ECHO: Ego-Centric modeling of Human-Object interactions

Ilya A. Petrov[1,2], Vladimir Guzov[1,2], Riccardo Marin[3,4], Emre Aksan[5], Xu Chen[5], Daniel Cremers[3,4], Thabo Beeler[5], and Gerard Pons-Moll[1,2,6]

[1]University of Tübingen, Germany [2] Tübingen AI Center, Germany
[3] Technical University of Munich, Germany
[4] Munich Center for Machine Learning, Germany [5] Google, Switzerland
[6] Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
https://virtualhumans.mpi-inf.mpg.de/echo/

Abstract. Modeling human-object interactions (HOI) from an egocentric perspective is a critical yet challenging task, particularly when relying on sparse signals from wearable devices like smart glasses and watches. We present ECHO, the first unified framework to jointly recover human pose, object motion, and contact dynamics solely from head and wrist tracking. To tackle the underconstrained nature of this problem, we introduce a novel tri-variate diffusion process with independent noise schedules that models the mutual dependencies between the human, object, and interaction modalities. This formulation allows ECHO to operate with flexible input configurations, making it robust to intermittent tracking and capable of leveraging partial observations. Crucially, it enables training on a combination of large-scale human motion datasets and smaller HOI collections, learning strong priors while capturing interaction nuances. Furthermore, we employ a smooth inpainting inference mechanism that enables the generation of temporally consistent interactions for arbitrarily long sequences. Extensive evaluations demonstrate that ECHO achieves state-of-the-art performance, significantly outperforming existing methods lacking such flexibility.

1 Introduction

Wearable sensors like smart glasses [78], rings [69], and wristbands [64] are becoming ubiquitous. Beyond enabling XR experiences, they serve as everyday companions that continuously monitor user activities. Consequently, developing algorithms to robustly perceive human-object interactions from such sparse signals is a key research challenge with far-reaching impact, unlocking applications in healthcare, personal assistants, entertainment, robotics, and spatial AI.

Recent work [26] demonstrates that it is possible to reconstruct users' motion in-the-wild from sparse sensors. Despite these advances, egocentric human-object interaction (HOI) remains underexplored. Existing methods [25,111] typically require RGB images, pre-scanned scenes, or specialized capturing suits, and rely on hand-crafted constraints, restricting scalability and generalization. While

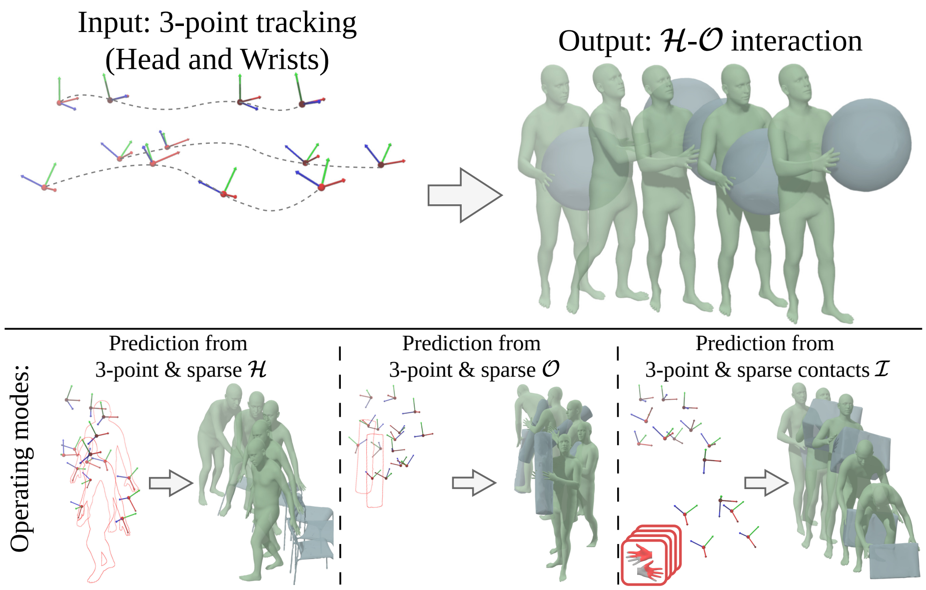

Fig. 1: ECHO. Inferring complex interactions from sparse wearable signals is challenging. ECHO is the first method to jointly recover full-body Human-Object Interaction sequences (top) solely from sparse 3-point tracking. Our flexible framework supports various inference modes (bottom), leveraging partial or intermittent observations (shown in red) of human pose, object trajectory, or contact dynamics.

learning-based approaches offer a promising alternative, current HOI datasets remain limited in scale and diversity, especially when compared to large human motion collections such as AMASS [62]. Hence, what is missing is a unified egocentric HOI method capable of learning from multiple modalities jointly.

We address this gap with ECHO (Fig. 1), the first method to model full-body HOI from head and hand tracking over arbitrary-length sequences. ECHO simultaneously predicts human pose, object trajectory, and contact dynamics. While anchored in 3-point tracking, it can additionally leverage *sparse* observations from any of these modalities (e.g., partial object tracking) to further constrain its predictions. The model is also robust to intermittent hand tracking, which is common in real-world scenarios, and has no constraints on the sequence length.

We achieve this through a unified framework with several innovations. The core of the method is a tri-variate diffusion formulation that effectively learns the inter-modal relationships between human, object, and contacts. This design enables training on a combination of datasets, including human-only motion collections such as AMASS [62] and HOI datasets like BEHAVE [6] and OMOMO [54]. By using these data sources within one generative framework, the model learns a strong human motion prior while simultaneously learning interaction-specific dynamics, something existing HOI modeling methods cannot achieve.

Crucially, our approach supports flexible conditioning, leveraging partial observations of any modality to ensure robustness against sensor noise and intermittent tracking. By explicitly modeling both human-environment and human-object contacts, we further enable self-supervised inference guidance to ensure physically plausible interactions. Finally, we extend the inference approach of HMD^2 [26] with a novel smooth inpainting that blends past and current predictions, enabling seamless, real-time processing of arbitrarily long sequences.

ECHO relies on minimal assumptions, making it flexible and adaptable to diverse settings. Its tri-variate diffusion formulation naturally accommodates additional modalities (e.g., human tracking from IMUs or object tracking from RGB), making ECHO a universal approach for egocentric HOI modeling. ECHO achieves state-of-the-art performance, surpassing competitors that lack the same flexibility. Through detailed evaluation, we demonstrate the effectiveness of our design choices, which we believe will be instrumental for further research in egocentric perception and interaction modeling.

Our key contributions are:

- We present the first method to reconstruct HOI from wearable 3-point tracking, jointly recovering human motion, object trajectory, and contact.
- We introduce a unified tri-variate diffusion model that supports flexible cross-modal conditioning and partial observations, is robust to sensor noise, and performs inference for arbitrary-length sequences.
- We achieve state-of-the-art performance and validate our design through extensive ablations. The code and trained model are available at https://github.com/ptrvilya/echo.

2 Related Work

2.1 Egocentric Motion Reconstruction

Using body-worn and head-mounted sensors for human motion reconstruction is an emerging research area. Early methods used body-worn cameras to recover hand positions [5,8,21,61,83,118] or full-body motion [1,43,48,57,58,82,92,108]. These approaches focus on joint angles, ignoring the user-scene relation.

Other works utilize body-worn sensors like EMs [45], EMGs [12], and most popularly, IMUs [35,41,66,94,95,107,114,115,130]; however, double-integrating acceleration for position induces drift.

HPS [27] advanced the field by fusing IMU tracking with camera localization for global pose estimation in large scenes. Follow-up works enabled scene scanning [15,116], scaled datasets [32,60,63,119,122], reduced sensor counts [13,38, 39,49,97,128,128], conditioned on past observations [4], and integrated biomechanical constraints [37]. Generative diffusion models enable realistic motion synthesis from underconstrained inputs [10, 18, 96]. For instance, EgoEgo [53] generates full-body motion from head trajectories, while LookOut [70] recovers head trajectories from in-the-wild video.

Recent works [11, 19, 26, 72, 85, 98, 113] utilize modalities like RGB, point clouds, and hand detections, exploring appearance modeling [22] and multi-modal fusion [33].

We similarly use sparse head and hand conditioning. However, unlike these methods, our approach models dynamic interactions with objects, enabling a more complete reconstruction of human activity in real-world settings.

2.2 Exocentric HOI Modeling

Early HOI reconstruction focused on human-centric modeling in static environments [29, 30, 34, 65, 90, 103, 124], including recent work on reconstructing and placing humans within scenes [46,109,117]. These methods do not model object displacement in the scenes, limiting their applicability in real-world scenarios.

More relevant are methods modeling dynamic HOI. Many condition on text or action labels [9, 17, 52, 55, 87, 105], which is flexible but lacks fine-grained control and detailed interactions. Such flexible conditioning is explored for human motion generation [50, 51, 68], but without modeling HOI. Others, like TRUMANS [40], condition on the scene, improving realism but requiring prior scene knowledge. Visual conditioning (RGB [2,14,20,67,93,99–102,121], multi-view [42,120], RGBD [6,36], or text and images [110]) boosts fidelity but relies on cameras, limiting scalability. I'M HOI [126] uses an object-mounted IMU, enhancing dynamics but requiring the instrumented object. Several methods use an AMASS motion prior externally to the HOI model via retrieval-and-optimization (InterDreamer [106]) or fine-tuning an AMASS-pretrained MDM (HOI-Diff [75]), rather than training jointly on motion-only and HOI data.

