EmbodMocap: In-the-Wild 4D Human-Scene Reconstruction for Embodied Agents

Wenjia Wang[1*] Liang Pan[1*] Huaijin Pi[1] Yuke Lou[1] Xuqian Ren[2] Yifan Wu[1]
Zhouyingcheng Liao[1] Lei Yang[3] Rishabh Dabral[4] Christian Theobalt[4] Taku Komura[1]

(*: Core contributor.)

[1]The University of Hong Kong [2]Tampere University
[3]The Chinese University of Hong Kong [4]Max-Planck Institute for Informatics

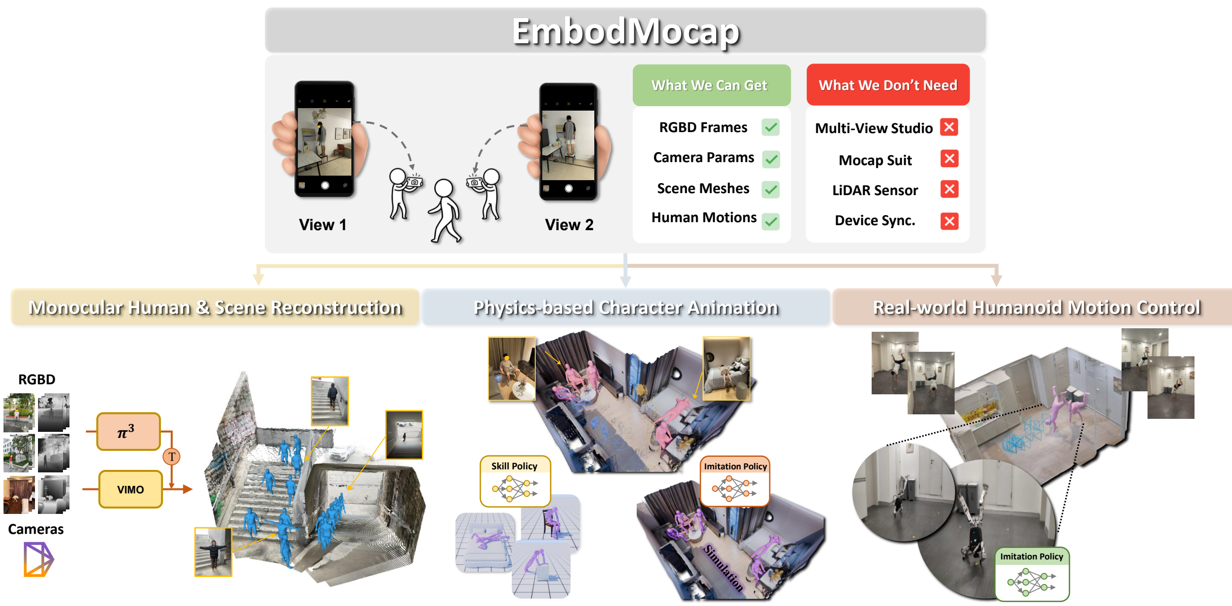

Figure 1. Introducing **EmbodMocap**, a portable and low-cost system for simultaneous 4D human and scene reconstruction, deployable anywhere using two moving iPhones. The dataset captured by EmbodMocap benefits three crucial embodied AI tasks: monocular human & scene reconstruction, physics-based character animation, and real-world humanoid motion control. Project page.

Abstract

Human behaviors in the real world naturally encode rich, long-term contextual information that can be leveraged to train embodied agents for perception, understanding, and acting. However, existing capture systems typically rely on costly studio setups and wearable devices, limiting the large-scale collection of scene-conditioned human motion data in the wild. To address this, we propose EmbodMocap, a portable and affordable data collection pipeline using two moving iPhones. Our key idea is to jointly calibrate dual RGB-D sequences to reconstruct both humans and scenes within a unified metric world coordinate frame. The proposed method allows metric-scale and scene-consistent capture in everyday environments without static cameras or markers, bridging human motion and scene geometry seamlessly. Compared with optical capture ground truth, we demonstrate that the dual-view setting exhibits a remarkable ability to mitigate depth ambiguity, achieving superior alignment and reconstruction performance over single iphone or monocular models. Based on the collected data, we empower three embodied AI tasks: monocular human-scene-reconstruction, where we fine-tune on feedforward models that output metric-scale, world-space aligned humans and scenes; physics-based character animation, where we prove our data could be used to scale human-object interaction skills and scene-aware motion tracking; and robot motion control, where we train a hu-

manoid robot via sim-to-real RL to replicate human motions depicted in videos. Experimental results validate the effectiveness of our pipeline and its contributions towards advancing embodied AI research.

1. Introduction

Embodied Artificial Intelligence (Embodied AI) aims to build agents that can perceive, understand, and act within real-world environments. Progress in this field relies on datasets that capture both human motion and the surrounding 3D scene, enabling physically grounded perception and action learning. Such scene-aware data allows modeling of realistic human–scene interactions, simulation of lifelike behaviors, and training of humanoids to operate seamlessly in complex environments. They serve as a foundation for advancing embodied reasoning and control across robotics, virtual reality, and computer vision.

However, collecting high-quality human–scene data remains difficult. Precise 3D motion and scene geometry cannot be automatically obtained from internet videos due to occlusions and depth ambiguity. Existing capture systems that provide high-quality human–scene data typically rely on multi-view camera rigs [12, 74], wearable motion suits [23, 35], or LiDAR scanners [7, 20], which are costly, complex, and limited to controlled studio environments. These constraints hinder scalable and scene-aware data acquisition, limiting the ability of embodied AI models to learn from natural human behavior in diverse indoor and outdoor environments.

In this paper, we propose EmbodMocap, an efficient and affordable framework for capturing metrically accurate 4D human and scene using only two iPhones. Our key idea is to jointly calibrate and optimize dual RGB-D inputs to reconstruct both humans and scenes within a unified world coordinate frame. Specifically, we first reconstruct the static scene from a single RGB-D sequence to define the world scale, then capture synchronized dual-view RGB-D videos of human motion, and finally perform geometric alignment and motion optimization to recover world-anchored human poses. In contrast to existing systems that rely on multi-camera rigs or wearable sensors, our approach achieves high-quality, scene-consistent reconstruction using only moving consumer devices. This design enables scalable, in-the-wild data collection that preserves precise human motion and authentic scene context, supporting realistic human–scene interaction modeling for embodied AI research.

Based on the data collected with EmbodMocap, we demonstrate the reliability and versatility of our capture pipeline through three representative applications. The first application verifies geometric consistency, where we fine-tune reconstruction models to jointly recover humans and scenes in world coordinates. The second validates physical realism, showing that the captured motions enable scalable training of physics-based character skills and scene-aware motion tracking. The third demonstrates embodied transferability, where our data support humanoid robot training through a sim-to-real motion tracking framework [27, 44]. These results highlight that EmbodMocap enables scalable and physically grounded data acquisition for embodied AI.

