatial and temporal invariance criteria for improving model performance, from which we derive a head motion conditioning parameterization that improves estimation by up to 18%. We also show how the bodies estimated by our system can improve the hands: the resulting kinematic and temporal constraints can reduce errors in noisy monocular estimates by 40%. Project page: https://egoallo.github.io

1. Introduction

Head-mounted devices are permeating the mainstream market, with applications in areas like virtual reality, content creation, and assistive technologies. For research in machine perception, these devices also present exciting opportunities in the form of egocentric sensing. Sensors from wearable devices provide abundant observations of 3D environments, while capturing the embodied perspective of human agents as they navigate and interact with the world around them.

What can we understand from these devices as their adoption widens? As a starting point, parallax-inducing egomotion provides excellent conditions for advances in 3D reconstruction and scene understanding [17, 65, 109]. Limiting perception to only the surrounding world, however, would neglect a crucial piece of the ego-sensory puzzle: the individual whose decisions shape the inputs. Capturing the wearer's actions and motion *in addition to* the scene promises to unlock applications across augmented and virtual reality, robotics, and general human behavior analysis, while unveiling action-tied semantics in the scene itself.

We therefore introduce EgoAllo, a system that uses egocentric inputs to estimate the wearer's actions in the allocentric frame of the world. This is a difficult task: while body parts like hands occasionally appear in egocentric image frames, most body parameters are never directly observed by these devices. Accurate estimates in the global frame also require harmony

between both pose *and* height parameters. These must be consistent with the egomotion and scene scale, aligning estimated feet to the ground and head to the sensed camera height.

We cast estimation as guided sampling from a human motion diffusion model. We consider two inputs: head poses from SLAM and visual hand observations from images. We then estimate body pose, height, and hand parameters, unlike most prior works in egocentric human motion estimation [6, 46] that focus on body pose.

We achieve this by training a head motion-conditioned diffusion model as a motion prior, and guiding its sampling to recover a hand-body sequence that aligns with image observations. Our results are enabled by a key insight: that the representation used for head pose conditioning is critical for accurate ego-sensed motion estimation. We study choices for this representation by (1) identifying desirable spatial and temporal invariance properties that are not fulfilled by existing systems, (2) using these properties to derive improved parameterizations for our motion prior, and (3) presenting a systematic evaluation that shows between 4.9% and 17.9% error difference when compared against prior work. Furthermore, we show how the resulting system can improve hand estimation, reducing errors by over 40% compared to noisy monocular estimates.

2. Related Work

3D human recovery from external visual inputs. A large body of work has addressed estimating the parameters of human body models like SCAPE [2] or SMPL and its variants [50, 60, 76] from third-person visual inputs, where human subjects are observed from the view of outside cameras. The majority of these works focus on extracting 3D representations from single images, for example by lifting 2D keypoint observations to 3D [55], via end-to-end regression [19, 30, 32, 38, 57, 59, 74], via optimization [18, 42, 60], or by exploiting synergies between regression and optimization [40]. When multiple frames are available in the form of a video, temporal context and tracking can also be incorporated [15, 33, 37, 61, 62, 68, 106]. The inputs (images) and outputs (human meshes) of many of these systems are similar to the egocentric setting addressed by EgoAllo, but egocentric devices present unique challenges because the body being estimated is typically *behind* the outwards-facing cameras used as input.

Priors for human motion. The primary challenge of ego-sensed human motion estimation is limited observability; a prior is required to resolve ambiguities. For human motion, these priors are typically framed as unconditional distributions over plausible human motions. These distributions can be represented either by modeling the physical constraints of our world [5, 48, 64, 70] or by learning generative models of human motion directly from data. For learning unconditional priors, classical data-driven approaches include fitting mixtures-of-Gaussians to 3D keypoint trajectories [24], while modern approaches include training variational autoencoders [36, 72] to model either autoregressive transitions [14, 49, 71] or full spatiotemporal sequences [21]. After training, these priors can be applied to estimation problems in iterative optimization frameworks [39, 71, 99]. EgoAllo is built on the same intuition as these methods, but follows previous work in ego-sensed motion estimation and uses a task-specific conditional prior.

Denoising diffusion for human motion. The core of EgoAllo is a denoising diffusion model [22, 66, 81] from which we can sample 3D human body motion. While diffusion models are primarily known for their success in text-conditioned image generation [75, 77], they have also enabled advances in human motion synthesis conditioned on modalities like text [34, 35, 110], music [1, 90], poses [34, 46], and object geometry [41, 45, 47]. EgoAllo adopts a similar conditional diffusion approach, while specifically studying the design of conditioning parameters used for ego-sensed human motion estimation. The iterative nature of denoising diffusion also enables guidance [10, 12, 29, 34, 84, 112], where denoising steps are steered to satisfy a desired objective. We use guidance to incorporate observations like visual hand pose observations during test-time.

Human motion from egocentric observations. EgoAllo builds on intuition from several prior works in egocentric sensing for human motion estimation. Many rely on fisheye cameras that place the wearer's body into the field of view [26, 73, 88, 88, 89, 92, 93, 95]. Other approaches rely on body-mounted cameras [80], simulation-based physical plausibility [52, 104, 105], body- and hand-mounted inertial sensors [44, 102, 103], hand-held controllers [6, 27, 28], and interaction cues from other humans [56]. Concurrent works have also used the Nymeria [53] dataset for egocentric motion with language description outputs [23], as well for online settings with scene geometry and CLIP [67] feature inputs [20]. Most relevantly, EgoEgo [46] demonstrates how human body poses can be estimated offline without body observability assumptions. The authors accomplish this by carefully integrating several components: a monocular SLAM system [86], a pose-conditioned gravity vector regression network, an optical flow feature-conditioned head orientation and scale regression network, and a head pose-conditioned body diffusion model. EgoAllo differs in both inputs—we study conditioning parameters computed from the metric SLAM poses provided by devices like Project Aria [82]—and outputs—we consider body height variation and hand poses.

Conditioning for ego-sensed poses. Prior works vary in how head pose information is parameterized and used as neural network input. AvatarPoser [27] and BoDiffusion [6] parameterize head pose as four components: world-frame orientation, orientation deltas, world-frame position, and world-frame position deltas. These works are focused on settings with VR controller input, and parameterize controller pose inputs the same way. EgoEgo [46]'s diffusion model uses only absolute head positions and orientations, but, similar to HuMoR [71],

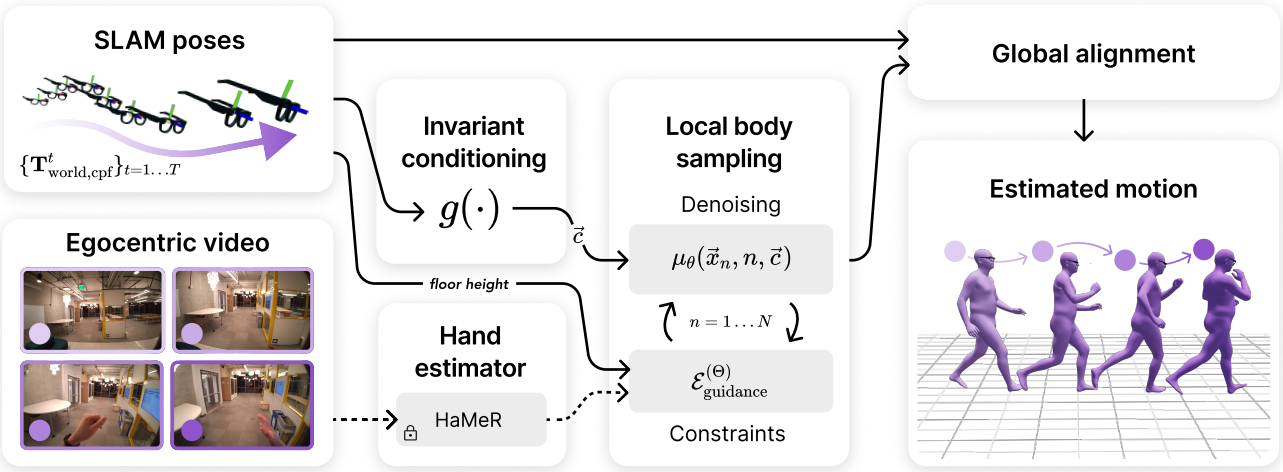


Figure 2. **Overview of components in EgoAllo.** We restrict the diffusion model to local body parameters $\{\Theta^t, \beta^t\}_t, \{\psi^t_j\}_{t,j}$ (Section 3.1.1). An invariant parameterization (Section 3.1.2) of SLAM poses is used to condition a diffusion model. These can be placed into the global coordinate frame via global alignment (Section 3.2.1) to input poses. When available, egocentric video is used for hand detection via HaMeR [63], which can be incorporated into samples via guidance (Section 3.2.2).

