WHAC: World-grounded Humans and Cameras

Wanqi Yin[⋆,1,2], Zhongang Cai[∗,1,3], Ruisi Wang[1], Fanzhou Wang[1], Chen Wei[1], Haiyi Mei[1], Weiye Xiao[1], Zhitao Yang[1], Qingping Sun[1], Atsushi Yamashita[2], Ziwei Liu[3], Lei Yang[1]

[1] SenseTime Research, [2] The University of Tokyo,
[3] S-Lab, Nanyang Technological University

Abstract. Estimating human and camera trajectories with accurate scale in the world coordinate system from a monocular video is a highly desirable yet challenging and ill-posed problem. In this study, we aim to recover expressive parametric human models (*i.e.*, SMPL-X) and corresponding camera poses jointly, by leveraging the synergy between three critical players: the world, the human, and the camera. Our approach is founded on two key observations. Firstly, camera-frame SMPL-X estimation methods readily recover absolute human depth. Secondly, human motions inherently provide absolute spatial cues. By integrating these insights, we introduce a novel framework, referred to as **WHAC**, to facilitate world-grounded expressive human pose and shape estimation (EHPS) alongside camera pose estimation, without relying on traditional optimization techniques. Additionally, we present a new synthetic dataset, **WHAC-A-Mole**, which includes accurately annotated humans and cameras, and features diverse interactive human motions as well as realistic camera trajectories. Extensive experiments on both standard and newly established benchmarks highlight the superiority and efficacy of our framework. The code and dataset are available on the homepage[1].

Keywords: Expressive Human Pose and Shape Estimation

1 Introduction

Expressive human pose and shape estimation (EHPS) has garnered considerable research attention due to its wide applications across the entertainment, fashion, and healthcare industries. Despite remarkable advancements in recent years, the majority of EHPS methods primarily focus on estimating parametric human models (*i.e.*, SMPL-X [29]) in the camera coordinate system. This approach falls short in dynamic situations where the camera and subject move concurrently. Estimating 3D trajectories in the world coordinate system (world-grounded) from 2D camera footage is challenging as the 3D-to-2D projection results in a loss of critical spatial information. Consequently, camera trajectories deduced are thus inherently "scaleless", and the depth of humans directly estimated from the camera perspective lacks validity.

⋆ Equal contributions.

[1] Homepage: https://wqyin.github.io/projects/WHAC/.

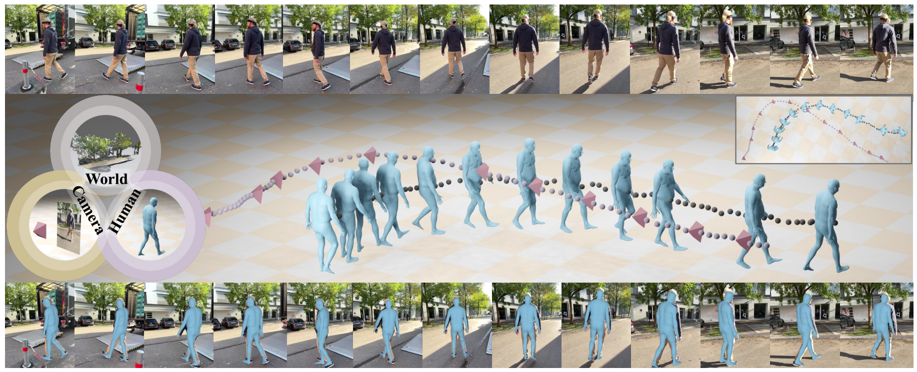

Fig. 1: WHAC synergizes human-camera (camera-frame SMPL-X estimation), camera-world (visual odometry), and human-world (our proposed MotionVelocimeter) modeling for constructing world-grounded human and camera trajectories.

In this work, we demonstrate the synergy between humans, cameras, and the world. First, existing camera-frame EHPS methods, although not specifically supervised to estimate human depth directly, can still accurately deduce the true depth. This only requires a reasonably accurate focal length that can be obtained from the video capture devices or estimated [18]. Second, root translation is a critical component of human motions, allowing the latter to serve as a strong prior after an association is learned. Hence, by analyzing human poses, one can make an informed estimation of the velocity of human movement. Building upon these insights, we present WHAC, a novel framework designed to jointly estimate expressive human models and camera movements using a monocular video. For any given input video, camera-frame SMPL-X parameters and a preliminary camera trajectory are first estimated using plug-and-play EHPS [5] and visual odometry (VO) [39] models. The human-camera relative positions are first deduced. These estimations are then utilized with VO estimations to canonicalize the sequences of human poses for accurate velocity estimation. Consequently, the scale of the camera trajectory can be recovered. It is noteworthy that WHAC pioneers whole-body, optimization-free estimation in a world-grounded context to recover human and camera trajectories jointly.

Moreover, the development of a new dataset becomes essential to more accurately assess model performance on world-grounded human motions and camera trajectories across a broader spectrum of scenarios. Recent studies have underscored the surprising efficacy of synthetic data [2, 5, 6, 41], thanks to its diversity and controllability. Inspired by these findings, we introduce WHAC-A-Mole, a comprehensive synthetic dataset for World-grounded Humans And Cameras with a rich collection of Animated subjects under MOving viewpoints in muLtiple Environments. WHAC-A-Mole features comprehensive motion sequences that include 1) interactive human activities from DLP-MoCap [3], 2) partner dances from DD100 [36], in addition to 3) the standard AMASS [25] motion repository. Notably, WHAC-A-Mole includes automatically generated camera trajectories that mimic cinematic filming techniques, such as *tracking shots* and *arc shots*, thereby offering a high level of realism.

We validate our WHAC on standard benchmarks and WHAC-A-Mole to obtain consistent performance gains compared to the state-of-the-art (SoTA) methods under both camera-frame and world-grounded settings. WHAC demonstrates a surprising capability to handle corner cases when motion-based and camera-based observations contradict, paving the way for potential applications.

In summary, our contributions are three-fold. First, we propose WHAC, a novel regression-based framework that capitalizes on human priors for the pioneering world-grounded EHPS method. Second, we contribute WHAC-A-Mole, a comprehensive benchmark with accurate human and camera annotations of diverse human activities. Third, our empirical evaluations underscore the superior performance of WHAC across multiple benchmarks.

