I'M HOI: Inertia-aware Monocular Capture of 3D Human-Object Interactions

Chengfeng Zhao[1] Juze Zhang[1,2,3] Jiashen Du[1] Ziwei Shan[1] Junye Wang[1]
Jingyi Yu[1] Jingya Wang[1] Lan Xu[1,†]
[1]ShanghaiTech University [2]Shanghai Advanced Research Institute, Chinese Academy of Sciences
[3]University of Chinese Academy of Sciences

{zhaochf2022,zhangjz,dujsh2022,shanzw2022,wangjy22022,yujingyi,wangjingya,xulan1}@shanghaitech.edu.cn

Figure 1. Taking a monocular RGB video and a single inertial measurement unit (IMU) sensor recording, our approach, I'm-HOI, efficiently and robustly captures challenging and dynamic human-object interactions (HOI), such as skateboarding.

Abstract

We are living in a world surrounded by diverse and "smart" devices with rich modalities of sensing ability. Conveniently capturing the interactions between us humans and these objects remains far-reaching. In this paper, we present I'm-HOI, a monocular scheme to faithfully capture the 3D motions of both the human and object in a novel setting: using a minimal amount of RGB camera and object-mounted Inertial Measurement Unit (IMU). It combines general motion inference and category-aware refinement. For the former, we introduce a holistic human-object tracking method to fuse the IMU signals and the RGB stream and progressively recover the human motions and subsequently the companion object motions. For the latter, we tailor a category-aware motion diffusion model, which is conditioned on both the raw IMU observations and the results from the previous stage under over-parameterization representation. It significantly refines the initial results and generates vivid body, hand, and object motions. Moreover, we contribute a large dataset with ground truth human and object motions, dense RGB inputs, and rich object-mounted IMU measurements. Extensive experiments demonstrate the effectiveness of I'm-HOI under a hybrid capture setting. Our dataset and code will be released to the community.

†Corresponding author

1. Introduction

Capturing human-object interactions (HOI) is essential to understanding how we humans connect with the surrounding world, with numerous applications in robotics, gaming, or VR/AR. Yet, an accurate and convenient solution remains challenging in the vision community.

For high-fidelity capture of human-object interactions, early high-end solutions [6, 9, 29] require dense cameras, while recent approaches [4, 10, 26, 70] require less RGB or RGBD video inputs (from 3 to 8 views). Yet, such a multi-view setting is still undesirable for consumer-level daily usage. Instead, the monocular method with more handiest captured devices is more attractive. Specifically, most recent methods [80, 82, 83, 101, 104] track the rigid and skeletal motions of objects and humans using a pre-scanned template or parametric model [46] from a single RGB video input. Yet, inherently due to the RGB-setting, they remain vulnerable to depth ambiguity and the occlusion between human and object, especially for handling challenging fast motions like skateboarding. In contrast, the Inertial Measurement Units (IMUs) serve as a rescue to provide motion information that is robust to occlusion. Actually, IMU-based motion capture is widely adopted in both industry [84] and academia [22, 60, 61, 90]. Recent methods [45, 53] further combine monocular RGB video and sparse IMUs, enabling lightweight and robust human motion capture. However, these schemes mostly focus on human-only scenarios and ignore the interacted objects. Moreover, compared to the sometimes tedious requirement of body-worn IMUs, it's more natural and convenient to attach the IMU sensor to the captured object, since IMUs have been widely integrated into daily objects like phones and smartwatches. Researchers surprisingly pay less attention to capturing human-object interactions with a minimal amount of RGB camera and IMU. The lack of motion data under rich interactions and modalities also constitute barriers to exploring such directions.

In this paper, we propose *I'm-HOI* – an inertia-aware and monocular approach for robustly tracking both the 3D human and object under challenging interactions (see Fig. 1). In stark contrast to prior arts, I'm-HOI adopts a lightweight and hybrid setting: a minimal amount of RGB camera and object-mounted IMU. Given the expected technological trend of mobile sensing as more and more RGB cameras and IMUs will be integrated into our surrounding devices, we believe that our approach will serve as a viable alternative to traditional human-object motion capture.

In I'm-HOI, our key idea is to adopt a two-stage paradigm to make full use of both the object-mounted IMU signals and the RGB stream, which consists of general motion inference and category-aware motion refinement. For the former stage, we introduce holistic human-object tracking in an end-to-end manner. Specifically, we generate human motions via a multi-scale CNN-based network for 3D keypoints, followed by an Inverse Kinematics (IK) optimization layer. To reason the companion object motions, we progressively fuse the human features with IMU measurements via object-orientated mesh alignment feedback. We also adopt a robust optimization to refine the tracked object pose and improve the overlay performance, especially when the object is invisible in the RGB input. For the second refinement stage, we propose to tailor the conditional motion diffusion models [20, 42] for utilizing category-level interaction priors. During training the diffusion model corresponding to a certain object, we treat the tracked motions and the raw IMU measurements from the previous stage as the condition information. We also adopt a novel over-parameterization representation with extra regularization designs to jointly consider the body, object, and especially hand regions during the denoising process. Thus, our refinement stage not only projects the initial human-object motions onto the category-specific motion manifold but also infills possible hand motions for vividly capturing human-object interactions. To train and evaluate our I'm-HOI, we contribute a large multi-modal dataset of human-object interactions, covering 295 interaction sequences with 10 diverse objects in total 892k frames of recording. We also provide ground truth body, hand, and 3D object meshes, with dense RGB inputs and rich object-mounted IMU measurements. To summarize, our main contributions include:

- We propose a multi-modal method to jointly capture human and object motions from a minimal amount of RGB camera and object-mounted IMU sensor.
- We adopt an efficient holistic human-object tracking method to progressively fuse the motion features, companion with a conditional diffusion model to refine and generate vivid interaction motions.
- We contribute a large dataset for human-object interactions, with rich RGB/IMU modalities and ground-truth annotations. Our data and model will be disseminated to the community.

2. Related Work

Monocular Human-centric Capture. Since the release of the parametric body model SMPL [46, 55, 65], there has been tremendous progress [5, 31, 35–39, 41, 49, 52, 54, 63, 72, 86, 97, 98, 100] in human motion capture from single RGB images and videos. However, reconstructing contextual human-object and human-scene interactions (HOI/HSI) from monocular input is far more challenging. The pioneer work PHOSA [101] proposes a purely optimization-based framework to estimate static human-object spatial arrangements relying on handcrafted contact heuristics. This approach is unscalable and error-prone to depth ambiguities. Benefited from emerging 3D interaction motion datasets [4, 7, 11, 14, 15, 21, 23, 25, 66, 81, 106], learning-and-optimization work [24, 82] has shown promising results by modeling human-object relative distance field in data-driven manner, followed by joint post-optimization. The state-of-the-art video-based method, VisTracker [83], further incorporates motion infilling techniques [93] to enable space-time coherent tracking. However, these approaches still suffer from unacceptable runtime costs and unsatisfying accuracy under complex interaction scenarios.