Another line of research generates interactions from partial information, such as past observations [24,104], object positions [7,47,54], or human pose [76,123]. Among these, TriDi [77] models the joint human-object-interaction distribution, but operates on static poses and trains on HOI-only data. ECHO extends the tri-variate formulation to arbitrary-length sequences and, crucially, trains from both motion-only (AMASS) and HOI data in a single model, building the robust motion prior that sparse egocentric conditioning requires.

2.3 Egocentric HOI Reconstruction Methods

Most methods that model human-scene interaction from egocentric data assume static environments [15, 27, 49, 116], and are therefore less relevant to our focus on dynamic interaction.

iReplica [25] is the only other method considering both humans and dynamic objects in egocentric scenarios. However, iReplica has significant limitations: it requires full-body IMUs, a pre-defined 3D scene, and a complex initialization process. Furthermore, it relies on hand-crafted heuristics for interaction modeling, whereas ECHO learns HOI dynamics from data. Recent methods like EgoGrasp [23] and WHOLE [112] recover interactions from egocentric video but focus solely on hands; Glove2Hand [125] synthesizes hand-object interaction from multi-modal sensing gloves, likewise modeling only the hands and requiring specialized worn hardware. Concurrent work IMU-HOI [56] reconstructs HOI from six body-worn IMUs plus an object-mounted IMU. Unlike ECHO, it requires a dense IMU setup and an IMU attached to the object, whereas ECHO operates from sparse three-point head-and-hand tracking and uninstrumented objects. ECHO is the first method to model full-body HOI under sparse ego-centric conditioning and over arbitrary-length sequences, handling objects of various types and sizes from wearable 3-point trackers, and remaining robust to intermittent hand tracking.

3 Method

In this section, we present ECHO, the first approach for joint human-object interaction modeling from head and wrist tracking. At the core of our method is a transformer-based diffusion model that predicts human motion $\mathcal{H}$, object motion $\mathcal{O}$, and contact sequence $\mathcal{I}$ from three-point conditioning. Subsequent sections introduce the representations for all modalities (Sec. 3.1) and describe our formulation of the diffusion process, architecture, and training (Sec. 3.2). Finally, we present the smooth inpainting that we adopt for online inference with arbitrary sequence lengths (Sec. 3.3).

3.1 Representation for Human, Object and Contact

Head-centric modeling. One of the key design choices for our method is the representation for the network's input and output modalities. Unlike sequence-level canonicalization [26, 53], we build on the per-frame variant of EgoAllo [113] by extending the representation to hands and objects. We choose to use the canonicalized head position on *each frame* as an anchor for the object's position. This per-frame definition maintains invariance to the global configuration of the sequence and allows the use of arbitrary chunks of the sequence for inference without explicitly canonicalizing each window. We illustrate this representation in Fig. 2 and discuss details in the following paragraphs.

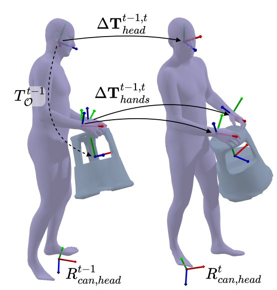

Fig. 2: Representation. ECHO operates in a per-frame head-centric coordinate system.

Object. For every object, we assume that its canonical mesh is given to the model as input. Hence, we represent the object as a sequence of its $SE(3)$ transformations in a head-centric coordinate frame, consisting of rotation and translation pairs w.r.t. the head at each frame $\boldsymbol{T_{\mathcal{O}}} = (R_{\mathcal{O}}, t_{\mathcal{O}})$. Following [129], we convert all rotations to $\mathbb{R}^6$.

Thus a sequence of N object poses is denoted as:

$$\mathcal{O} = \{\boldsymbol{T_{\mathcal{O}}}^{1..N}\} \tag{1}$$

To represent the object's class and geometry, we encode the class label in a one-hot encoded vector $\boldsymbol{y}_{\mathcal{O}}$, and extract a feature vector $\boldsymbol{f}_{\mathcal{O}} \in \mathbb{R}^{1024}$ from the canonicalized object vertices $\boldsymbol{V}_{\mathcal{O}}$ using PointNext [79]. The resulting pair $\mathcal{C}_{\mathcal{O}} = (\boldsymbol{y}_{\mathcal{O}}, \boldsymbol{f}_{\mathcal{O}})$ is used as a global conditioning for ECHO.

Human. To represent the human body, SMPL-X [73] is a natural choice. SMPL-X is a parametric body model that can be seen as a function $SMPL(\boldsymbol{T}_{\mathcal{H}}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\psi})$ of root position and orientation $\boldsymbol{T}_{\mathcal{H}} \in SE(3)$, pose $\boldsymbol{\theta}$, shape $\boldsymbol{\beta}$, and facial parameters $\boldsymbol{\psi}$. The function $SMPL$ maps parameters to posed vertices $\boldsymbol{V}_{\mathcal{H}} \in \mathbb{R}^{10475\times 3}$ of the predefined template mesh. The pose vector $\boldsymbol{\theta}$ is a concatenation of parameters for body, hands, eyes, and jaw poses in axis-angle format. As our method focuses on realistic full-body interactions, we model only the body pose $\boldsymbol{\theta}_{\mathcal{H}} \in \mathbb{R}^{21\times 6}$ part of the pose vector and assume known shape parameters $\boldsymbol{\beta}$; for the rest of the manuscript, we simplify the notation to $SMPL(\boldsymbol{T}_{\mathcal{H}}, \boldsymbol{\theta}_{\mathcal{H}})$. We follow standard practice and infer $\boldsymbol{T}_{\mathcal{H}}$ via aligning the head joint of the SMPL-X model with the head position from three-point tracking.

Hence, we define a sequence of human motion with N frames as:

$$\mathcal{H} = \{\boldsymbol{\theta}_{\mathcal{H}}^{1..N}\} \tag{2}$$

Contacts. Explicit contact modeling is crucial for robust interaction reconstruction, providing strong training signals, enabling self-supervised guidance, and allowing flexible control via sparse conditioning. We model contact as a continuous modality $\mathcal{I}$, encompassing both human-object and human-ground interactions. For human-object contact, we compute the shortest distance $d(p, \boldsymbol{V}_{\mathcal{O}})$ from sampled SMPL-X surface points $\boldsymbol{P}_c$ to the object. To fit the diffusion framework and avoid instability, we map distances to $[0, 1]$ using sigmoid:

$$\boldsymbol{c}_{\mathcal{I}}^{\mathrm{HOI}} = \{\sigma\left(\alpha \cdot (\tau_c - d(p, \boldsymbol{V}_{\mathcal{O}}))\right) \mid p \in \boldsymbol{P}_c \subset \boldsymbol{V}_{\mathcal{H}}\} \tag{3}$$

where σ is the sigmoid function, τ_c is the distance threshold, and α controls decay sharpness. Similarly, we compute human-environment contact $\boldsymbol{c}_{\mathcal{I}}^{\mathrm{Env}}$ for lower body joints based on velocity and ground proximity [81, 113]. The final contact vector is $\boldsymbol{c}_{\mathcal{I}} = \{\boldsymbol{c}_{\mathcal{I}}^{\mathrm{HOI}}, \boldsymbol{c}_{\mathcal{I}}^{\mathrm{Env}}\}$, and the sequence is defined as:

$$\mathcal{I} = \{\boldsymbol{c}_{\mathcal{I}}^{1..N}\} \tag{4}$$

Egocentric conditioning. ECHO extends the conditioning formulation of EgoAllo [113] by incorporating relative hand transformations at each frame into the method's conditioning. The method is conditioned on canonicalized head and both hands orientations $\boldsymbol{R}_{\mathrm{can,head}}^t, \boldsymbol{R}_{\mathrm{can,hands}}^t$, head-to-floor distance h_{head}^t, and relative head and hand transformations. The relative head transformation between the current and previous frame $\Delta\boldsymbol{T}_{\mathrm{head}}^{t-1,t} \in SE(3)$ is computed as:

$$\Delta\boldsymbol{T}_{\mathrm{head}}^{t-1,t} = (\boldsymbol{T}_{\text{world, head}}^{t-1})^{-1} \cdot \boldsymbol{T}_{\text{world, head}}^{t} \tag{5}$$

where $\boldsymbol{T}_{\text{world, head}}^{t}$ is the transformation of the head at time t. Similarly, we compute the relative transformation for the hands $\Delta\boldsymbol{T}_{\mathrm{hands}}^{t-1,t}$. We represent the transformation as $\mathbb{R}^6$ rotation and $\mathbb{R}^3$ translation before passing it to the network.

Thus, egocentric conditioning for a sequence $\mathcal{E}$ is defined as:

$$\mathcal{E} = \{[\Delta\boldsymbol{T}_{\mathrm{head}}^{t-1,t}, \boldsymbol{R}_{\mathrm{can,head}}^t, h_{\mathrm{head}}^t, \Delta\boldsymbol{T}_{\mathrm{hands}}^{t-1,t}, \boldsymbol{R}_{\mathrm{can,hands}}^t]^{1..N}\} \tag{6}$$

3.2 ECHO model

ECHO models the joint distribution of human motion $\mathcal{H}$, object motion $\mathcal{O}$, and contact sequence $\mathcal{I}$, conditioned on the three-point tracking $\mathcal{E}$ and object features $\mathcal{C}_{\mathcal{O}}$. Below, we formulate a three-variate diffusion process that is used to model the three modalities within one network and provide details on the underlying network's architecture.