in implementation defines a per-sequence canonical coordinate frame to ensure that all input trajectories passed to the model are aligned with the same initial xy position and forward direction. In our work, we will refer to this as *sequence canonicalization*. Finally, EgoPoser [28] proposes a similar scheme that aligns initial positions for both head pose and controller pose inputs. We propose an alternative to these parameterizations that is motivated by the robustness and generalization benefits of invariance, as observed in prior work for designing both representations [9, 51, 78, 94, 101, 108, 113] and neural network architectures [7, 8, 11, 13, 31, 43, 87, 97, 98, 107]. Specifically, we introduce in Section 3.1.2 a parameterization that is invariant to both spatial and temporal shifts.

3. Method

We study the problem of using sensors from an egocentric device to estimate the actions of the wearer in an allocentric coordinate frame. We assume a flat floor and two egocentric inputs: poses from the device's SLAM system and camera images.

Our system uses head pose information to condition a diffusion-based prior over body pose and height, and incorporates visual hand observations during sampling. This allows it to benefit from both 3D human motion capture datasets [54], which are used for the motion prior, and from large-scale image datasets [63], which are used for hand estimates.

3.1. Ego-conditioned motion diffusion

Notation: we use $\mathbf{T}_{\text{A,B}} = (\mathbf{R}_{\text{A,B}}, \mathbf{p}_{\text{A,B}})$ to denote an SE(3) transform to frame A from frame B, composed of rotation ($\mathbf{R}_{\text{A,B}}$) and position ($\mathbf{p}_{\text{A,B}}$) terms. Temporal steps t are superscripted and diffusion noise steps n are subscripted. $\vec{x}^t_0$ thus refers to the t-th timestep of a clean ($n=0$) human motion sequence.

Given an observation window of T timesteps, EgoAllo's motion prior is a diffusion model that aims to capture the distribution of human motions $\vec{x}_0 = \{\vec{x}_0^{\,1}, \ldots, \vec{x}_0^{\,T}\}$ conditioned on head pose encodings $\vec{c} = \{\vec{c}^{\,1}, \ldots, \vec{c}^{\,T}\}$. For each timestep t, we represent human motion in the form of SMPL-H [50, 76] model parameters $\{\mathbf{T}^t_{\text{world,root}}, \Theta^t, \beta\}$: root transforms $\mathbf{T}^t_{\text{world,root}} \in \text{SE}(3)$, where the person's root frame is located at their pelvis, axis-angle local joint rotations $\Theta^t \in \mathbb{R}^{51\times 3}$, and time-invariant shape $\beta \in \mathbb{R}^{16}$.

Dependencies between local joint rotations, body size variation, and global motion make this learning task a challenging one. Our key insight is that this difficulty can be reduced by designing parameterizations with desirable invariance properties. Spatial and temporal invariances allow the model to focus on the essential structure of motion, without being affected by irrelevant shifts in position or time.

3.1.1 Diffusion output representation

As output, we sample joint rotations, body shapes, and binary contact predictions $\vec{x}^t_0 = \{\Theta^t, \beta^t, \psi^t_{j=1\ldots 21}\}$, where body shape β^t is supervised to be equal for all timesteps and ψ^t_j is a per-joint contact indicator. Notably, these parameters are all local—we discuss how outputs can be placed into the allocentric coordinate frame in Section 3.2.1.

We choose this output set for three main reasons. (1) Body shape encodes the wearer's height, which is critical for grounding in the metric-scale geometry of the scene. This is rarely considered by prior work: with the exception of [28], which is focused on tracking with controller input, existing

methods [6, 27, 46] otherwise produce outputs using a fixed "mean" human shape. (2) Contact predictions enable losses for common problems like foot skating, which are discussed in Section 3.2.2. (3) Finally, local bodies are invariant to the global coordinate frame. As we discuss next, the conditioning parameterization for the model can therefore also be invariant to arbitrary transformations along the floor plane.

3.1.2 Invariant conditioning

The goal of our conditioning representation is to map raw SLAM poses (head motion) to a parameterization that is amenable to learning for the diffusion model.

Raw inputs. To capture the head motion at each time step, we assume as input poses of a *central pupil frame* (CPF), which the SLAM systems of devices like Project Aria can provide with millimeter-level accuracy [82]. For time $1...T$, we reparameterize these poses for conditioning using a function g:

$$\mathbf{T}^t_{\text{world,cpf}} = (\mathbf{R}^t_{\text{world,cpf}}, \mathbf{p}^t_{\text{world,cpf}}) \in \text{SE}(3), \tag{1}$$

$$\{\vec{c}^{\,1}, ..., \vec{c}^{\,T}\} = g(\{\mathbf{T}^1_{\text{world,cpf}}, ..., \mathbf{T}^T_{\text{world,cpf}}\}). \tag{2}$$

The CPF frame differs from prior works that condition on a coordinate frame attached to the SMPL human model's "head joint" [6, 27, 28, 46]. The offset between this head joint and the device pose depends on the head shape captured by β^t, and is thus difficult to precompute in our setting.

To encode absolute height, we assume that the world frame's $+z$-axis faces upwards, and that the ground is located at $z=0$. Ground parameters are directly available in the training data [54]; at test time, we can also extract these parameters from sparse SLAM points via RANSAC (Appendix A.1).

Invariance goals. As discussed in Section 2, prior work varies in how the function g is implemented. To understand how choices impact learning, we propose two invariance properties for head motion representations. Each reduces representational redundancy, which eases the learning problem.

Invariance 1 (Spatial) *Global transformations along the floor plane should not affect a person's local motion. Given* $\mathbf{T}_{xy} \in SE(3)$ *restricted to the XY plane,* g *should fulfill* $g(\{\mathbf{T}_{xy}\mathbf{T}^t_{world,cpf}\}_t) = g(\{\mathbf{T}^t_{world,cpf}\}_t)\ \forall\ \mathbf{T}_{xy}$.

Invariance 2 (Temporal) *Head motion representations for a given body motion should be independent of location within a temporal window. This can be expressed as temporal shift equivariance. Let* $\vec{c}^{\,t}$ *be as defined in Equation 2. For any shift* δ *such that* $\{\vec{c}^{\,1}_{shift}, ..., \vec{c}^{\,T}_{shift}\} = g(\{\mathbf{T}^{1+\delta}_{world,cpf}, ..., \mathbf{T}^{T+\delta}_{world,cpf}\})$, g *should satisfy* $\vec{c}^{\,t}_{shift} = \vec{c}^{\,t+\delta}$ *for overlapping timesteps.*

No parameterization used by existing work satisifies both of these properties. The sequence canonicalization approach of EgoEgo [46] achieves spatial invariance (Invariance 1), but inserts a sequence-wide dependency on the first timestep of each window that results in a violation of Invariance 2. The absolute poses and pose deltas used by [6, 27] satisfy Invariance 2, but not Invariance 1. Finally, the relative positions considered by [28] are neither spatially nor temporally invariant.