Inertial and Multi-modal Motion Capture. Complementary to vision-based methods, human motion capture using inertial measurement units (IMUs) has also been extensively studied. Previous commercial solutions [51, 84] can capture accurate and detailed motion with dense sensors. Since the exploration of SIP [78], data-driven methods under sparse sensors configurations [22, 27, 76, 90, 91] have been developed to achieve real-time performance, deriving consumer-level products [69]. To address the limitations

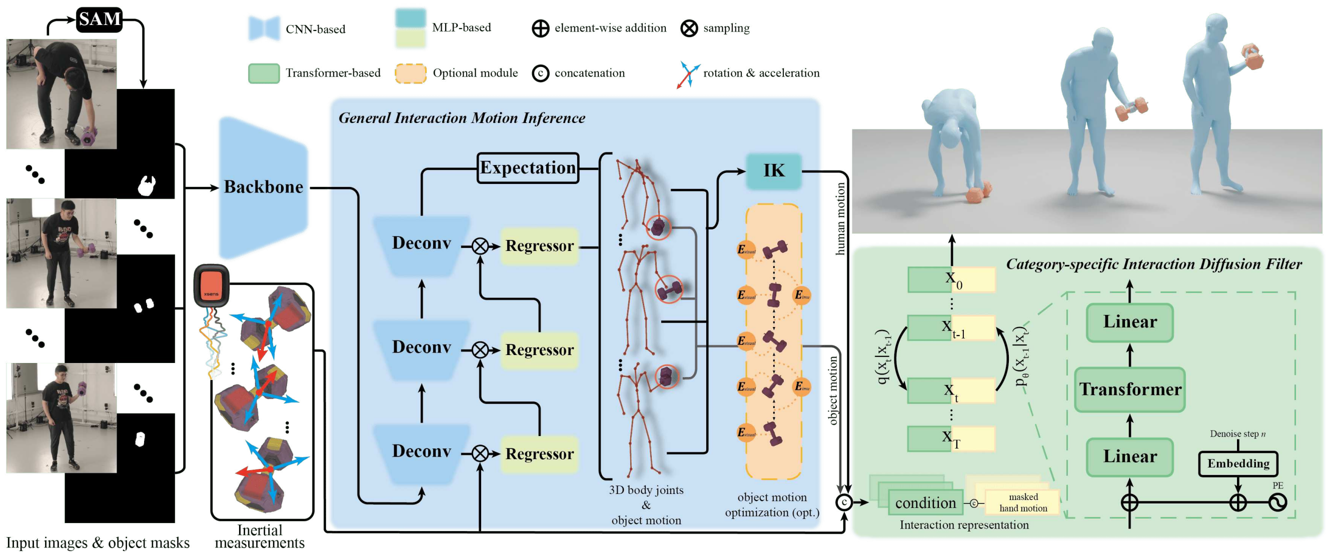

Figure 2. The pipeline of I'm-HOI. Assuming video and inertial measurements input, our approach consists of a general interaction motion inference module (Sec. 3.1) and a category-specific interaction diffusion filter (Sec. 3.2) to capture challenging interaction motions.

of single-modal systems, multi-modal approaches [17] fuse inertial signals with RGB [12, 19, 30, 45, 47, 48, 53, 59, 60, 75, 77, 79, 107], RGBD [18, 109], ego-view [92], and LiDAR [64] references, achieving balanced local pose estimation and global localization. In this work, we extend the multi-sensor fusion strategy to 3D human-object interactions capture, which is beneficial for both accuracy and efficiency.

Object-specific Interaction Prior. Human motion priors have been demonstrated crucial for realism in the capture and synthesis by multiple modeling methodologies, including predefined kinematic structure [2, 110], GMM [55], GAN [3, 13, 31], VAE [33, 57, 63, 102], MLP [74] and more cutting-edge, diffusion models [20, 32, 67, 68, 73, 94]. Additionally, context-aware human motion synthesis [50, 103, 108] and scene placement generation [88, 89] successfully extract contextual prior knowledge from data. More recent work [43, 56, 85] modeled dynamic interaction patterns but ignored object category-level distribution differences. Other methods [40, 105] focus on specific interactions with chairs, which are static and lack diversity. This work aims to learn object category-specific interaction prior to model dynamic interaction distributions between human and diverse objects.

3. Method

We present a new paradigm for 3D human-object interactions capture in a lightweight and hybrid setting: utilizing a minimal amount of RGB camera and object-mounted Inertial Measurement Unit (IMU). As illustrated in Figure 2, we propose a general interaction motion inference module (Sec. 3.1) to jointly recover human-object spatial arrangements in an end-to-end fashion. An category-specific interaction diffusion filter (Sec. 3.2) is tailored to refine capture results from the former with the learned object category-level prior.

3.1. General Interaction Motion Inference

Current vision-based methods [82, 83] typically adhere to fitting-learning-optimization framework, which we have observed to be susceptible to substantial or prolonged human-object occlusions, inefficient in inference time and limited in generalization capabilities, as discussed in Sec. 5.2. In contrast, we treat the object as an additional body joint and propose to estimate human-object spatial arrangements holistically and end-to-end. An optional optimization procedure can be incorporated to enhance capture accuracy further.

Preprocessing. Given a monocular image sequence $\boldsymbol{I} \in \mathbb{R}^{T\times h\times w\times 3}$, we first segment human and object mask $\boldsymbol{S}_h, \boldsymbol{S}_o \in \mathbb{R}^{T\times h\times w\times 1}$ separately using SAM [34]. Following that, a pre-trained ResNet-34 [16] image encoder is adopted to extract image feature from stacked RGB image and object mask. After that, We take the raw inertial rotation $\boldsymbol{Q} \in \mathbb{R}^{T\times 6}$, acceleration $\boldsymbol{A} \in \mathbb{R}^{T\times 3}$ and normalized object template $\mathcal{O}$, combined with $\boldsymbol{I}, \boldsymbol{S}_o$ as our network input. Our approach outputs human shape $\boldsymbol{\beta} \in \mathbb{R}^{T\times 10}$ and pose $\boldsymbol{\theta} \in \mathbb{R}^{T\times 3N_{J_b}}$, object rotation $\boldsymbol{R}_o \in \mathbb{R}^{T\times 6}$ and translation $\boldsymbol{T}_o \in \mathbb{R}^{T\times 3}$. Here, $T = 64$ is the sequence length, $h\times w$ is the resolution of images and masks, $N_{J_b} = 22$ is the number of body joints. We adopt standard SMPL model [46] for human motion representation.

End-to-end Holistic Human-Object Tracking. We first introduce a multi-scale CNN-based network to jointly detect 3D human body joints $\boldsymbol{J} \in \mathbb{R}^{T\times 3N_{J_b}}$ and the object center. Leveraging the extracted image feature, we feed it into a series of deconvolution layers followed by a final convolution layer to reconstruct 3D keypoint heatmaps. The 3D keypoint positions are then determined by the expectation of each heatmap [71], with all keypoints canonicalized to root-relative representation except for the root joint. Commonly used combination of 3D keypoints and 2D reprojection loss $\mathcal{L}_{\text{kp3d}} + \lambda_{\text{j2d}}\mathcal{L}_{\text{j2d}}$ is utilized to train the CNN network, and simultaneously fine-tune the image extractor. Subsequent to 3D keypoint estimation, we employ and fine-tune an off-the-shelf pre-trained inverse kinematics layer [100] by $\mathcal{L}_{\text{twist}}$ to recover human pose $\boldsymbol{\theta}$ and shape $\boldsymbol{\beta}$ based on $\hat{\boldsymbol{J}}$. Please refer to [41, 82, 83, 100] for detailed loss functions.