2 Related Works

2.1 Expressive Human Pose and Shape Estimation (EHPS)

EHPS captures body, face, and hands from monocular images or videos, typically through parametric human models (*e.g.*, SMPL-X [29]). Early optimization-based method [29] fits SMPL-X models on 2D keypoints, and was soon outperformed by regression-based methods that were trained on a large amount of paired data. Two-stage methods estimate body parameters first, then hand/face parameters from crop-out image patches [8, 10, 20, 27, 28, 33, 45, 48]. Recently, OSX [22] proposes the one-stage paradigm that estimates body, hand, and face with shared features. This paradigm shift simplifies the pipeline and led to the first foundation model SMPLer-X [5] that achieved unprecedented generalization ability across key benchmarks. However, despite their success, these methods estimate parametric humans in the camera coordinate, lacking information on the global trajectory especially when the camera is moving.

2.2 World-grounded Recovery of Humans and Cameras

Estimation of human trajectory in world coordinate system typically requires a multi-camera setup [4, 7, 12, 13, 16, 30, 47], additional wearable devices (*e.g.*, IMU [11, 26] or electromagnetic sensors [17]). Methods that require only a single camera often rely on other assumptions: Yu *et al*. [43] needs the scene to be provided by the user and Luvizon *et al*. [24] assumes a static camera. D&D [21], GLAMR [44], and TRACE [37] estimate global human trajectories from single-frame poses or image features. However, camera and human rotation have a coupled effect on camera-frame global orientation estimation, which leads to ambiguity. Liu *et al*. [23] leverages Structure-from-Motion (SfM) [34] to reconstruct both camera and human trajectories and adjust human's scale to match the camera's, which may not reflect the absolute scale. Recently, SLAHMR leverages SLAM [38] and human motion prior [32] in the optimization to recover

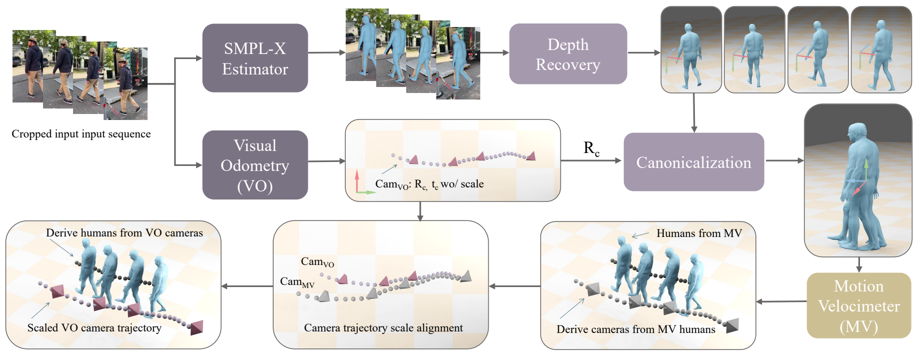

Fig. 2: Overview of **WHAC**. SMPL-X estimator extracts camera-frame SMPL-X with dummy depth [5], which is recovered in Sec. 3.2. The scaleless camera trajectory estimated by VO [39] is then used to canonicalize the human trajectory to estimate its velocity and thus scale in Sec. 3.3. A camera trajectory is then derived for alignment and scale recovery, which subsequently updates the human trajectory in Sec. 3.4.

humans and the camera. However, the process is computationally expensive and takes excessively long to complete. PACE [19] also leverages VO [39] for camera pose estimation and a faster human motion to significantly accelerate the optimization process but is still time-consuming. WHAM [35] is the first regression-based work in the domain that features real-time performance. It takes camera estimation (angular velocity) as the input and estimates human parameters in the camera frame and human trajectory in the world frame through separate branches, while the camera trajectory is not recovered. Our WHAC aims to recover both human and camera trajectories in the world coordinate with accurate scales.

3 Methodology

Recovering accurate 3D dimensions from 2D observations is an ill-posed problem: a small object at a close range may appear the same as a large object at a far range. In this section, we aim to address two ambiguities with priors that are surprisingly effective, but not thoroughly utilized in existing EHPS works: the parametric humans themselves.

3.1 Preliminaries

Problem Formulation. We aim to estimate human and camera pose sequences in the world coordinate system. The humans are represented by SMPL-X parameters: global orientation $\theta_{go}^w \in \mathbb{R}^{1\times3}$, translation $t_h^w \in \mathbb{R}^{1\times3}$, body pose $\theta_b \in \mathbb{R}^{21\times3}$, left hand pose $\theta_{lh} \in \mathbb{R}^{15\times3}$, right hand pose $\theta_{rh} \in \mathbb{R}^{15\times3}$, jaw pose $\theta_j \in \mathbb{R}^{1\times3}$, body shape $\beta \in \mathbb{R}^{10}$ and facial expression $\phi \in \mathbb{R}^{10}$. The cameras are represented by world-frame rotation $R_c^w \in \mathbb{R}^{1\times3}$ and translation $t_c^w \in \mathbb{R}^{1\times3}$. In this work, the superscript indicates the coordinate system (*i.e.*, w for world and c for camera).

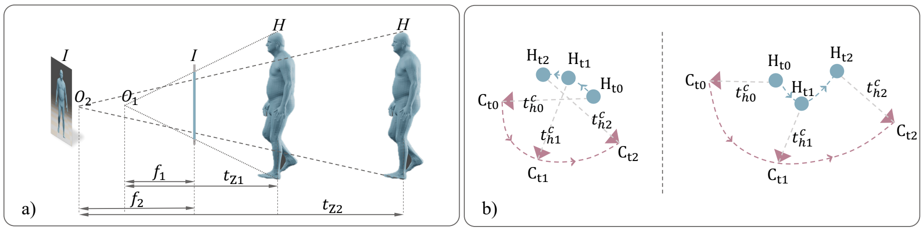

Fig. 3: a) Different pairs of focal length f and t_z can correspond to the same image. b) Human trajectories H derived from camera trajectories C of different scales can be vastly different in both shape and direction, despite that the same camera-frame human root translations t_h^c are used.