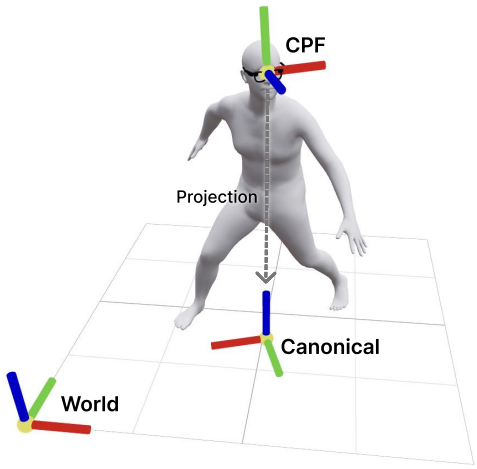


Figure 3. **Locally canonicalized coordinate frames.** We compute our invariant conditioning parameterization (Equation 4) using transformations computed from three coordinate frames. Following [82], the CPF has the z-axis forward. Following HuMoR [71], the world and canonical z-axes point up. Canonical frames are computed by projecting the CPF frame origin to the ground plane, then aligning the canonical y-axis to the CPF forward direction.

Invariant conditioning. We propose a formulation for g that achieves both invariance properties by locally canonicalizing head motion with respect to the floor at each timestep. We build on the relative motion of the CPF frame at each time t, which respects both Invariance 1 and 2:

$$\Delta\mathbf{T}^{t-1,t}_{\text{cpf}} = (\mathbf{T}^{t-1}_{\text{world,cpf}})^{-1}\mathbf{T}^t_{\text{world,cpf}}. \tag{3}$$

Importantly, the translation component of this transformation is in the local frame. This is distinct from world-frame position deltas [6, 27, 28], which still violate Invariance 1.

Relative transforms alone, however, do not encode information relative to the scene or floor: full trajectories can even be flipped upside down without impacting $\Delta\mathbf{T}^{t-1,t}_{\text{cpf}}$. We therefore propose to ground relative motion to the floor plane with a transformation between the CPF frame and a *per-timestep* canonical frame, which is computed by projecting the CPF frame to the floor. This encodes head height and orientation. Our full representation then becomes:

$$\vec{c}^{\,t} = \underbrace{\left\{\Delta\mathbf{T}^{t-1,t}_{\text{cpf}},\quad (\mathbf{T}^t_{\text{world,canonical}})^{-1}\mathbf{T}^t_{\text{world,cpf}}\right\}}_{\text{Invariant implementation of } g(\cdot)}. \tag{4}$$

We visualize an example of a canonical frame in Figure 3. Canonical frames are positioned by projecting the CPF origin to the floor plane; given standard bases $\mathbf{e}_{\{x,y,z\}}$, we compute:

$$\mathbf{p}^t_{\text{world,canonical}} = \begin{bmatrix}\mathbf{e}_x & \mathbf{e}_y & \vec{0}\end{bmatrix}^\top \mathbf{p}^t_{\text{world,cpf}}. \tag{5}$$

For orientation, we align the canonical frame's local z-axis parallel to the world z-axis and its local y-axis toward the "forward" direction $\vec{v}^t$ of the CPF frame. With $\mathbf{R}_z(\cdot):\mathbb{R}\rightarrow \mathrm{SO}(3)$ constructing a z-axis rotation and $\mathbf{e}_{\{x,y,z\}}$ again as standard bases, we compute this as:

$$\vec{v}^t = \mathbf{R}^t_{\text{world,cpf}}\mathbf{e}_z, \tag{6}$$

$$\mathbf{R}^t_{\text{world,canonical}} = \mathbf{R}_z\left(-\arctan2\left(\mathbf{e}_x^\top \vec{v}^{\,t}, \mathbf{e}_y^\top \vec{v}^{\,t}\right)\right)\mathbf{R}^t_{\text{world,cpf}}. \tag{7}$$

This canonical frame definition is an important departure from prior work. While EgoEgo [46] and HuMoR [71] use similar canonical frames, they only compute one per sequence. Instead, we compute Equations 5 and 7 at every timestep. This enables floor plane grounding without sacrificing Invariance 2.

3.2. Estimation via sampling

We use our local body representation and invariant conditioning strategies to train a motion prior in the form a denoising diffusion model [22]. Given diffusion step $n = N \dots 1$, we follow [69] and approximate the denoising process as:

$$p_\theta(\vec{x}_{n-1}|\vec{x}_n,\vec{c}) = \mathcal{N}(\mu_\theta(\vec{x}_n,n,\vec{c}),\sigma_n^2\mathbf{I}), \tag{8}$$

where a transformer [91] μ_θ is trained to predict the posterior mean from noised sample $\vec{x}_n$ and conditioning $\vec{c}$. With noise-dependent weight term w_n, the loss can be written as:

$$\min_\theta \ \mathbb{E}_{\vec{x}_0}\mathbb{E}_{n\sim\mathcal{U}}\left[w_n\|\mu_\theta(\vec{x}_n,n,\vec{c}) - \vec{x}_0\|^2\right]. \tag{9}$$

After training, we estimate human motions by following DDIM [83] for sampling. The final EgoAllo sampling procedure includes several additional components: a global alignment phase, guidance losses for physical constraints and visual hand observations, and a path fusion [3] approach for longer sequence lengths. We describe these below.

3.2.1 Global alignment

To place sampled bodies into the allocentric coordinate system, we compute the absolute pose of the SMPL-H root as:

$$\mathbf{T}^t_{\text{world,root}} = \mathbf{T}^t_{\text{world,cpf}}\mathbf{T}^{(\Theta^t,\beta^t)}_{\text{cpf,root}}, \tag{10}$$

where $\mathbf{T}^{(\Theta^t,\beta^t)}_{\text{cpf,root}}$ computes the transform between the root of the human and their CPF frame for a given set of local pose and shape parameters. Similar processes are applied in [6, 27, 28]. In contrast to directly outputting absolute body transformations from the diffusion model [46], this guarantees exact alignment between estimates and the input SLAM sequences.

3.2.2 Guidance losses

Our diffusion model learns a distribution of human motion conditioned on the central pupil frame motion. At test time, we incorporate constraints from physical priors and visual hand observations via guidance [10, 29, 112]. Similar to [34, 45], we accomplish this by applying losses to the joint rotations $\Theta = \{\Theta^1,\dots,\Theta^T\}$ predicted by $\mu_\theta(\vec{x}_n,n,\vec{c})$. We treat the body shape β^t and contacts $\psi^t_{j=1\dots21}$ as fixed and optimize over body and finger pose to minimize hand observation, skating, and prior costs with a Levenberg-Marquardt optimizer:

$$\mathcal{E}^{(\Theta)}_{\text{guidance}} = \mathcal{E}^{(\Theta)}_{\text{hands}} + \mathcal{E}^{(\Theta)}_{\text{skate}} + \mathcal{E}^{(\Theta)}_{\text{prior}}. \tag{11}$$

We begin by running HaMeR on the egocentric image corresponding to each timestep t. When detected, this produces 3D hand estimates in the form of MANO [76] joint parameters and camera-centric 3D hand keypoints $\hat{\mathbf{p}}^t_{\text{camera,j}}$ for hand joint set $j\in\mathcal{H}$. Optionally, wrist and palm poses can also be estimated using Project Aria's Machine Perception Services [82]. With each subcripted λ indicating a scalar weighting term, we have:

$$\mathcal{E}^{(\Theta)}_{\text{hands}} = \lambda_{\text{hands3D}}\mathcal{E}^{(\Theta)}_{\text{hands3D}} + \lambda_{\text{reproj}}\mathcal{E}^{(\Theta)}_{\text{reproj}}. \tag{12}$$

