MultiPhys: Multi-Person Physics-aware 3D Motion Estimation

Nicolas Ugrinovic*[1,2] Boxiao Pan[2] Georgios Pavlakos[3] Despoina Paschalidou[2]
Bokui Shen[2] Jordi Sanchez-Riera[1] Francesc Moreno-Noguer[1] Leonidas Guibas[2]
[1]Institut de Robotica i Informatica Industrial, CSIC-UPC, Barcelona, Spain
[2]Stanford University [3]UT Austin

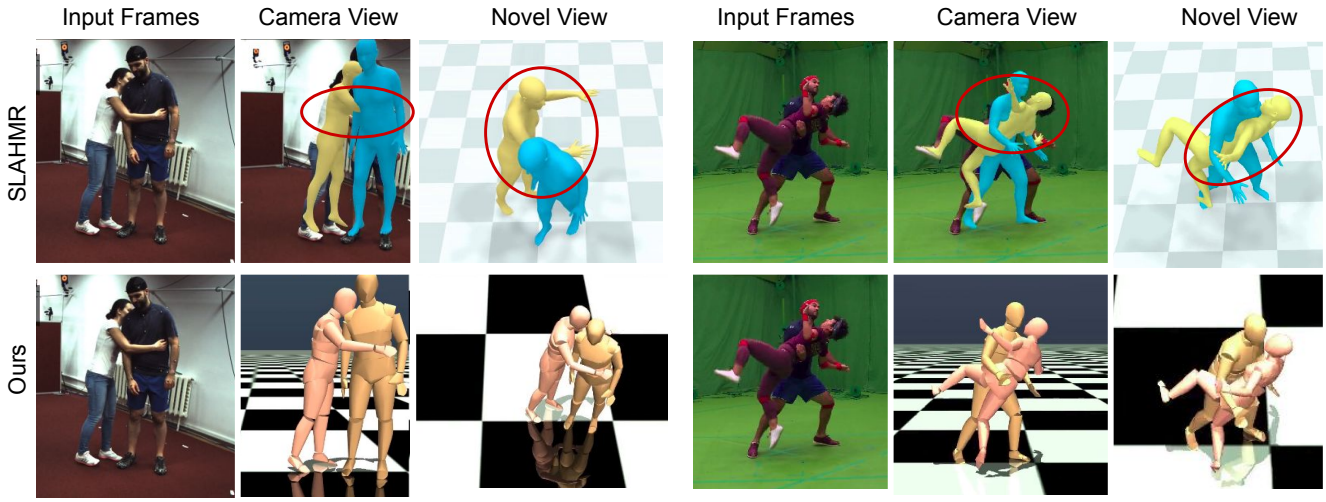


Figure 1. **MultiPhys enables recovering multi-person 3D motion in a physically-aware manner.**. State-of-the-art methods (SLAHMR [39], top row) for multi-person motion recovery mostly rely on kinematic approaches, which typically ignore physical constraints, such as body penetration. Note that while individual poses are kinematically coherent, their spatial placement is suboptimal, resulting in significant penetration errors. MultiPhys (bottom row) incorporates physics constraints into the reconstruction process, yielding more physically plausible results.

Abstract

We introduce MultiPhys, a method designed for recovering multi-person motion from monocular videos. Our focus lies in capturing coherent spatial placement between pairs of individuals across varying degrees of engagement. MultiPhys, being physically aware, exhibits robustness to jittering and occlusions, and effectively eliminates penetration issues between the two individuals. We devise a pipeline in which the motion estimated by a kinematic-based method is fed into a physics simulator in an autoregressive manner. We introduce distinct components that enable our model to harness the simulator's properties without compromising the accuracy of the kinematic estimates. This results in final motion estimates that are both kinematically coherent and physically compliant. Extensive evaluations on three challenging datasets characterized by substantial inter-person interaction show that our method significantly reduces errors associated with penetration and foot skating, while performing competitively with the state-of-the-art on motion accuracy and smoothness. Results and code can be found in our project page.

1. Introduction

In recent years, significant advancements have been made in recovering human motion from monocular RGB videos [4, 18, 19, 39, 42, 44]. While most of these works focus on videos of a single person, estimating motions for multiple people, especially those interacting, becomes significantly more challenging. This challenge primarily arises due to severe occlusion during close interactions, leading

*Work done during internship at Stanford.

to multiple body parts being invisible for extended periods. Fig. 1 illustrates an example of this, causing previous state-of-the-art [39] to produce motions with heavy inter-person penetrations. Additionally, previous methods also suffer from problems such as ground penetration and foot skating [4, 39, 42]. We argue that this is due to a lack of physics modeling.

In contrast, prior works in single-person motion estimation have explored incorporating physics into the process [18, 19, 44]. These methods typically employ a physics simulator and train a motion policy to generate motion that complies with physical constraints while imitating input observations. However, extending such methods to the scenario of multiple people presents significant challenges. Due to severe occlusion, detected 2D keypoints become less reliable and methods that rely on them are prone to fail [19, 44]. This raises the question of what input representation should be used.

To this end, we propose a framework, dubbed MultiPhys, that employs a physics simulation engine [34] to recover motion for multiple interacting people in a physics-plausible manner. Instead of relying on detected 2D keypoints in a feedforward model [19, 44], we initialize our approach with the output from SLAHMR [39], which performs global optimization on the entire sequence. However, since SLAHMR is physics-agnostic, its outputs may be noisy, particularly concerning the spatial placement of the bodies. This also results in inter-person penetrations. We devise a pipeline in which these preliminary body poses are fed into the physics simulator in an autoregressive fashion, aiming to obtain physically compliant motion estimates. We observe, however, that naively feeding these poses makes it difficult for the policy to generate the control signal to drive the simulation, which results in motion degradation. This happens especially when dealing with highly dynamic motions, as they present larger pose displacements between consecutive frames. As a result, the policy struggles to "catch up" with the reference signal. To counter this, we design an iterative refinement procedure and observe that the input poses are better matched while remaining physically compliant.