Camera-frame SMPL-X Estimation typically omits absolute depth estimation. Hence, primary metrics (*e.g.*, PA-MPJPE, MPJPE, and PVE) all perform root alignment. We employ SMPLer-X [5], a strong foundation model that demonstrates accurate estimation of human pose and shapes. We add additional GRUs before the prediction heads and finetune the model to better capture the temporal cues.

Visual Odometry (VO) typically provides high-quality R_c^w; the trajectory formed by t_c^w is scaleless but accurate in shape. We adopt DPVO [39] and follow the standard practice to define the first camera frame of the input video as the world coordinate system.

3.2 Recovering Camera-space Human Root Translation

Mainstream EHPS methods [5, 22, 27] recover parametric humans in the camera space and adopt a weak perspective camera model, which considers all points to be at the same depth away from the camera.

$$\begin{bmatrix} f^* & 0 & 0 \\ 0 & f^* & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} t_x + \delta x \\ t_y + \delta y \\ t_z + \delta z \end{bmatrix} \approx \begin{bmatrix} f^* & 0 & 0 \\ 0 & f^* & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} t_x + \delta x \\ t_y + \delta y \\ t_z \end{bmatrix}, \tag{1}$$

where f^* indicates the focal length in NDC (Normalized Device Coordinate) space (image pixel coordinates are normalized into $[-1, 1]$). $t_h^c = (t_x, t_y, t_z)$ is the root translation in the camera frame, $(\delta x, \delta y, \delta z)$ is the relative translation of a point relative to the root. Hence, the projected 2D point (NDC space) is written as:

$$\begin{bmatrix} u^* \\ v^* \end{bmatrix} = \begin{bmatrix} f^*(t_x + \delta x)/t_z \\ f^*(t_y + \delta y)/t_z \end{bmatrix} = \begin{bmatrix} s(t_x + \delta x) \\ s(t_y + \delta y) \end{bmatrix}, \tag{2}$$

where s is the scale parameter. The SMPL-X estimator also predicts camera parameters (s, t_x, t_y) to reproject SMPL-X joints on the image plane. Hence, we obtain the relationships between focal length, depth, and scale:

$$s = \frac{f^*}{t_z} = \frac{2}{I} \times \frac{f}{t_z} \Rightarrow t_z = \frac{2}{I} \times \frac{f}{s}, \tag{3}$$

where f is the focal length in pixels. I is the resolution (side pixel length) of the input crop to the SMPL-X estimator.

We point out that although these camera-frame methods do not supervise the human root depth t_z, by training the model to produce a scale s that overlays SMPL-X accurately back on the image plane, the model implicitly learns human root depth t_z that is coupled with focal length f as illustrated in Fig. 3a). A dummy focal length of 5,000 is often used [5,22,27], however, this leads to unrealistic human root depth t_z. We highlight that accurate intrinsic parameters are accessible from many devices, and our empirical results show using the diagonal pixel length [18] also yields satisfactory results as shown in Tab. 8.

3.3 Estimating World-frame Human Motions for Scale Extraction

In recent years, there has been a plethora of high-quality optical motion capture datasets that become available, covering a wide range of human activities. Previous art [35] estimates human global trajectory from 2D keypoint observations, which may not capture subtle 3D information. Hence, we propose to learn absolute scale from 3D human motions.

First, for a SMPL-X sequence of K frames, estimated in the camera coordinate system, we compute the 3D joint coordinates. Specifically, for each frame:

$$J^c = M(\theta^c_{go}, t^c, \theta_b, \theta_{lh}, \theta_{rh}, \theta_j, \beta, \phi), J^c \in \mathbb{R}^{K\times 15\times 3}, \tag{4}$$

where M is the SMPL-X parametric model. We select 15 joints (14 LSP [15] joints and the pelvis) from the original 55 joints. We then compute the joints in the world coordinate system:

$$J^w = T^{vo} \times J^c, T^{vo} = [R^w_c | t^w_c], \tag{5}$$

where T^{vo} is the visual odometry's estimation (camera-to-world transformation). Note that t^w_c does not have a valid scale.

To facilitate the training, we standardize the input: we define a canonical frame where the human is root-aligned with zero global orientation. We then compute the canonical transformation T^{cano} by using the first (0^{th}) frame's rotation and offset the translation to zero:

$$T^{cano} = [R^{cano} | t^{cano}], R^{cano} = (R^w_{c,0} \times \theta^c_{go,0})^{-1} = (\theta^c_{go,0})^{-1}, t^{cano} = -p^w_0, \tag{6}$$

where $R^w_{c,0}$ is 0^{th} camera rotation in the world frame estimated from visual odometry, $\theta^c_{go,0}$ is 0^{th} global orientation estimated in the camera space, p^w_0 is the pelvis joint of J^w_0. Note $R^w_{c,0}$ is an identity matrix I_3 as the 0^{th} camera frame is defined as the world frame. All joints are then canonicalized as $J^{cano} = T^{cano} \times J^w$.

MotionVelocimeter then estimates per-frame velocity in the canonical space:

$$V^{cano} = \text{MotionVelocimeter}(J^{cano}), \tag{7}$$

where the velocity is then de-canonicalized back to the world frame:

$$V^{w} = (T^{cano})^{-1} \times V^{cano}, \tag{8}$$

with V^w, we can reconstruct the human trajectory with scale in the world coordinate system. MotionVelocimeter only requires a simple architecture that we include in the Supplementary Material.

3.4 Recovering Scaled Human and Camera Trajectories

As we obtain human trajectory t_h^w with absolute scale, one possible way is to align the human trajectory derived from the VO-estimated camera trajectory using camera-frame human root translation to t_h^w. However, Fig. 3b) shows that such alignment is problematic as the human trajectory derived from scaleless camera trajectory may be invalid. Hence, we propose to transfer the scale to the camera trajectory in two steps. First, we derive a camera trajectory from the human trajectory:

$$T^{w}_{c,derived} = (T^{cano})^{-1} \times T^{cano}_{h} \times (T^{c}_{h})^{-1}, \tag{9}$$

$$T^{w}_{c,derived} = [R^{w}_{c,derived} | t^{w}_{c,derived}]. \tag{10}$$

This derived camera trajectory already has an accurate scale with a good shape. However, we find that the camera trajectory estimated by VO has a better, more robust shape because it can leverage visual cues that are much denser than human motion cues. In this light, we perform Umeyama's method [40] (shown as $\xrightarrow{U}$) to align the VO-estimated camera trajectory with the human-derived camera trajectory $t^w_{c,derived}$ while discard $R^w_{c,derived}$ and keeping the camera rotation R^w_c:

$$t^{w}_{c,final} = t^{w}_{c} \xrightarrow{U} t^{w}_{c,derived}. \tag{11}$$

Hence, we then update human trajectory by deriving it from the aligned camera trajectory $t^w_{c,final}$:

$$T^{w}_{h,final} = [R^{w}_{h,final} | t^{w}_{h,final}] = T^{w}_{c,final} \times T^{c}_{h}, T^{w}_{c,final} = [R^{w}_{c} | t^{w}_{c,final}]. \tag{12}$$

As a result, we obtain human trajectory $t^w_{h,final}$ and camera trajectory $t^w_{c,final}$, both in the world coordinate system and with absolute scales.