For object tracking, we get an initially-posed object through $\mathcal{C}(\boldsymbol{Q})\mathcal{O} + \hat{\boldsymbol{T}}_o$ with estimated object translation and raw rotation data from the object-mounted IMU sensor, where $\mathcal{C}(\cdot)$ is the mapping from 6D rotation representation [111] to rotation matrix. In order to eliminate systematic biases in $\boldsymbol{Q}$ and correct inaccurate $\hat{\boldsymbol{T}}_o$ under occlusions, we attach one MLP-based regressor to each intermediate image feature grid [97, 98], which forms a feedback loop to estimate corrective increment of object motion progressively. In the i-th loop, we first uniformly sample $N_S = 400$ vertices on the posed object mesh $\mathcal{C}(\hat{\boldsymbol{R}}_o^{(i)})\mathcal{O} + \hat{\boldsymbol{T}}_o^{(i)}$. Then, we project sampled vertices onto the i-th feature map to obtain object mesh-aligned feature, which is subsequently fed into the i-th regressor to predict $\Delta\hat{\boldsymbol{R}}_o^{(i)}$ and $\Delta\hat{\boldsymbol{T}}_o^{(i)}$. Particularly, $\hat{\boldsymbol{R}}_o^{(0)} = \mathcal{C}(\boldsymbol{Q})$. The training loss of the feedback network group is defined as $\mathcal{L}_{\text{maf}} = \lambda_{\text{occ-sil}}\mathcal{L}_{\text{occ-sil}} + \lambda_{\text{area}}\mathcal{L}_{\text{area}}$, where $\mathcal{L}_{\text{occ-sil}}$ is the occlusion-aware silhouette loss proposed in [101]. We find that better results can be achieved with an augmented silhouette area loss:

$$\mathcal{L}_{\text{area}} = \frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=0}^{N_F-1} ||\sum \mathcal{D}(\hat{\boldsymbol{R}}_{o,t}^{(i)}\mathcal{O} + \hat{\boldsymbol{T}}_{o,t}^{(i)}) - \sum \boldsymbol{S}_{o,t}||_2^2, \quad (1)$$

where $\mathcal{D}(\cdot)$ refers to differentiable rendering function [28] and $N_F = 3$ is the number of feedback iterations. Here, we re-define $\hat{\boldsymbol{R}}_o = \hat{\boldsymbol{R}}_o^{N_F-1}$ and $\hat{\boldsymbol{T}}_o = \hat{\boldsymbol{T}}_o^{N_F-1}$.

Overall, our training loss of the end-to-end inference module is:

$$\mathcal{L} = \mathcal{L}_{\text{kp3d}} + \lambda_{\text{j2d}}\mathcal{L}_{\text{j2d}} + \mathcal{L}_{\text{twist}} + \mathcal{L}_{\text{maf}}. \quad (2)$$

It's noteworthy that human mask $\boldsymbol{S}_h$ is only required in $\mathcal{L}_{\text{occ-sil}}$ during training but not when testing.

Robust and Lightweight Optimization. To improve object tracking precision, especially in invisible cases, we propose an ***optional*** optimization module. In addition to visual cues, we further constraint object rotation and trajectory to inertial measurements. The energy function is formulated as:

$$\mathcal{E} = \mathcal{E}_{\text{visual}} + w_{\text{imu}}\mathcal{E}_{\text{imu}}. \quad (3)$$

Specifically, $\mathcal{E}_{\text{visual}}$ minimizes the discrepancy between the rendering result and the segmentation of object: $\mathcal{E}_{\text{visual}} = \frac{1}{T}\sum_{t=0}^{T-1}\sum ||\mathcal{D}(\hat{\boldsymbol{R}}_{o,t}\mathcal{O} + \hat{\boldsymbol{T}}_{o,t}) - \boldsymbol{S}_{o,t}||_2^2$. At the same time, $\mathcal{E}_{\text{imu}}$ regularizes object motion temporally:

$$\mathcal{E}_{\text{imu}} = \frac{1}{T-1}\sum_{t=1}^{T-1}||(\hat{\boldsymbol{T}}_{o,t-1} + \hat{\boldsymbol{T}}_{o,t+1} - 2\hat{\boldsymbol{T}}_{o,t}) - 0.5\boldsymbol{A}_t^2||_2^2 + \frac{1}{T}\sum_{t=0}^{T-1}||\hat{\boldsymbol{R}}_{o,t} - \boldsymbol{Q}_t||_2^2. \quad (4)$$

3.2. Category-specific Interaction Diffusion Filter

In the second refinement stage, a category-specific interaction motion filter is proposed to *(i)* project capture results from the preceding stage onto the manifold; *(ii)* infill hand motions conditioned on body-object interaction motions.

Interaction Representation. We propose a novel over-parameterization interaction representation containing human, object motion and raw inertial measurements. At timestamp t and noise level n, $\boldsymbol{x}_t^n \in \mathbb{R}^{486}$ consists of body-hand joint positions $\boldsymbol{j}_{h,t} \in \mathbb{R}^{156}$ and rotations $\boldsymbol{\theta}_{h,t} \in \mathbb{R}^{312}$; object translation $\boldsymbol{j}_{o,t} \in \mathbb{R}^3$ and rotation $\boldsymbol{\theta}_{o,t} \in \mathbb{R}^6$; inertial rotation $\boldsymbol{q}_t \in \mathbb{R}^6$ and free acceleration signal $\boldsymbol{a}_t \in \mathbb{R}^3$. We use 6D representation [111] for all the rotation data, and the 52 joints body-hand model SMPL-H [65] is adopted. The target interaction motion is represented as $\boldsymbol{x}_0$.

Conditional Diffusion Denoising Process. Given $\boldsymbol{x}_0$, forward diffusion process adds Gaussian noise iteratively along an N-step Markov chain. For each noising step n, the noised interaction motion is drawn from conditional probability distribution determined by a pre-defined schedule $\{\alpha_n\}_{n=1}^N$:

$$q(\boldsymbol{x}_{1:T}^n|\boldsymbol{x}_{1:T}^{n-1}) = \mathcal{N}(\sqrt{\alpha_n}\boldsymbol{x}_{1:T}^{n-1}, (1-\alpha_n)\mathcal{I}). \quad (5)$$

In reverse process, we formulate condition information as a tuple $\boldsymbol{c} = (\boldsymbol{j}_{h_b}, \boldsymbol{j}_o, \boldsymbol{\theta}_{h_b}, \boldsymbol{\theta}_o, \boldsymbol{q}, \boldsymbol{a}) \in \mathbb{R}^{216}$ and concatenate it with masked hand motion $\boldsymbol{m} = (\boldsymbol{j}_{h_h}, \boldsymbol{\theta}_{h_h}) \in \mathbb{R}^{270}$, where $\boldsymbol{j}_{h_b} \in \mathbb{R}^{66}$, $\boldsymbol{\theta}_{h_b} \in \mathbb{R}^{132}$ represents body-only joint positions and rotations. We follow [62] to predict $\boldsymbol{x}_0$ itself as $\hat{\boldsymbol{x}}_\phi(\boldsymbol{x}_n, n, \boldsymbol{c})$, where ϕ represents the parameters of neural network. The training loss is $L1$-norm simple objective [20, 42]:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{\boldsymbol{x}_0,n}||\hat{\boldsymbol{x}}_\phi(\boldsymbol{x}_n, n, \boldsymbol{c}) - \boldsymbol{x}_0||_1. \quad (6)$$