In summary, our contributions are:

- **EmbodMocap**: A portable capture framework that jointly calibrates and optimizes dual moving RGB-D cameras (iPhones) to reconstruct metrically accurate, world-anchored human motions and static scenes without multi-camera setups, mocap suits, or controlled environments.
- **A multi-modal dataset**: A collection of high-quality, scene-aware human motion data captured with EmbodMocap across diverse real-world environments, enabling scalable training for embodied AI.

We validate the effectiveness of our method and dataset through experiments in monocular human-scene reconstruction, physics-based character animation, and sim-to-real humanoid control, demonstrating their utility across key embodied AI tasks.

2. Related Work

Datasets for 4D Human & Scene Capture. Early motion datasets, such as AMASS [11, 36], focus on pure human motion, unifying multiple motion capture sources into a large-scale repository. While invaluable for studying human motion, these datasets lack the 3D scene context essential for understanding human–scene interactions. Recent 4D datasets, like PROX [12], RICH [20], and EgoBody [74], combine scanned 3D scenes with motion capture using multi-view camera systems, while EMDB [23] and SPLOPER4D [7], employ IMUs or electromagnetic sensors for motion recording in large-scale environments. Nymeria [35] extends this further with Project Aria glasses and optical marker-based systems for wide-area motion capture. However, these approaches face notable limitations: marker-based and multi-camera systems are expensive and restricted to small studio environments, while IMU and EM-based methods, though more flexible, require extensive manual alignment and post-processing to synchronize motion with 3D scenes. And the wearable devices will influence the human appearance in RGB images. In contrast, our approach uses minimal equipment, operates in diverse environments without static camera setups, and avoids wearable devices, preserving the naturalness of RGB images for authentic human–scene interaction capture. Table 1 compares these datasets.

Monocular Human & Scene Reconstruction. Early works [4, 9, 22, 25, 42] on RGB-based human mesh recovery focus on reconstructing 3D pose and shape but of-

Table 1. Comparison of 4D Human & Scene datasets based on different features.

Datasets	Publication	Device					Outcome		
		Mocap Suit	Scanner	Static Cam.	Dyna. Cam.	Total Cost($)	Mesh	Dyna.Anno.	Outdoor
PROX [12]	ICCV2019	-	Structure Sensor	Kinetic-One	-	2K	✓	✗	✗
RICH [20]	CVPR 2022	-	Leica RTC360	6-8×Cameras	1×Camera	20K+	✓	✓	✓
EgoBody [74]	ECCV2022	-	1×IPhone	5×Azure Kinect	Hololens2	9K	✓	✓	✗
SLOPER4D [7]	CVPR2023	Noitom PN+NUC11	Ouster-os1 LiDAR	-	DJI-Action2+TLS	20K	✓	✓	✓
EMDB [23]	ICCV 2023	EM Sensors	-	-	1×IPhone	15K	✗	✓	✓
Nymeria [35]	ECCV2024	2×XSens+Aria Wistband	-	-	2×Project Aria	60K+	✗	✓	✓
EmbodMocap	-	-	1×IPhone	-	2×IPhone	1K	✓	✓	✓

ten ignore scene context [59] or camera information [26, 62], leading to inconsistencies under camera motion. Recent methods address this by combining motion cues [73], SLAM or visual odometry [54, 64, 72], and human motion priors [53, 73] to recover global trajectories in world coordinates.

Emerging models move toward jointly reconstructing humans and 3D scenes with spatial intelligence models [60, 61]. For example, HSFM [38] combines Dust3R [61] with multi-view correspondence to jointly recover human meshes, scene point clouds, and camera parameters from multi-cameras. HAMSt3R [49] integrates DensePose [10] and multi-view scene reconstruction in one model, with an optimization to get human poses, while JOSH [29] uses MASt3R-SLAM [39] and joint optimization to achieve globally consistent 4D human-scene reconstructions. Human3R [5] introduces a unified, feed-forward framework for online 4D human-scene reconstruction, jointly recovering multi-person SMPL-X bodies and dense scene point clouds in a global world frame from monocular videos. Crisp [66] presents a contact-guided Real2Sim pipeline that recovers simulatable human motion and scene geometry by fitting compact planar primitives and leveraging human-scene contact cues to hallucinate occluded interaction surfaces. This trend emphasizes the simultaneous prediction of human motion and scene geometry, which further requires multi-model data pairs with high-quality annotations. In our paper, we propose a monocular human & scene reconstruction pipeline combined with 2 feedforward models, and finetuned it on our proposed dataset to prove the efficiency of our paired data.

Training Humanoid from Video Data. Recent advances in physics-based animation and reinforcement learning enable humanoid agents to perform realistic and physically consistent motions using control policies learned from marker-based motion capture data. These methods have shown strong realism in tasks like motion tracking [32, 44], locomotion [33, 45, 46], and human–scene interaction [41, 63], and have been extended to real-world applications in motion tracking [16, 18, 21], locomotion [17], and scene interaction [3, 15]. However, marker-based methods require dedicated studios, expensive hardware, and extensive manual effort, making them costly and hard to scale. Adapting captured motions to new scenes or robot morphologies also demands complex retargeting and re-simulation. To address this, recent works like VideoMimic [2], ASAP [18], and HDMI [67] train humanoid control directly from in-the-wild video data. By using monocular motion capture methods such as TRAM [64] and GVHMR [53], they estimate human motion from videos and retarget it to virtual humanoids for training in physical simulators. This video-driven paradigm leverages diverse real-world data but struggles with capturing complex skills or scene geometries due to occlusion and depth ambiguities. In this paper, we propose a method for high-precision human motion and scene reconstruction that overcomes these limitations.

3. Proposed Capture System

We aim to capture metrically accurate human motion and scene geometry using only two iPhones. As shown in Fig. 2, our capture process consists of four sequential stages that progressively reconstruct and align the scene, cameras, and human motion within a unified world coordinate frame. We first reconstruct a metrically accurate static scene and establish the world reference using a single iPhone RGB-D sequence (Sec. 3.1). Then, we use two synchronized iPhones to record dual-view RGB-D videos of human motion and extract per-frame camera poses and human priors with off-the-shelf perception models (Sec. 3.2). Next, we align the dual-view camera trajectories to the reconstructed scene through a combination of COLMAP registration and multi-view geometric optimization (Sec. 3.3). Finally, we refine the SMPL parameters by triangulating dual-view 2D keypoints into 3D space and optimizing human poses and translations in the world coordinate system (Sec. 3.4).