4 WHAC-A-Mole Dataset

We highlight that WHAC-A-Mole combines fine-crafted automatic camera movements with varied characters animated with diverse, high-quality motion sequences to generate a dataset with accurate camera and SMPL-X annotations.

Table 1: Dataset Comparison. #Inst.: number of human instances (crops). #Seq.: number of video sequences. R/S: Real or Synthetic. Multi.: multiperson scenes. Track.: track ID labels. HHI: human-human interaction motions. *: Based on the released version. †: EgoSet. ⋄: typically short (<100 frames) clips.

Dataset	#Inst.	#Seq.	R/S	Multi.	Track.	Contact	HHI	Camera	Human
3DPW [26]	74.6K	60	R	✓	×	×	✓	Moving	SMPL
RICH [14]	476K	141	R	✓	✓	✓	✓	Mixed	SMPL
HCM* [19]	5379	21	S	×	N.A.	×	N.A.	Moving	SMPL
EMDB [17]	109K	81	R	×	N.A.	×	N.A.	Moving	SMPL
EgoBody† [46]	175K	125	R	✓	✓	×	✓	Moving	SMPL-X
BEDLAM [2]	951K	10.4K⋄	S	✓	✓	×	×	Mixed	SMPL-X
SynBody [41]	2.7M	27K⋄	S	✓	✓	×	×	Static	SMPL-X
WHAC-A-Mole	1.46M	2434	S	✓	✓	✓	✓	Moving	SMPL-X

The dataset is constructed with the advanced human data synthesis toolbox XR-Feitoria [9]. It leverages SMPL-XL (a layered extension of SMPL-X) to create virtual humans with diverse body shapes, clothing, and accessories. We follow SynBody [41] in the scene setup, subject creation, and placement. We further improve the data synthesized in two ways: diverse motion sources (Sec. 4.1) and camera trajectory generation (Sec. 4.2). In Tab. 1, we compare WHAC-A-Mole with popular video-based benchmarks with both camera and human annotations. WHAC-A-Mole features a competitive scale of training instances and video sequences, multiperson scenes with track IDs, contact labels, accurate camera pose and SMPL-X annotations. We split WHAC-A-Mole by motion sequence into 80%:20% for training and testing. Examples of WHAC-A-Mole are visualized in Fig. 4.

4.1 Interactive Human Motions

AMASS [25] is a popular motion repository, widely used by existing synthetic datasets [2, 19, 41]. However, AMASS only contains single-person motions. As a result, synthetic data is captured in virtual scenes populated with unrelated single-person motions, typically scattered sparsely to avoid collision. However, close human interactions are common in daily life, and difficult to solve. In this light, we select two latest motion datasets that contain comprehensive interactive human motions. First, DD100 [36], a duet dance motion capture dataset that includes near two hours of partner dances of 10 different genres. Second, DLP-MoCap [3], a motion capture dataset containing daily interactions between two subjects. Since SMPL-XL models are fully compatible with SMPL-X body pose sequences, we animate virtual characters with a combination of AMASS, DD100, and DLP-MoCap.

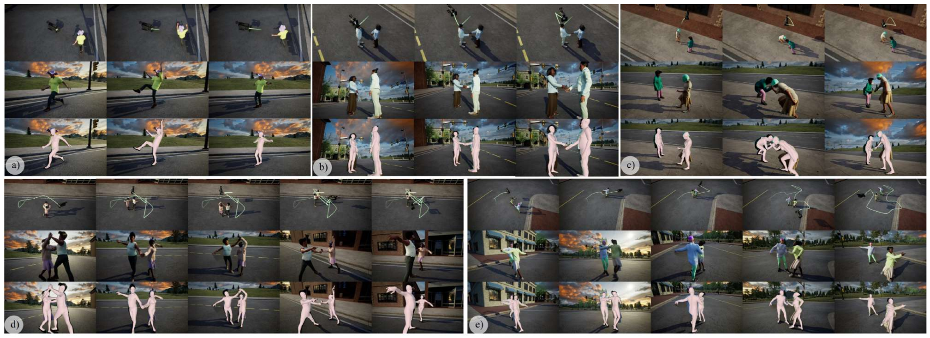

Fig. 4: Visualization of WHAC-A-Mole sample sequences, animated with a) AMASS, b-c) DLP-MoCap, and d-e) DD100. In each sample, the first row depicts the overview (note the camera trajectory shown in bright rays), and the second and the third rows show the camera view and overlaid SMPL-X annotations.

4.2 Camera Trajectory Generation

To better model the camera movement, we adopt the representation in [31] to define the camera in a human-centric spherical coordinate system (r_c, θ_c, ϕ_c), in which the r_c represents the distance from the camera to the character, while the polar angle θ_c and the azimuthal angle ϕ_c define the angle between the camera's looking direction and the character's facing direction. Therefore, given a character's location (x_{ch}, y_{ch}, z_{ch}) and facing direction (θ_{ch}, ϕ_{ch}), the camera's location in the world space is

$$(x_c, y_c, z_c) = (x_{ch}, y_{ch}, z_{ch}) + r_c(\sin(\theta)\cos(\phi), \sin(\theta)\sin(\phi), \cos(\theta)), \tag{13}$$

where the $\theta = (\theta_c + \theta_{ch}) \mod 2\pi$, the $\phi = (\phi_c + \phi_{ch}) \mod 2\pi$, and the camera's rotation is thereby calculated by restricting the camera look at the (x_{ch}, y_{ch}, z_{ch}). In WHAC-A-Mole, we design two types of shot scales including the *medium shot* and the *full shot*, which respectively use the location of the *neck* and the *pelvis* as the character's location (x_{ch}, y_{ch}, z_{ch}). For the motion sequences that consist of multiple characters, the (x_{ch}, y_{ch}, z_{ch}) and the (θ_{ch}, ϕ_{ch}) are derived from the average of the locations and the facing directions of all the characters. Based on the human-centric spherical coordinate system (r_c, θ_c, ϕ_c), we design different keyframe-setting strategies to simulate five common camera movements below. Detailed implementations for camera movements are included in the Supplementary Material.