We evaluate MultiPhys on three challenging datasets. Our method performs competitively with previous state-of-the-arts on pose accuracy and smoothness, while significantly reducing errors on inter-body and ground penetration as well as foot skating. Specifically, compared to SLAHMR [39], we improve the (inter-body) penetration score by more than 7 times across all datasets, and the ground penetration score by 30, 5, and 1.35 times on the three datasets, respectively. We also reduce skating by 35% and 137% on the Hi4D and ExPI datasets, respectively.

In summary, our contributions are (1) A physics simulator-based framework for multi-person 3D motion estimation in a physically plausible manner. To the best of our knowledge, our method is the first that incorporates a physics simulator for multi-person 3D motion estimation; (2) Extensive evaluation showing that our method achieves significantly better results on physics-related metrics, while performing on par with prior works w.r.t. pose accuracy and smoothness.

2. Related work

Human mesh recovery from videos. Methods that reconstruct human mesh from videos [11, 13] build upon earlier works that focus on single images [10], incorporating temporal coherence to enhance motion reconstruction. Consequently, these methods are able to recover smooth and plausible human motion. While these and other regression methods that follow the same line of work [1, 17, 24] make valuable contributions, they often lack the ability to recover global trajectory – a crucial aspect for a comprehensive understanding of human motion. Simply extending these methods to videos with multiple people leads to spatially incoherent distribution of human meshes. To mitigate this issue, recent approaches focus on estimating global trajectories from per-frame local human poses [16, 41, 42]. Others take a step further by incorporating motion cues and additional constraints to more faithfully track the global trajectory [14, 32, 39]. TRACE [32] uses optical flow cues to track human motions. SLAHMR [39] and PACE [14] employ SLAM to compute dynamic camera parameters, refining them in an optimization stage. These methods also include constraints such as motion priors and contact with an estimated ground plane. Despite their impressive results, they often overlook fundamental physical constraints governing human motion in the real world, such as gravity and collisions with other individuals. To address these limitations, we propose to enforce these constraints explicitly by leveraging a full-featured physics engine.

Physics-based approaches. Trying to cope with the limitation that most human mesh recovery methods do not include physic constraints, some works explicitly incorporate physics notions. HuMoR [26], while not a physics simulator-based approach, models the human motion dynamics using a probabilistic generative prior that is later used in an optimization framework to recover motion. In the optimization stage, some losses for foot skating and velocity are applied to force physical compliance. However, once again, these are applied as soft constraints. Others [6, 30] take a step further and incorporate a physics simulator into their pipeline to reconstruct single-person motion. However, since most full-featured physics simulators are not differentiable, alternative optimization methods must be explored. For instance, [6] utilizes an evolutionary algorithm for refining poses, which can be challenging to optimize and result in extended inference times. To address this, [5, 29]

use differentiable physics. While this is a promising approach, differentiable simulators either require specific formulations for each problem or are often simplistic, lacking the features of non-differentiable simulators. Finally, human motion capture methods from sparse IMUs turn to utilizing physical simulators [15, 38] or apply soft physical constraints [8, 46]. All these approaches focus on a single person, whereas our method handles multiple people.

RL approaches. Works that use Reinforcement Learning (RL) to reconstruct human motion include [18, 19, 25, 43, 44]. Typically, these approaches involve learning a control policy in a simulation environment to govern a humanoid agent. Subsequently, another policy is often learned to generate reference motion, operating in conjunction with the first policy. These methods often leverage fully-featured simulators and utilize more realistic humanoid models compared to other works. Our work builds upon [19], originally designed for a single person.

Multi-person human mesh recovery. Several approaches address multi-person mesh recovery, each targeting specific challenges within this problem, such as occlusion [12, 31], depth ambiguity [35], or accurate spatial placement [9, 33]. A limited number of methods have concentrated on modeling contacts and addressing penetration issues [2, 3, 9, 22]. While these existing works focus on recovering human mesh solely from static images, our proposed method takes a step further by estimating multiple humans from videos while focusing on addressing the ground and inter-person penetration, as well as foot skating issues. There exist other methods that use multi-view cameras for capturing multi-person motion [45], which is out of the scope of this work.

3. Method

In the following, we describe the steps we take to correct an initial kinematic motion estimate to be physically plausible. We first introduce the necessary background on SLAHMR in Sec. 3.1. We then formalize our problem in Sec. 3.2. Next, we present an overview of our pipeline in Sec. 3.3. Finally in Sec. 3.4, we describe in detail the core part of our method, which is the physics-aware correction.

3.1. Background on SLAHMR [39]

SLAHMR [39] offers the state-of-the-art solution for multi-person motion estimation in a global coordinate frame. SLAHMR proposes a multi-stage optimization-based pipeline, where it optimizes a number of objectives as soft constraints. These include reprojection, motion smoothness, foot skating, ground contacts, pose and motion priors. This method is relatively robust to occlusion, thanks to the pose [23] and motion [26] priors it leverages. However, since it does not explicitly model physics, the estimated motion often manifests severe inter-person and ground penetration, as well as foot skating. In this work, we take the motion estimate from SLAHMR as the initialization and correct it to be physically plausible via our proposed method.