Background A diffusion process in the context of generative neural networks is divided into two phases. The forward phase progressively adds noise to an original data sample, while the backward phase uses a learned model to recover the sample from noise. We adopt the formulation of Denoising Diffusion Probabilistic Model (DDPM) [31] in our work with a modification following [80], predicting the original sample with the neural network instead of predicting the added noise. To achieve this, we parametrize the reverse process by a denoising neural network $\mathcal{D}_\psi$ that is trained to recover the original sample $\mathbf{z}^0$ from the noisy sample $\mathbf{z}^{\mathcal{T}}$ at denoising step $\mathcal{T}$ given the condition c. Defining for brevity $\mathbb{E}_p \equiv \mathbb{E}_{\mathbf{z}^0 \sim p_{data}}$, $\mathbb{E}_{\mathcal{T}} \equiv \mathbb{E}_{\mathcal{T} \sim \mathcal{U}\{0,\ldots,T\}}$, and $\mathbb{E}_q \equiv \mathbb{E}_{\mathbf{z}^{\mathcal{T}} \sim q(\mathbf{z}^{\mathcal{T}}|\mathbf{z}^0)}$ we obtain the training objective (the full definition of forward and backward processes is included in the Sup. Mat.):

$$\min_{\psi} \mathbb{E}_p \, \mathbb{E}_{\mathcal{T}} \, \mathbb{E}_q \, \|\mathcal{D}_\psi(\mathbf{z}^{\mathcal{T}}; c, \mathcal{T}) - \mathbf{z}^0\|_2. \tag{7}$$

The original formulation of the diffusion model [31,86] focuses on generating single-modality data, e.g., images. Inspired by TriDi [77], we formulate a tri-variate diffusion process for HOI modeling, proposing a new formulation that diffuses motion sequences.

Tri-variate diffusion with independent schedules. We formulate a three-variate diffusion process for human motion $\mathcal{H}$, object trajectory $\mathcal{O}$, and sequence of contacts $\mathcal{I}$, denoting the corresponding sets of denoising steps as $\mathcal{T}_{\mathcal{H}}, \mathcal{T}_{\mathcal{O}}, \mathcal{T}_{\mathcal{I}}$. We define:

$$\begin{aligned} \mathbb{E}_p &\equiv \mathbb{E}_{(\mathcal{H}^0,\mathcal{O}^0,\mathcal{I}^0) \sim p(\mathcal{H},\mathcal{O},\mathcal{I}|\mathcal{E})}, \\ \mathbb{E}_{\mathcal{T}} &\equiv \mathbb{E}_{(\mathcal{T}_{\mathcal{H}},\mathcal{T}_{\mathcal{O}},\mathcal{T}_{\mathcal{I}}) \sim \mathcal{U}\{0,\ldots,T\}^{N\times 3}}, \\ \mathbb{E}_q &\equiv \mathbb{E}_{\mathcal{H}^{\mathcal{T}_{\mathcal{H}}} \sim q(\mathcal{H}^{\mathcal{T}_{\mathcal{H}}}|\mathcal{H}^0), \mathcal{O}^{\mathcal{T}_{\mathcal{O}}} \sim q(\mathcal{O}^{\mathcal{T}_{\mathcal{O}}}|\mathcal{O}^0), \mathcal{I}^{\mathcal{T}_{\mathcal{I}}} \sim q(\mathcal{I}^{\mathcal{T}_{\mathcal{I}}}|\mathcal{I}^0)} \end{aligned} \tag{8}$$

The key feature of this formulation is that ECHO diffuses the three modalities following three independent time schedules, allowing them to vary (e.g., one can provide tracking information for the human in addition to three-point conditioning and predict the object motion and contact sequence corresponding to it). The main minimization objective for parameters ψ of a model ECHO_ψ is:

$$\mathbb{E}_p \mathbb{E}_{\mathcal{T}} \mathbb{E}_q \|\mathrm{ECHO}_\psi(\mathcal{H}^{\mathcal{T}_{\mathcal{H}}}, \mathcal{O}^{\mathcal{T}_{\mathcal{O}}}, \mathcal{I}^{\mathcal{T}_{\mathcal{I}}}; \mathcal{T}_{\mathcal{H}}, \mathcal{T}_{\mathcal{O}}, \mathcal{T}_{\mathcal{I}}; \mathcal{C}_{\mathcal{O}}, \mathcal{E}) - (\mathcal{H}^0, \mathcal{O}^0, \mathcal{I}^0)\|_2 \tag{9}$$

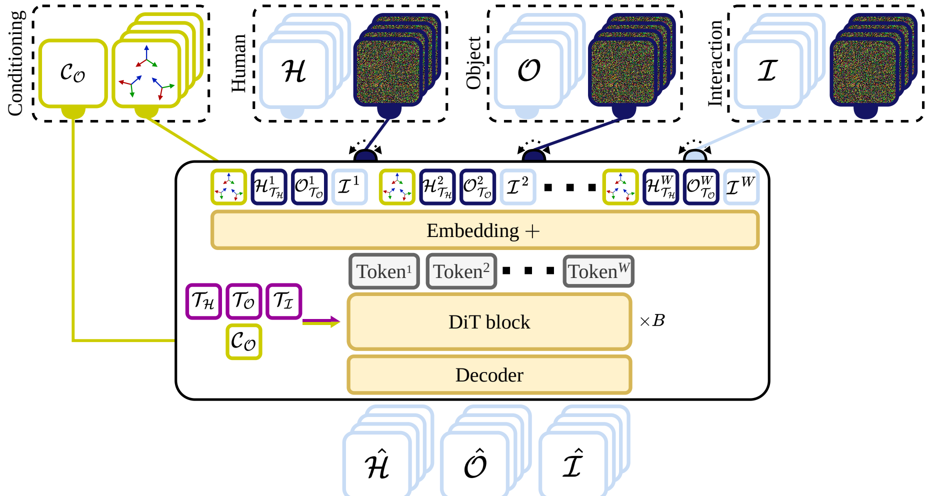

Fig. 3: ECHO overview. ECHO requires just head and hand tracking and object class, to predict Human, Object, and Interaction. The input tokens are composed of **condition**, and of either **observed modality**, or **noise** for $\mathcal{H}, \mathcal{O}$, and $\mathcal{I}$. For every modality, we use a unique denoising **step**. Our model allows flexible input configuration. In the example above we use contacts $\mathcal{I}$ as an **additional input** to the network, that infers the **other modalities** $\mathcal{H}$ and $\mathcal{O}$, matching the extended condition.

Universal multi-modal architecture. We build the ECHO denoising network on the Diffusion Transformer (DiT) [74] with rotary positional embeddings [88], adapting it to our multi-modal setting (Fig. 3). The model takes as input the sequences $\mathcal{H}$, $\mathcal{O}$, and $\mathcal{I}$ with varying noise levels, along with egocentric $\mathcal{E}$ and object $\mathcal{C}_\mathcal{O}$ conditioning, and the denoising step for each modality $\{\mathcal{T}_\mathcal{H}, \mathcal{T}_\mathcal{O}, \mathcal{T}_\mathcal{I}\}$. A key advantage of our design is the use of independent noise schedules for each modality, unlike standard approaches that use a single schedule. This enables flexible conditioning: by setting the noise level to zero for known modalities (e.g., observed human motion), the model effectively predicts the remaining components (e.g., object motion and contacts) consistent with the input. When no interaction data is available, it generates realistic motion based solely on egocentric conditioning. Furthermore, this formulation naturally handles partial observations, allowing ECHO to leverage sparse signals – such as intermittent object tracking or partial human pose from IMUs – to constrain the generation. Ultimately, the method outputs the full human-object interaction sequence $\{\hat{\mathcal{H}}, \hat{\mathcal{O}}, \hat{\mathcal{I}}\}$.

Training. During training, each modality ($\mathcal{H}$, $\mathcal{O}$, $\mathcal{I}$) is either diffused or provided as clean condition. To ensure stability, we sample from the 2^3 combinations of noisy/denoised states and synchronize the noise level across all diffused modalities, instead of sampling independent noise levels for each modality. This effectively trains the model to reconstruct missing modalities from the observed ones. To improve generalization, we simulate intermittent tracking by randomly

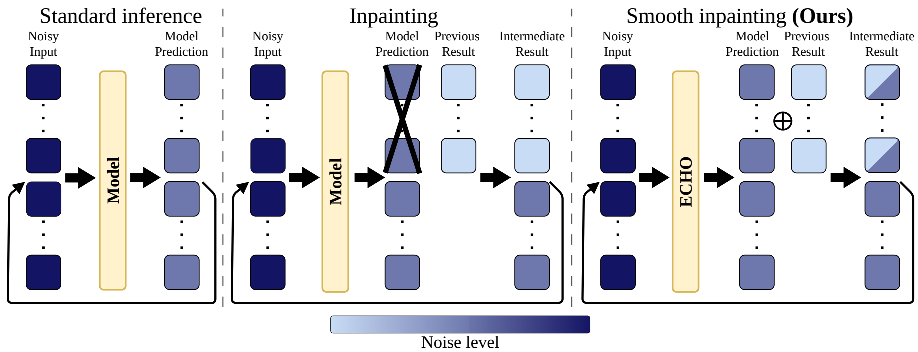

Fig. 4: Comparison of inference strategies. Standard per-window inference (left) ignores the context of the past predictions. Inpainting (middle) uses past prediction as condition but drops new predictions for the overlapping region. Our smooth inpainting (right) blends past and current predictions in the overlapping region on every diffusion step, ensuring seamless transitions.

dropping hand and object conditioning. Experiments confirm that ECHO remains robust to moderate tracking degradation.