3.1. Stage I: Scene Reconstruction

In this stage, we aim to reconstruct a metrically accurate, Z-up scene mesh that serves as the reference world coordinate system. We first use a single iPhone to capture an RGB-D video of the scene, along with synchronized IMU data. The recorded data are processed by the SpectacularAI SDK (SAI) [1], which automatically selects keyframes according to the accumulated camera translation and esti-

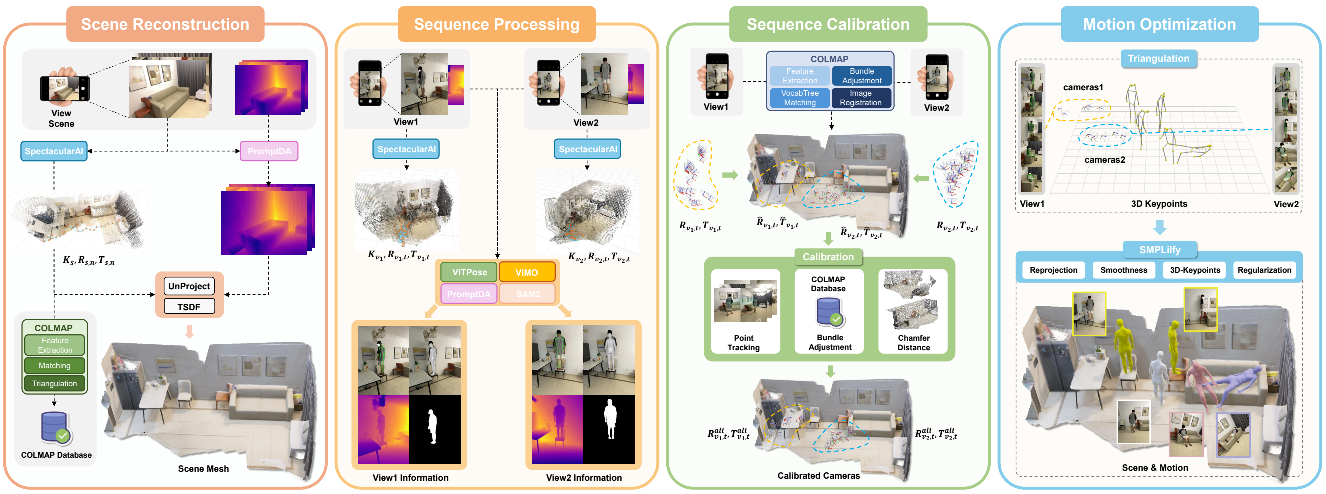

Figure 2. EmbodMocap: We propose an affordable dataset capture and processing system. From left to right, the four stages (Stage-I to Stage-IV) illustrate our core logic: leveraging high-quality camera matrices provided by SpectacularAI [1] and aligning sequence coordinates to the scene's world frame. For detailed explanations, please refer to Sec. 3.

mates corresponding camera parameters $(\boldsymbol{K}_s, \boldsymbol{R}_{s,n}, \boldsymbol{T}_{s,n})$ in Z-up world coordinates with metric scale. These trajectories establish a consistent world frame for all subsequent stages. Based on the recovered poses, we refine the iPhone LiDAR depth maps using PromptDA [28], unproject them into 3D space, and integrate the point clouds through TSDF fusion [6] to obtain a dense and metrically accurate global mesh $\mathcal{M}_g$. Note that the depth maps are truncated based on a threshold determined by the effective range of the iPhone's depth sensor. Specifically, we use a threshold of 3.5m for indoor scenes and 5m for outdoor scenes. We further apply lightweight post-processing such as outlier removal and small-component filtering to clean the mesh. Finally, we extract SIFT features from the same SAI keyframes and run COLMAP [50] with fixed camera parameters to build a sparse structure database. This database preserves the metric scale and serves as a reference for registering dual-view sequences in later stages.

3.2. Stage II: Sequence Processing

After reconstructing the static scene in Stage I, we proceed to capture and process dual-view human motion sequences within the same environment. In this stage, we use two iPhones to record synchronized RGB-D videos of a performer moving inside the reconstructed scene, with each device providing an independent camera coordinate system. The goal is to convert these raw dual-view videos into temporally aligned and metrically consistent per-frame human and camera information, which will serve as the foundation for subsequent calibration and motion optimization.

Firstly, we use SAI to obtain per-frame calibrated cameras for each view. Let v denote the view index ($v \in \{v_1, v_2\}$), and let t denote index time. For each view independently, SAI provides intrinsics and extrinsics $(\boldsymbol{K}_v, \boldsymbol{R}_{v,t}, \boldsymbol{T}_{v,t})$ for every decoded frame $\boldsymbol{I}_{v,t}$ in the native coordinate system of that view.

Next, we extract human-related information using several off-the-shelf models: (i) YOLO [55] for person detection and proposal pruning; (ii) ViTPose [70] for 2D human keypoints with confidence scores; (iii) SAM2 [48] for person segmentation masks; (iv) PromptDA [28] to refine dual-view depths; and (v) VIMO [64] for camera space SMPL parameters. Finally, we employ a laser pointer cue for frame-level synchronization between the two camera streams. By identifying the frame index where the laser dot disappears, we temporally align both videos and slice all associated image, depth, and parameter data accordingly. This process yields synchronized dual-view RGB-D sequences with calibrated camera trajectories and per-frame human priors, providing clean inputs for subsequent sequence calibration.

3.3. Stage III: Sequence Calibration

After obtaining the static scene reconstruction in Stage 3.1 and the dual-view camera trajectories in Stage 3.2, the next step is to align all coordinate systems into a unified world frame. At this point, we have three separate coordinate systems: one for the reconstructed scene and two for each iPhone camera trajectory estimated by SAI. Since the dual-view coordinate systems differ from the scene coordinate system only by rigid transformations, our goal is to optimize these 2 rigid transformations to unify the dual-view coordinates into the same metric, gravity-aligned world frame. The optimization process is sensitive to the initial values; therefore, it is necessary to first obtain a good initial estimate for the rigid transformations.

Get Initial Transformation from COLMAP. We register each dual-view sequence to the sparse COLMAP model constructed in Stage 3.1 using the known intrinsics K_v and background-only SIFT features $\mathcal{F}_v$, extracted from images with human regions removed. Matches are established through a trained vocabulary tree [51], and images are registered against the sparse COLMAP model to obtain COLMAP camera poses $(\hat{\boldsymbol{R}}_{v,t}, \hat{\boldsymbol{T}}_{v,t})$ in the same metric, gravity-aligned world coordinates as the scene.

To obtain the initial rigid transformation aligning the SAI camera trajectories $\boldsymbol{T}_{v,t}$ with their COLMAP counterparts $\hat{\boldsymbol{T}}_{v,t}$, we solve for an offset transformation $(s^{\text{off}}, \boldsymbol{R}^{\text{off}}, \boldsymbol{T}^{\text{off}})$ by minimizing:

$$\min_{s^{\text{off}}, \boldsymbol{R}^{\text{off}}, \boldsymbol{T}^{\text{off}}} \sum_{t=1}^{N} \left\| \hat{\boldsymbol{T}}_{v,t} - (s^{\text{off}} \boldsymbol{R}^{\text{off}} \boldsymbol{T}_{v,t} + \boldsymbol{T}^{\text{off}}) \right\|_2^2, \tag{1}$$

where N is the number of frames. After centering the trajectories, we solve this minimization problem using singular value decomposition (SVD).