The 3D objective $\mathcal{E}^{(\Theta)}_{\text{hands3D}}$ minimizes the distance between the detected hand parameters and the corresponding SMPL-H hand parameters, in terms of wrist pose and local joint rotations. With Π_K as projection with camera intrinsics K, $\mathbf{p}^{(\Theta^t)}_{world,j}\in\mathbb{R}^3$ as the world position for joint j at time t, and $\mathbf{T}_{\text{camera,cpf}}$ from the device calibration, the reprojection loss is:

$$\mathcal{E}^{(\Theta)}_{\text{reproj}} = \sum_{t,j\in\mathcal{H}} ||\Pi_K(\mathbf{p}^{(\Theta^t)}_{\text{camera},j}) - \Pi_K(\hat{\mathbf{p}}^t_{\text{camera},j})||_2^2, \tag{13}$$

$$\mathbf{p}^{(\Theta^t)}_{\text{camera},j} = \mathbf{T}_{\text{camera,cpf}}(\mathbf{T}^t_{\text{world,cpf}})^{-1}\mathbf{p}^{(\Theta^t)}_{\text{world},j}. \tag{14}$$

To reduce foot skating, we use contact predictions to apply a skating loss [71, 99] for each time t and joint j:

$$\mathcal{E}^{(\Theta)}_{\text{skate}} = \sum_{t,j}\lambda_{\text{skate}}||\frac{1}{2}(\psi^t_j + \psi^{t-1}_j)(\mathbf{p}^t_{world,j} - \mathbf{p}^{t-1}_{world,j})||_2^2. \tag{15}$$

Finally, we minimize a prior term $\mathcal{E}^{(\Theta)}_{\text{prior}}$. This term penalizes deviations between each constrained rotation matrix Θ^t and the original output rotations $\hat{\Theta}^t$ from our denoiser $\mu_\theta(\vec{x}_n,n,\vec{c})$, in terms of both joint rotation and rotational velocity.

3.2.3 Sequence length extrapolation

For longer sequences at test time, we draw on existing methods in compositional generation for both image [3, 111] and human motion [4, 79] diffusion models. We train our motion prior using subsequences of up to length 128; when input observations exceed this length at test time, we split into windows with a 32-timestep overlap between neighbors. We then run our model $\mu_\theta(\vec{x}_n,\vec{c},n)$ on windows in parallel. Diffusion paths for overlapping regions are fused following MultiDiffusion [3] after each denoising step.

Conditioning	Seqlen	Invariance 1 / 2	MPJPE ↓	% Diff	PA-MPJPE ↓	% Diff	GND ↑
`EgoAllo` (Eq. 4)	32	✔ / ✔	129.8±1.1	—	109.8±1.1	—	0.98±0.00
`Absolute+Local Relative`	32	P / ✔	133.0±1.1	2.4%	113.6±1.2	3.4%	0.95±0.00
`Absolute+Global Deltas` [6, 27]	32	✗ / ✔	136.2±1.1	4.9%	118.3±1.2	7.7%	0.93±0.01
`Sequence Canonicalization` [46]	32	✔ / ✗	153.1±1.5	17.9%	128.7±1.5	17.1%	0.76±0.01
`Absolute`	32	✗ / ✔	159.9±1.2	23.2%	141.0±1.3	28.4%	0.89±0.01
`EgoAllo` (Eq. 4)	128	✔ / ✔	119.7±1.3	—	101.1±1.3	—	1.00±0.00
`Absolute+Local Relative`	128	P / ✔	124.5±1.3	4.0%	104.9±1.4	3.8%	1.00±0.00
`Absolute+Global Deltas` [6, 27]	128	✗ / ✔	127.4±1.3	6.4%	109.8±1.4	8.6%	0.99±0.00
`Sequence Canonicalization` [46]	128	✔ / ✗	134.0±1.8	11.9%	112.1±1.6	10.9%	0.88±0.02
`Absolute`	128	✗ / ✔	148.3±1.5	23.9%	131.2±1.6	29.8%	0.96±0.01

Table 1. **Motion prior conditioning comparison.** We train and evaluate otherwise identical models using four conditioning parameterizations on AMASS [54] test set sequences, using sequences of length 32 and 128. Parameterizations vary in their spatial (1) and temporal (2) invariance properties, which we loosely classify as following completely (✔), partially (P), or not at all (✗). The conditioning parameterization used by EgoAllo reduces errors by almost 18% compared to the sequence canonicalization approach used by the most relevant related work [46].

4. Experiments

We conduct a series of experiments to evaluate EgoAllo's conditioning parameterization, body estimation accuracy, and hand estimation performance.

Training. To train EgoAllo models used in our experiments, we need sequences containing human body and hand pose parameters, body shapes, and device SLAM poses $\mathbf{T}^t_{\text{world,cpf}}$. Similar to prior work [6, 27, 46], we train EgoAllo using AMASS [54] with synthesized device poses. We annotate train split sequences by anchoring a central pupil frame between vertices corresponding to the left and right pupils in the blend skinned mesh, and at train time sample sequences between length 32 and 128.

Evaluation. We evaluate with four datasets. We use AMASS [54], RICH [25], and Aria Digital Twins (ADT) [58] for body estimation evaluation, and EgoExo4D [16] for hand estimation evaluation. AMASS and RICH do not include egocentric data; we annotate these with synthetic device poses using the same procedure we use for training. ADT and EgoExo4D both include egocentric images and SLAM poses captured using Project Aria glasses [82], which we use directly.

Metrics. To quantify performance, we report four metrics: **(1) MPJPE** is a world-frame mean per-joint position error (millimeters). **(2) PA-MPJPE** is the Procrustes-aligned mean per-joint position error in millimeters, where joint positions are aligned on a per-timestep basis before error are computed. **(3) GND** is a grounding metric, designed in response to a phenomena where ego-sensed humans "float" above the ground. Given a human body trajectory, this metric contains a simple binary indicator of whether the feet of the human ever touch the ground plane. **(4)** $\mathbf{T}_{\text{head}}$ is the average SMPL head joint position error in millimeters.

4.1. Body estimation

In our first set of experiments, we evaluate body estimation from only device SLAM poses, without considering images or hands. This setting allows us to isolate the advantages of our body motion prior, while facilitating direct comparison against methods that do not consider hands.

4.1.1 Invariant conditioning evaluation

We begin by evaluating the importance of the spatial and temporal invariance criteria discussed in Section 3.1.2. We do this by comparing five implementations of the conditioning g: **(1)** **`EgoAllo`** is the final invariant representation that we propose in Equation 4. **(2)** **`Absolute+Local Relative`** appends absolute poses with the relative pose deltas written in Equation 3. **(3)** **`Absolute+Global Deltas`** appends absolute poses with relative orientation and the world-frame position deltas used by [6, 27]. **(4)** **`Sequence Canonicalization`** uses the alignment approach implemented by [46], which violates temporal invariance. **(5)** **`Absolute`** naively conditions on absolute poses, which violate spatial invariance.

We train conditional diffusion models with otherwise identical architecture using each parameterization, and then evaluate on the AMASS [54] test set. Metrics and percent differences compared to EgoAllo are reported in Table 1.

Overall, we find that the choice of conditioning parameterization makes a dramatic impact on estimation accuracy. We observe accuracy improve consistently as invariance properties are incorporated into the representation. Compared to `EgoAllo`, `Absolute` conditioning increases MPJPE by over 23% for both shorter (length 32) and longer (length 128) sequences. Compared to `EgoAllo`, `SeqCanonical` conditioning increases MPJPE by nearly 18% for length 32 sequences and 12% for length 128 sequences.