The objective function is a weighted sum of six terms: reconstruction losses for each diffused modality, object trajectory smoothness, human joint error, and a foot skating penalty (details in Sup. Mat.). Given the limited scale of HOI datasets (OMOMO, BEHAVE), we augment training with the large-scale AMASS [62] dataset to learn a robust human motion prior. For AMASS samples, we replace object conditioning $\mathcal{C}_{\mathcal{O}}$ with learnable tokens, signaling the model to ignore object interactions; ablations confirm the benefits of this joint training.

3.3 Inference

Smooth inpainting To generate temporal sequences, recent methods [113] employ a sliding window approach. Unlike offline methods [10, 91], this approach doesn't require the full sequence to be present at the start, enabling online inference. However, sliding windows require post-processing to stitch results. HMD^2 [26] addresses this via inpainting-based inference, conditioning new windows on past predictions. However, it discards new predictions in the overlapping region, relying solely on past data. Our idea is to take the inpainting-based inference one step further by incorporating smooth blending between past and new predictions, which we call *smooth inpainting* (Fig. 4). It replaces predictions in the overlapping region with a weighted average of past and current predictions, allowing for a gradual transition between windows. This enables ECHO to autoregressively generate arbitrarily long sequences with smooth transitions. Additionally, the overlap size can be tuned to balance context length and inference latency for real-time applications.

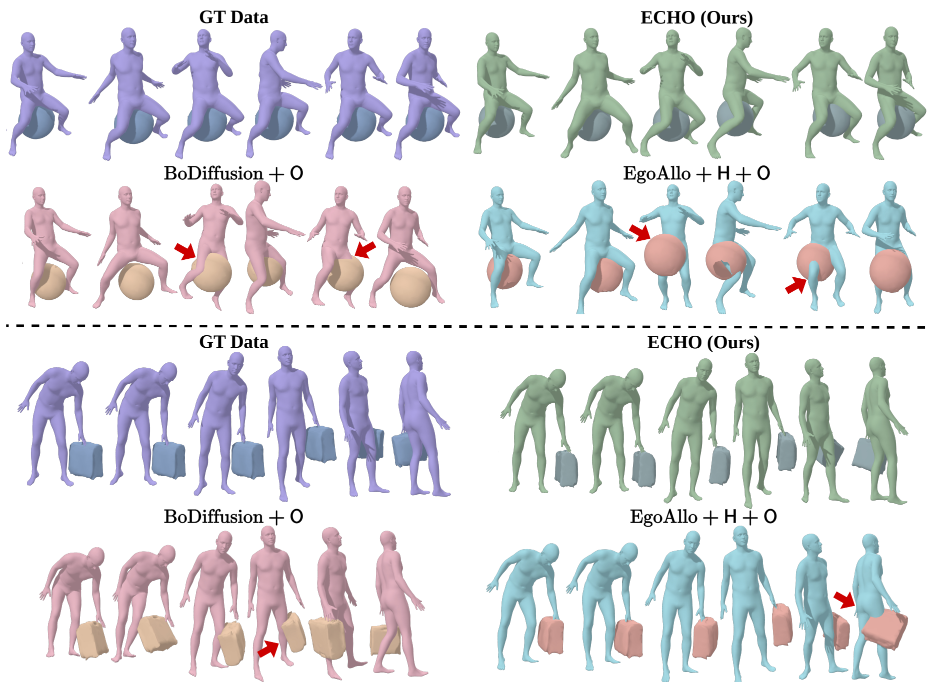

Fig. 5: Qualitative results of ECHO. Our method accurately reconstructs human-object interactions across diverse scenarios. In contrast, competing methods often fail to capture correct contact dynamics, leading to artifacts such as object penetration or floating. For dynamic visualizations, please refer to the supplementary video.

Inference guidance To further enhance the quality of generated interactions, we adopt a classifier-based guidance [16] approach at inference time. Based on the guidance function's value, the network's prediction is updated at each step of the denoising process. We formulate the guidance loss to ensure that the predicted human and object meshes align with the predicted contact vector $\widehat{\boldsymbol{c}}_{\mathcal{I}}$. The function, therefore, includes two terms: one for human-object contact and one for foot-floor contact.

4 Experiments

Implementation details. The ECHO model comprises 57.7M parameters and is trained using the AdamW optimizer [59] for 300k steps (approximately 30 hours) with a batch size of 256 and a learning rate of $5e^{-4}$ on a single NVIDIA RTX 5090 GPU. For inference, we use 100 DDPM steps to balance generation quality and latency. With a 30-frame overlap, ECHO processes a 60-frame window in 640 ms without guidance and 980 ms with guidance on an RTX 5090 GPU. This corresponds to a throughput of approximately 46 FPS and 30 FPS, respectively, enabling operation at low latency of 640–980 ms per window, *i.e.*,

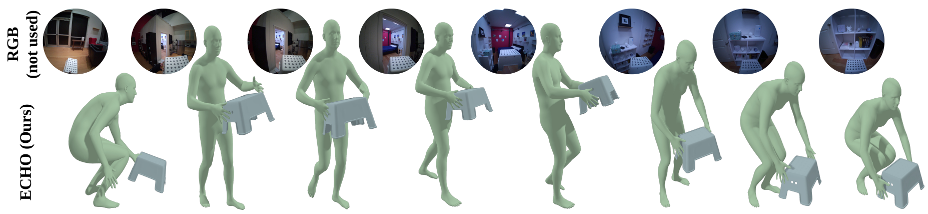

Fig. 6: Qualitative results of ECHO. We demonstrate generalization to novel motion and objects from the Aria Digital Twin [71]; RGB is included for reference.

∼21–33 ms amortized per frame. Visualizations are generated using aitviewer [44] and Blendify [28].

Datasets. We train and evaluate our method on the union of the BEHAVE [6], OMOMO [54], and AMASS [62] datasets. We follow the official train-test splits for BEHAVE and OMOMO, and the EgoAllo [113] split for AMASS. To use the SMPL-X body model in ECHO, we convert BEHAVE sequences from the SMPL+H [84] format using code from [89]. We downsample all sequences to 30 fps for consistency, as the BEHAVE data is limited to this frame rate. During training, we sample windows of size $W = 60$. For evaluation, we perform continuous inference on full sequences.

Metrics. To evaluate human motion reconstruction, we use the Mean Per-Joint Position Error (MPJPE, in cm), Mean Per-Joint Velocity Error (MPJVE [128]), and Foot Contact (FC) score [53,113]. For object reconstruction, we compute the vertex-to-vertex error E_{v2v} [76,104] (cm), center error E_c [76] (cm), and contact accuracy $\text{Acc}_{\mathcal{I}}$ [77]. Detailed metric definitions are provided in the Supp. Mat.

4.1 Comparison with baselines

Baselines. ECHO is the first approach for the end-to-end modeling of egocentric human-object interactions. To evaluate its performance, we select several existing end-to-end egocentric motion modeling methods as baselines and extend them for HOI modeling. While some recent methods employ multi-stage training and inference, adapting them to HOI modeling would require significant architectural modifications. We compare against **BoDiffusion** [10], which builds on DiT [74] to process concatenated input tracking and noisy motion. During training, we canonicalize each window to a head-centric space to ensure a fair comparison. To evaluate HOI prediction, we extend BoDiffusion to include object modeling, referring to this baseline as **BoDiffusion**+O. We modify the network to process object poses and conditioning by concatenating object transformations to the human motion and egocentric tokens, and appending object shape embeddings and class information to the global conditioning. We also construct an HOI baseline on top of **EgoAllo** [113] by integrating object pose prediction. Since EgoAllo is originally conditioned only on head tracking, we extend it to support

Table 1: Comparison with baselines on BEHAVE and OMOMO. ECHO demonstrates better performance for human-object interaction modeling and competitive motion modeling quality.