For gravity alignment, $\boldsymbol{R}^{\text{off}}$ is constrained to rotations about the z-axis, ensuring proper alignment of SAI trajectories with the COLMAP coordinate system.

Calibration via Multiple Constraints. While the rigid transformations obtained in the previous step provide coarse alignment between the two camera trajectories and the reconstructed scene, this initialization alone is not sufficient to achieve accurate synchronization and metric consistency. To further refine the calibration, we jointly optimize all alignment parameters by introducing multiple geometric and photometric constraints across views. Specifically, we optimize the per-view global offsets R_v^{off} (constrained to z-axis rotations) and T_v^{off}, using the initial alignment as the starting value. The aligned camera extrinsics are:

$$\boldsymbol{R}_{v,t}^{\text{ali}} = \boldsymbol{R}_v^{\text{off}} \boldsymbol{R}_{v,t}, \quad \boldsymbol{T}_{v,t}^{\text{ali}} = \boldsymbol{R}_v^{\text{off}} \boldsymbol{T}_{v,t} + \boldsymbol{T}_v^{\text{off}}. \tag{2}$$

The optimization minimizes a composite loss of point tracking loss, Chamfer distance, and bundle adjustment loss to ensure spatial consistency between views and the global reconstruction.

$$\mathcal{L}_{\text{calib}} = \lambda_{\text{track}} \mathcal{L}_{\text{track}} + \sum_v \lambda_{\text{ch}} d_{\text{Chamfer}} + \sum_v \lambda_{\text{ba}} \mathcal{L}_{\text{ba},v}. \tag{3}$$

where each loss is defined in the rest of this seciton.

Through VGGT tracking, a subset of keyframes is selected, yielding accurate dual-view pixel tracking results in the human masks region. The tracked human surface 2D pixel coordinates $\boldsymbol{q}_{v,t}^{(i)}$, along with their corresponding depth values $d_{v,t}^{(i)}$, are back-projected into the world frame:

$$\boldsymbol{Q}_{v,t}^{(i)} = d_{v,t}^{(i)} \boldsymbol{R}_{v,t}^{\top\text{ali}} \boldsymbol{K}_v^{-1} \begin{bmatrix} \boldsymbol{q}_{v,t}^{(i)} \\ 1 \end{bmatrix} + \boldsymbol{R}_{v,t}^{\top\text{ali}} \boldsymbol{T}_{v,t}^{\text{ali}}, \tag{4}$$

To enforce track consistency between views, the following loss is minimized:

$$\mathcal{L}_{\text{track}} = \frac{1}{\sum_{v,t} |\mathcal{Q}_{v,t}|} \sum_t \sum_i \tilde{w}_t^{(i)} \left\| \boldsymbol{Q}_{1,t}^{(i)} - \boldsymbol{Q}_{2,t}^{(i)} \right\|_2^2, \tag{5}$$

where $\boldsymbol{Q}_{1,t}^{(i)}$ and $\boldsymbol{Q}_{2,t}^{(i)}$ are the 3D back-projected coordinates of the i-th point from view 1 and view 2, respectively. The weights $\tilde{w}_t^{(i)}$ are used to control the contribution of each point based on its tracking confidence. Here $\tilde{w}_t^{(i)} = \min(w_{1,t}^{(i)}, w_{2,t}^{(i)})$ combines the VGGT confidence scores for the same point across views. The Chamfer distance term d_{Chamfer} aligns local pointclouds $\boldsymbol{\mathcal{P}}_v$ $(v \in \{v_1, v_2\})$ with the global reconstruction $\boldsymbol{\mathcal{P}}_{\text{g}}$ sampled from $\mathcal{M}_{\text{g}}$ in Sec. 3.1, where $\boldsymbol{\mathcal{P}}_v$ is obtained by reconstructing the scene using the method from Sec. 3.1 with humans cropped by masks. The Chamfer distance is formally defined as:

$$\begin{aligned} d_{\text{Chamfer}}(\boldsymbol{\mathcal{P}}_v, \boldsymbol{\mathcal{P}}_{\text{g}}) = \frac{1}{|\boldsymbol{\mathcal{P}}_v|} \sum_{\boldsymbol{p}_v \in \boldsymbol{\mathcal{P}}_v} \min_{\boldsymbol{p}_{\text{g}} \in \boldsymbol{\mathcal{P}}_{\text{g}}} \|\boldsymbol{p}_v - \boldsymbol{p}_{\text{g}}\|_2^2 \\ + \frac{1}{|\boldsymbol{\mathcal{P}}_{\text{g}}|} \sum_{\boldsymbol{p}_{\text{g}} \in \boldsymbol{\mathcal{P}}_{\text{g}}} \min_{\boldsymbol{p}_v \in \boldsymbol{\mathcal{P}}_v} \|\boldsymbol{p}_{\text{g}} - \boldsymbol{p}_v\|_2^2. \end{aligned} \tag{6}$$

Finally, $\mathcal{L}_{\text{ba},v}$ $(v \in \{v_1, v_2\})$ ensures reprojection consistency for persistent matches, where the points are obtained from COLMAP image registration:

$$\mathcal{L}_{\text{ba},v} = \frac{1}{|M_v|} \sum_{(t,j) \in M_v} \left\| \boldsymbol{x}_{v,t,j} - \pi(\boldsymbol{K}_v, \boldsymbol{R}_{v,t}^{\text{ali}}, \boldsymbol{T}_{v,t}^{\text{ali}}, \boldsymbol{X}_j) \right\|_2^2. \tag{7}$$

We solve Eq. (3) using the Adam [24] optimizer with gradient clipping. For yaw-only updates, R_v^{off} is parameterized by a single z-axis angle to preserve gravity alignment.

3.4. Stage IV: Motion Optimization

After obtaining calibrated dual-view trajectories and a unified scene coordinate system in Stage 3.3, we further refine the human reconstruction results to achieve accurate and temporally consistent body motions in the world frame. At this stage, both camera poses and scene geometry are fixed, allowing us to focus on optimizing the human parameters. We first triangulate dual-view 2D keypoints into world-space 3D keypoints, which serve as reliable geometric constraints across views. Then, we optimize the SMPL parameters using these triangulated 3D keypoints to recover precise body poses and translations under the unified world coordinate system.

3D Keypoint Triangulation. To triangulate the 3D keypoints $\boldsymbol{Y}_{t,j}$ from their 2D projections $\{y_{v,t,j}\}$, we estimate the 3D position by minimizing the weighted reprojection error across all views:

$$\min_{\boldsymbol{Y}_{t,j}} \sum_{v=1}^{V} c_{v,t,j} \left\| \boldsymbol{y}_{v,t,j} - \boldsymbol{P}_v \boldsymbol{Y}_{t,j} \right\|_2^2, \tag{8}$$

where $\boldsymbol{P}_v = \boldsymbol{K}_v[\boldsymbol{R}_{v,t} \mid \boldsymbol{T}_{v,t}]$ is the camera projection matrix for the v-th view. The problem can be formulated as a weighted least squares optimization. Using SVD, $\boldsymbol{Y}_{t,j}$ is obtained as the right singular vector corresponding to the smallest singular value of $\boldsymbol{A}$.