AMASS [54]

Method	Seq	MPJPE ↓	PA-MPJPE ↓	GND ↑	$\mathbf{T}_{\text{head}}$ ↓
EgoAllo	32	129.8±1.1	109.8±1.1	0.98±0.00	6.4±0.1
NoShape	32	138.1±1.1	118.8±1.1	0.94±0.01	44.7±0.4
EgoEgo	32	184.0±1.5	158.6±1.6	0.81±0.01	45.2±1.0
VAE+Opt	32	199.5±1.3	191.4±1.4	0.49±0.01	78.0±1.5
EgoAllo	128	119.7±1.3	101.1±1.3	1.0±0.00	6.2±0.1
NoShape	128	128.1±1.3	110.3±1.4	0.98±0.01	44.6±0.7
EgoEgo	128	167.4±2.1	145.8±2.0	0.92±0.01	54.9±1.9
VAE+Opt	128	205.3±2.6	192.3±2.8	0.75±0.02	67.8±3.1

RICH [25]

Method	Seq	MPJPE ↓	PA-MPJPE ↓	GND ↑	$\mathbf{T}_{\text{head}}$ ↓
EgoAllo	32	193.7±3.4	174.8±3.6	0.95±0.01	8.8±0.2
NoShape	32	200.9±3.3	183.3±3.6	0.73±0.02	44.9±0.9
EgoEgo	32	215.4±3.9	192.9±4.0	0.73±0.02	56.2±2.9
VAE+Opt	32	352.0±6.7	354.8±6.5	0.59±0.02	319.3±11.6
EgoAllo	128	176.2±5.6	160.1±5.9	0.96±0.02	8.9±0.3
NoShape	128	185.7±5.5	169.9±5.8	0.82±0.03	45.8±1.6
EgoEgo	128	207.8±6.9	187.8±6.8	0.88±0.03	66.5±5.4
VAE+Opt	128	319.8±10.1	323.8±10.5	0.75±0.04	274.4±17.6

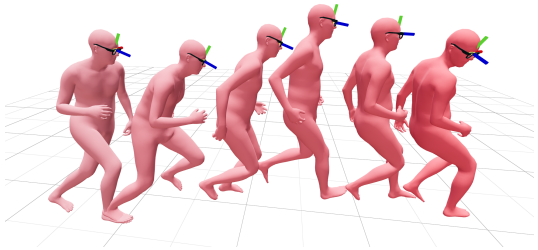
Aria Digital Twins [58]

Method	Seq	MPJPE ↓	PA-MPJPE ↓	GND ↑	$\mathbf{T}_{\text{head}}$ ↓
EgoAllo	32	173.5±1.1	146.1±1.1	0.88±0.01	-
NoShape	32	178.5±1.1	153.0±1.1	0.89±0.01	-
EgoEgo	32	212.5±1.4	181.3±1.6	0.64±0.01	-
VAE+Opt	32	284.9±1.6	283.9±1.9	0.63±0.01	-
EgoAllo	128	155.1±1.6	129.3±1.6	0.94±0.01	-
NoShape	128	163.7±1.6	140.0±1.6	0.96±0.01	-
EgoEgo	128	182.6±2.3	153.9±2.6	0.73±0.02	-
VAE+Opt	128	290.8±3.8	282.5±4.4	0.7±0.02	-

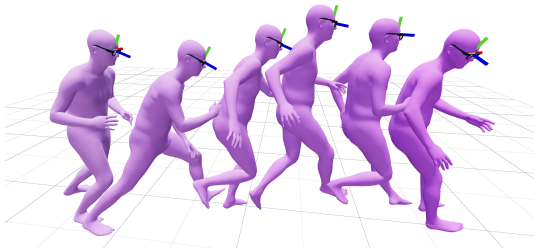
Table 2. **Body estimation performance, compared against a baseline without shape prediction, EgoEgo [46], and VAE+Opt [71, 99].** We exclude the $\mathbf{T}_{\text{head}}$ metric for ADT because the Biomech57 head joints used by ADT are not directly comparable to the SMPL-H head joints used by our model.

4.1.2 Comparisons against baselines

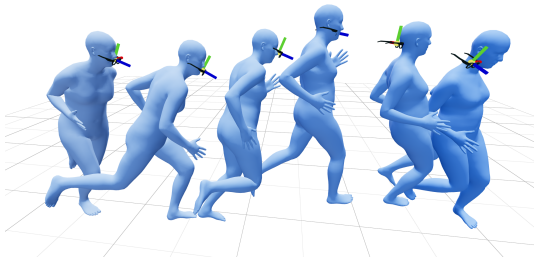
To further study EgoAllo's body estimation quality, we compare against three baselines. **(1) NoShape.** First, NoShape refers to a variation of EgoAllo that turns off shape estimation, and thus cannot estimate the wearer's height. **(2) EgoEgo.** We also compare against the human motion diffusion model from EgoEgo [46]. This is similar to EgoAllo, but considers only the SMPL "mean" body shape and uses canonicalized coordinates for conditioning and as model output. **(3) VAE+Opt.** Finally, we compare against an approach based on the SLAHMR [99] framework for human motion estimation from exocentric video. A key advantage of SLAHMR is that it uses an unconditional motion prior [71] in an optimization framework. It can therefore be adapted to new settings without re-training—we keep the same body pose and shape variables as the original pipeline, but replace the exocentric keypoint [68] cost with an egocentric CPF pose alignment cost.



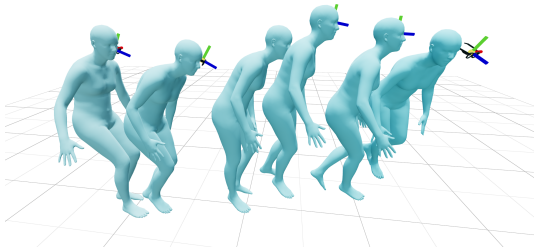
(a) Ground-truth



(b) EgoAllo



(c) EgoEgo [46]



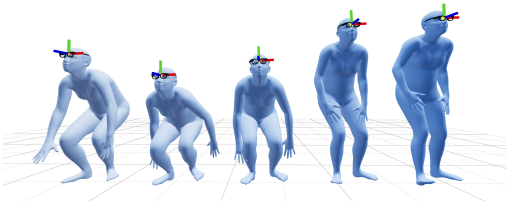
(d) VAE+Opt

Figure 4. **Egocentric human motion estimation for a running sequence.** We show the ground-truth, an output from EgoAllo, and outputs from two baselines. The glasses CAD model is placed at the conditioning transformation $\mathbf{T}_{\text{world,cpf}}$.

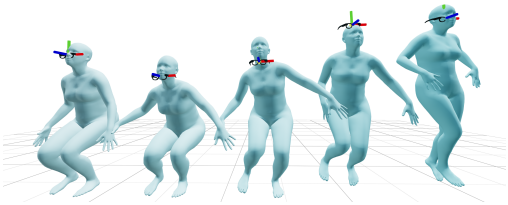
Due to differences in problem formulation, many existing methods for egocentric human motion estimation are difficult to directly compare. This is particularly true when they have different inputs, such as fisheye cameras [88, 92, 93], wrist-mounted sensors [44], or handheld controller poses [6, 27, 28]. Additionally, prior works like EgoEgo [46] do not incorporate vision inputs for hand estimation. For fairness, we restrict all

(a) Ground-truth

(b) EgoAllo



(c) EgoEgo [46]



(d) VAE+Opt

Figure 5. **Egocentric human motion estimation for a squatting sequence.** The contents of this figure mirror Figure 4, but use spatial shifts to visualize different timesteps within a sequence.

methods in this section to only CPF or head pose as input.

EgoAllo improves body motion estimates. We report metrics in Table 2 and visualize example outputs in Figures 4 and 5. We find that EgoAllo enables significant estimation improvements across all datasets, including 20∼30% accuracy improvements over EgoEgo for both shorter and longer evaluation sequences. We found shape estimation critical for producing metric-scale, grounded estimates of human body motion, with the head aligned to input SLAM poses and the feet planted on the observed ground plane. This is evident in qualitative results, improved grounding metrics, and in the 6∼7% MPJPE gap between EgoAllo and the NoShape ablation.

VAE optimization converges poorly. Optimization-based estimation approaches have been effective for settings with keypoint costs [71, 99], but we found convergence difficult in our less constrained setting. In Table 2, we observe poor generalization: VAE+Opt performs competitively on the AMASS test set, but performance deteriorates dramatically when evaluating on RICH or Aria Digital Twins. VAE+Opt outputs in Figure 4 also look overly smoothed, without the same expressiveness as the conditional predictions of EgoAllo or EgoEgo [46]. This highlights the advantage of using a conditional diffusion model problem for this estimation problem.