3.2. Problem Formulation

Given a monocular video $I_{1:T}$ consisting of T frames that have N interacting people, our goal is to recover the motion in world coordinates in a *physically plausible* manner, denoted as $\mathbf{Q}^i = \{\mathbf{q}_0^i, \mathbf{q}_1^i, \dots, \mathbf{q}_T^i\}$, for all people $i = 1 \dots N$. Each pose is parameterized following the SMPL-H [27] body model, which consists of the global orientation $\Phi_t^i \in \mathbb{R}^3$, body pose $\Theta_t^i \in \mathbb{R}^{22\times 3}$, body shape $\beta^i \in \mathbb{R}^{16}$, and root translation $\Gamma_t^i \in \mathbb{R}^3$. That is:

$$\mathbf{q}_t^i = \{\Phi_t^i, \Theta_t^i, \beta^i, \Gamma_t^i\} \tag{1}$$

Throughout our experiments, $N = 2$, but note that our method does not have a fundamental limit on N and can in theory work with an arbitrary number of interacting people.

3.3. Method Overview

An overview of our method is presented in Fig. 2. We use a set of poses estimated by SLAHMR [39], denoted by $\widetilde{\mathbf{q}}_t^i$, as the initial estimates, which are later corrected to be physically plausible using a physics simulator. We denote these corrected poses as $\mathbf{q}_t^i$. Here each pose is for the i-th person at timestep t.

The motion estimated from SLAHMR $\widetilde{\mathbf{q}}_t^i$ is also represented with the SMPL-H model:

$$\widetilde{\mathbf{q}}_t^i = \{\widetilde{\Phi}_t^i, \widetilde{\Theta}_t^i, \beta^i, \widetilde{\Gamma}_t^i\} \tag{2}$$

Note that the body shape β is the same for all timesteps and we directly keep it from the SLAHMR estimates.

To enforce the physical constraints missing from SLAHMR, we leverage a full-featured physics simulator (Mujoco [34]). Inside the simulation, we represent each person as a humanoid agent consisting of different body parts and joints with actuators over those joints. We create one humanoid for each person. The creation process follows SimPoE [44].

To prevent the simulated character from losing track of the input motion, we re-purpose the Universal Humanoid Controller (UHC) [19], whose goal was to imitate a set of target poses while producing signals to drive the simulated character. In the original paper, the target poses input to the UHC are parameterized by the proposed Multi-step Projection Gradients (MPG), which link the 2D observations to 3D simulated body poses via gradients of the 2D reprojection loss. In our case, we re-purpose the UHC to take the SLAHMR-estimated poses $\widetilde{\mathbf{q}}_t^i$ as input. Note that the UHC also receives as input the body shape, thus different bodies can be controlled. This is especially important when working with multiple interacting people.

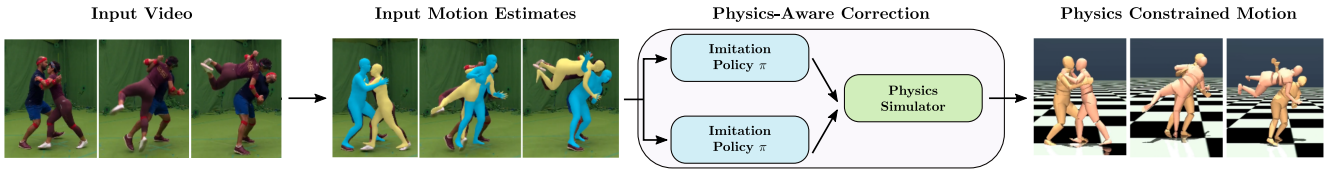


Figure 2. **MultiPhys Pipeline.** Given an input video with multiple people (left), we first obtain initial kinematic estimates of the camera poses and 3D human motion using SLAHMR [39]. Using these initial motion estimates, our proposed framework corrects them and makes them physically plausible (right).

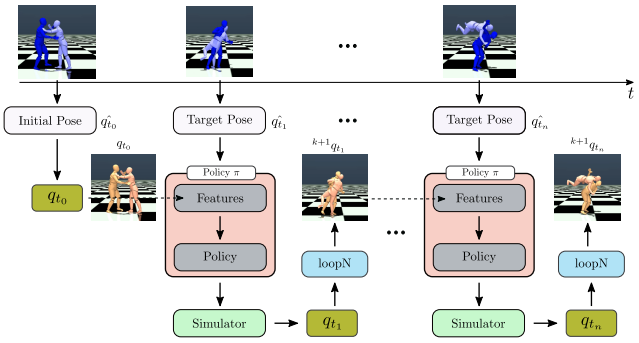


Figure 3. **Physics-aware Correction Module**. We use the policy π to control the humanoid agents with the initial kinematic poses. We simulate all agents simultaneously in order to apply physics-based constraints to the reconstructed motion. The policy computes features from both the current state of the simulation and the target pose to later generate the action signal a that controls the agents. We place our *loop-N* component between target poses $\tilde{\mathbf{q}}^i_{t+1}$ that correspond to each video frame.

3.4. Physics-Aware Correction

A detailed diagram of our physics-aware correction is presented in Fig. 3. For this stage, our formulation is the following. Inside the simulation, we define the 3D human pose as $\mathbf{q}^i_t$. At each timestep, we have access to each agent's state $\mathbf{s}^i_t \triangleq (\mathbf{q}^i_t, \dot{\mathbf{q}}^i_t)$ which is the combination of the 3D pose and velocity.

In order to drive the humanoid agents inside the simulation to mimic the kinematic poses and thus correct them to be physically plausible, we use an imitation policy similar to [19]. This policy is modeled as a Markov Decision Process (MDP) following the standard formulation in physics-based character control. This process is defined as a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma \rangle$ of states, actions, transition dynamics, reward function, and discount factor. The state $s \in \mathcal{S}$, reward $r \in \mathcal{R}$, and the transition dynamics $\mathcal{T}$ are determined by the physics simulator, while the action $a \in \mathcal{A}$ is given by the control policy $\pi(\mathbf{a}^i_t | \mathbf{s}^i_t, \tilde{\mathbf{q}}^i_{t+1}, \beta^i)$. We employ Proximal Policy Optimization [28] (PPO) to find the the optimal control policy π that maximizes the expected discounted reward $\mathbb{E}[\sum_{t=1}^{T} \gamma^{t-1} r_t]$.