Inspired by [63], components inside of the over-parameterization representation have mutual constraints. We accordingly introduce four regularization terms:

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{off}}\mathcal{L}_{\text{off}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{consist}}\mathcal{L}_{\text{consist}} + \lambda_{\text{imu}}\mathcal{L}_{\text{imu}}. \quad (7)$$

Specifically, $\mathcal{L}_{\text{off}}$ enforces the predicted object center to lie in a small region determined by the distance offsets relative to 52 body-hand joints as:

$$\mathcal{L}_{\text{off}} = \frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=0}^{N_J}||(\hat{\boldsymbol{j}}_{o,t} - \hat{\boldsymbol{j}}_{h,t}^{(i)}) - (\boldsymbol{j}_{o,t} - \boldsymbol{j}_{h,t}^{(i)})||_1. \quad (8)$$

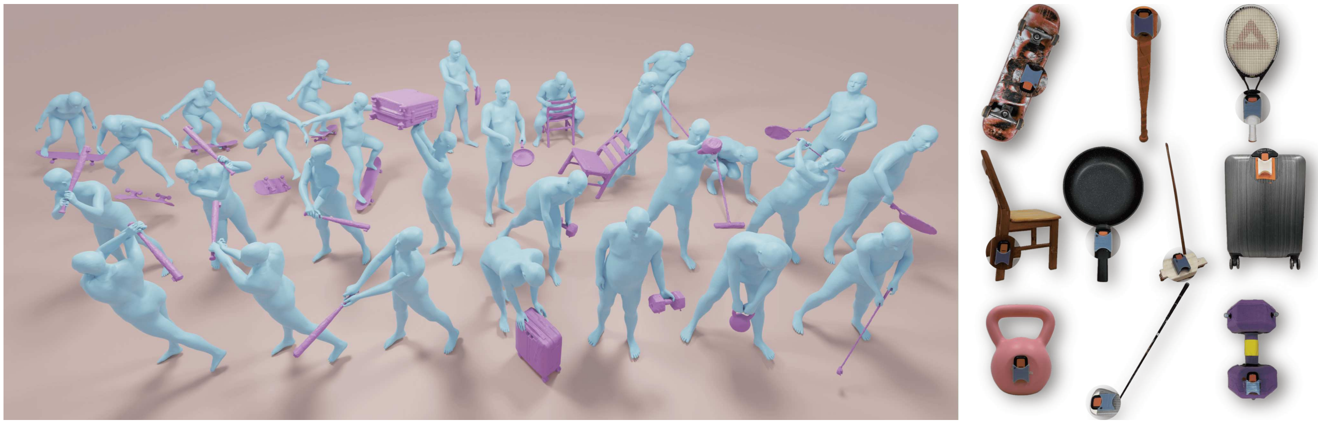

Figure 3. We exhibit selected highlights of IMHD^2 on the left side, and 10 well-scanned objects on the right side. In total, our dataset comprises 295 sequences and captures approximately 892k frames of data.

We then constraint the reproduced human body-hand joints consistent with the body model skinned from predicted joint rotations:

$$\mathcal{L}_{\text{consist}} = \frac{1}{T}\sum_{t=0}^{T-1} ||\hat{\boldsymbol{j}}_{h,t} - \mathcal{J}(\mathcal{M}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}_{h,t}))||_1, \tag{9}$$

where $\mathcal{M}(\cdot)$ refers to forward function of SMPL-H model and $\mathcal{J}(\cdot)$ is the joint regressor. In order to temporally smooth human motion, the velocity term $\mathcal{L}_{\text{vel}}$ is formulated as:

$$\mathcal{L}_{\text{vel}} = \frac{1}{T-1}\sum_{t=1}^{T-1} ||(\hat{\boldsymbol{j}}_t - \hat{\boldsymbol{j}}_{t-1}) - (\boldsymbol{j}_t - \boldsymbol{j}_{t-1})||_1, \tag{10}$$

where $\boldsymbol{j}_t = [\boldsymbol{j}_{h,t}, \boldsymbol{j}_{o,t}]$. Finally, $\mathcal{L}_{\text{imu}}$ guides the generated object poses and the trajectory conform to IMU measurements, which improves robustness under invisible scenarios:

$$\mathcal{L}_{\text{imu}} = \mathcal{L}_{\text{rot}} + \frac{1}{T-1}\sum_{t=1}^{T-1} \mathcal{L}_{\text{acc},t}. \tag{11}$$

Wherein, object rotation is directly regularized by:

$$\mathcal{L}_{\text{rot}} = \frac{1}{T}\sum_{t=0}^{T-1} ||\hat{\boldsymbol{\theta}}_{o,t} - \boldsymbol{q}_t||_1, \tag{12}$$

and the trajectory is constrained to acceleration through:

$$\mathcal{L}_{\text{acc},t} = ||(\hat{\boldsymbol{j}}_{o,t} - \hat{\boldsymbol{j}}_{o,t-1} + \frac{\boldsymbol{a}_t\tau^2}{2}) - (\boldsymbol{j}_{o,t+1} - \boldsymbol{j}_{o,t})||_1, \tag{13}$$

where τ is the time interval between two consecutive frames. It's noteworthy that, related work [83] simulates such second-order constraints in a pseudo manner [96] to eliminate mutation of first-order signals. In contrast, we incorporate acceleration explicitly [45, 64].

4. Dataset

To train and evaluate our I'm-HOI, we collect an ***I**nertial and **M**ulti-view **H**ighly **D**ynamic human-object interactions **D**ataset* (***IMHD***^{2}), consisting of human, object motions, inertial measurements and object 3D scans.

Capture Preparations. A high-end multi-view camera system consisting of 32 Z CAMs [95] was set up to capture 4K videos at 60 fps. Simultaneously, two Xsens DOT IMU sensors [84] mounted on the object and the leg of performer were used to record object inertia and align timestamps at 60 Hz respectively. We invited 15 subjects (13 males, 2 females) to participate in 10 different interaction scenarios. Sequence-level textual guidance was provided for each capture split to ensure reasonable and meaningful interactions. Each split lasted from half a minute to one minute. We conducted visual-inertial system calibration once per ten minutes to eliminate disturbances caused by magnetic field changes.