Shape estimation evaluation. To better understand the shape estimation characteristics of EgoAllo, we compare against against the "mean" shape used by EgoEgo and the NoShape ablation. On the AMASS test set, we find: EgoAllo slightly improves overall shape (19mm → 18mm mean vertex-to-vertex error) and produces much better height (52mm → 32mm mean height error), but is not able to generalize in terms of body weight (5kg → 8kg mean weight error). The body shape is inferred from the wearer's head pose, which intuitively provides strong height constraints but is less correlated with weight. Accurate height is key for proper scene placement, as reflected by both the MPJPE and GND metrics.

4.2. Hand estimation

To evaluate evaluate hands estimated by EgoAllo, we run HaMeR on the segment of the EgoExo4D [16] validation set that is labeled with 3D hand pose keypoints. We quantitatively compare four hand estimation methods in Table 4.2. In **(1) HaMeR** [63], we use HaMeR out-of-the-box on undistorted egocentric RGB images. We do not assume bounding boxes as input; instead, we follow the HaMeR demo code and compute crops using ViTPose [96]. When multiple hands are detected for a single side, we naively take the first one. **(2) EgoAllo-Mono** refers our full system, which uses the same monocular HaMeR hand estimates for guidance. In **(3) EgoAllo-Wrist3D**, we augment our system with wrist pose estimates from Project Aria's Machine Perception Services [82]—unlike HaMeR, which assumes monocular input, this uses a pair of SLAM cameras that are unique to Project Aria. Finally, **(4) EgoAllo-NoReproj** removes the reprojection term (Equation 14) from EgoAllo-Mono. Instead, hand guidance is done directly using the 3D wrist poses predicted by HaMeR. For fairness across settings, we compute metrics only on timesteps where HaMeR estimates are available.

Quantitative results are provided in Table 3. While HaMeR's local poses (PA-MPJPE) are slightly better, EgoAllo's hand-body estimation significantly improves how well hands are estimated in the world coordinate system. Compared to HaMeR, EgoAllo drops MPJPE from 237.90mm → 131.45mm. Incorporating more accurate wrist pose estimates (EgoAllo-Wrist3D) offers a practical solution for further improvements: 131.45mm → 60.08mm. Reprojection-based guidance is also more robust: EgoAllo-NoReproj outputs are the worst in both MPJPE and PA-MPJPE.

Qualitatively, we observed that high hand estimation errors in naive monocular estimation with HaMeR are explained by a combination of detection failures and monocular ambiguities. Even when detections succeed, the scale and distance of

Figure 6. **Body estimation improves hand estimation.** We show raw outputs from HaMeR [63] in blue and hand-body estimations from EgoAllo in purple. *Top:* improved scene interaction during touchscreen operation with EgoAllo-Mono. We know a priori that the fingers are contacting the screen in this sequence. *Bottom:* qualitative examples from EgoExo [16] evaluation, showing the differences between monocular hands and EgoAllo-Wrist3D estimates.

Method	MPJPE ↓	PA-MPJPE ↓
HaMeR	237.90±1.89	13.04±1.89
EgoAllo-Mono	131.45±0.39	14.71±0.39
EgoAllo-Wrist3D	60.08±0.26	14.38±0.26
EgoAllo-NoReproj	143.20±0.42	14.75±0.42

Table 3. **Hand estimation errors in millimeters.** EgoAllo's hand-body estimation can constrain and resolve ambiguities in noisy outputs from HaMeR, which we observe can reduce MPJPE for hands by over 40%.

monocular HaMeR estimates are often incorrect or flicker in between frames. Incorporating these hands via guidance with our diffusion motion prior encourages final outputs that obey the kinematic and smoothness constraints imposed by plausible body motion—we provide examples of HaMeR estimates rendered jointly with EgoAllo outputs in Figure 6, where we note realistic elbow and shoulder poses.

5. Dicussion

Limitations and future work. While the core contributions of EgoAllo are general, the current implementation of our system has a few limitations that we hope to explore in future work. First, diffusion model guidance is a test-time optimization process that depends on hyperparameters and incurs a runtime cost. In the future, it may be possible to bootstrap using outputs from our model to train a feedforward model that avoids this step. Success for hand guidance also still depends on reasonable monocular hand estimates. Estimation can therefore fail as a result of errors like left/right flipping or spurious detections, which we found causes high errors in our Table 3 hand estimation metrics. Finally, we also train only on AMASS [54], which includes floor planes but no detailed scene geometry. As a result, our method will fail if we cannot detect an approximate floor plane; this is most likely in settings like hills or staircases. In the future, we hope to extend our insights to data with more detailed scene information, which concurrent work has highlighted the usefulness of in informing human body estimation [20].

Conclusion. We presented EgoAllo, a system for estimating human motion using sensors from head-mounted devices. Our method takes advantage of motion capture data [54] and an off-the-shelf visual hand estimator [63] to jointly estimate human body pose, height, and hand parameters. EgoAllo highlights the importance of spatial and temporal invariance in conditioning for this problem, while demonstrating how estimated bodies can be used to improve hand estimation.

Acknowledgments. This project was funded in part by NSF:CNS-2235013 and IARPA DOI/IBC No. 140D0423C0035. YM acknowledges support from the joint Simons Foundation-NSF DMS grant #2031899, the ONR grant N00014-22-1-2102, the NSF grant #2402951, and partial support from TBSI, InnoHK, and the University of Hong Kong. JM was supported by ONR MURI N00014-21-1-2801. BY is supported by the National Science Foundation Graduate Research Fellowship Program under Grant DGE 2146752. The authors would also like to thank Hongsuk Choi, Michael Taylor, Tyler Bonnen, Songwei Ge, Chung Min Kim, and Justin Kerr for insightful technical discussion and suggestions, as well as Jiaman Li for helpful answers to questions about EgoEgo.

References

[1] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–20, 2023. 2

[2] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. ACM New York, NY, USA, 2005. 2

[3] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023. 5, 14

[4] German Barquero, Sergio Escalera, and Cristina Palmero. Seamless human motion composition with blended positional encodings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 457–469, 2024. 5

[5] Marcus A Brubaker, David J Fleet, and Aaron Hertzmann. Physics-based person tracking using the anthropomorphic walker. *International journal of computer vision*, 87(1-2): 140–155, 2010. 2

[6] Angela Castillo, Maria Escobar, Guillaume Jeanneret, Albert Pumarola, Pablo Arbeláez, Ali Thabet, and Artsiom Sanakoyeu. Bodiffusion: Diffusing sparse observations for full-body human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4221–4231, 2023. 2, 4, 5, 6, 7

[7] R Qi Charles, Hao Su, Mo Kaichun, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 77–85. IEEE, 2017. 3

[8] Haiwei Chen, Shichen Liu, Weikai Chen, Hao Li, and Randall Hill. Equivariant point network for 3d point cloud analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14514–14523, 2021. 3

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[10] Hai Ci, Mingdong Wu, Wentao Zhu, Xiaoxuan Ma, Hao Dong, Fangwei Zhong, and Yizhou Wang. Gfpose: Learning 3d human pose prior with gradient fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4800–4810, 2023. 2, 5

[11] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016. 3

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[13] Haiwen Feng, Peter Kulits, Shichen Liu, Michael J Black, and Victoria Fernandez Abrevaya. Generalizing neural human fitting to unseen poses with articulated se (3) equivariance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7977–7988, 2023. 3

[14] 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*, pages 225–239. Wiley Online Library, 2020. 2

[15] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa*, and Jitendra Malik*. Humans in 4D: Reconstructing and tracking humans with transformers. In *Int. Conf. Comput. Vis.*, 2023. 2

[16] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. *arXiv preprint arXiv:2311.18259*, 2023. 6, 8, 9, 14

[17] Qiao Gu, Zhaoyang Lv, Duncan Frost, Simon Green, Julian Straub, and Chris Sweeney. Egolifter: Open-world 3d segmentation for egocentric perception. *arXiv preprint arXiv:2403.18118*, 2024. 1