The connection between the kinematic estimates and the simulation is through the policy π, as shown in Fig. 3. We drive the simulation with the kinematic poses $\tilde{\mathbf{q}}^i_t$ by inputting them into the policy. In this way, the agent in the simulation will follow these poses as the reference. We use the same policy for each agent in the scene. Each agent is controlled independently, mimicking what happens in reality (*i.e*. each person can move independently in the world), but they all reside in the same simulation. By having different agents sharing the simulation simultaneously, we can directly impose physical restrictions between them, *e.g*. they cannot penetrate each other. This results in physically compliant estimates. This formulation also has practical implications that we do not need to train the policy on multi-person datasets. Instead, it is trained only on the large-scale single-person MoCap dataset, AMASS [20].

For motion sequences where the persons are moving a lot and especially when they undergo extreme poses (*e.g*. in the ExPI [37] dataset), we observe that it is more difficult for the policy to imitate the reference poses. This happens because the policy was trained with a specific set of actions and poses. For any action that has different dynamics and overall different distribution from the training data, the policy is not able to completely match the target pose. Given that the policy's formulation follows the form of an auto-regressive system, this error tends to accumulate quickly, leading to noticeable final errors. To remediate this, we devise an iterative strategy for the agent to slowly get closer to the reference pose. Throughout the paper, we dub this strategy *loop-N* . As stated before, the policy samples an action $\mathbf{a}^i_t$ for each agent i at timestep t given the current simulation state of each agent ($\mathbf{s}^i_t$) and reference pose ($\tilde{\mathbf{q}}^i_{t+1}$). In normal operation, the updated 3D pose taken from the simulation output once the action is applied is defined by:

$$\mathbf{q}^i_{t+1} = \mathcal{T}(\mathbf{q}^i_t, \mathbf{a}^i_t). \tag{3}$$

To help $\mathbf{q}^i_{t+1}$ match the reference pose, we iteratively update the simulation state for N_l iterations while keeping the reference pose fixed until it gets close enough to the it. Let k be the current iteration while keeping the reference pose fixed, where $k = \{1, 2, ..., N_l\}$. For every iteration, we sample a new action $\pi({}^k\mathbf{a}^i_t | {}^k\mathbf{s}^i_t, \tilde{\mathbf{q}}^i_{t+1}, \beta^i)$ that will drive the current pose closer to the reference pose. For the updates inside "*loop-N* ", we redefine Eq. 3 in terms of k:

$${}^{k+1}\mathbf{q}^i_t = \mathcal{T}({}^k\mathbf{q}^i_t, {}^k\mathbf{a}^i_t). \tag{4}$$

After N_l iterations, we keep only the last update, which should be closer to the reference pose. Specifically, we make $\mathbf{q}_{t+1}^i = {}^{N_l}\mathbf{q}_t^i$. We repeat this process for every timestep in the sequence. Once we project all the sequences into a physically plausible space, it is trivial to convert these poses back to the SMPL representation.

4. Experiments

We start by describing the datasets and the metrics we evaluate our method on. We then compare to baseline methods in Sec. 4.1. Next, we ablate the *loop-N* component in Sec. 4.2. For additional implementation details, we refer the reader to our supplementary material.

Datasets. We carefully choose the evaluation datasets to be those that have significant inter-person interaction, where purely kinematics-based approaches tend to fail. To this end, we evaluate our method on three datasets with increasing levels of interaction. **CHI3D** [2] contains mild interactions, while more intense interactions in **Hi4D** [40], and significant interaction and occlusion in **ExPI** [37]. ExPI also features extreme poses and highly dynamic motion. There are other datasets containing multiple interacting people, such as MuPoTS-3D [21], ShakeFive2 [36], and MultiHuman [45]. However, they do not contain close interactions in the same amount as the three above.

CHI3D contains 127 motion sequences for each of the 5 pairs of subjects (3 train, 2 test) interacting in everyday actions such as posing (for a photograph), pushing, hugging, etc. CHI3D is captured with cameras from four different views. Each motion sequence is annotated with the action label together with ground truth 3D poses in a world coordinate system in the SMPL format.

The Hi4D dataset is captured with up to 8 cameras at different locations. It contains 100 short motion sequences with close interaction and high contact ratio between the subjects performing diverse actions, such as hugging, posing, dancing, and playing sports. It includes 20 unique pairs of participants.

ExPI is the most challenging dataset that contains subjects performing 16 complicated two-person dance routines and, thus, presents highly dynamic sequences with a high amount of contact. Each of the aerials is repeated five times and in total, it contains 115 motion sequences. The data is collected with 60+ synchronized cameras and a motion capture system.