- **Arc shot** adds equally-spaced keyframes to rotate the camera around the character horizontally or vertically, with the controllable angular velocity.
- **Push shot** adds equally-spaced keyframes and moves the camera towards the character with adjustable camera speed.
- **Pull shot** is opposite to the *push shot* and moves the camera further away from the character. Continuous pushing and pulling is commonly used when filming dances.

- Tracking shot follows the character and maintains the relative position between the camera and the character. A new keyframe of the *tracking shot* is added when the overlap ratio of the character's bounding box in the current frame and the last keyframe is greater than a threshold.
- Pan shot rotates the camera horizontally to keep the camera looking at the character, therefore it is another way to make the camera follow the character, and it shares the same rule with the *tracking shot* to add a new keyframe.

Rather than assigning a specific camera movement to an entire human motion sequence, our pipeline automatically combines several types of camera movements into one motion sequence to increase the variety of camera movements. For example, when capturing static motions (whose longest edge of the bounding box formed by the (x_{ch}, y_{ch}, z_{ch}) across all frames is less than a threshold λ_{bbox}) or interactive motions, we combine the horizontal and the vertical *arc shots* with the random *pull* or *push shots* to rotate the camera around the characters as well as transiting smoothly between different shot angles, such as high-angle, low-angle or eye-level, and pushing in or pulling out the distance between the camera and the characters to increase the rhythm of the camera movement. For the motions with long-distance movements, we combine the *tracking shots* and the *pan shots* to follow the character. If the character's facing direction is stable (*i.e.*, that the rotation angle from the character's facing direction in the last keyframe to the current keyframe is less than a threshold λ_{angle}), we use the *tracking shot*. Otherwise, we use the *pan shot*. This rule effectively smooths the camera's movement, especially when the character turns dramatically.

5 Experiments

We evaluate WHAC on both camera-frame and world-grounded benchmarks to compare its parametric human recovery abilities with existing SoTA methods. Due to space constraints, we include inference speed comparison, more visualizations on trajectory reconstruction, and more qualitative results in the Supplementary Material.

5.1 Implementation Details

We finetune SMPLer-X-B [5] with EgoBody, 3DPW, and EMDB for camera-frame estimation of SMPL-X parameters. WHAC-A-Mole (with motions from AMASS, DD100, and DLP-MoCap), 3DPW, EMDB, and RICH are used to train the MotionVelocimeter. More details are in the Supplementary Material.

5.2 Datasets

In addition to our proposed WHAC-A-Mole, mainstream benchmarks for human pose and shape estimation with parametric human labels are used. **EgoBody** [46] includes 125 sequences of 36 subjects in 15 indoor scenes, featuring 3D human motions interacting with scenes. We study the EgoSet that is captured

Table 2: World-frame evaluation on **WHAC-A-Mole** (DD subset). *: adapted to world-grounded evaluation. H-AS and C-AS: the closer to 1.0, the better.

	PA-MPJPE↓	W-MPJPE↓	WA-MPJPE↓	H-ATE↓	H-AS	C-ATE↓	C-AS
OSX* [22] + DPVO [39]	92.3	1123.2	413.2	202.64	0.6	**0.5**	6.1
SMPLer-X-B* [5] + DPVO [39]	83.4	958.2	362.5	156.7	0.6	**0.5**	6.1
WHAC (GT Gyro)	**78.6**	435.0	**211.2**	**118.2**	**0.9**	**0.5**	**1.3**
WHAC	**78.6**	**434.9**	**211.2**	**118.2**	**0.9**	**0.5**	**1.3**

by a head-mounted camera; 2) **3DPW** [26], a popular dataset with 60 sequences captured by an iPhone, featuring diverse human activities in outdoor scenes; 3) **EMDB** [17] provides 58 minutes of motion data of 10 subjects in 81 indoor and outdoor scenes. Notably, it contains a subset, EMDB 2, that contains global trajectories of humans and cameras. 4) **RICH** [14] consists of 142 multi-view videos with 22 subjects and 5 scenes with 6-8 fixed cameras. RICH is not used for evaluation as the cameras are static.

5.3 Evaluation Metrics

For camera-frame human recovery, we use the standard Mean Per Joint Position Error (**MPJPE**), Procrustes-aligned MPJPE (**PA-MPJPE**), Per Vertex Error (**PVE**) in millimeters (mm), and Acceleration error (**Accl.**) in m/s^2. Note that these metrics are evaluated after root alignment between estimated and ground truth parametric humans, thus not considering discrepancy in translation estimation. In this light, we also report **T-MPJPE** [1] and similarly **T-PVE**, which are variants of MPJPE and PVE that includes translation estimation to reflect the accuracy of depth estimation in the camera space.

For world-frame human/camera recovery, we follow previous works [19,35,42] to split human motion sequences with global trajectory into 100-frame segments. The segments are Procrustes-aligned to the ground truth for MPJPE computation: **W-MPJPE**$_{100}$ if the first two frames are used in the alignment or **WA-MPJPE**$_{100}$ if the entire segment. To evaluate the quality of trajectory, we extend Average Trajectory Error (ATE) [19] to **C-ATE** and **H-ATE** for camera and human respectively, which are computed after Procrustes-alignment of estimated and ground truth trajectories. All metrics are in millimeters (mm). We also report respective Alignment Scales (AS) used in the alignment for the camera (**C-AS**) and human (**H-AS**) and values closer to 1.0 indicate more accurate scale estimation.