World-Space SMPLify. Start from initial shape $\boldsymbol{\beta}_0$ and body pose $\boldsymbol{\theta}_t^{\mathrm{b},0}$ in Sec. 3.2, our World Frame SMPLify [30] jointly optimizes shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$, per-frame pose $\boldsymbol{\theta}_t = \{\boldsymbol{\theta}_t^{\mathrm{g}}, \boldsymbol{\theta}_t^{\mathrm{b}}\} \in \mathbb{R}^{72}$ and root translation $\boldsymbol{\gamma}_t \in \mathbb{R}^3$ by minimizing:

$$\mathcal{L}_{\text{SMPLify}} = \mathcal{L}_{\text{3D}} + \mathcal{L}_{\text{smooth}} + \mathcal{L}_{\text{prior}} + \mathcal{L}_{\text{reproj}} \tag{9}$$

We use a two-stage optimization phase to ensure the smoothness and alignment with the original dual views. For the first stage, we only fit the body shape and transition, and for the second stage we fit all the parameters.

4. Evaluation

In this section, we aim to prove the effectness of our optimization pipeline. We compare ours dual view optimization pipeline with the monocular model, single-view only and the optical captured ground truth.

4.1. Comparison on Capture Methods

Direct comparison in optical mocap studio. To evaluate the accuracy of dual view capture system, we set up furniture in a mocap studio and use a Vicon system to capture ground truth human motion. Two photographers record dual-view videos of the actor with iPhones, while the actor performs basic motions(see Fig. 3, zoom in). We record 5 sequences of one participant with 9420 frames in total. We compare the errors against optical mocap GT of: monocular model GVHMR, our dual-view optimization, and our single-view version(v1 and v2). For the single-view version, we calibrate the actor coordinates to the scene coordinates system using COLMAP and optimize the motion with reprojection, smooth, and prior losses. The optical mocap results are fitted to SMPLX parameters by Mosh [31] and synchronized to dual-view parameters with foot contact keyframs. Results are compared in chunk sizes of 100, 500, and 1000. Our dual-view method outperforms the monocular model and single-view optimization by a large margin. As the chunk length increases, our advantage becomes increasingly evident. (see Tab. 2)

Table 2. Comparision among monocular model, single view optimization, with dual view optimization(ours)

Method	chunk=100		chunk=500		chunk=1000		RTE↓
	WA-MPJPE↓	W-MPJPE↓	WA-MPJPE↓	W-MPJPE↓	WA-MPJPE↓	W-MPJPE↓	
GVHMR	66.56	123.44	124.61	333.34	179.47	593.79	1.85
Single-View V1	124.68	218.22	233.06	489.11	297.83	768.31	2.71
Single-View V2	108.31	211.83	231.41	357.22	338.42	762.80	3.65
Dual View	**56.61**	**72.86**	**76.90**	**99.75**	**119.45**	**169.11**	**1.13**

The advantage of dual-view over single-view lies in two key aspects: 1)dual-view effectively addresses occlusion and self-occlusion of body joints, 2)it handles the challenging alignment of actor motion coordinates to the scene coordinates. The COLMAP estimates the camera locations for the images but suffers from depth ambiguity in the camera's facing direction. Using a single iPhone results in large errors in the depth direction. In contrast, using two iPhones enables pixel-wise dense correspondence(see Eq. (5)), which ensures the rigid transformation between the two cameras during the optimization, and resolves the depth ambiguity in each view. ***This enables a good localization of human trajectories in the scene coordinate system automatically.*** Our dual view could achieve a calibration accuracy to the scene of about 5cm (human touching table in the figure), while the single view is over 30cm, measured in MeshLab by putting markers on the ground for the actor's start and end positions.

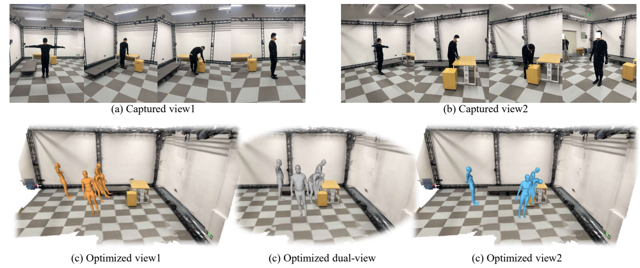

Figure 3. Our dual view *vs.* single view results in optical studio.

5. Downstream Tasks

In this section, we validate our capture pipeline's effectiveness across three key applications. In Sec. 5.1, we propose a monocular human & scene reconstruction pipeline and finetune it with our captured RGBD, cameras, and SMPL annotations. In Sec. 5.2, we train several human-object interaction skills and scene-aware motion tracking with our captured motion & scene. In Sec. 5.3, we train a humanoid in simulator and deploy it to real-world robot.

5.1. Monocular Human & Scene Reconstruction

Motivation. We propose a data scheme combining RGBD data from dynamic cameras with camera and human motion parameters to train monocular human and scene reconstruction models. As no feedforward model exists, we establish a baseline using π^3[65] for SLAM and VIMO[64] for metric-scale human motion reconstruction from monocular videos.

Implementation. To process long sequences, videos are divided into overlapping chunks, with π^3 estimating camera parameters and local point maps per chunk. Adjacent chunks are aligned using Procrustes alignment, and scale/transformations are recursively applied for global consistency. Metric scale is determined as the median ratio of SMPL to π^3 depth values. SMPL predictions are then transformed to metric world space. For details, refer to

Supp. Mat. We fine-tuned two π^3 variants Tab. 3 by adding LoRA [19] layers to the camera and point decoders, supervised with the original π^3 loss. For VIMO, we froze the encoder and finetuned the decoder with MSE loss on SMPL parameters. A human mask was used to limit supervision to the human region due to our dataset's smaller range.

Metrics. We evaluate motion and trajectory accuracy on global coordinates using EMDB (subset 2)[23], featuring extended sequences with ground-truth trajectories and meshes. Consistent with prior work[54, 64], each sequence is split into 100-frame chunks, and 3D joint errors are measured using W-MPJPE (aligning the first two frames) and WA-MPJPE (aligning the entire segment), both in millimeters. Additionally, Root Translation Error (RTE) is reported as a percentage (%), normalized by total displacement after rigid alignment (excluding scaling).