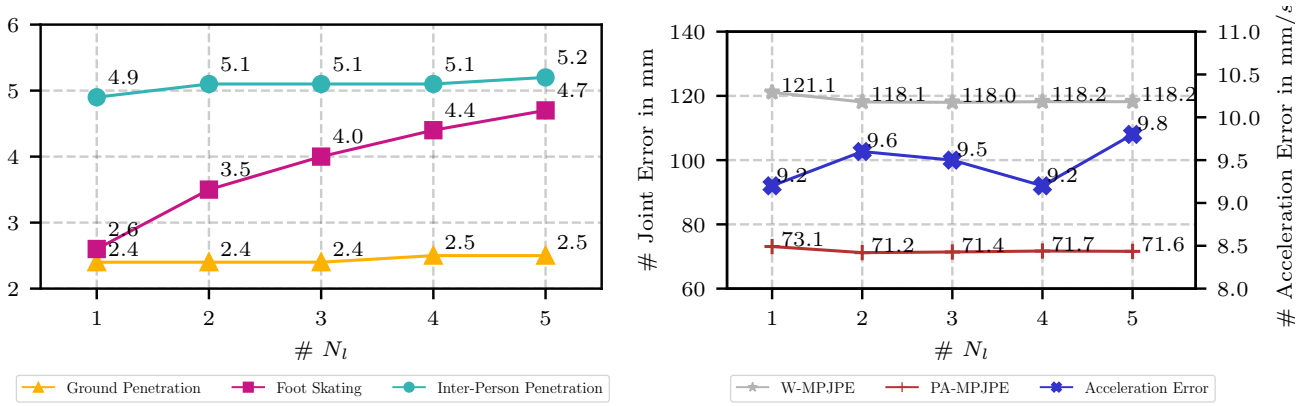
Evaluation metrics. We report a variety of metrics in two categories: (1) pose metrics, which measure both the pose accuracy and smoothness, and (2) physics metrics, which measure the physics plausibility of the motion sequence. For the first category, we follow SLAHMR [39] and report the World PA First - MPJPE (*W-MPJPE*), which reports the MPJPE [7] after aligning the *first* coordinate frames of the prediction and the ground truth, as well as *PA-MPJPE (joint)*, which reports the MPJPE error after *jointly* aligning the predictions of all the people in *all* frames with the ground truth poses. We also report the acceleration error (*Acc. Error*), which measures the acceleration difference between the ground truth and the estimated motion. In practice, it serves as a measure for motion jittering. For the second category, we report the foot skating (*Skating*) and ground penetration (*Gnd Pen.*) metrics following [19, 26]. We also report the amount of inter-person penetration (*Pen.*). Specifically, we compute the signed distance function (SDF) values for each person and evaluate each penetrating vertex of one person into the other. We then report the cumulative values of this, *i.e.* we report the sum of SDF values for all penetrating vertices averaged *per person* for the entire sequence. We report the results in meters, except the *Pen.* metric which is in mm. Please see the Supp. Mat. for more details on the evaluation protocol.

4.1. Comparison with Baselines

To the best of our knowledge, there is no previous approach that uses a physics simulator to estimate 3D motion for multiple people. We hence extend EmbPose [19] to take 2D keypoints as input and operate in the simulation environment for all people at the same time. Specifically, we extend their model to have N identical but separate branches (one per person). Each branch produces an initial estimate using the kinematics-based policy π_{KIN}, and an action signal to drive the simulation with the policy π. We also adapt the simulation to include N humanoid agents with different poses and body shapes that move independently but simultaneously. We refer to the adapted method as EmbPose-MP. Additionally, we compare to SLAHMR [39] as the state-of-the-art multi-person motion estimation method, which is purely kinematics-based.

We report the results in Tab. 1. On Hi4D and ExPI, our method outperforms SLAHMR on all physics-based metrics, while inferior in terms of acceleration. This is because SLAHMR directly optimizes for smooth motions, which results in lower acceleration errors.

On CHI3D, our method outperforms both SLAHMR and EmbPose-MP in terms of pose metrics. These results are encouraging as they show that enforcing physics compliance can not only enhance the physical plausibility of the estimated motion, but can also lead to more accurate poses. Pose improvement happens especially in cases where people have both feet in contact with the ground. In the simulation, the ground is taken as a hard constraint and thus when the agent moves, it cannot penetrate it, resulting in more realistic poses. This effect has also been reported before [30, 44]. The physics simulation also helps improve the poses in cases where the two people are in close proximity, allowing for better spatial placement as body meshes cannot penetrate. This is reflected mostly in the W-MPJPE metric,



(a) Effect on Physics-based Metrics

(b) Effect on Pose Metrics

Figure 4. **Effect of *loop-N* component for different values of N_l.** We study the effect of different values of $N_l = \{1, ..., 5\}$ on both (a) physics and (b) pose metrics. We report Inter-Person Penetration (measured in m.), the Ground Penetration (measured in mm), the Floor Skating (measured in mm), the W-MPJPE and PA-MPJPE (measured in mm) and the Acceleration Error (measured in mm/s^2). We choose $N_l = 2$ for the rest of the experiments as it provides a good balance between physics and pose metrics, see Sec. 4.2. Note that we scale Pen. metric by a factor of 1/10 to fit the graph. To see the table for these numbers refer to the Supp. Mat.

and happens when people are, e.g., touching or hugging, as shown in Fig. 5 and Fig. 6. Thus, our method improves the motion estimation in these cases by correcting the inter-person penetrations, which often exist in SLAHMR poses.

Results on CHI3D. For CHI3D, we see that our system outperforms SLAHMR both in penetration (Pen. and Gnd Pen.) and pose metrics. However, for this dataset, SLAHMR has a better skating score, which we hypothesize is due to the fact that CHI3D contains less dynamic motions than the others, and most of the time both people keep their feet on the ground (as opposed to, *e.g.* in ExPI). This results in better ground plane estimation making it easier for SLAHMR to deal with skating.

Results on Hi4D and ExPI. For Hi4D, we see that EmbPose-MP does better on penetration metrics than our model, while the pose metrics are worse. This is due to poor estimation of both global spatial placement and individual poses. Because EmbPose-MP cannot handle inter-person occlusion, poses where the people are close together are not well captured by it and tend to be estimated farther away from each other, when they should in fact be closer together or in contact. As a consequence, penetrations are naturally less likely to occur.

In contrast, our model is able to capture people in close proximity while not breaking the laws of physics (see Fig. 5 and Fig. 6). This is reflected in better pose while still achieving good penetration metrics. Note that penetration reduces drastically in comparison to the purely kinematic baseline which recovers poses accurately but presents high penetration. Moreover, skating and ground penetration are corrected (see Gnd. Pen. and Skating metrics) *w.r.t.* the kinematic method. The baseline has better skating scores than our model due to our *loop-N* which slightly introduces skating as explained in Sec. 4.2. For both datasets, we observe that ground penetration is worse for the baseline as it presents erratic estimated poses and also jittery motion. For these datasets, we observe pose metric values close to the ones obtained with SLAHMR but with slight differences. This is caused mainly by the type of motion present in each dataset. Hi4D and ExPI, in contrast to CHI3D, contain more dynamic motion, which in some cases can be harder to match for the simulated agent.