Data Processing. Given multi-view videos, we reproduced human motions in SMPL-H format [65] using an open-source toolbox [1]. To accurately track object pose in a 3D scene, we manually annotated single key-frame segmentation in all views and broadcasted it to the entire sequence [8, 34, 44, 87]. Subsequently, we optimized Euclidean transformations, which precisely define object motions, by fitting reprojected silhouette to multi-view masks. For object geometries, we utilized a public application [58] to obtain 3D scan templates. In terms of inertial signals, we adopted primitive rotation data $\boldsymbol{R}_s$ in matrix form and transformed raw acceleration $\boldsymbol{a}_{raw}$ in sensor coordinate to free acceleration $\boldsymbol{a}_{free}$ in global coordinate through $\boldsymbol{a}_{free} = \boldsymbol{R}_s\boldsymbol{a}_{raw} - \boldsymbol{g}$, where $\boldsymbol{g} = [0, 0, 9.81]^{\mathrm{T}}$ is the gravitational acceleration.

5. Experiments

In this section, we first introduce the datasets and metrics used for training and evaluation. We then provide a comprehensive comparison between our approach with baseline methods. We also perform extensive ablation studies to demonstrate the effectiveness of pivot components in our network design and the necessity of the IMU modality.

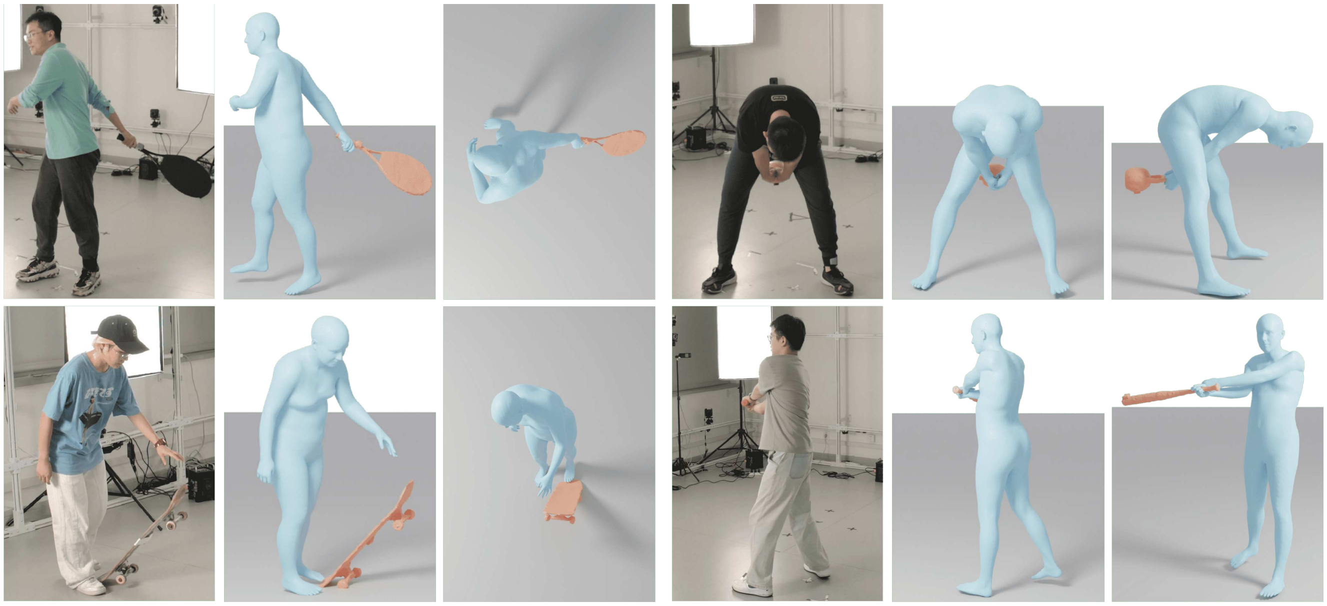

Figure 4. Qualitative 3D capturing results of I'm-HOI on IMHD^2 dataset. Each sample includes an RGB image input, captured motion from camera view, and side-view visualization.

	IMHD^2 (fast)				BEHAVE (slow to medium) [4]				InterCap (slow) [23]			
	CD (per-frame)		CD (10s)		CD (per-frame)		CD (10s)		CD (per-frame)		CD (10s)	
Method	smpl	object	smpl	object	smpl	object	smpl	object	smpl	object	smpl	object
PHOSA [101]	29.20	20.26	41.26	56.80	12.86	26.90	27.01	59.08	11.20	20.57	24.16	43.06
CHORE [82]	14.20	16.81	24.32	31.76	5.55	10.02	18.33	20.32	7.12	12.59	16.11	21.05
VisTracker [83]	19.96	23.28	17.02	18.10	**5.25**	8.04	7.81	8.49	6.76	10.32	9.35	11.38
Ours	**6.50**	**6.93**	**5.36**	**8.53**	5.26	**7.43**	**5.65**	**4.82**	**5.66**	**8.92**	**5.81**	**7.14**

Table 1. Quantitative comparison was conducted with several baselines on both human and object tracking accuracy.

5.1. Datasets and Evaluation Metrics

We train I'm-HOI using BEHAVE [4], InterCap [23] and IMHD^2, and evaluate it on five datasets which also include HODome [99] and CHAIRS [25]. We adhere to the official train-test data partitioning of BEHAVE and InterCap, which is established by VisTracker [83]. Given the relatively slow inference speeds of baselines [82, 83, 101], we curate partial yet representative data from IMHD^2, HODome, and CHAIRS to construct sub test sets for thorough evaluation.

Evaluation Metrics.

- **Per-frame Chamfer Distance** (cm) [82] computes the chamfer distance between predicted human and object mesh with the ground truth respectively after holistic Procrustes alignment for every single frame.
- **Sliding Window Chamfer Distance** (cm) [83] computes the chamfer distance in the same way but performing holistic Procrustes alignment on the combined mesh of 10-second results with the ground truth.

5.2. Comparison

Results. As shown in Table 1, I'm-HOI consistently outperforms the baselines on several datasets, especially on IMHD^2 which is characterized by fast interaction motions, with a large margin around 15cm. We visualize qualitative results in Figure 5, I'm-HOI captures better human-object spatial arrangements, including both relative pose and position. In addition, our approach shows better robustness than baselines under severe occlusions.

Generalization. To assess the generalization capabilities, we evaluate the performance of purely optimization-based method PHOSA [101], learning-and-optimization methods CHORE[82] and VisTracker[83], as well as our proposed approach trained on BEHAVE[4] and InterCap[23], across HODome[99] and CHAIRS[25]. As shown in Table 2, I'm-HOI generalizes better than the baselines by a large margin and achieves more balanced performance between per-frame and sequential results. Furthermore, Figure 5 demonstrates