Results. We present 3 variants in Tab. 3: the proposed baseline with the original checkpoints from π^3 [65] and VIMO [64], fine-tuning only VIMO, and fine-tuning both π^3 and VIMO. The results demonstrate that our approach significantly improves the accuracy of VIMO, as we provide paired high-quality real-world RGB sequences and ground truth SMPL parameters. Additionally, leveraging our high-quality RGB-D data and camera parameter pairs, π^3's ability to predict in the world coordinate system also shows improvement. Our pipeline demonstrates good performance on large-scale, real-world videos; see demo images in the Supplement.

Table 3. Comparison of Finetuned Models on EMDB Benchmarks

Finetuned		EMDB		
Pi3	VIMO	WA-MPJPE↓	W-MPJPE↓	RTE↓
✗	✗	83.56	229.04	1.78
✗	✓	82.89	222.93	1.73
✓	✓	**82.21**	**220.65**	**1.71**

5.2. Physics-based Character Animation

5.2.1. Human Object Interaction Skill Training

Motivation. We train several human-object interaction skills to demonstrate the physical realism of our approach and the scalability of our capture framework to new interaction skills. We aim to prove the efficiency and quality superiority of our framework over optical capture and monocular estimation methods.

Implementation. Following [41, 45, 63], we train physical character policies use goal-conditioned reinforcement learning to formulate character control as a Markov Decision Process (MDP) defined by states, actions, transition dynamics, a reward function r, and a discount factor γ. The reward $r_t \in \mathcal{R}$ is calculated by a style reward r_t^{style} [45] and a task reward r_t^{task}. The policies are trained to maximize the expected discounted return: $J(\pi) = \mathbb{E}_{p(\tau|\pi)}\left[\sum_{t=0}^{T-1} \gamma^t r_t\right]$, where T is the episode length, $\gamma \in [0, 1]$ is the discount factor, and r_t is the reward at time step t. We use the widely adopted Proximal Policy Optimization (PPO) algorithm [52] to train the control policy model.

Following [14, 41, 63], we train a set of human object interaction skills in simulator [37], including *follow*, *climb*, *sit*, and *lie*. These common interaction skills are designed to guide the character's root joint to reach specific target positions in 3D environments while maintaining physically realistic and motion divisty. We train these four common skills on 3 different input data: optical captured, which are collected from AMASS [36] and SAMP [13] following TokenHSI [41]; ours, by segmenting the reconstructed motions into skill clips; monocular, by using the motion predicted by GVHMR [53] which is commonly used in humanoid reference motion prediction[18, 67], segmented with the same temporal slices as ours. We also train 2 extra interaction skills which have not been implemented in previous physics-based human object interaction papers: Prone and Support. We will illustrate the observation, reward designs, and the training details of each skill in Supp.Mat.

Metrics. We follow [13, 68] that uses *Success Rate* and *Contact Error* as the main metrics to measure the quality of interactions quantitatively. Success Rate records the percentage of trials that humanoids successfully complete the contact within a certain threshold. We follow [14, 40, 68] in setting the thresholds for various actions: 20cm for Sit, Follow, and Climb; 30cm for Lie and Prone; and 10cm for Support. For Support, the error is defined as the distance from the object surface center to the hand center, while also taking into account the distance between the two feet. Please see details in Supp.Mat. We evaluate motion diversity using Average Pairwise Distance (APD) [8], which measures the average pairwise distance between joint rotations and positions in generated samples. Higher APD values indicate greater diversity.

Results. We can find in Tab. 4, for skills such as Follow, Climb, and Sit, the inherent difficulty is relatively low, and all three data settings achieve good results, very close to 100%. Although the quality of our data is slightly inferior to optically captured data, we provide more variety of task completion trajectories and motion diversities, which contribute to improve task performance. To prove this, we ablate on skills trained with different data proportions. 1X and 2X indicate the ratio of the number of clips relative to the optical capture data. On the 4 common skills, we observe a general trend where increased data amount leads to improvements in success rate, contact error, and APD metrics.

We also implement 2 extra skills, Prone and Support, demonstrate the versatility of our data collection pipeline. First, these new skills highlight the ability of our approach to generalize to novel interaction tasks. Second, the Support skill significantly increases the level of difficulty. Unlike

Table 4. Comparison of data duration, Success Rate, Contact Error, and APD for different skills among 3 data settings.

Task	Data	Clips	Duration (min)	Rate (%) ↑	Error (cm) ↓	APD ↑
Follow	Optical Mocap	12	1.59	**99.9**	**6.0**	**20.17 ± 0.19**
	Ours 1X	12	1.48	**99.9**	6.7	18.42 ± 0.22
	Ours 2X	24	3.06	99.7	6.8	18.45 ± 0.17
	Ours Full	148	22.43	99.8	6.2	19.69 ± 0.32
	Monocular	148	22.43	98.0	7.2	19.85 ± 0.39
Climb	Optical Mocap	7	0.28	**99.9**	2.7	22.03 ± 0.30
	Ours 1X	7	0.54	99.8	**1.8**	**22.77 ± 0.29**
	Ours 2X	14	0.97	**99.9**	**1.8**	20.72 ± 0.30
	Ours Full	21	1.54	**99.9**	**1.8**	22.22 ± 0.27
	Monocular	21	1.54	99.2	1.8	21.34 ± 0.38
Sit	Optical Mocap	20	4.08	98.0	5.5	**16.07 ± 0.39**
	Ours 1X	20	2.11	99.8	5.4	14.35 ± 0.27
	Ours 2X	40	4.47	**99.9**	5.1	14.46 ± 0.24
	Ours Full	80	8.05	**99.9**	**4.7**	15.90 ± 0.51
	Monocular	80	8.05	98.4	5.7	15.80 ± 0.51
Lie	Optical Mocap	10	2.52	89.0	**17.5**	**8.76 ± 0.14**
	Ours 1X	10	0.99	85.3	20.2	7.43 ± 0.10
	Ours 2X	20	2.32	86.3	19.8	8.27 ± 0.06
	Ours Full	39	4.25	**89.4**	18.8	8.57 ± 0.10
	Monocular	39	4.25	81.2	21.0	8.14 ± 0.10
Prone	Ours Full	3	0.26	**75.4**	**16.5**	**17.58 ± 0.69**
	Monocular	3	0.26	71.2	16.5	16.18 ± 0.30
Support	Ours Full	8	0.97	**66.0**	**4.9**	**21.08 ± 0.59**
	Monocular	8	0.97	20.6	6.4	20.94 ± 0.48

other tasks, where a humanoid only needs to walk or offload the full body weight onto furniture surface, Support requires the hands to bear the weight of the body while the feet remain close together, demanding much higher accuracy in reference motion generation. This experiment shows that our approach outperforms monocular estimation methods by a large margin, particularly for high-difficulty interaction skills. The success rate trained on monocular estimated motions degrades to only 20% in Tab. 4.