4.2. Ablation Study

We perform an ablation to study the effect of the *loop-N* component in our system, whose results are reported in Fig. 4. We ablate on different values of N_l and report the performance of: (i) the basic version of our approach (Loop1) where we use kinematic estimates plus the physics simulator and (ii) our method with *loop-N* variant for different values of N_l, where $N_l > 1$. The measured metrics are: Gnd Pen., Skating, Pen., W-MPJPE, PA-MPJPE, and Acc. Error. In Fig. 4, we show plots of the metrics in two groups: physics in Fig. 4a and pose in Fig. 4b to better analyze the effects and trends on these when N_l is changed. We choose to perform our ablation study on Hi4D as it is the most representative among the datasets. It includes both mild and dynamic motion and at the same time poses where people are very close spatially.

Our *loop-N* component, which composes the full system, helps the simulated poses to better match the kinematic reference poses especially for highly dynamic motions such as the ones present in ExPI, see Sec. 3.4. We see that with a correctly chosen value of N_l, the policy is able to better match the reference poses.

The *loop-N* component gets the simulated poses closer

	Method	Pen.↓	Gnd Pen.↓	Skating↓	Acc. Error↓	W-MPJPE↓	PA-MPJPE (joint)↓
CHI3D	SLAHMR [39]	139.3	4.4	**1.0**	**6.5**	177.1	83.5
	EmbPose-MP [19]	40.2	**2.6**	2.8	7.7	214.7	96.5
	Ours	**18.7**	3.2	2.7	7.4	**174.7**	**80.4**
Hi4D	SLAHMR [39]	367.3	12.2	4.9	**6.9**	121.6	**69.1**
	EmbPose-MP [19]	**39.8**	3.8	**1.3**	12.7	148.8	92.9
	Ours	51.1	**2.4**	3.5	9.6	**118.1**	71.2
ExPI	SLAHMR [39]	567.3	18.6	5.4	**8.2**	263.3	**159.1**
	EmbPose-MP [19]	92.1	0.9	**1.9**	27.7	386.4	207.6
	Ours	**73.0**	**0.6**	2.3	17.1	**250.9**	164.3

Table 1. **Comparison with the state of the art.** We report various metrics on CHI3D [2], Hi4D [40], and ExPI [37] datasets. Pose metrics are W-MPJPE and PA-MPJPE (joint) in mm. See Sec. 4.1.

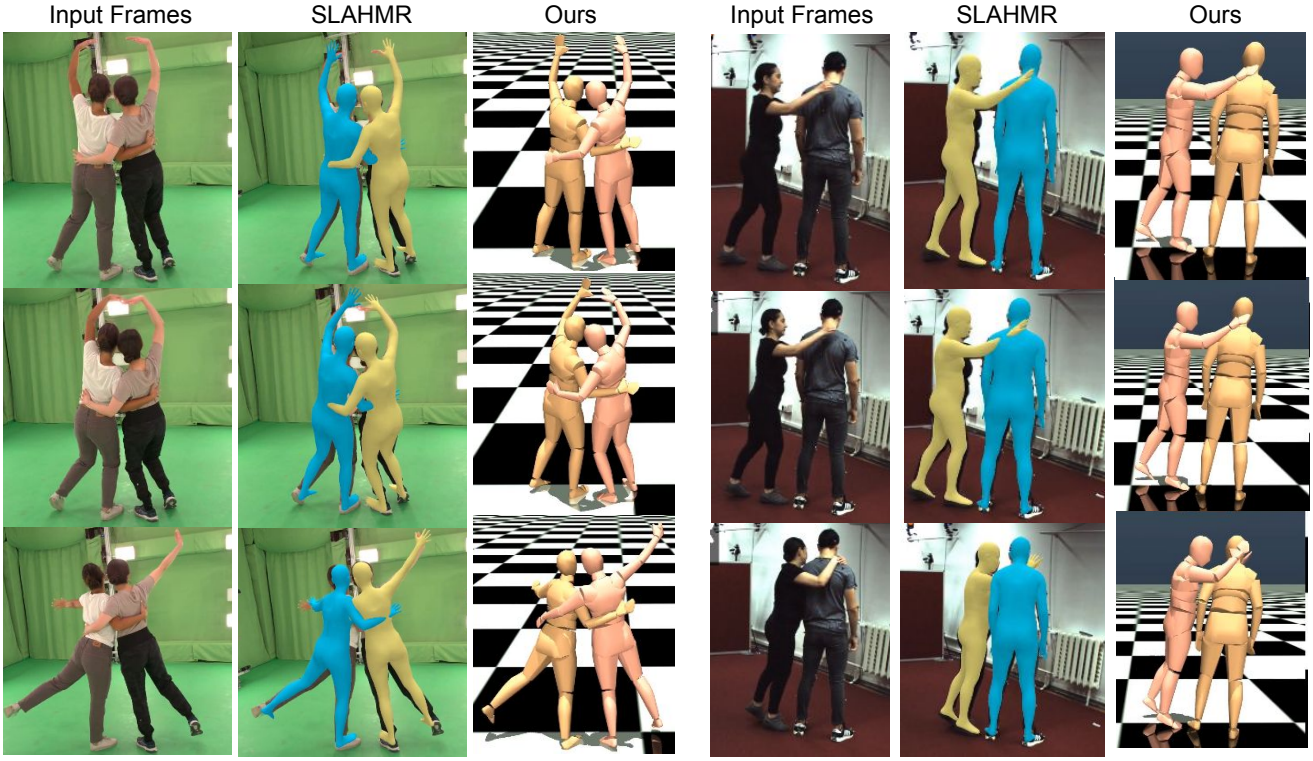


Figure 5. **Qualitative results of the proposed approach.** The first three columns (from left to right) are from Hi4D [40] and the other three are from CHI3D [2]. Each row corresponds to one frame of the same sequence. The columns compare the resulting poses at each frame using SLAHMR [39] and our method. In these cases of close inter-person interaction, the estimated motion from SLAHMR often has severe inter-person penetrations, while our method is able to eliminate these penetrations through physics-aware correction.