5.4 World-grounded Benchmarks

We evaluate on EMDB2 [17] and WHAC-A-Mole in Tab. 2 and Tab. 3. WHAC-A-Mole provides expressive human (*i.e.*, SMPL-X), with accurately annotated camera motions. Since no existing EHPS methods produce SMPL-X in the world coordinate system and a strictly fair comparison is not plausible, we build the

Table 3: World-frame evaluation on **EMDB2**. *: adapted to world-grounded evaluation. H-AS and C-AS: the closer to 1.0, the better.

	PA-MPJPE↓	W-MPJPE↓	WA-MPJPE↓	H-ATE↓	H-AS	C-ATE↓	C-AS
GLAMR [44]	56.0	756.1	286.2	-	-	-	-
SLAHMR [42]	61.5	807.4	336.9	207.8	1.9	-	-
WHAM [35] (GT Gyro)	41.9	436.4	165.9	83.2	1.5	-	-
OSX-L* [22] + DPVO [39]	99.9	1186.2	458.8	235.4	2.3	**14.8**	5.1
SMPLer-X-B* [5] + DPVO [39]	42.5	930.1	375.8	200.6	2.0	**14.8**	5.1
WHAC (GT Gyro)	**39.4**	392.5	143.1	**75.8**	**1.1**	**14.8**	1.5
WHAC	**39.4**	**389.4**	**142.2**	76.7	**1.1**	**14.8**	**1.4**

Table 4: Results of camera-frame methods on **EgoBody (EgoSet)** with SMPL-X ground truths. PVE variants are measured for whole-body (SMPL-X) methods only.

	PA-MPJPE↓	PA-PVE-all↓	PVE-all↓	PVE-hand↓	PVE-face↓	Accl.↓
GLAMR [44]	114.3	-	-	-	-	173.5
SLAHMR [42]	79.1	-	-	-	-	25.8
Hand4Whole [27]	71.0	59.8	127.6	48.0	41.2	27.2
OSX-L [22]	66.5	54.6	115.7	50.5	41.0	24.7
SMPLer-X-B [5]	47.1	40.7	72.7	43.7	32.4	18.9
WHAC	**46.9**	**39.0**	**64.7**	**41.0**	**26.3**	**11.6**

first benchmark by making two adaptations to the SoTA methods (OSX [22] and SMPLer-X [5]): camera-frame translation estimation and visual odometry. Implementation details are included in the Supplementary Material. It is noted that methods that achieve good results on EMDB still struggle on WHAC-A-Mole, which can be attributed to the more challenging scenarios of WHAC-A-Mole (involving hard poses, diverse interactions, occlusions, and complicated camera movements). We hope WHAC-A-Mole can serve as a useful foundation for future world-grounded EHPS research.

Moreover, we compare WHAC with both body-only methods (GLAMR [44], SLAHMR [42], and WHAM [35]) and whole-body methods (OSX-L [22] and SMPLer-X [5]) on EMDB2, where WHAC achieves best performance, even surpassing body-only methods that are native to EMDB's SMPL annotations.

5.5 Camera-space Benchmarks

In Tab. 4, it is shown that WHAC outperforms existing SoTAs. We highlight that 1) WHAC archives immense T-PVE-all improvement, which captures absolute depth estimation from humans to cameras. This is because WHAC formulates the subject distance to the camera. 2) With temporal information embedded in the EHPS module, WHAC attains substantial reductions in acceleration error (Accl.) compared to previous single-frame SoTAs. Moreover, temporal cues also lead to significant performance gains in hand and face estimation. In Tab. 5, we further evaluate WHAC on EMDB and 3DPW, where the plausibility of camera-

Table 5: More camera-frame evaluations on **EMDB1** and **3DPW**. Compared to existing mainstream EHPS methods, WHAC recovers meaningful human depths (T-PVE) and achieves lower acceleration errors (Accl.).

Method	EMDB1 [17]				3DPW [26]			
	PA-PVE↓	PVE↓	T-PVE↓	Accl.↓	PA-PVE↓	PVE↓	T-PVE↓	Accl.↓
Hand4Whole [27]	99.5	143.1	36851.8	34.2	81.7	124.7	30279.0	31.0
OSX-L [22]	93.3	134.0	45526.0	30.3	76.9	117.8	38472.2	24.9
SMPLer-X-B [5]	68.2	99.3	41298.0	24.4	**62.6**	95.6	32532.0	24.8
WHAC	**61.0**	**91.2**	**140.2**	**18.4**	62.8	**91.9**	**260.8**	**20.3**

Table 6: Results of camera-frame methods on **WHAC-A-Mole** (DD subset). WHAC is on par with SMPLer-X but produces a lower acceleration error.

	PA-MPJPE↓	PA-PVE-all↓	PVE-all↓	PVE-hand↓	PVE-face↓	Accl.↓
OSX-L [22]	94.4	92.3	167.9	85.5	89.7	42.3
SMPLer-X-B [5]	83.4	82.0	138.0	**75.7**	75.2	50.3
WHAC	**78.6**	**78.1**	**130.1**	80.2	**70.8**	**36.6**

frame human translation estimation and the significance of temporal modeling are validated again.

Albeit WHAC-A-Mole is mainly designed for world-grounded evaluation of human and camera pose sequences, we evaluate WHAC's performance under the camera-frame setting on WHAC-A-Mole in Tab. 6. Similar to previous experiments, it is observed that WHAC is on par with SMPLer-X with better performance on the acceleration error.

5.6 Ablation Study

We evaluate the necessity of the key components in Tab. 7 on EMDB2 [17]. It is observed that using visual odometry alone (body trajectory depends on estimated camera trajectory) leads to accurate camera trajectory shape (lowest camera trajectory error) but lacks accurate scale (alignment scale is far from 1.0). Using MotionVelocimeter alone (camera trajectory depends on estimated body trajectory), results in very accurate scale recovery but less accurate in camera trajectory. WHAC leverages the scale recovery ability of MotionVelocimeter and visual odometer, achieving high-quality body and camera trajectories with only a slight decline in scale accuracy.

5.7 Visualization

We highlight that WHAC is the first regression-based, whole-body method that simultaneously predicts camera and human trajectories. We highlight that the camera provides supplementary cues to human motions. In Fig. 5 a) and b), we test two corner cases where the human motion itself can be misleading: when

Table 7: Ablation on key components. DPVO represents visual odometry, MV represents MotionVelocimeter.