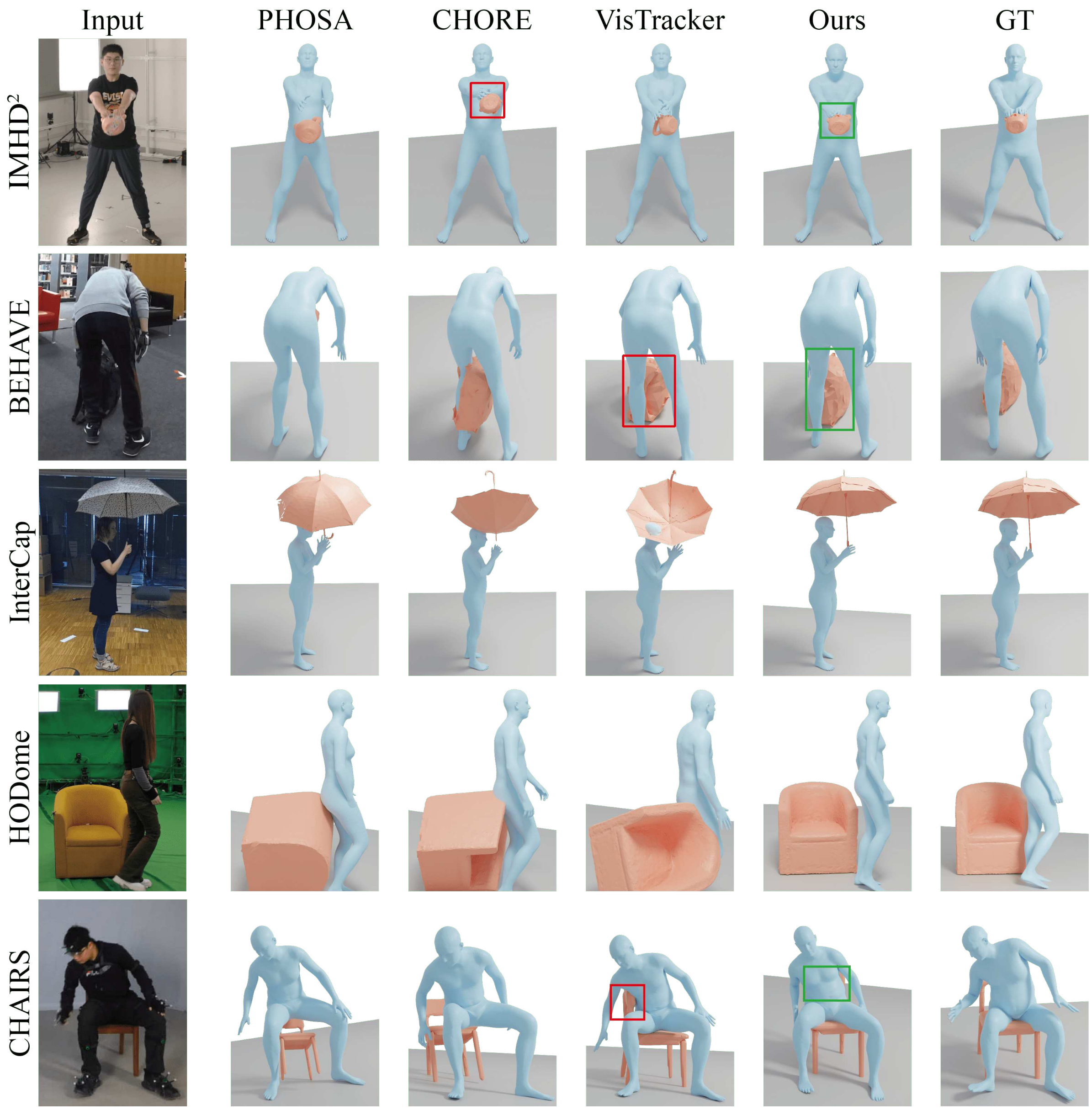

Figure 5. Qualitative comparison results. I'm-HOI outperforms the baselines and generalizes well to new datasets.

the adaptability to diverse scenarios of I'm-HOI.

Runtime Cost. We conduct a comparative analysis of the inference efficiency across different methods using a specific sequence from InterCap dataset [23]. Among the methods evaluated, the purely optimization-based framework PHOSA [101] takes the longest inference time which is approximately 2 minutes per frame. CHORE [82] speeds up to 1 minute per frame, while VisTracker [83] further reduces the time cost to 20 seconds. Notably, I'm-HOI requires only about 0.5 seconds per frame for the complete pipeline. It is worth mentioning that omitting the optional optimization module could lead to additional enhancements in efficiency.

5.3. Ablation Study

Extensive ablation studies are conducted on IMHD^2 to evaluate our network architecture design and IMU modality.

Network Architecture. Table 3 shows the performance of models with and without the mesh alignment feedback (maf.), optimization module (optim.) and diffusion filter (filter.). It is demonstrated that maf. improves per-frame object tracking results and optim. brings better temporal consistency. In addition, filter. further corrects human-object spatial arrangements onto the learned interaction manifold. Compared to the naive implementation, the full pipeline of

Dataset	Method	CD (per-frame) smpl	CD (per-frame) object	CD (10s) smpl	CD (10s) object
HODome [99]	PHOSA [101]	34.41	29.70	60.15	54.98
	CHORE [82]	23.18	16.18	43.35	31.64
	VisTracker [83]	11.87	19.86	32.77	34.53
	Ours	**8.19**	**9.05**	**12.07**	**15.31**
CHAIRS [25]	PHOSA [101]	35.26	28.35	43.17	37.67
	CHORE [82]	19.10	36.13	16.71	52.95
	VisTracker [83]	17.42	23.31	15.23	16.85
	Ours	**9.55**	**9.91**	**6.34**	**7.85**

Table 2. Quantitative evaluations of generalization ability.

Module			CD (per-frame)		CD (10s)	
maf.	optim.	filter.	smpl	object	smpl	object
✗	✗	✗	8.42	15.12	13.88	27.85
✓	✗	✗	8.02	9.73	10.31	19.87
✗	✗	✓	37.35	35.14	48.90	65.03
✓	✓	✗	7.16	7.42	7.25	10.75
✓	✗	✓	7.52	8.80	8.56	12.41
✓	✓	✓	**6.50**	**6.93**	**5.36**	**8.53**

Table 3. Quantitative evaluation of network architecture design.

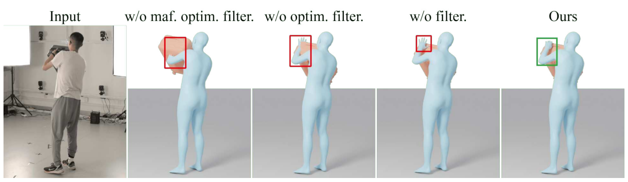

Figure 6. Qualitative evaluation of our network architecture. The figure illustrates the effectiveness of each key design.

I'm-HOI performs 4 times better. Figure 6 illustrates that inaccurate prediction is progressively corrected when maf. and optim. are applied. Also, the hand motion generated by filter. makes the capture result more vivid and realistic.

Input Modality. We experiment on several different baselines with IMU modality input by adding an additional inertial optimization term described in Equation 4 to their pipelines. The qualitative results shown in Figure 7 clearly demonstrate the improvements in object pose estimation of the baselines after introducing the IMU modality, compared to Figure 5. The quantitative performance reported in Table 4 shows a decent increase in the performance of baselines, compared to the statistics in Table 1. Additionally, we observe that our approach achieves better results when the IMU modality is involved, especially for object tracking. Furthermore, Table 4 shows that naively incorporating the IMU modality input into baselines is unable to maximize its benefits, which further verifies the effectiveness of our network design.

5.4. Limitation

The proposed I'm-HOI is the first trial to explore challenging 3D human-object interactions capture using a minimal amount of RGB camera and object-mounted IMU sensor.