to the kinematic reference pose, however, the best value for N_l to ensure a better match, depends on the particular motion. Nevertheless, we observe that $N_l = 2$ works well for all the datasets, striking a good balance between pose and physics-based metrics. In the curves in Fig. 4a, we see that Acc Error., Gnd Pen. and Pen. almost stay constant with small variations when N_l changes. This is different for Skating as it increases with N_l. This is due to the fact that as N_l increases, the pose change between frames is potentially larger, thus for a given value of Gnd Pen., increases Skating. In Fig. 4b, we see that for $N_l = 2$ the pose errors decrease, however, for values where $N_l > 2$ the error starts to increase. The results shown in Tab. 1 are calculated using $N_l = 2$.

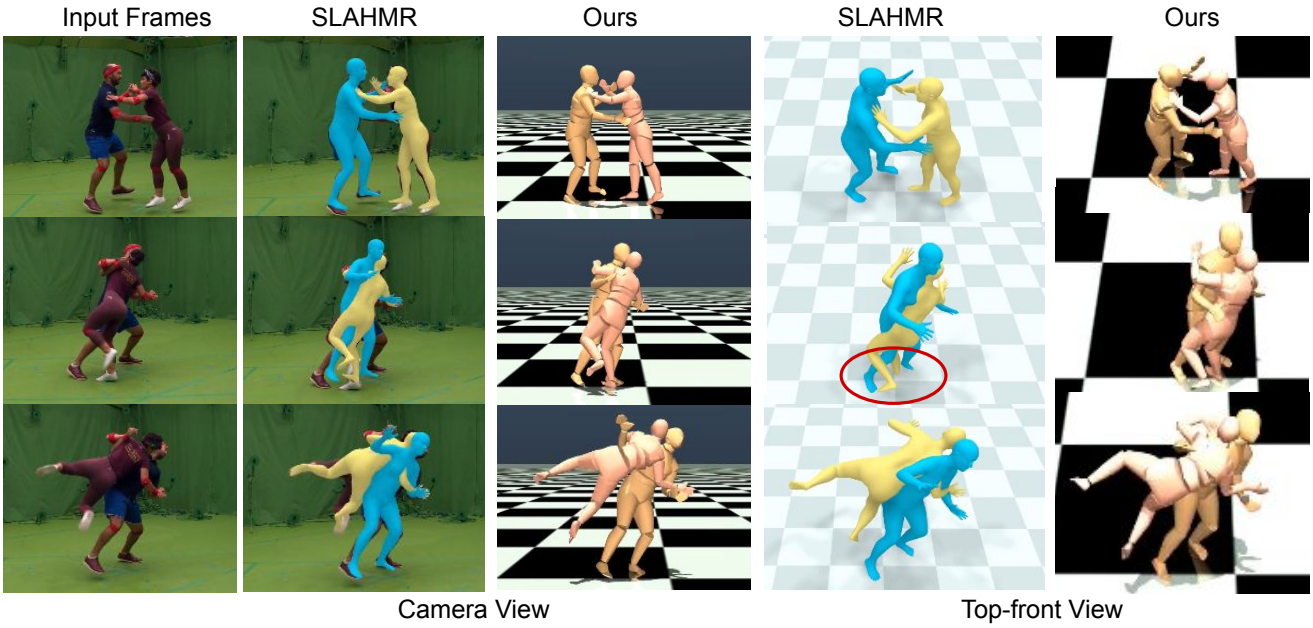


Figure 6. **Effect of the physical constraints on spatial placement.** One key advantage of our method is that complying with physical constraints (e.g., penetration between bodies) helps to improve the spatial placement of the bodies. Here we show results for motion estimated with both the kinematic approach SLAHMR [39] and our system. See how the bodies from the kinematic poses overlap and penetrate the ground (red circle in the figure) leading to unrealistic spatial placement. Our method eliminates these penetrations both *w.r.t.* the body and the ground.

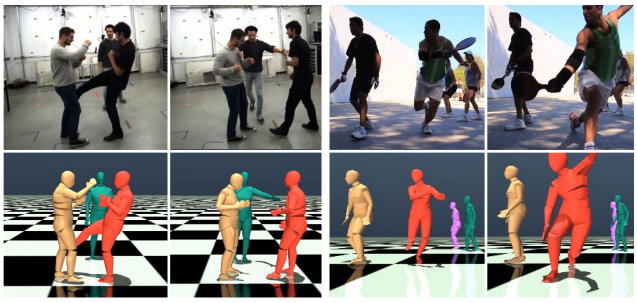


Figure 7. **Additional results** on videos with three (left) and four (right) people.

4.3. Scaling to More People

There is no inherent limitation preventing our model from scaling to scenes with more people. However, most datasets that contain close interactions only consider two people since that is the unit of such interactions. Current datasets with more people do not capture close interactions. To showcase our method's capability, we apply it to videos with more than two people (see Fig. 7). Note that our model reliably captures the human's pose and spatial placement also in these scenarios.

5. Discussion

We propose a method for recovering physically plausible 3D human motion from a monocular RGB video, and in particular for two interacting people. Our approach leverages a fully-featured physics simulator to add constraints to the motion estimation process and to force it to follow the laws of physics. This allows us to improve the realism of the estimated motion by both avoiding penetration between human bodies and with the ground plane, while also improving the pose in terms of spatial placement. This is corroborated by our experiments on three challenging datasets.