Method	Human			Object			
BEHAVE	MPJPE↓	MPJVE↓	FC↑	$E_{v2v}\downarrow$	$E_c\downarrow$	Rot. Diff.↓	$\text{Acc}_{\mathcal{I}}\uparrow$
Data	-	-	0.73	-	-	-	-
BoDiffusion [10]+O	$8.3^{\pm 0.2}$	$10.1^{\pm 0.3}$	$0.82^{\pm 0.04}$	$44.2^{\pm 1.2}$	$29.9^{\pm 1.2}$	$\underline{110.9}^{\pm 4.4}$	$\underline{91.8}^{\pm 0.4}$
EgoAllo [113]+H+O	$\underline{7.6}^{\pm 0.1}$	$\underline{8.6}^{\pm 0.2}$	$\mathbf{0.95}^{\pm 0.01}$	$\underline{39.1}^{\pm 1.1}$	$\underline{22.5}^{\pm 0.7}$	$111.6^{\pm 3.8}$	$91.7^{\pm 0.3}$
ECHO **(Ours)**	$\mathbf{6.8}^{\pm 0.1}$	$\mathbf{7.5}^{\pm 0.1}$	$\underline{0.94}^{\pm 0.00}$	$\mathbf{33.5}^{\pm 0.5}$	$\mathbf{20.1}^{\pm 0.3}$	$\mathbf{92.9}^{\pm 2.2}$	$\mathbf{93.1}^{\pm 0.2}$
OMOMO	MPJPE↓	MPJVE↓	FC↑	$E_{v2v}\downarrow$	$E_c\downarrow$	Rot. Diff.↓	$\text{Acc}_{\mathcal{I}}\uparrow$
Data	-	-	0.96	-	-	-	-
BoDiffusion [10]+O	$7.6^{\pm 0.4}$	$8.6^{\pm 0.3}$	$\mathbf{0.98}^{\pm 0.01}$	$33.2^{\pm 1.9}$	$22.2^{\pm 1.7}$	$\underline{94.9}^{\pm 7.4}$	$\underline{96.2}^{\pm 0.3}$
EgoAllo [113]+H+O	$\underline{6.6}^{\pm 0.1}$	$\underline{7.3}^{\pm 0.1}$	$\underline{0.95}^{\pm 0.01}$	$\underline{30.8}^{\pm 0.9}$	$\underline{18.3}^{\pm 0.5}$	$98.2^{\pm 5.3}$	$96.5^{\pm 0.2}$
ECHO **(Ours)**	$\mathbf{6.0}^{\pm 0.1}$	$\mathbf{6.1}^{\pm 0.1}$	$0.93^{\pm 0.01}$	$\mathbf{26.5}^{\pm 1.1}$	$\mathbf{15.2}^{\pm 0.3}$	$\mathbf{86.5}^{\pm 6.7}$	$\mathbf{96.9}^{\pm 0.1}$

Table 2: Quality of motion generation. ECHO outperforms the baselines on the AMASS dataset, demonstrating that our joint HOI formulation effectively learns a strong human motion prior. The significant drop in performance for **NoAMASS** confirms the importance of training on large-scale motion data.

Method	AMASS		
	MPJPE↓	MPJVE↓	$\text{FC}\uparrow^{1.0}$
BoDiffusion [10]+O	$11.4^{\pm 0.3}$	$14.3^{\pm 0.3}$	1.0
EgoAllo [113]+H+O	$\underline{8.9}^{\pm 0.1}$	$\underline{11.6}^{\pm 0.1}$	1.0
ECHO **(Ours)** - NoAMASS	$43.1^{\pm 0.1}$	$40.3^{\pm 0.1}$	0.8
ECHO **(Ours)**	$\mathbf{7.4}^{\pm 0.1}$	$\mathbf{8.6}^{\pm 0.2}$	1.0

hand conditioning, following the official implementation. We refer to this method as **EgoAllo**+H+O. For fair comparison, we train ECHO and all baselines on the union of AMASS, BEHAVE, and OMOMO. Additionally, both baselines predict object pose in a head-centric space, matching ECHO, to ensure a consistent coordinate system.

HOI generation. We evaluate HOI reconstruction quality on the BEHAVE and OMOMO test sets, reporting mean and variance across three runs in Tab. 1. ECHO significantly outperforms BoDiffusion+O in both human and object predictions, and performs on par with EgoAllo+H+O in human motion while surpassing it in object prediction. This superior object tracking demonstrates the efficacy of our tri-variate modeling. Qualitative results (Fig. 5) confirm this advantage: while the baselines often fail to maintain realistic contact, resulting in artifacts like penetration or floating, ECHO generates physically plausible interactions. We further demonstrate generalization on an Aria Digital Twin [71] sequence in Fig. 6; see the supplementary video for dynamic visualizations.

Table 3: Evaluation of ECHO with noise simulation. We demonstrate the robustness of ECHO to intermittent hand tracking by randomly dropping a percentage of the input. The model maintains stable performance even with significant missing hand tracking data, confirming its resilience to sensor noise.

%	BEHAVE			
	Human		**Object**	
	MPJPE↓	MPJVE↓	E_{v2v} ↓	E_c ↓
0	$6.8^{\pm 0.1}$	$7.5^{\pm 0.1}$	$33.5^{\pm 0.5}$	$20.1^{\pm 0.3}$
25	$6.9^{\pm 0.1}$	$7.5^{\pm 0.2}$	$33.8^{\pm 0.8}$	$20.6^{\pm 0.5}$
50	$7.0^{\pm 0.2}$	$7.8^{\pm 0.2}$	$33.9^{\pm 1.0}$	$20.8^{\pm 0.6}$
75	$7.6^{\pm 0.3}$	$10.3^{\pm 0.5}$	$34.6^{\pm 1.3}$	$21.2^{\pm 1.0}$
90	$9.3^{\pm 0.5}$	$10.3^{\pm 0.6}$	$36.8^{\pm 2.0}$	$24.6^{\pm 1.9}$

Motion generation. While ECHO is designed for human-object interactions, robust human motion reconstruction is also essential. We evaluate motion quality by comparing ECHO against the baselines on the AMASS dataset. Results are reported in Tab. 2 (mean and variance across three runs). Our method outperforms both baselines, demonstrating the effectiveness of our joint modeling formulation. Notably, performance significantly degrades when ECHO is trained without the AMASS dataset (**NoAMASS**), highlighting the importance of large-scale motion data for learning a strong human motion prior. For fair comparison, all models (except (**NoAMASS**)) were trained on the union of the three datasets.

4.2 Noisy conditioning and sparse input

Egocentric human-object interactions are captured using diverse technologies and settings, often involving data streams (e.g., IMUs, head-mounted cameras) that provide additional, albeit often sparse and noisy, information. To study the robustness and versatility of ECHO under such conditions, we simulate two scenarios: noisy hand tracking and access to additional sparse tracking information. To simulate intermittent hand tracking, we randomly drop a percentage of hand tracking data from the input while keeping head tracking intact. Tab. 3 demonstrates that ECHO is robust to missing hand tracking, showing significant performance degradation only when 75% or more of the data is missing.

While tracking objects from an egocentric camera remains challenging [3, 127], the object's location and orientation can often be determined for a few frames. For human motion, it is also possible to obtain partial tracking from RGB-based pose estimation or IMU suits. Such information provides important cues to reduce uncertainty, and ECHO is designed to leverage these opportunities. We test this capability by simulating a scenario where either the human or object modality is only partially observed. Tab. 4 shows that additional constraints significantly improve the accuracy of the partially observed modality, while predictions for the unobserved modality improve slightly.

Table 4: Evaluation of ECHO with sparse tracking. We demonstrate the versatility of ECHO by providing additional sparse tracking information alongside egocentric conditioning. Providing partial information for one modality (Human or Object) significantly improves its reconstruction quality and helps regularize the other.

Mode	%	BEHAVE Human MPJPE↓	BEHAVE Human MPJVE↓	BEHAVE Object E_{v2v} ↓	BEHAVE Object E_c ↓
ECHO		$6.82^{\pm0.08}$	$7.50^{\pm0.11}$	$33.46^{\pm0.50}$	$20.13^{\pm0.26}$
Sparse $\mathcal{H}$ available	*10*	$5.84^{\pm0.10}$	$6.27^{\pm0.12}$	$33.44^{\pm0.73}$	$20.12^{\pm0.38}$
	25	$4.84^{\pm0.09}$	$5.09^{\pm0.12}$	$33.42^{\pm0.74}$	$20.00^{\pm0.38}$
	50	$3.71^{\pm0.07}$	$3.80^{\pm0.10}$	$33.29^{\pm0.72}$	$19.81^{\pm0.37}$
	100	-	-	$32.79^{\pm0.65}$	$19.41^{\pm0.31}$
Sparse $\mathcal{O}$ available	*10*	$6.81^{\pm0.10}$	$7.48^{\pm0.12}$	$27.20^{\pm0.70}$	$16.55^{\pm0.39}$
	25	$6.81^{\pm0.09}$	$7.47^{\pm0.11}$	$19.77^{\pm0.78}$	$12.69^{\pm0.42}$
	50	$6.79^{\pm0.09}$	$7.45^{\pm0.11}$	$10.75^{\pm0.64}$	$7.93^{\pm0.38}$
	100	$6.69^{\pm0.08}$	$7.37^{\pm0.10}$	-	-

Table 5: Ablation study on BEHAVE. Evaluating the impact of ECHO components proves the usefulness of guidance, smooth inpainting, usage of three modalities, and training with AMASS data.

Method	BEHAVE Human MPJPE↓	BEHAVE Human MPJVE↓	BEHAVE Object E_{v2v} ↓	BEHAVE Object E_c ↓
ECHO	$6.8^{\pm0.1}$	$7.5^{\pm0.1}$	$33.5^{\pm0.5}$	$20.1^{\pm0.3}$
NoGuide	$6.8^{\pm0.1}$	$7.5^{\pm0.1}$	$33.6^{\pm0.5}$	$20.3^{\pm0.4}$
Inpaint w/o smooth	$6.9^{\pm0.1}$	$7.6^{\pm0.1}$	$33.7^{\pm0.5}$	$20.4^{\pm0.3}$
($\mathcal{H}$, $\mathcal{O}$)	$8.1^{\pm0.1}$	$9.0^{\pm0.1}$	$34.4^{\pm0.3}$	$20.7^{\pm0.1}$
NoAMASS	$8.7^{\pm0.1}$	$9.6^{\pm0.1}$	$34.7^{\pm0.2}$	$21.8^{\pm0.2}$

4.3 Ablation

We conduct an ablation study to analyze the effectiveness of our method's components. All models are trained in the same setting, varying only the ablated component. First, we evaluate the model without inference-time guidance (**NoGuide**), observing that it primarily affects object prediction quality on noisier data (i.e., BEHAVE). Results on BEHAVE are presented in Tab. 5; OMOMO results are in the Supplementary Material. **Inpaint w/o smooth** performs inference using standard inpainting, without smooth blending between past and current predictions. The performance drop without smooth inpainting highlights its importance for generating temporally consistent interactions. The model without the interaction modality (i.e., using only $(\boldsymbol{\mathcal{H}}, \boldsymbol{\mathcal{O}})$) exhibits substantially worse quality for both human and object motion, suggesting that contacts play a crucial role in linking these modalities. Training without the AMASS dataset (**NoAMASS**) leads to a significant decrease in human motion modeling quality, once again highlighting the importance of large-scale datasets for learning human motion priors.