5.2.2. Scene-aware Motion Tracking

Motivation. Recent works [33, 34, 46, 56–58, 71] suggest that solving complex tasks requires pre-training on large-scale human motion data via motion tracking objectives, in order to obtain reusable and generalizable skill priors. However, existing motion tracking frameworks are mainly built for human-only [32] or single-object interaction [69] scenarios, primarily because current public datasets are concentrated in these settings. We argue that motion tracking pre-training on diverse 3D scenes is equally important, as it also provides rich priors—such as navigation, interaction, and long-horizon task execution. In this work, we mitigate this gap by: 1) proposing a scene-aware motion tracking framework, and 2) supporting it with high-fidelity paired 3D human-scene data captured by our EmbodMocap system.

Implementation. We extend MimicKit [43] by incorporating the height map into the observation space to achieve scene-aware tracking (details in the Supp. Mat.). For training, we use four 3D scenes, each containing several minutes of motion clips, and train one policy per scene to track all the motion clips in that scene.

Metrics. Policies are evaluated using a success rate metric: an episode is initialized from a random frame and run

Table 5. Quantitative evaluation of scene-aware motion tracking and dataset statistics across four 3D scenes.

Scene	Clips	Duration (min)	Status	Rate (%)	Eps. Len. (s)
a	14	12.31	Succ.	87.2	9.97 ± 0.21
			Fail.	12.8	3.94 ± 2.10
b	6	3.62	Succ.	96.7	9.99 ± 0.12
			Fail.	3.3	4.16 ± 2.38
c	12	7.87	Succ.	95.9	9.98 ± 0.17
			Fail.	4.1	5.43 ± 2.18
d	7	5.06	Succ.	90.4	9.96 ± 0.21
			Fail.	9.6	4.44 ± 1.92

for 10s, and is considered successful if tracking exceeds 8s. For each scene, 3,072 episodes are used to compute average success, failure rates, and episode length statistics.

Results. The quantitative results in Tab. 5 demonstrate that our data is simulation-ready, enabling the training of scene-aware tracking policies with high success rates. Please see the qualitative results in the Supplement.

5.3. Real-world Humanoid Robot Control

Motivation. Learning from human videos [2, 47, 67] has emerged as a crucial paradigm for humanoid robots to learn motor skills at scale. In this section, we demonstrate how EmbodMocap contributes to this paradigm by enabling accurate reconstruction of humans and their interacting 3D environments from videos, while preserving accurate contact information.

Implementation. We capture videos of humans performing ground-contact-rich motions, including locomotion and challenging cartwheels that require precise hand-ground contact. EmbodMocap is then used for real-to-sim reconstruction. The produced motions are used to train a single tracking policy via sim-to-real RL with domain randomization using BeyondMimic [27].

Results. We deploy the policy on a real-world High Torque Hi humanoid robot with 21 joint DoF and a height of 80cm. Please see the demo images in the Supplement.

6. Conclusion

We propose EmbodMocap, a portable and affordable framework for capturing high-quality 4D human & scene data using only two iPhones. Our method enables scalable, metrically accurate reconstruction of human motion and scenes mesh in diverse real-world environments. We directly compare in optical capture studios, and prove the superiority in solving body occlusion and sequence coordinate alignment of our dual-view design. Through downstream applications in monocular human-scene reconstruction, physics-based character animation, and humanoid robot motion control, we demonstrate the effectiveness and scalability of our approach. By lowering the barrier for embodied AI research, EmbodMocap opens new opportunities for real-world applications.

7. Acknowledge

This work was partially funded by the Innovation and Technology Commission of the HKSAR Government under the ITSP-Platform grants (Ref: ITS/335/23FP, ITS/469/24FP).

References

[1] Spectacular ai sdk. https://www.spectacularai.com, 2021. 3, 4

[2] Arthur Allshire, Hongsuk Choi, Junyi Zhang, David McAllister, Anthony Zhang, Chung Min Kim, Trevor Darrell, Pieter Abbeel, Jitendra Malik, and Angjoo Kanazawa. Visual imitation enables contextual humanoid control. *arXiv:2505.03729*, 2025. 3, 8

[3] Qingwei Ben, Feiyu Jia, Jia Zeng, Junting Dong, Dahua Lin, and Jiangmiao Pang. Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit. *arXiv:2502.13013*, 2025. 3

[4] 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 *ECCV*, 2016. 2

[5] Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Pons-Moll Gerard. Human3r: Everyone everywhere all at once. *arXiv:2510.06219*, 2025. 3

[6] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *CGIT*, 1996. 4

[7] Yudi Dai, YiTai Lin, XiPing Lin, Chenglu Wen, Lan Xu, Hongwei Yi, Siqi Shen, Yuexin Ma, and Cheng Wang. Sloper4d: A scene-aware dataset for global 4d human pose estimation in urban environments. In *CVPR*, 2023. 2, 3

[8] Zhiyang Dou, Xuelin Chen, Qingnan Fan, Taku Komura, and Wenping Wang. C· ase: Learning conditional adversarial skill embeddings for physics-based characters. In *SIGGRAPH*, 2023. 7

[9] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 2

[10] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 3

[11] Félix G. Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. 2020. 2

[12] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *ICCV*, 2019. 2, 3

[13] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *ICCV*, 2021. 7

[14] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. In *SIGGRAPH*, 2023. 7

[15] Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. *arXiv:2406.08858*, 2024. 3

[16] Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Learning human-to-humanoid real-time whole-body teleoperation. *arXiv:2403.04436*, 2024. 3

[17] Tairan He, Wenli Xiao, Toru Lin, Zhengyi Luo, Zhenjia Xu, Zhenyu Jiang, Jan Kautz, Changliu Liu, Guanya Shi, Xiaolong Wang, Linxi Fan, and Yuke Zhu. Hover: Versatile neural whole-body controller for humanoid robots. *arXiv:2410.21229*, 2024. 3

[18] Tairan He, Jiawei Gao, Wenli Xiao, Yuanhang Zhang, Zi Wang, Jiashun Wang, Zhengyi Luo, Guanqi He, Nikhil Sobanbabu, Chaoyi Pan, Zeji Yi, Guannan Qu, Kris Kitani, Jessica Hodgins, Linxi "Jim" Fan, Yuke Zhu, Changliu Liu, and Guanya Shi. Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills. *arXiv preprint arXiv:2502.01143*, 2025. 3, 7

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 7

[20] 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 *CVPR*, 2022. 2, 3

[21] Mazeyu Ji, Xuanbin Peng, Fangchen Liu, Jialong Li, Ge Yang, Xuxin Cheng, and Xiaolong Wang. Exbody2: Advanced expressive humanoid whole-body control. *arXiv:2412.13196*, 2024. 3

[22] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2

[23] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. Emdb: The electromagnetic database of global 3d human pose and shape in the wild. In *ICCV*, 2023. 2, 3, 7

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. 5

[25] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 2

[26] Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J. Black. SPEC: Seeing people in the wild with an estimated camera. In *ICCV*, 2021. 3

[27] Qiayuan Liao, Takara E Truong, Xiaoyu Huang, Guy Tevet, Koushil Sreenath, and C Karen Liu. Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion. *arXiv preprint arXiv:2508.08241*, 2025. 2, 8