While our method unlocks many new possibilities for generating more realistic and physically plausible motions, some problems remain to be addressed. When 2D keypoint detectors fail, the kinematic-based motion estimation on which we rely as initialization also fails, leading to a decline of pose accuracy. Developing more powerful and robust 2D keypoint detector is hence important in further improving the performance of our method. Please see supplement for failure cases. Another interesting direction is to develop a two-person interaction prior to guide the physics-aware correction process.

Acknowledgments

Despoina Paschalidou is supported by the Swiss National Science Foundation under grant number P500PT_206946. Leonidas Guibas is supported by a Vannevar Bush Faculty Fellowship. This work is partially supported by projects SMARTGAZEII CPP2021-008760 and MoHuCo PID2020-120049RB-I00.

References

[1] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1964–1973, 2021. 2

[2] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3, 5, 7

[3] Mihai Fieraru, Mihai Zanfir, Teodor Szente, Eduard Bazavan, Vlad Olaru, and Cristian Sminchisescu. Remips: Physically consistent 3d reconstruction of multiple interacting people under weak supervision. In *Advances in Neural Information Processing Systems*, 2021. 3

[4] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1, 2

[5] Erik Gärtner, Mykhaylo Andriluka, Erwin Coumans, and Cristian Sminchisescu. Differentiable dynamics for articulated 3d human motion reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[6] Erik Gärtner, Mykhaylo Andriluka, Hongyi Xu, and Cristian Sminchisescu. Trajectory optimization for physics-based reconstruction of 3d human pose from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[7] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristia Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. In *IEEE. trans. PAMI*, 2014. 5

[8] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *European conference on computer vision*, pages 443–460. Springer, 2022. 3

[9] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[10] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[11] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5614–5623, 2019. 2

[12] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1715–1725, June 2022. 3

[13] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

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

[15] Sunmin Lee, Sebastian Starke, Yuting Ye, Jungdam Won, and Alexander Winkler. Questenvsim: Environment-aware simulated motion tracking from sparse sensors. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–9, 2023. 3

[16] Jiefeng Li, Siyuan Bian, Chao Xu, Gang Liu, Gang Yu, and Cewu Lu. D&d: Learning human dynamics from dynamic camera. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[17] Zhengyi Luo, S Alireza Golestaneh, and Kris M Kitani. 3d human motion estimation via motion compression and refinement. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2

[18] Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. In *Advances in Neural Information Processing Systems*, 2021. 1, 2, 3

[19] Zhengyi Luo, Shun Iwase, Ye Yuan, and Kris Kitani. Embodied scene-aware human pose estimation. In *Advances in Neural Information Processing Systems*, 2022. 1, 2, 3, 4, 5, 7

[20] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 4

[21] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018. 5

[22] Lea Müller, Vickie Ye, Georgios Pavlakos, Michael Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3d social interaction from images. *arXiv preprint arXiv:2306.09337*, 2023. 3

[23] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. Osman, Dimitrios Tzionas, and Michael J. Black. 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 (CVPR)*, 2019. 3

[24] Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Human mesh recovery from multiple shots. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[25] Xue Bin Peng, Angjoo Kanazawa, Jitendra Malik, Pieter Abbeel, and Sergey Levine. Sfv: Reinforcement learning of physical skills from videos. *ACM Trans. Graph.*, 37(6), Nov. 2018. 3

[26] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. Proceedings of

the IEEE/CVF International Conference on Computer Vision (ICCV), 2021. 2, 3, 5

[27] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 2022. 3

[28] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 4

[29] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, Patrick Pérez, and Christian Theobalt. Neural monocular 3d human motion capture with physical awareness. *ACM Transactions on Graphics*, 40(4), aug 2021. 2

[30] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics*, 39(6), dec 2020. 2, 5

[31] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021. 3

[32] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J. Black. 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 (CVPR)*, 2023. 2

[33] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J. Black. Putting people in their place: Monocular regression of 3D people in depth. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 3

[34] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012. 2, 3

[35] Nicolas Ugrinovic, Adria Ruiz, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Body size and depth disambiguation in multi-person reconstruction from single images. 2021. 3

[36] Coert van Gemeren, Ronald Poppe, and Remco C. Veltkamp. Spatio-temporal detection of fine-grained dyadic human interactions. In Mohamed Chetouani, Jeffrey Cohn, and Albert Ali Salah, editors, *Human Behavior Understanding*, 2016. 5

[37] Guo Wen, Bie Xiaoyu, Alameda-Pineda Xavier, and Moreno-Noguer Francesc. Multi-person extreme motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4, 5, 7

[38] Alexander Winkler, Jungdam Won, and Yuting Ye. Questsim: Human motion tracking from sparse sensors with simulated avatars. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–8, 2022. 3

[39] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023. 1, 2, 3, 4, 5, 7, 8

[40] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 5, 7

[41] Ri Yu, Hwangpil Park, and Jehee Lee. Human dynamics from monocular video with dynamic camera movements. *ACM Trans. Graph.*, 40(6), 2021. 2

[42] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[43] Ye Yuan and Kris Kitani. Residual force control for agile human behavior imitation and extended motion synthesis. In *Advances in Neural Information Processing Systems*, 2020. 3

[44] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. Simpoe: Simulated character control for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 3, 5

[45] Yuxiang Zhang, Zhe Li, Liang An, Mengcheng Li, Tao Yu, and Yebin Liu. Light-weight multi-person total capture using sparse multi-view cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3, 5

[46] Xiaozheng Zheng, Zhuo Su, Chao Wen, Zhou Xue, and Xiaojie Jin. Realistic full-body tracking from sparse observations via joint-level modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14678–14688, 2023. 3