5 Conclusions

In this work, we introduced ECHO, the first unified framework for modeling human-object interactions solely from sparse egocentric tracking. Central to our approach is a novel tri-variate diffusion formulation with independent noise schedules that jointly models human pose, object motion, and contact dynamics. This design not only captures the complex interdependencies between the user and the object but also enables flexible inference, allowing the model to adapt to intermittent tracking and partial observations. Crucially, it facilitates training on a combination of large-scale human motion datasets and smaller HOI collections, learning strong priors while capturing interaction nuances. Additionally, our smooth inpainting inference mechanism ensures temporally consistent generation for sequences of arbitrary length. Extensive evaluations demonstrate that ECHO significantly outperforms state-of-the-art methods, offering a robust and versatile solution for egocentric HOI reconstruction.

Limitations and future work. Despite its strong performance, ECHO has limitations that open exciting directions for future research. First, our current model focuses on object interactions and environmental contacts with the ground. Incorporating sophisticated constraints (*e.g.*, those coming from dynamic surroundings) would be crucial for consistent long-term motion in complex scenes. Second, although robust for various object types, the absence of fine-grained finger tracking limits the reconstruction of dexterous interactions with small objects (e.g., pens, scissors). Integrating additional modalities, such as egocentric RGB video, could help resolve these fine motor details. Finally, extending our framework to support visual conditioning, alongside the collection of currently absent RGB-based egocentric HOI datasets, represents a promising avenue for achieving comprehensive egocentric perception.

Acknowledgments Special thanks to Nikita Kister and Berna Kabadayi for the helpful discussions. This work is funded by the Deutsche Forschungsgemeinschaft - 409792180 (EmmyNoether Programme, project: Real Virtual Humans). G. Pons-Moll is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting I. A. Petrov. R. Marin has been supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 101109330. The project was made possible by funding from the Carl Zeiss Foundation. The computational resources for this project were provided by the Google Cloud grant. This work was supported by the European Research Council (ERC) Advanced Grant SIMULACRON and by the GNI Project "AI4Twinning".

References

1. Akada, H., Wang, J., Shimada, S., Takahashi, M., Theobalt, C., Golyanik, V.: Unrealego: A new dataset for robust egocentric 3d human motion capture. In: European Conference on Computer Vision. pp. 1–17. Springer (2022) 3
2. Antić, D., Budria, A., Paschalidis, G., Dwivedi, S.K., Tzionas, D.: Lexis: Latent proximal interaction signatures for 3d hoi from an image. arXiv preprint arXiv:2604.20800 (2026) 4
3. Banerjee, P., Shkodrani, S., Moulon, P., Hampali, S., Han, S., Zhang, F., Zhang, L., Fountain, J., Miller, E., Basol, S., et al.: Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7061–7071 (2025) 13
4. Barquero, G., Bertsch, N., Marramreddy, M., Chacón, C., Arcadu, F., Rigual, F., He, N.S., Palmero, C., Escalera, S., Ye, Y., et al.: From sparse signal to smooth motion: Real-time motion generation with rolling prediction models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1850–1860 (2025) 3
5. Bhatnagar, B.L., Singh, S., Arora, C., Jawahar, C.: Unsupervised learning of deep feature representation for clustering egocentric actions. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17. pp. 1447–1453 (2017) 3
6. Bhatnagar, B.L., Xie, X., Petrov, I.A., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Behave: Dataset and method for tracking human object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15935–15946 (2022) 2, 4, 11
7. Braun, J., Christen, S., Kocabas, M., Aksan, E., Hilliges, O.: Physically plausible full-body hand-object interaction synthesis. In: 2024 International Conference on 3D Vision (3DV). pp. 464–473. IEEE (2024) 4
8. Cao, C., Zhang, Y., Wu, Y., Lu, H., Cheng, J.: Egocentric gesture recognition using recurrent 3d convolutional neural networks with spatiotemporal transformer modules. 2017 IEEE International Conference on Computer Vision (ICCV) (2017) 3
9. Cao, Y., Pan, L., Han, K., Wong, K.Y.K., Liu, Z.: AvatarGO: Zero-shot 4d human-object interaction generation and animation. In: The Thirteenth International Conference on Learning Representations (2025) 4
10. Castillo, A., Escobar, M., Jeanneret, G., Pumarola, A., Arbeláez, P., Thabet, A., Sanakoyeu, A.: Bodiffusion: Diffusing sparse observations for full-body human motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4221–4231 (2023) 3, 9, 11, 12
11. Chi, S., Huang, P.H., Sachdeva, E., Ma, H., Ramani, K., Lee, K.: Estimating ego-body pose from doubly sparse egocentric video data. Advances in Neural Information Processing Systems **37**, 55178–55203 (2024) 3
12. Chiquier, M., Vondrick, C.: Muscles in action. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22091–22101 (2023) 3
13. Dai, P., Zhang, Y., Liu, T., Fan, Z., Du, T., Su, Z., Zheng, X., Li, Z.: Hmd-poser: On-device real-time human motion tracking from scalable sparse observations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 874–884 (2024) 3
14. Dai, S., Li, W., Sun, H., Huang, H., Ma, C., Huang, H., Xu, K., Hu, R.: Interfusion: Text-driven generation of 3d human-object interaction. In: European Conference on Computer Vision. pp. 18–35. Springer (2024) 4

15. Dai, Y., Lin, Y., Wen, C., Shen, S., Xu, L., Yu, J., Ma, Y., Wang, C.: Hsc4d: Human-centered 4d scene capture in large-scale indoor-outdoor space using wearable imus and lidar. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6792–6802 (2022) 3, 4
16. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021) 10
17. Diller, C., Dai, A.: Cg-hoi: Contact-guided 3d human-object interaction generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19888–19901 (2024) 4
18. Du, Y., Kips, R., Pumarola, A., Starke, S., Thabet, A., Sanakoyeu, A.: Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 481–490 (2023) 3
19. Escobar, M., Puentes, J., Forigua, C., Pont-Tuset, J., Maninis, K.K., Arbelaez, P.: Egocast: Forecasting egocentric human pose in the wild. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5831–5841. IEEE (2025) 3
20. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: Arctic: A dataset for dexterous bimanual hand-object manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12943–12954 (2023) 4
21. Fathi, A., Farhadi, A., Rehg, J.M.: Understanding egocentric activities. In: 2011 international conference on computer vision. pp. 407–414. IEEE (2011) 3
22. Feng, H., Ma, W., Gao, Q., Zheng, X., Xue, N., Xu, H.: Stratified avatar generation from sparse observations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 153–163 (2024) 3
23. Fu, H., Wang, W., Qiao, X., Yang, S., Liu, Z., Zhao, B.: Egograsp: World-space hand-object interaction estimation from egocentric videos. arXiv preprint arXiv:2601.01050 (2026) 4
24. Ghosh, A., Dabral, R., Golyanik, V., Theobalt, C., Slusallek, P.: Imos: Intent-driven full-body motion synthesis for human-object interactions. In: Computer Graphics Forum. vol. 42, pp. 1–12. Wiley Online Library (2023) 4
25. Guzov, V., Chibane, J., Marin, R., He, Y., Saracoglu, Y., Sattler, T., Pons-Moll, G.: Interaction replica: Tracking human-object interaction and scene changes from human motion. In: International Conference on 3D Vision (3DV) (March 2024) 1, 4
26. Guzov, V., Jiang, Y., Hong, F., Pons-Moll, G., Newcombe, R., Liu, C.K., Ye, Y., Ma, L.: Hmd^2: Environment-aware motion generation from single egocentric head-mounted device. In: International Conference on 3D Vision (3DV) (March 2025) 1, 3, 5, 9
27. Guzov, V., Mir, A., Sattler, T., Pons-Moll, G.: Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2021) 3, 4
28. Guzov, V., Petrov, I.A., Pons-Moll, G.: Blendify – python rendering framework for blender. arXiv preprint arXiv:2410.17858 (2024) 11
29. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3d human pose ambiguities with 3d scene constraints. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2282–2292 (2019) 4