[28] Haotong Lin, Sida Peng, Jingxiao Chen, Songyou Peng, Jiaming Sun, Minghuan Liu, Hujun Bao, Jiashi Feng, Xiaowei Zhou, and Bingyi Kang. Prompting depth anything for 4k resolution accurate metric depth estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17070–17080, 2025. 4

[29] Zhizheng Liu, Joe Lin, Wayne Wu, and Bolei Zhou. Joint optimization for 4d human-scene reconstruction in the wild. *arXiv:2501.02158*, 2025. 3

[30] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *TOG*, 2015. 6

[31] Matthew M. Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 33(6):220:1–220:13, 2014. 6

[32] Zhengyi Luo, Jinkun Cao, Kris Kitani, Weipeng Xu, et al. Perpetual humanoid control for real-time simulated avatars. In *ICCV*, 2023. 3, 8

[33] Zhengyi Luo, Jinkun Cao, Josh Merel, Alexander Winkler, Jing Huang, Kris Kitani, and Weipeng Xu. Universal humanoid motion representations for physics-based control. *arXiv:2310.04582*, 2023. 3, 8

[34] Zhengyi Luo, Jinkun Cao, Sammy Christen, Alexander Winkler, Kris Kitani, and Weipeng Xu. Grasping diverse objects with simulated humanoids. *arXiv:2407.11385*, 2024. 8

[35] Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *ECCV*, 2024. 2, 3

[36] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, 2019. 2, 7

[37] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv:2108.10470*, 2021. 7

[38] Lea Müller, Hongsuk Choi, Anthony Zhang, Brent Yi, Jitendra Malik, and Angjoo Kanazawa. Reconstructing people, places, and cameras. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21948–21958, 2025. 3

[39] Riku Murai, Eric Dexheimer, and Andrew J Davison. Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16695–16705, 2025. 3

[40] Liang Pan, Jingbo Wang, Buzhen Huang, Junyu Zhang, Haofan Wang, Xu Tang, and Yangang Wang. Synthesizing physically plausible human motions in 3d scenes. In *3DV*, 2024. 7

[41] Liang Pan, Zeshi Yang, Zhiyang Dou, Wenjia Wang, Buzhen Huang, Bo Dai, Taku Komura, and Jingbo Wang. Tokenhsi: Unified synthesis of physical human-scene interactions through task tokenization. In *CVPR*, 2025. 3, 7

[42] 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 *CVPR*, 2019. 2

[43] Xue Bin Peng. Mimickit: A reinforcement learning framework for motion imitation and control, 2025. 8

[44] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *TOG*, 2018. 2, 3

[45] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *TOG*, 2021. 3, 7

[46] Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. *TOG*, 2022. 3, 8

[47] Ri-Zhao Qiu, Shiqi Yang, Xuxin Cheng, Chaitanya Chawla, Jialong Li, Tairan He, Ge Yan, David J Yoon, Ryan Hoque, Lars Paulsen, et al. Humanoid policy˜ human policy. *arXiv preprint arXiv:2503.13441*, 2025. 8

[48] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv:2408.00714*, 2024. 4

[49] Sara Rojas, Matthieu Armando, Bernard Ghanem, Philippe Weinzaepfel, Vincent Leroy, and Gregory Rogez. Hamst3r: Human-aware multi-view stereo 3d reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5027–5037, 2025. 3

[50] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 4

[51] Johannes Lutz Schönberger, True Price, Torsten Sattler, Jan-Michael Frahm, and Marc Pollefeys. A vote-and-verify strategy for fast spatial verification in image retrieval. In *ACCV*, 2016. 5

[52] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 7

[53] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia*, 2024. 3, 7

[54] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *CVPR*, 2024. 3, 7

[55] Juan Terven, Diana-Margarita Córdova-Esparza, and Julio-Alejandro Romero-González. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. *Machine learning and knowledge extraction*, 2023. 4

[56] Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. Maskedmimic: Unified physics-based character control through masked motion inpainting. *TOG*, 2024. 8

[57] Chen Tessler, Yifeng Jiang, Erwin Coumans, Zhengyi Luo, Gal Chechik, and Xue Bin Peng. Maskedmanipulator: Versatile whole-body control for loco-manipulation. *arXiv preprint arXiv:2505.19086*, 2025.

[58] Andrea Tirinzoni, Ahmed Touati, Jesse Farebrother, Mateusz Guzek, Anssi Kanervisto, Yingchen Xu, Alessandro Lazaric, and Matteo Pirotta. Zero-shot whole-body humanoid control

via behavioral foundation models. In *The Thirteenth International Conference on Learning Representations*. 8

[59] Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J. Black. DECO: Dense estimation of 3D human-scene contact in the wild. In *ICCV*, 2023. 3

[60] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025. 3

[61] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 3

[62] Wenjia Wang, Yongtao Ge, Haiyi Mei, Zhongang Cai, Qingping Sun, Yanjun Wang, Chunhua Shen, Lei Yang, and Taku Komura. Zolly: Zoom focal length correctly for perspective-distorted human mesh reconstruction. In *ICCV*, 2023. 3

[63] Wenjia Wang, Liang Pan, Zhiyang Dou, Jidong Mei, Zhouyingcheng Liao, Yuke Lou, Yifan Wu, Lei Yang, Jingbo Wang, and Taku Komura. Sims: Simulating stylized human-scene interactions with retrieval-augmented script generation. *ICCV.*, 2025. 3, 7

[64] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *ECCV*. Springer, 2024. 3, 4, 6, 7

[65] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Scalable permutation-equivariant visual geometry learning, 2025. 6, 7

[66] Zihan Wang, Jiashun Wang, Jeff Tan, Yiwen Zhao, Jessica K. Hodgins, Shubham Tulsiani, and Deva Ramanan. Contact-guided real2sim from monocular video with planar scene primitives. In *ICLR*, 2026. 3

[67] Haoyang Weng, Yitang Li, Nikhil Sobanbabu, Zihan Wang, Zhengyi Luo, Tairan He, Deva Ramanan, and Guanya Shi. Hdmi: Learning interactive humanoid whole-body control from human videos. *arXiv:2509.16757*, 2025. 3, 7, 8

[68] Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts. In *ICLR*, 2024. 7

[69] Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. Intermimic: Towards universal whole-body control for physics-based human-object interactions. *arXiv:2502.20390*, 2025. 8

[70] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *NeurIPS*, 2022. 4

[71] Heyuan Yao, Zhenhua Song, Yuyang Zhou, Tenglong Ao, Baoquan Chen, and Libin Liu. Moconvq: Unified physics-based motion control via scalable discrete representations. *ACM Transactions on Graphics (TOG)*, 43(4):1–21, 2024. 8

[72] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *CVPR*, 2023. 3

[73] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *CVPR*, 2022. 3

[74] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *ECCV*. Springer, 2022. 2, 3