[18] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *Int. Conf. Comput. Vis.*, pages 1381–1388. IEEE, 2009. 2

[19] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10884–10894, 2019. 2

[20] Vladimir Guzov, Yifeng Jiang, Fangzhou Hong, Gerard Pons-Moll, Richard Newcombe, C Karen Liu, Yuting Ye, and Lingni Ma. Hmd^2: Environment-aware motion generation from single egocentric head-mounted device. *arXiv preprint arXiv:2409.13426*, 2024. 2, 9

[21] Chengan He, Jun Saito, James Zachary, Holly Rushmeier, and Yi Zhou. Nemf: Neural motion fields for kinematic animation. In *NeurIPS*, 2022. 2

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 5

[23] Fangzhou Hong, Vladimir Guzov, Hyo Jin Kim, Yuting Ye, Richard Newcombe, Ziwei Liu, and Lingni Ma. Egolm: Multi-modal language model of egocentric motions. *arXiv preprint arXiv:2409.18127*, 2024. 2

[24] Nicholas Howe, Michael Leventon, and William Freeman. Bayesian reconstruction of 3d human motion from single-camera video. *Advances in neural information processing systems*, 12, 1999. 2

[25] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13274–13285, 2022. 6, 7, 14

[26] Hao Jiang and Vamsi Krishna Ithapu. Egocentric pose estimation from human vision span. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10986–10994. IEEE, 2021. 2

[27] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *European conference on computer vision*, pages 443–460. Springer, 2022. 2, 4, 5, 6, 7

[28] Jiaxi Jiang, Paul Streli, Manuel Meier, Andreas Fender, and Christian Holz. Egoposer: Robust real-time ego-body pose estimation in large scenes. *arXiv preprint arXiv:2308.06493*, 2023. 2, 3, 4, 5, 7

[29] Zhongyu Jiang, Zhuoran Zhou, Lei Li, Wenhao Chai, Cheng-Yen Yang, and Jenq-Neng Hwang. Back to optimization: Diffusion-based zero-shot 3d human pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6142–6152, 2024. 2, 5

[30] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *2021 International Conference on 3D Vision (3DV)*, pages 42–52. IEEE, 2021. 2

[31] Angjoo Kanazawa, Abhishek Sharma, and David Jacobs. Locally scale-invariant convolutional neural networks. *arXiv preprint arXiv:1412.5104*, 2014. 3

[32] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7122–7131, 2018. 2

[33] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5614–5623, 2019. 2

[34] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2151–2162, 2023. 2, 5

[35] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8255–8263, 2023. 2

[36] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[37] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5253–5263, 2020. 2

[38] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11127–11137, 2021. 2

[39] Muhammed Kocabas, Ye Yuan, Pavlo Molchanov, Yunrong Guo, Michael J. Black, Otmar Hilliges, Jan Kautz, and Umar Iqbal. Pace: Human and motion estimation from in-the-wild videos. In *3DV*, 2024. 2

[40] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Int. Conf. Comput. Vis.*, pages 2252–2261, 2019. 2

[41] Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. Nifty: Neural object interaction fields for guided human motion synthesis. *arXiv preprint arXiv:2307.07511*, 2023. 2

[42] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6050–6059, 2017. 2

[43] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995. 3

[44] Jiye Lee and Hanbyul Joo. Mocap everyone everywhere: Lightweight motion capture with smartwatches and a head-mounted camera. *arXiv preprint arXiv:2401.00847*, 2024. 2, 7

[45] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. *arXiv preprint arXiv:2312.03913*, 2023. 2, 5

[46] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 17142–17151, 2023. 2, 4, 5, 6, 7, 8, 14

[47] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023. 2

[48] Zongmian Li, Jiri Sedlar, Justin Carpentier, Ivan Laptev, Nicolas Mansard, and Josef Sivic. Estimating 3d motion and forces of person-object interactions from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8640–8649, 2019. 2

[49] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. Character controllers using motion vaes. *ACM Transactions on Graphics (TOG)*, 39(4):40–1, 2020. 2

[50] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866, 2023. 2, 3

[51] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 3

[52] Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. *Advances in Neural Information Processing Systems*, 34:25019–25032, 2021. 2

[53] Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. *arXiv preprint arXiv:2406.09905*, 2024. 2

[54] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5442–5451, 2019. 3, 4, 6, 7, 9, 14

[55] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Int. Conf. Comput. Vis.*, pages 2640–2649, 2017. 2

[56] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9890–9900, 2020. 2

[57] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, pages 484–494. IEEE, 2018. 2

[58] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20133–20143, 2023. 6, 7, 14

[59] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 459–468, 2018. 2

[60] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10975–10985, 2019. 2

[61] Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Human mesh recovery from multiple shots. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1485–1495, 2022. 2

[62] Georgios Pavlakos, Ethan Weber, Matthew Tancik, and Angjoo Kanazawa. The one where they reconstructed 3d humans and

environments in tv shows. In *Eur. Conf. Comput. Vis.*, pages 732–749. Springer, 2022. 2

[63] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *arxiv*, 2023. 3, 8, 9

[64] Xue Bin Peng, Angjoo Kanazawa, Jitendra Malik, Pieter Abbeel, and Sergey Levine. Sfv: Reinforcement learning of physical skills from videos. *ACM Transactions On Graphics (TOG)*, 37(6):1–14, 2018. 2

[65] Chiara Plizzari, Shubham Goel, Toby Perrett, Jacob Chalk, Angjoo Kanazawa, and Dima Damen. Spatial cognition from egocentric video: Out of sight, not out of mind. *arXiv preprint arXiv:2404.05072*, 2024. 1

[66] Ryan Po, Wang Yifan, Vladislav Golyanik, Kfir Aberman, Jonathan T Barron, Amit H Bermano, Eric Ryan Chan, Tali Dekel, Aleksander Holynski, Angjoo Kanazawa, et al. State of the art on diffusion models for visual computing. *arXiv preprint arXiv:2310.07204*, 2023. 2

[67] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[68] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people by predicting 3d appearance, location and pose. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2740–2749, 2022. 2, 7

[69] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 5

[70] Davis Rempe, Leonidas J Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 71–87. Springer, 2020. 2

[71] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11488–11499, 2021. 2, 4, 5, 7, 8

[72] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014. 2

[73] Helge Rhodin, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt. Egocap: egocentric marker-less motion capture with two fisheye cameras. *ACM Transactions on Graphics (TOG)*, 35(6):1–11, 2016. 2

[74] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3433–3441, 2017. 2

[75] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[76] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 2, 3, 5

[77] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[78] Silvio Savarese and Li Fei-Fei. 3d generic object categorization, localization and pose estimation. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE, 2007. 3

[79] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 5

[80] Takaaki Shiratori, Hyun Soo Park, Leonid Sigal, Yaser Sheikh, and Jessica K Hodgins. Motion capture from body-mounted cameras. In *ACM SIGGRAPH 2011 papers*, pages 1–10. ACM New York, NY, USA, 2011. 2

[81] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2

[82] Kiran Somasundaram, Jing Dong, Huixuan Tang, Julian Straub, Mingfei Yan, Michael Goesele, Jakob Julian Engel, Renzo De Nardi, and Richard Newcombe. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 2, 4, 5, 6, 8

[83] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5

[84] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[85] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 14

[86] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021. 2

[87] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018. 3

[88] Denis Tome, Patrick Peluse, Lourdes Agapito, and Hernan Badino. xr-egopose: Egocentric 3d human pose from an hmd camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7728–7738, 2019. 2, 7

[89] Denis Tome, Thiemo Alldieck, Patrick Peluse, Gerard Pons-Moll, Lourdes Agapito, Hernan Badino, and Fernando De la Torre. Selfpose: 3d egocentric pose estimation from a headset mounted camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[90] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 448–458, 2023. 2

[91] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5, 14

[92] Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, and Christian Theobalt. Estimating egocentric 3d human pose in global space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11500–11509, 2021. 2, 7