30. Hassan, M., Ghosh, P., Tesch, J., Tzionas, D., Black, M.J.: Populating 3d scenes by learning human-scene interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14708–14718 (2021) 4
31. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 7
32. Hollidt, D., Streli, P., Jiang, J., Haghighi, Y., Qian, C., Liu, X., Holz, C.: Egosim: An egocentric multi-view simulator and real dataset for body-worn cameras during motion and activity. Advances in Neural Information Processing Systems **37**, 106607–106627 (2024) 3
33. Hong, F., Guzov, V., Kim, H.J., Ye, Y., Newcombe, R., Liu, Z., Ma, L.: Egolm: Multi-modal language model of egocentric motions. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5344–5354 (2025) 3
34. Huang, S., Wang, Z., Li, P., Jia, B., Liu, T., Zhu, Y., Liang, W., Zhu, S.C.: Diffusion-based generation, optimization, and planning in 3d scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16750–16761 (2023) 4
35. Huang, Y., Kaufmann, M., Aksan, E., Black, M.J., Hilliges, O., Pons-Moll, G.: Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **37**(6), 185:1–185:15 (nov 2018) 3
36. Huang, Y., Taheri, O., Black, M.J., Tzionas, D.: InterCap: Joint markerless 3D tracking of humans and objects in interaction from multi-view RGB-D images. International Journal of Computer Vision (IJCV) (2024) 4
37. Jiang, J., Streli, P., Luo, X., Gebhardt, C., Holz, C.: Manikin: biomechanically accurate neural inverse kinematics for human motion estimation. In: European Conference on Computer Vision. pp. 128–146. Springer (2024) 3
38. Jiang, J., Streli, P., Meier, M., Holz, C.: Egoposer: Robust real-time egocentric pose estimation from sparse and intermittent observations everywhere. In: European Conference on Computer Vision. pp. 277–294. Springer (2024) 3
39. Jiang, J., Streli, P., Qiu, H., Fender, A., Laich, L., Snape, P., Holz, C.: Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In: European conference on computer vision. pp. 443–460. Springer (2022) 3
40. Jiang, N., Zhang, Z., Li, H., Ma, X., Wang, Z., Chen, Y., Liu, T., Zhu, Y., Huang, S.: Scaling up dynamic human-scene interaction modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1737–1747 (2024) 4
41. Jiang, Y., Ye, Y., Gopinath, D., Won, J., Winkler, A.W., Liu, C.K.: Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022) 3
42. Jiang, Y., Jiang, S., Sun, G., Su, Z., Guo, K., Wu, M., Yu, J., Xu, L.: Neuralhofusion: Neural volumetric rendering under human-object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6155–6165 (2022) 4
43. Kang, T., Lee, K., Zhang, J., Lee, Y.: Ego3dpose: Capturing 3d cues from binocular egocentric views. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–10 (2023) 3
44. Kaufmann, M., Vechev, V., Mylonopoulos, D.: aitviewer (7 2022), https://github.com/eth-ait/aitviewer 11

45. Kaufmann, M., Zhao, Y., Tang, C., Tao, L., Twigg, C., Song, J., Wang, R., Hilliges, O.: Em-pose: 3d human pose estimation from sparse electromagnetic trackers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11510–11520 (2021) 3
46. Kister, N., YM, P., Sárándi, I., Wang, J., Khoreva, A., Pons-Moll, G.: Inhabit: Leveraging image foundation models for scalable 3d human placement. arXiv preprint arXiv:2604.19673 (2026) 4
47. Kulkarni, N., Rempe, D., Genova, K., Kundu, A., Johnson, J., Fouhey, D., Guibas, L.: Nifty: Neural object interaction fields for guided human motion synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 947–957 (2024) 4
48. Lee, J., Xu, W., Richard, A., Wei, S.E., Saito, S., Bai, S., Wang, T.L., Sung, M., Kim, T.K., Saragih, J.: Rewind: Real-time egocentric whole-body motion diffusion with exemplar-based identity conditioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7095–7104 (2025) 3
49. Lee, J., Joo, H.: Mocap everyone everywhere: Lightweight motion capture with smartwatches and a head-mounted camera. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1091–1100 (2024) 3, 4
50. Li, C., Chibane, J., He, Y., Pearl, N., Geiger, A., Pons-Moll, G.: Unimotion: Unifying 3d human motion synthesis and understanding. In: 2025 International Conference on 3D Vision (3DV). pp. 240–249. IEEE (2025) 4
51. Li, C., Xie, X., Cao, Y., Geiger, A., Pons-Moll, G.: FrankenMotion: Part-level human motion generation and composition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2026) 4
52. Li, J., Clegg, A., Mottaghi, R., Wu, J., Puig, X., Liu, C.K.: Controllable human-object interaction synthesis. In: European Conference on Computer Vision. pp. 54–72. Springer (2024) 4
53. Li, J., Liu, K., Wu, J.: Ego-body pose estimation via ego-head pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17142–17151 (2023) 3, 5, 11
54. Li, J., Wu, J., Liu, C.K.: Object motion guided human motion synthesis. ACM Transactions on Graphics (TOG) **42**(6), 1–11 (2023) 2, 4, 11
55. Li, Q., Wang, J., Loy, C.C., Dai, B.: Task-oriented human-object interactions generation with implicit neural representations. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3035–3044 (2024) 4
56. Lin, L., Xia, S., Lai, Z., Sun, L., Yang, J., Pei, L.: Imu-hoi: A symbiotic framework for coherent human-object interaction and motion capture via contact-conscious inertial fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 42901–42910 (2026) 4
57. Liu, Y., Yang, J., Gu, X., Chen, Y., Guo, Y., Yang, G.Z.: Egofish3d: Egocentric 3d pose estimation from a fisheye camera via self-supervised learning. IEEE Transactions on Multimedia **25**, 8880–8891 (2023) 3
58. Liu, Y., Yang, J., Gu, X., Guo, Y., Yang, G.Z.: Egohmr: Egocentric human mesh recovery via hierarchical latent diffusion model. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 9807–9813. IEEE (2023) 3
59. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 10

60. Ma, L., Ye, Y., Hong, F., Guzov, V., Jiang, Y., Postyeni, R., Pesqueira, L., Gamino, A., Baiyya, V., Kim, H.J., et al.: Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In: European Conference on Computer Vision. pp. 445–465. Springer (2024) 3
61. Ma, M., Fan, H., Kitani, K.M.: Going deeper into first-person activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1894–1903 (2016) 3
62. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: International Conference on Computer Vision. pp. 5442–5451 (Oct 2019) 2, 9, 11
63. de Mayo, M., Cremers, D., Pire, T.: The monado slam dataset for egocentric visual-inertial tracking. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 13111–13118. IEEE (2025) 3
64. Meta EMG Wearable Technology (accessed February 7, 2026), https://www.meta.com/en-gb/emerging-tech/emg-wearable-technology 1
65. Mir, A., Puig, X., Kanazawa, A., Pons-Moll, G.: Generating continual human motion in diverse 3d scenes. In: 2024 International Conference on 3D Vision (3DV). pp. 903–913. IEEE (2024) 4
66. Mollyn, V., Arakawa, R., Goel, M., Harrison, C., Ahuja, K.: Imuposer: Full-body pose estimation using imus in phones, watches, and earbuds. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–12 (2023) 3
67. Nam, H., Jung, D.S., Moon, G., Lee, K.M.: Joint reconstruction of 3d human and object via contact-based refinement transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10218–10227 (2024) 4
68. Nazarenus, E., Li, C., He, Y., Xie, X., Lenssen, J.E., Pons-Moll, G.: Actionplan: Future-aware streaming motion synthesis via frame-level action planning. In: European Conference on Computer Vision. Springer (2026) 4
69. OpenRing: world's first Open-source Smart Ring (accessed February 7, 2026), https://o-ring.tech/ 1
70. Pan, B., Harley, A.W., Engelmann, F., Liu, C.K., Guibas, L.J.: Lookout: Real-world humanoid egocentric navigation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24977–24988 (2025) 3
71. Pan, X., Charron, N., Yang, Y., Peters, S., Whelan, T., Kong, C., Parkhi, O., Newcombe, R., Ren, Y.C.: Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20133–20143 (2023) 11, 12
72. Patel, C., Nakamura, H., Kyuragi, Y., Kozuka, K., Niebles, J.C., Adeli, E.: Uniegomotion: A unified model for egocentric motion reconstruction, forecasting, and generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10318–10329 (2025) 3
73. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 10975–10985 (2019) 6
74. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023) 8, 11