Method	WA-MPJPE↓	H-ATE↓	C-ATE↓	C-AS
DPVO	376.0	177.8	**14.8**	5.10
MV	233.2	129.9	134.1	**1.10**
MV + DPVO	**142.2**	**76.7**	**14.8**	1.40

Table 8: Ablation on intrinsic sources. A reasonable intrinsic drastically improve root translation estiamtion on EMDB2.

Intrinsics	T-MPJPE↓	W-MPJPE↓	WA-MPJPE↓
Dummy(5,000)	36020.4	6239.9	604.6
Assumed [18]	179.7	391.2	144.0
GT Intrinsics	**100.3**	**389.4**	**142.2**

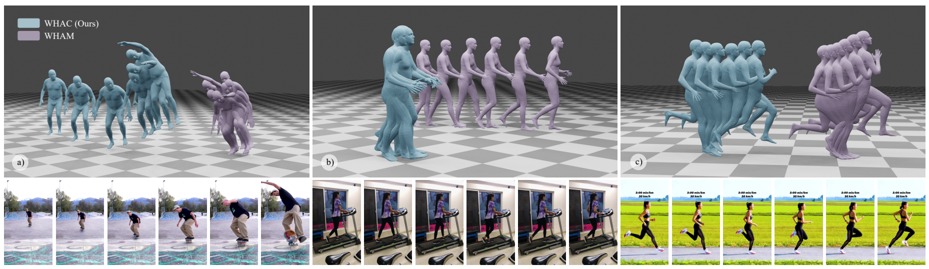

Fig. 5: Visualization on in-the-wild hard cases. WHAC leverages human-camera-scene collaboration to resolve cases where motion prior alone would fail: a) Skateboarding and b) Treadmill. c) WHAC can also handle fast cases.

human pose appears stationary but there is root movement in the world coordinates (*e.g.*, skateboarding), and when human pose clearly indicates motion but there is no root movement in the world coordinates (*e.g.*, running on a treadmill). Our formulation considers both motion and camera cues to predict the correct trajectories whereas WHAM fails, which leverages foot contact and locking but no camera information in human's global trajectory estimation. We also show complicated scenarios on c) in-the-wild video from TikTok, which features a fast-moving subject. Our MotionVelocimeter can estimate reasonable root movement, whereas WHAM's contact estimation and foot-locking results in a floating subject. More visualizations are included in the Supplementary Material.

6 Conclusion

In conclusion, we present WHAC, the pioneering regression-based EHPS method that jointly recovers human motions and camera trajectories in the world coordinate system. Moreover, our WHAC-A-Mole serves as a useful benchmark for the evaluation of world-grounded EHPS methods. WHAC achieves SoTA performance on both standard benchmarks and our proposed WHAC-A-Mole, demonstrating strong potentials for downstream applications.

Limitations. WHAC-A-Mole includes a rich collection of multiperson scenarios that may require special algorithm designs to tackle close interaction and occlusions, which WHAC lacks. We leave this to the future work.

Potential negative societal impact. WHAC may be used for unwarranted surveillance as it recovers human trajectories in the world frame.

References

1. Bazavan, E.G., Zanfir, A., Zanfir, M., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Hspace: Synthetic parametric humans animated in complex environments. arXiv preprint arXiv:2112.12867 (2021) 11
2. Black, M.J., Patel, P., Tesch, J., Yang, J.: Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8726–8737 (2023) 2, 8
3. Cai, Z., Jiang, J., Qing, Z., Guo, X., Zhang, M., Lin, Z., Mei, H., Wei, C., Wang, R., Yin, W., et al.: Digital life project: Autonomous 3d characters with social intelligence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 582–592 (2024) 2, 8
4. Cai, Z., Ren, D., Zeng, A., Lin, Z., Yu, T., Wang, W., Fan, X., Gao, Y., Yu, Y., Pan, L., et al.: Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In: Proceedings of the European Conference on Computer Vision. pp. 557–577. Springer (2022) 3
5. Cai, Z., Yin, W., Zeng, A., Wei, C., Sun, Q., Yanjun, W., Pang, H.E., Mei, H., Zhang, M., Zhang, L., Loy, C.C., Yang, L., Liu, Z.: Smpler-x: Scaling up expressive human pose and shape estimation. In: Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 11454–11468. Curran Associates, Inc. (2023) 2, 3, 4, 5, 6, 10, 11, 12, 13
6. Cai, Z., Zhang, M., Ren, J., Wei, C., Ren, D., Lin, Z., Zhao, H., Yang, L., Loy, C.C., Liu, Z.: Playing for 3d human recovery. arXiv preprint arXiv:2110.07588 (2021) 2
7. Cheng, W., Chen, R., Fan, S., Yin, W., Chen, K., Cai, Z., Wang, J., Gao, Y., Yu, Z., Lin, Z., et al.: Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19982–19993 (2023) 3
8. Choutas, V., Pavlakos, G., Bolkart, T., Tzionas, D., Black, M.J.: Monocular expressive body regression through body-driven attention. In: Proceedings of the European Conference on Computer Vision. pp. 20–40. Springer (2020) 3
9. Contributors, X.: Openxrlab synthetic data rendering toolbox. https://github.com/openxrlab/xrfeitoria (2023) 8
10. Feng, Y., Choutas, V., Bolkart, T., Tzionas, D., Black, M.J.: Collaborative regression of expressive bodies using moderation. In: Proceedings of the International Conference on 3D Vision. pp. 792–804. IEEE (2021) 3
11. Guzov, V., Mir, A., Sattler, T., Pons-Moll, G.: Human positioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4318–4329 (2021) 3
12. Hasler, N., Rosenhahn, B., Thormahlen, T., Wand, M., Gall, J., Seidel, H.P.: Markerless motion capture with unsynchronized moving cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 224–231. IEEE (2009) 3
13. Huang, B., Shu, Y., Zhang, T., Wang, Y.: Dynamic multi-person mesh recovery from uncalibrated multi-view cameras. In: Proceedings of the International Conference on 3D Vision. pp. 710–720. IEEE (2021) 3
14. Huang, C.H.P., Yi, H., Höschle, M., Safroshkin, M., Alexiadis, T., Polikovsky, S., Scharstein, D., Black, M.J.: Capturing and inferring dense full-body human-scene