[93] Jian Wang, Zhe Cao, Diogo Luvizon, Lingjie Liu, Kripasindhu Sarkar, Danhang Tang, Thabo Beeler, and Christian Theobalt. Egocentric whole-body motion capture with fisheyevit and diffusion-based motion refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 777–787, 2024. 2, 7

[94] Laurenz Wiskott and Terrence J Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770, 2002. 3

[95] Weipeng Xu, Avishek Chatterjee, Michael Zollhoefer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. Mo 2 cap 2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. *IEEE transactions on visualization and computer graphics*, 25(5):2093–2101, 2019. 2

[96] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose+: Vision transformer foundation model for generic body pose estimation. *arXiv preprint arXiv:2212.04246*, 2022. 8

[97] Jingyun Yang, Congyue Deng, Jimmy Wu, Rika Antonova, Leonidas Guibas, and Jeannette Bohg. Equivact: Sim(3)-equivariant visuomotor policies beyond rigid object manipulation, 2023. 3

[98] Jingyun Yang, Zi-ang Cao, Congyue Deng, Rika Antonova, Shuran Song, and Jeannette Bohg. Equibot: Sim (3)-equivariant diffusion policy for generalizable and data efficient learning. *arXiv preprint arXiv:2407.01479*, 2024. 3

[99] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 21222–21232, 2023. 2, 5, 7, 8

[100] Brent Yi, Michelle Lee, Alina Kloss, Roberto Martín-Martín, and Jeannette Bohg. Differentiable factor graph optimization for learning smoothers. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021. 14

[101] Brent Yi, Weijia Zeng, Sam Buchanan, and Yi Ma. Canonical factors for hybrid neural fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3414–3426, 2023. 3

[102] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 2

[103] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Vladislav Golyanik, Shaohua Pan, Christian Theobalt, and Feng Xu. Egolocate: Real-time motion capture, localization, and mapping with sparse body-mounted sensors. *arXiv preprint arXiv:2305.01599*, 2023. 2

[104] Ye Yuan and Kris Kitani. 3d ego-pose estimation via imitation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 735–750, 2018. 2

[105] Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time pd control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10082–10092, 2019. 2

[106] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2

[107] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017. 3

[108] Fangneng Zhan, Lingjie Liu, Adam Kortylewski, and Christian Theobalt. General neural gauge fields. *arXiv preprint arXiv:2305.03462*, 2023. 3

[109] Daiwei Zhang, Gengyan Li, Jiajie Li, Mickaël Bressieux, Otmar Hilliges, Marc Pollefeys, Luc Van Gool, and Xi Wang. Egogaussian: Dynamic scene understanding from egocentric video with 3d gaussian splatting. *arXiv preprint arXiv:2406.19811*, 2024. 1

[110] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 2

[111] Qinsheng Zhang, Jiaming Song, Xun Huang, Yongxin Chen, and Ming-Yu Liu. Diffcollage: Parallel generation of large content with diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10188–10198. IEEE, 2023. 5

[112] Siwei Zhang, Qianli Ma, Yan Zhang, Sadegh Aliakbarian, Darren Cosker, and Siyu Tang. Probabilistic human mesh recovery in 3d scenes from egocentric views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7989–8000, 2023. 2, 5

[113] Zhengdong Zhang, Arvind Ganesh, Xiao Liang, and Yi Ma. Tilt: Transform invariant low-rank textures. *International journal of computer vision*, 99:1–24, 2012. 3

A. Appendix

A.1. Floor height estimation

A key requirement of our algorithm is SLAM poses that can be situated relative to the floor. While floor heights are provided in our training data, they are not directly available on real-world data. We found that a simple RANSAC-based algorithm works well on real-world data from Project Aria [58]. We filter SLAM points by confidence, then use RANSAC to find a z-value with that best fits a plane. Example floor plane outputs using scenes from the EgoExo4D [16] dataset are shown in Figure 7.

A.2. Sequence length evaluation

At test time, EgoAllo follows MultiDiffusion [3] for extrapolating to sequences that are longer than the 128 that we use for training. To evaluate that this works, we filter out test sequences shorter than 256 frames and then evaluate both EgoAllo and EgoEgo [46] with subsequences of size 32, 128, and 256. We report metrics on these sequences in Table 4. Both EgoAllo and EgoEgo include windowing strategies for handling longer sequences; unlike prior work, however, we find that the EgoAllo system improves even after test set sequence lengths surpass the training set sequence length.

Seqlen	32	128	256
EgoAllo	149.3	130.3	127.9
EgoEgo	187.7	173.8	184.3

Table 4. **How does MPJPE change with sequence length?**

A.2.1 Biomech57 evaluation details

The majority of our evaluation data (AMASS [54] and RICH [25]) is provided directly using SMPL conventions. Because our model outputs SMPL-H parameters, this makes computation of joint error metrics straightforward.

The one exception is the Aria Digital Twins dataset [58], which we use for quantitative body metrics. Each device wearer in the Aria Digital Twins dataset is recorded via an Optitrack motion capture system, which records 57 joint locations (30 hand joints, 27 body joints) following the Biomech57 joint template. To evaluate our method on ADT, we match and compare the common major joints between the two templates. We manually corresponded each of the 57 joints between Biomech57 and the standard SMPL-H joint conventions. While the majority of these have 1:1 correspondences—feet, knees, hips, shoulders, elbows, wrist, and finger joints, for example, are consistently defined—we mask out others like the head and collar bone joints that are misaligned.

A.3. Implementation details

We use a transformer [91] architecture with rotary positional embeddings [85] for EgoAllo's denoising model $\mu_\theta(\vec{x}_n, \vec{c}, n)$.

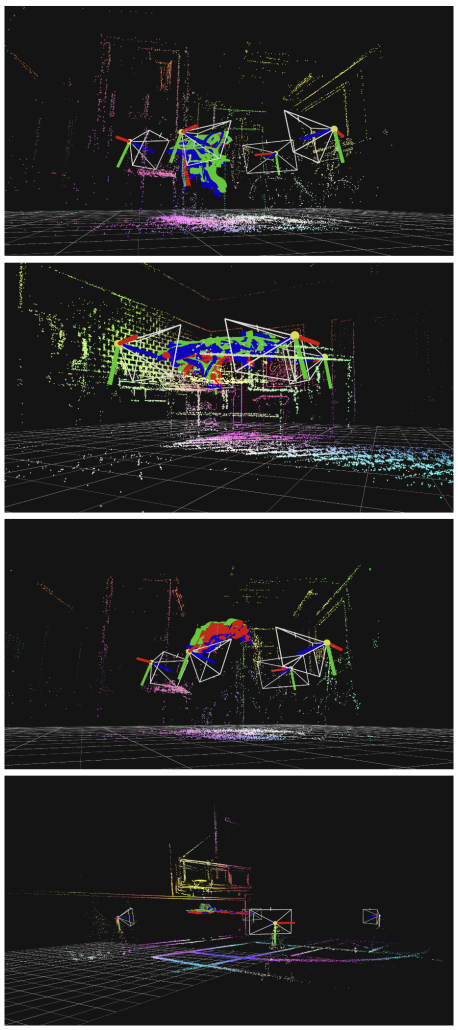


Figure 7. **Floor height examples.** Point cloud-derived floor height examples on the EgoExo4D dataset.

Latent representations $\vec{z}_c$ for conditioning sequences $\vec{c}$ are computed using six transformer blocks, each containing a self-attention layer followed by a 2-layer MLP. Skip connections are used throughout. The denoised output is produced by using six additional transformer blocks that take $\vec{x}_n$ as input, while conditioning on $\vec{z}_c$ via cross-attention. All hidden dimensions are set to 512. For guidance, we use infrastructure for sparse nonlinear least squares implemented in [100]. For additional details, we refer to our code release.

A.4. Runtimes

Without guidance, each denoising step takes around 0.05 seconds on an RTX 4090; our experiments use 30 DDIM steps for sampling. The guidance optimizer can typically converge in around 0.2 seconds on GPU. Runtimes are reported on length-128 sequences.