75. Peng, X., Xie, Y., Wu, Z., Jampani, V., Sun, D., Jiang, H.: Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2878–2888 (2025) 4
76. Petrov, I.A., Marin, R., Chibane, J., Pons-Moll, G.: Object pop-up: Can we infer 3d objects and their poses from human interactions alone? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023) 4, 11
77. Petrov, I.A., Marin, R., Chibane, J., Pons-Moll, G.: Tridi: Trilateral diffusion of 3d humans, objects, and interactions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025) 4, 7, 11
78. Project Aria (accessed February 7, 2026), https://www.projectaria.com/glasses/ 1
79. Qian, G., Li, Y., Peng, H., Mai, J., Hammoud, H., Elhoseiny, M., Ghanem, B.: Pointnext: Revisiting pointnet++ with improved training and scaling strategies. Advances in neural information processing systems **35**, 23192–23204 (2022) 5
80. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022) 7
81. Rempe, D., Birdal, T., Hertzmann, A., Yang, J., Sridhar, S., Guibas, L.J.: Humor: 3d human motion model for robust pose estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11488–11499 (2021) 6
82. Rhodin, H., Richardt, C., Casas, D., Insafutdinov, E., Shafiei, M., Seidel, H.P., Schiele, B., Theobalt, C.: Egocap: egocentric marker-less motion capture with two fisheye cameras. ACM Transactions on Graphics (TOG) **35**(6), 1–11 (2016) 3
83. Rogez, G., Supancic, J.S., Ramanan, D.: First-person pose recognition using egocentric workspaces. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4325–4333 (2015) 3
84. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6) (Nov 2017) 11
85. Shin, S., Pahuja, A., Richard, A., Kitani, K., Saragih, J., Chen, Y., Xu, W., Halilaj, E., Bagautdinov, T.: Egomdm: Diffusion-based human motion synthesis from sparse egocentric sensors. In: Thirteenth International Conference on 3D Vision (2026) 3
86. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015) 7
87. Song, W., Zhang, X., Li, S., Gao, Y., Hao, A., Hou, X., Chen, C., Li, N., Qin, H.: Hoianimator: Generating text-prompt human-object animations using novel perceptive diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 811–820 (2024) 4
88. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024) 8
89. Sárándi, I., Pons-Moll, G.: Neural localizer fields for continuous 3d human pose and shape estimation. Advances in Neural Information Processing Systems (NeurIPS) (2024) 11
90. Tang, J., Wang, J., Ji, K., Xu, L., Yu, J., Shi, Y.: A unified diffusion framework for scene-aware human motion estimation from sparse signals. In: Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21251–21262 (2024) 4
91. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. In: The Eleventh International Conference on Learning Representations (2023) 9
92. Tome, D., Alldieck, T., Peluse, P., Pons-Moll, G., Agapito, L., Badino, H., De la Torre, F.: Selfpose: 3d egocentric pose estimation from a headset mounted camera. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(6), 6794–6806 (2020) 3
93. Tripathi, S., Chatterjee, A., Passy, J.C., Yi, H., Tzionas, D., Black, M.J.: Deco: Dense estimation of 3d human-scene contact in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8001–8013 (2023) 4
94. Vlasic, D., Adelsberger, R., Vannucci, G., Barnwell, J., Gross, M., Matusik, W., Popović, J.: Practical motion capture in everyday surroundings. ACM transactions on graphics (TOG) **26**(3), 35–es (2007) 3
95. Von Marcard, T., Rosenhahn, B., Black, M.J., Pons-Moll, G.: Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In: Computer graphics forum. vol. 36, pp. 349–360. Wiley Online Library (2017) 3
96. Wang, J., Cao, Z., Luvizon, D., Liu, L., Sarkar, K., Tang, D., Beeler, T., Theobalt, C.: 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. pp. 777–787 (2024) 3
97. Winkler, A., Won, J., Ye, Y.: Questsim: Human motion tracking from sparse sensors with simulated avatars. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–8 (2022) 3
98. Xia, S., Zhang, Y., Su, Z., Zheng, X., Lv, Z., Wang, G., Zhang, Y., Wu, Q., Chu, L., Pei, L.: Envposer: Environment-aware realistic human motion estimation from sparse observations with uncertainty modeling. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1839–1849 (2025) 3
99. Xie, X., Bhatnagar, B.L., Lenssen, J.E., Pons-Moll, G.: Template free reconstruction of human-object interaction with procedural interaction generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10003–10015 (2024) 4
100. Xie, X., Bhatnagar, B.L., Pons-Moll, G.: Chore: Contact, human and object reconstruction from a single rgb image. In: European Conference on Computer Vision (ECCV). Springer (October 2022) 4
101. Xie, X., Bhatnagar, B.L., Pons-Moll, G.: Visibility aware human-object interaction tracking from single rgb camera. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2023) 4
102. Xie, X., Lenssen, J.E., Pons-Moll, G.: Intertrack: Tracking human object interaction without object templates. In: International Conference on 3D Vision 2025 (2025) 4
103. Xu, L., Zhou, Y., Yan, Y., Jin, X., Zhu, W., Rao, F., Yang, X., Zeng, W.: Regennet: Towards human action-reaction synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1759–1769 (2024) 4
104. Xu, S., Li, Z., Wang, Y.X., Gui, L.Y.: Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14928–14940 (2023) 4, 11

105. Xu, S., Schulter, S., Ziyadi, M., He, X., Fei, X., Wang, Y.X., Gui, L.Y.: Interprior: Scaling generative control for physics-based human-object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23516–23527 (2026) 4
106. Xu, S., Wang, Y.X., Gui, L., et al.: Interdreamer: Zero-shot text to 3d dynamic human-object interaction. Advances in Neural Information Processing Systems **37**, 52858–52890 (2024) 4
107. Xu, V., Gao, C., Hoffmann, H., Ahuja, K.: Mobileposer: Real-time full-body pose estimation and 3d human translation from imus in mobile consumer devices. In: Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology. pp. 1–11 (2024) 3
108. Xu, W., Chatterjee, A., Zollhoefer, M., Rhodin, H., Fua, P., Seidel, H.P., Theobalt, C.: $\mathrm{Mo}^2\mathrm{Cap}^2$: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. IEEE Transactions on Visualization and Computer Graphics pp. 1–1 (2019) 3
109. Yalandur Muralidhar, P., Xue, Y., Xie, X., Kostyrko, M., Pons-Moll, G.: Physic: Physically plausible 3d human-scene interaction and contact from a single image (2025) 4
110. Yang, J., Niu, X., Jiang, N., Zhang, R., Huang, S.: F-hoi: Toward fine-grained semantic-aligned 3d human-object interactions. In: European Conference on Computer Vision. pp. 91–110. Springer (2024) 4
111. Yang, Y., Zhai, W., Wang, C., Yu, C., Cao, Y., Zha, Z.J.: Egochoir: Capturing 3d human-object interaction regions from egocentric views. Advances in Neural Information Processing Systems **37**, 54529–54557 (2024) 1
112. Ye, Y., Li, J., Rong, R., Liu, C.K.: Whole: World-grounded hand-object lifted from egocentric videos. arXiv preprint arXiv:2602.22209 (2026) 4
113. Yi, B., Ye, V., Zheng, M., Li, Y., Müller, L., Pavlakos, G., Ma, Y., Malik, J., Kanazawa, A.: Estimating body and hand motion in an ego-sensed world. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7072–7084 (2025) 3, 5, 6, 9, 11, 12
114. Yi, X., Zhou, Y., Habermann, M., Shimada, S., Golyanik, V., Theobalt, C., Xu, F.: Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13167–13178 (2022) 3
115. Yi, X., Zhou, Y., Xu, F.: Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. ACM Transactions On Graphics (TOG) **40**(4), 1–13 (2021) 3
116. Yin, H., Liu, B., Kaufmann, M., He, J., Christen, S., Song, J., Hui, P.: Egohdm: A real-time egocentric-inertial human motion capture, localization, and dense mapping system. ACM Transactions on Graphics (TOG) **43**(6), 1–12 (2024) 3, 4
117. YM, P., Xue, Y., Chen, Y., Kister, N., Sárándi, I., Pons-Moll, G.: Graft: Geometric refinement and fitting transformer for human scene reconstruction (2026) 4
118. Yonemoto, H., Murasaki, K., Osawa, T., Sudo, K., Shimamura, J., Taniguchi, Y.: Egocentric articulated pose tracking for action recognition. In: International Conference on Machine Vision Applications (MVA) (2015) 3
119. Yoon, T., Han, Y., Ji, S., Park, J., Kim, S., Kwon, T., Kim, H.S.: Egoxtreme: A dataset for robust object pose estimation in egocentric views under extreme conditions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 40696–40706 (2026) 3

120. Zhang, J., Luo, H., Yang, H., Xu, X., Wu, Q., Shi, Y., Yu, J., Xu, L., Wang, J.: Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8834–8845 (2023) 4
121. Zhang, J., Zhang, J., Song, Z., Shi, Z., Zhao, C., Shi, Y., Yu, J., Xu, L., Wang, J.: Hoi-mˆ 3: Capture multiple humans and objects interaction within contextual environment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 516–526 (2024) 4
122. Zhang, S., Ma, Q., Zhang, Y., Qian, Z., Kwon, T., Pollefeys, M., Bogo, F., Tang, S.: Egobody: Human body shape and motion of interacting people from head-mounted devices. In: European conference on computer vision. pp. 180–200. Springer (2022) 3
123. Zhang, X., Bhatnagar, B.L., Starke, S., Petrov, I., Guzov, V., Dhamo, H., Pérez-Pellitero, E., Pons-Moll, G.: Force: Physics-aware human-object interaction. In: 2025 International Conference on 3D Vision (3DV). pp. 1473–1486. IEEE (2025) 4
124. Zhang, X., Starke, S., Guzov, V., Zhang, Z., Pellitero, E.P., Pons-Moll, G.: Scenic: Scene-aware semantic navigation with instruction-guided control. arXiv preprint arXiv:2412.15664 (2024) 4
125. Zhang, X., Kou, Z., Qin, C., Huang, M., Ristani, E., Kumar, A., Chen, L., He, K., Boularias, A., Guan, L.: Glove2hand: Synthesizing natural hand-object interaction from multi-modal sensing gloves. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1829–1840 (2026) 4
126. Zhao, C., Zhang, J., Du, J., Shan, Z., Wang, J., Yu, J., Wang, J., Xu, L.: I'm hoi: Inertia-aware monocular capture of 3d human-object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 729–741 (2024) 4
127. Zhao, Y., Ma, H., Kong, S., Fowlkes, C.: Instance tracking in 3d scenes from egocentric videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21933–21944 (2024) 13
128. Zheng, X., Su, Z., Wen, C., Xue, Z., Jin, X.: Realistic full-body tracking from sparse observations via joint-level modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14678–14688 (2023) 3, 11
129. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5745–5753 (2019) 5
130. Zuo, C., Wang, Y., Zhan, L., Guo, S., Yi, X., Xu, F., Qin, Y.: Loose inertial poser: Motion capture with imu-attached loose-wear jacket. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2209–2219 (2024) 3