Modality	CD (per-frame) smpl	CD (per-frame) object	CD (10s) smpl	CD (10s) object
PHOSA+imu	28.41	18.60	39.19	38.44
CHORE+imu	12.98	11.92	22.09	23.31
VisTracker+imu	15.87	14.61	14.68	12.82
Ours w/o imu	10.58	14.49	6.55	15.23
Ours	**6.50**	**6.93**	**5.36**	**8.53**

Table 4. Quantitative evaluations on input modality configurations.

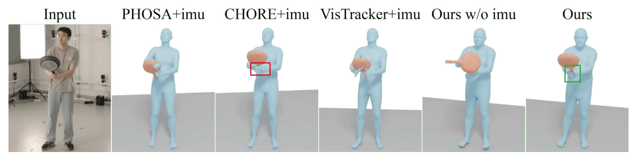

Figure 7. Qualitative evaluation of the IMU modality. This figure shows the importance of inertial measurements input.

However, it still has limitations. Firstly, our method relies on pre-scanned object templates and manual manipulation of sensor-template coordinate alignment. Additionally, our method is restricted to rigid object tracking. Extending this method to articulated or even deformable objects in a template-free framework is promising.

6. Conclusion

We have presented a novel and monocular scheme to faithfully capture the 3D motions of human-object interactions, using a minimal amount of RGB camera and object-mounted IMU. Our general motion inference stage progressively fuses the IMU signals and the RGB stream via holistic and end-to-end tracking, which efficiently recovers the human motions and subsequently the companion object motions via mesh alignment feedback. Our category-aware motion diffusion further treats the previous results as conditions and jointly considers the body, object, and especially hand regions during the denoising process with an over-parameterization representation. It encodes category-aware motion priors, so as to significantly improve the tracking accuracy and generate vivid hand motions. Our experimental results demonstrate the effectiveness of I'm-HOI for faithfully capturing human and object motions in a lightweight setting. As more and more sensors like RGB cameras or IMUs will be integrated into our surrounding world, we believe that our approach and dataset will serve as a critical step towards hybrid human-object motion capture, with many potential applications in robotics, embodied AI, or VR/AR.

Acknowledgement. This work was supported by National Key R&D Program of China (2022YFF0902301), Shanghai Local college capacity building program (22010502800). We also acknowledge support from Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI).

References

[1] Easymocap - make human motion capture easier. Github, 2021.

[2] Ijaz Akhter and Michael J Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1446–1455, 2015.

[3] Emad Barsoum, John Kender, and Zicheng Liu. HP-GAN: probabilistic 3d human motion prediction via GAN. *CoRR*, 2017.

[4] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022.

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision – ECCV 2016*, pages 561–578, Cham, 2016. Springer International Publishing.

[6] Derek Bradley, Tiberiu Popa, Alla Sheffer, Wolfgang Heidrich, and Tamy Boubekeur. Markerless garment capture. In *ACM SIGGRAPH 2008 papers*, pages 1–9. 2008.

[7] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qizhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *ECCV*. 2020.

[8] Ho Kei Cheng and Alexander G. Schwing. XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, 2022.

[9] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (TOG)*, 34 (4):69, 2015.

[10] Mingsong Dou, Philip Davidson, Sean Ryan Fanello, Sameh Khamis, Adarsh Kowdle, Christoph Rhemann, Vladimir Tankovich, and Shahram Izadi. Motion2fusion: Real-time volumetric performance capture. *ACM Trans. Graph.*, 36(6): 246:1–246:16, 2017.

[11] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 12943–12954, 2023.

[12] Andrew Gilbert, Matthew Trumble, Charles Malleson, Adrian Hilton, and John Collomosse. Fusing visual and inertial sensors with semantics for 3d human pose estimation. *International Journal of Computer Vision*, 127(4): 381–397, 2019.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[14] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision*, pages 2282–2292, 2019.

[15] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. Stochastic scene-aware motion prediction. In *Proceedings of the International Conference on Computer Vision 2021*, 2021.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Yannan He, Anqi Pang, Xin Chen, Han Liang, Minye Wu, Yuexin Ma, and Lan Xu. Challencap: Monocular 3d capture of challenging human performances using multi-modal references. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11400–11411, 2021.

[18] Thomas Helten, Meinard Muller, Hans-Peter Seidel, and Christian Theobalt. Real-time body tracking with one depth camera and inertial sensors. In *Proceedings of the IEEE international conference on computer vision*, pages 1105–1112, 2013.

[19] Roberto Henschel, Timo Von Marcard, and Bodo Rosenhahn. Accurate long-term multiple people tracking using video and body-worn imus. *IEEE Transactions on Image Processing*, 29:8476–8489, 2020.

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

[21] 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, pages 13264–13275, Piscataway, NJ, 2022. IEEE.

[22] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics (TOG)*, 37(6):1–15, 2018.

[23] Yinghao Huang, Omid Tehari, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3d tracking of humans and objects in interaction. In *German Conference on Pattern Recognition*. Springer, 2022.

[24] Chaofan Huo, Ye Shi, Yuexin Ma, Lan Xu, Jingyi Yu, and Jingya Wang. Stackflow: Monocular human-object reconstruction by stacked normalizing flow with offset. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 902–910. International Joint Conferences on Artificial Intelligence Organization, 2023. Main Track.

[25] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9365–9376, 2023.

[26] Yuheng Jiang, Suyi Jiang, Guoxing Sun, Zhuo Su, Kaiwen Guo, Minye Wu, Jingyi Yu, and Lan Xu. Neuralhofusion:

Neural volumetric rendering under human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6155–6165, 2022.

[27] Yifeng Jiang, Yuting Ye, Deepak Gopinath, Jungdam Won, Alexander W. Winkler, and C. Karen Liu. Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation. In *SIGGRAPH Asia 2022 Conference Papers*, 2022.

[28] Justin Johnson, Nikhila Ravi, Jeremy Reizenstein, David Novotny, Shubham Tulsiani, Christoph Lassner, and Steve Branson. Accelerating 3d deep learning with pytorch3d. In *SIGGRAPH Asia 2020 courses*, pages 1–1. 2020.

[29] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[30] Tomoya Kaichi, Tsubasa Maruyama, Mitsunori Tada, and Hideo Saito. Resolving position ambiguity of imu-based human pose with a single rgb camera. *Sensors*, 20(19):5453, 2020.

[31] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regognition (CVPR)*, 2018.

[32] 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.

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

[34] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[35] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[36] 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 (ICCV)*, pages 11127–11137, 2021.

[37] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019.

[38] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[39] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11605–11614, 2021.

[40] 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, 2023.

[41] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3383–3393, 2021.

[42] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17142–17151, 2023.

[43] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Trans. Graph.*, 42(6), 2023.

[44] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[45] Han Liang, Yannan He, Chengfeng Zhao, Mutian Li, Jingya Wang, Jingyi Yu, and Lan Xu. Hybridcap: Inertia-aid monocular capture of challenging human motions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2): 1539–1548, 2023.

[46] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, 2015.

[47] Charles Malleson, Andrew Gilbert, Matthew Trumble, John Collomosse, Adrian Hilton, and Marco Volino. Real-time full-body motion capture from video and imus. In *2017 International Conference on 3D Vision (3DV)*, pages 449–457. IEEE, 2017.