contact. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13274–13285 (2022) 8, 11
15. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: Proceedings of the British Machine Vision Conference. pp. 1–11. British Machine Vision Association (2010) 6
16. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3334–3342 (2015) 3
17. Kaufmann, M., Song, J., Guo, C., Shen, K., Jiang, T., Tang, C., Zárate, J.J., Hilliges, O.: Emdb: The electromagnetic database of global 3d human pose and shape in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14632–14643 (2023) 3, 8, 11, 13
18. Kissos, I., Fritz, L., Goldman, M., Meir, O., Oks, E., Kliger, M.: Beyond weak perspective for monocular 3d human pose estimation. In: Proceedings of the European Conference on Computer Vision. pp. 541–554. Springer (2020) 2, 6, 14
19. Kocabas, M., Yuan, Y., Molchanov, P., Guo, Y., Black, M.J., Hilliges, O., Kautz, J., Iqbal, U.: Pace: Human and camera motion estimation from in-the-wild videos. In: Proceedings of the International Conference on 3D Vision. pp. 397–408. IEEE (2024) 4, 8, 11
20. Li, J., Bian, S., Xu, C., Chen, Z., Yang, L., Lu, C.: Hybrik-x: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery. arXiv preprint arXiv:2304.05690 (2023) 3
21. Li, J., Bian, S., Xu, C., Liu, G., Yu, G., Lu, C.: D &d: Learning human dynamics from dynamic camera. In: Proceedings of the European Conference on Computer Vision. pp. 479–496. Springer (2022) 3
22. Lin, J., Zeng, A., Wang, H., Zhang, L., Li, Y.: One-stage 3d whole-body mesh recovery with component aware transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21159–21168 (2023) 3, 5, 6, 11, 12, 13
23. Liu, M., Yang, D., Zhang, Y., Cui, Z., Rehg, J.M., Tang, S.: 4d human body capture from egocentric video via 3d scene grounding. In: Proceedings of the International Conference on 3D Vision. pp. 930–939. IEEE (2021) 3
24. Luvizon, D.C., Habermann, M., Golyanik, V., Kortylewski, A., Theobalt, C.: Scene-aware 3d multi-human motion capture from a single camera. In: Computer Graphics Forum. vol. 42, pp. 371–383. Wiley Online Library (2023) 3
25. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5442–5451 (2019) 2, 8
26. von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: Proceedings of the European Conference on Computer Vision. pp. 601–617 (2018) 3, 8, 11, 13
27. Moon, G., Choi, H., Lee, K.M.: Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2308–2317 (2022) 3, 5, 6, 12, 13
28. Pang, H.E., Cai, Z., Yang, L., Tao, Q., Wu, Z., Zhang, T., Liu, Z.: Towards robust and expressive whole-body human pose and shape estimation. In: Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 17330–17344. Curran Associates, Inc. (2023) 3

29. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10975–10985 (2019) 1, 3
30. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9054–9063 (2021) 3
31. Rao, A., Jiang, X., Guo, Y., Xu, L., Yang, L., Jin, L., Lin, D., Dai, B.: Dynamic storyboard generation in an engine-based virtual environment for video production. In: ACM SIGGRAPH 2023 Posters, pp. 1–2 (2023) 9
32. 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) 3
33. Rong, Y., Shiratori, T., Joo, H.: Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1749–1759 (2021) 3
34. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4104–4113 (2016) 3
35. Shin, S., Kim, J., Halilaj, E., Black, M.J.: Wham: Reconstructing world-grounded humans with accurate 3d motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2070–2080 (2024) 4, 6, 11, 12
36. Siyao, L., Gu, T., Yang, Z., Lin, Z., Liu, Z., Ding, H., Yang, L., Loy, C.C.: Duolando: Follower gpt with off-policy reinforcement learning for dance accompaniment. In: Proceedings of the Twelfth International Conference on Learning Representations (2023) 2, 8
37. Sun, Y., Bao, Q., Liu, W., Mei, T., Black, M.J.: Trace: 5d temporal regression of avatars with dynamic cameras in 3d environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8856–8866 (2023) 3
38. Teed, Z., Deng, J.: Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. Advances in Neural Information Processing Systems **34**, 16558–16569 (2021) 3
39. Teed, Z., Lipson, L., Deng, J.: Deep patch visual odometry. Advances in Neural Information Processing Systems **36** (2024) 2, 4, 5, 11, 12
40. Umeyama, S.: Least-squares estimation of transformation parameters between two point patterns. IEEE Transactions on Pattern Analysis & Machine Intelligence **13**(04), 376–380 (1991) 7
41. Yang, Z., Cai, Z., Mei, H., Liu, S., Chen, Z., Xiao, W., Wei, Y., Qing, Z., Wei, C., Dai, B., Wu, W., Qian, C., Lin, D., Liu, Z., Yang, L.: Synbody: Synthetic dataset with layered human models for 3d human perception and modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20282–20292 (October 2023) 2, 8
42. Ye, V., Pavlakos, G., Malik, J., Kanazawa, A.: Decoupling human and camera motion from videos in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21222–21232 (2023) 11, 12
43. Yu, R., Park, H., Lee, J.: Human dynamics from monocular video with dynamic camera movements. ACM Transactions on Graphics **40**(6), 1–14 (2021) 3

44. Yuan, Y., Iqbal, U., Molchanov, P., Kitani, K., Kautz, J.: Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11038–11049 (2022) 3, 12
45. Zhang, H., Tian, Y., Zhang, Y., Li, M., An, L., Sun, Z., Liu, Y.: Pymaf-x: Towards well-aligned full-body model regression from monocular images. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) 3
46. 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: Proceedings of the European Conference on Computer Vision. pp. 180–200. Springer (2022) 8, 10
47. Zhang, Y., An, L., Yu, T., Li, X., Li, K., Liu, Y.: 4d association graph for realtime multi-person motion capture using multiple video cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1324–1333 (2020) 3
48. Zhou, Y., Habermann, M., Habibie, I., Tewari, A., Theobalt, C., Xu, F.: Monocular real-time full body capture with inter-part correlations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4811–4822 (2021) 3