[48] Charles Malleson, John Collomosse, and Adrian Hilton. Real-time multi-person motion capture from multi-view video and imus. *International Journal of Computer Vision*, 128(6):1594–1611, 2020.

[49] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4), 2017.

[50] Aymen Mir, Xavier Puig, Angjoo Kanazawa, and Gerard Pons-Moll. Generating continual human motion in diverse 3d scenes. In *International Conference on 3D Vision (3DV)*, 2024.

[51] Noitom. Noitom Motion Capture Systems. `https://www.noitom.com/`, 2015.

[52] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. In *International Conference on 3D Vision (3DV)*, 2018.

[53] Shaohua Pan, Qi Ma, Xinyu Yi, Weifeng Hu, Xiong Wang, Xingkang Zhou, Jijunnan Li, and Feng Xu. Fusing monocular images and sparse imu signals for real-time human motion capture. In *SIGGRAPH Asia 2023 Conference Papers*, New York, NY, USA, 2023. Association for Computing Machinery.

[54] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[55] 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 *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[56] Ilya A. Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Object pop-up: Can we infer 3d objects and their poses from human interactions alone? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023.

[57] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *International Conference on Computer Vision (ICCV)*, 2021.

[58] Polycam. 3D CAPTURE, FOR EVERYONE. `https://poly.cam/`, 2023.

[59] Gerard Pons-Moll, Andreas Baak, Thomas Helten, Meinard Müller, Hans-Peter Seidel, and Bodo Rosenhahn. Multisensor-fusion for 3d full-body human motion capture. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 663–670. IEEE, 2010.

[60] Gerard Pons-Moll, Andreas Baak, Juergen Gall, Laura Leal-Taixe, Meinard Mueller, Hans-Peter Seidel, and Bodo Rosenhahn. Outdoor human motion capture using inverse kinematics and von mises-fisher sampling. In *2011 International Conference on Computer Vision*, pages 1243–1250. IEEE, 2011.

[61] Jose Luis Ponton, Haoran Yun, Andreas Aristidou, Carlos Andujar, and Nuria Pelechano. Sparseposer: Real-time full-body motion reconstruction from sparse data. *ACM Transactions on Graphics*, 43(1):1–14, 2023.

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

[63] 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 (ICCV)*, pages 11488–11499, 2021.

[64] Yiming Ren, Chengfeng Zhao, Yannan He, Peishan Cong, Han Liang, Jingyi Yu, Lan Xu, and Yuexin Ma. Lidar-aid inertial poser: Large-scale human motion capture by sparse inertial and lidar sensors. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2337–2347, 2023.

[65] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017.

[66] Manolis Savva, Angel X. Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. Pigraphs: Learning interaction snapshots from observations. *ACM Trans. Graph.*, 35(4), 2016.

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

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

[69] SONY. Mobile Motion Capture "mocopi". `https://www.sony.net/Products/mocopi-dev/en/`, 2023.

[70] Guoxing Sun, Xin Chen, Yizhang Chen, Anqi Pang, Pei Lin, Yuheng Jiang, Lan Xu, Jingya Wang, and Jingyi Yu. Neural free-viewpoint performance rendering under complex human-object interactions. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

[71] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European conference on computer vision (ECCV)*, pages 529–545, 2018.

[72] Vince Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3d human body shape and pose prediction. In *British Machine Vision Conference (BMVC)*, 2018.

[73] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023.

[74] Garvita Tiwari, Dimitrije Antic, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision (ECCV)*, 2022.

[75] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *Proceedings of 28th British Machine Vision Conference*, pages 1–13. University of Surrey, 2017.

[76] Tom Van Wouwe, Seunghwan Lee, Antoine Falisse, Scott Delp, and C Karen Liu. Diffusion inertial poser: Human motion reconstruction from arbitrary sparse imu configurations. *arXiv preprint arXiv:2308.16682*, 2023.

[77] Timo Von Marcard, Gerard Pons-Moll, and Bodo Rosenhahn. Human pose estimation from video and imus. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1533–1547, 2016.

[78] Timo Von Marcard, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer Graphics Forum*, pages 349–360. Wiley Online Library, 2017.

[79] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, 2018.

[80] Xi Wang, Gen Li, Yen-Ling Kuo, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Reconstructing action-conditioned human-object interactions using commonsense knowledge priors. In *International Conference on 3D Vision (3DV)*, 2022.

[81] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[82] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision (ECCV)*. Springer, 2022.

[83] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single rgb camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[84] XSENS. Xsens Technologies B.V. `https://www.xsens.com/`, 2011.

[85] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. InterDiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, 2023.

[86] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics (TOG)*, 37(2):27:1–27:15, 2018.

[87] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos, 2023.

[88] Hongwei Yi, Chun-Hao P. Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies, and Michael J. Black. Human-aware object placement for visual environment reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, 2022.

[89] Hongwei Yi, Chun-Hao P. Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J. Black. MIME: Human-aware 3D scene generation. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 12965–12976, 2023.

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

[91] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[92] 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. *ACM Transactions on Graphics (TOG)*, 42(4), 2023.

[93] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[94] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. PhysDiff: Physics-guided human motion diffusion model. In *IEEE International Conference on Computer Vision (ICCV)*, 2023.

[95] Z CAM. Professional Cinema Camera. `https://www.z-cam.com/e2/`, 2023.

[96] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. Smoothnet: A plug-and-play network for refining human poses in videos. In *European Conference on Computer Vision*, pages 625–642. Springer, 2022.

[97] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

[98] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[99] Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In *CVPR*, 2023.

[100] Juze Zhang, Ye Shi, Yuexin Ma, Lan Xu, Jingyi Yu, and Jingya Wang. Ikol: Inverse kinematics optimization layer for 3d human pose and shape estimation via gauss-newton differentiation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

[101] Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *European Conference on Computer Vision (ECCV)*, 2020.

[102] Siwei Zhang, Yan Zhang, Federica Bogo, Pollefeys Marc, and Siyu Tang. Learning motion priors for 4d human body capture in 3d scenes. In *International Conference on Computer Vision (ICCV)*, 2021.

[103] Siwei Zhang, Yan Zhang, Federica Bogo, Pollefeys Marc, and Siyu Tang. Learning motion priors for 4d human body capture in 3d scenes. In *International Conference on Computer Vision (ICCV)*, 2021.

[104] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7376–7385, 2020.

[105] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. 2022.

[106] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J. Black, and Siyu Tang. Generating 3d people in scenes without people. In *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[107] Zhe Zhang, Chunyu Wang, Wenhu Qin, and Wenjun Zeng. Fusing wearable imus with multi-view images for human

pose estimation: A geometric approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2200–2209, 2020.

[108] Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, , and Siyu Tang. Synthesizing diverse human motions in 3d indoor scenes. In *International conference on computer vision (ICCV)*, 2023.

[109] Zerong Zheng, Tao Yu, Hao Li, Kaiwen Guo, Qionghai Dai, Lu Fang, and Yebin Liu. Hybridfusion: Real-time performance capture using a single depth sensor and sparse imus. In *European Conference on Computer Vision (ECCV)*, 2018.

[110] Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei. Deep kinematic pose regression. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 186–201. Springer, 2016.

[111] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019.