# Supplementary Material for: "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

In this supplementary material, we first present implementation details in Sec. 1. Later, we elaborate on the details of each dataset and present statistics about their content (Sec. 2). In Sec. 3, we complement the main paper by presenting an extended ablation study for CHI3D [1] and ExPI [6] datasets where we also discuss the nuances for each. Finally, in Sec. 4, also complementing the main text, we thoroughly explain the results we obtained for each dataset and shown in Tab. 1 (which is the same table found in the main text).

We strongly encourage the reviewer to look at our **supplementary video**. There, we provide a supplementary video showcasing various results for the datasets utilized in this work. Alongside the results for our method, we also include results from SLAHMR [7] and the baseline referred to as EmbPose-MP, an extension of [3] tailored for multiple people. Additionally, we include some illustrative instances of failure cases.

## 1. Implementation Details

In our implementation, we use the MuJoCo [5] physics simulator To initialize the kinematic motion, we use SLAHMR [7] with its default values. For the first optimization stage, 30 iterations are ran with $\lambda_{data} = 0.001$. For the second stage, we optimize for for 60 iterations and use $\lambda_{smooth} = 5$, $\lambda_{\beta} = 0.05$, $\lambda_{pose} = 0.04$. Finally for the last stage, we use $\lambda_{CVAE} = 0.075$, $\lambda_{skate} = 100$, and $\lambda_{con} = 10$. For the imitation policy $\pi$ we use the universal humanoid controller (UHC) [3]. The UHC from [3] extends the policy proposed in [2] to humanoids with different body shapes. It is important for us to capture different body shapes. The imitation policy is trained on the training split of the AMASS dataset [4] where motion sequences that involve human-object interactions are removed.

*Work done during internship at Stanford.

## 2. Datasets Details

In this section, we elaborate on the specifics of each dataset employed in this work. In particular, we argue that given the nature of each, they present different types of motions. CHI3D contains mild motions of interacting people. Hi4D contains more dynamic motions with more action variety than CHI3D. Finally, ExPI contains highly dynamic and fast motions.

**CHI3D.** This dataset contains 127 motion sequences for each of the 5 pairs of subjects (3 train, 2 test) interacting in everyday activities where people are in contact with each other. Only the training motion sequences have publicly accessible ground truth and the test annotations are hidden from the public. For these reasons, we use and analyze the train data split. Note that we do not train or fine-tune any components to build our framework or the baselines used, as such, using the train split is acceptable. The dataset contains a total of 381 sequences (*i.e.* from 3 different subject pairs), equivalent to 21 minutes of motion sequences. The data is captured with 4 cameras that provide different views. In this work, we use the videos from one randomly sampled camera view to test our system. The activities included are: kick, push, grab, posing, holding hands, handshake, hit, and hug. In Fig. 1, we present the statistics statistics related to the amount of data for each action present in the dataset. The three activities with most data are: pushing, hitting and grabbing. The three activities with less data are: posing, holding hands, and handshake. Based on the nature of these actions, they are mostly short sequences of around 6 seconds. The actions are performed by actors in a motion capture studio. As such and confirmed by visually inspecting the dataset, the motions are relatively slow and do not include extreme or uncommon poses.

**Hi4D.** The dataset contains 100 short motion sequences with close interactions and high contact between the subjects performing diverse actions with 11K frames in total and more than 6K frames of them with physical contact. Contact annotations across the whole dataset cover over

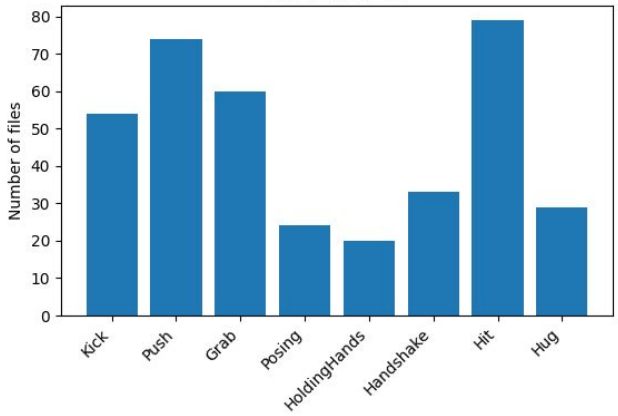



Figure 1. **Statistics for CHI3D[1].** We present statistics on the frequency of each action that is contained in the motion sequences. CHI3D, in general, presents mild dynamic motions, this is reflected on the nature of the actions.

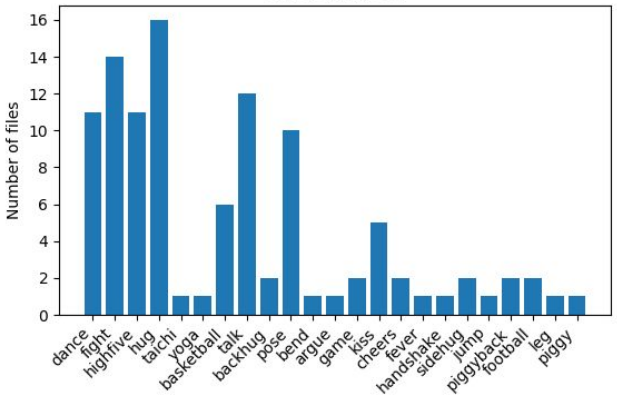



Figure 2. **Statistics for Hi4D[8].** We present statistics on the frequency of each action that is contained in the motion sequences. Hi4D, in general, presents more dynamic motions than CHI3D, this is reflected on the nature of the actions.

95% of the parametric human body. It includes 20 unique pairs of participants with varying body shapes. The activities included are: dance, fight, highfive, hug, taichi, yoga, basketball, talk, backhug, pose, bend, argue, game, kiss, cheers, fever, handshake, sidehug, jump, piggyback, football, leg, piggy. In Fig. 2 we see statistics of the amount of data for each action present in the dataset. The three activities with the most sequences are: hug, fight, and dance. The activities with fewer sequences are: taichi, yoga, bend, argue, (checking for) fever, handshake, jump, leg, and piggy. Similar to CHI3D, all these motions are mostly short sequences of around 6 seconds. Given the nature of these actions and confirmed by visually inspecting the dataset, the motions are more dynamic than CHI3D but less than ExPI. This means that there are more sequences with fast motion. In general, the motion include mostly common poses with some exceptions such as taichi, piggyback, and yoga. In summary, the Hi4D dataset presents dynamic motion and common everyday poses.

**ExPI.** This dataset contains subjects performing 16 different complicated two-person dance routines, specifically, Lindy Hop aerial steps. An aerial (or air step) is a dance move where one's feet leave the floor. The term is used to denote a wide range of special and unusual dance moves, including dips, slides, and tricks. Each of the aerials is repeated five times and in total, the dataset contains 115 short sequences of around 6 seconds. Tab. 3 shows the different dance routines performed by each pair of dancers, here we can have a sense of the statistics of the datasets. The motions that appear the most are at the top segment of the table, and the rest are only performed by one of the couples. Lindy Hop aerial steps are by definition a set of unusual dance moves where one of the subjects' feet does not hold contact with the floor. Hence, these motions are highly dynamic and often involve heavy contact between the dancers.

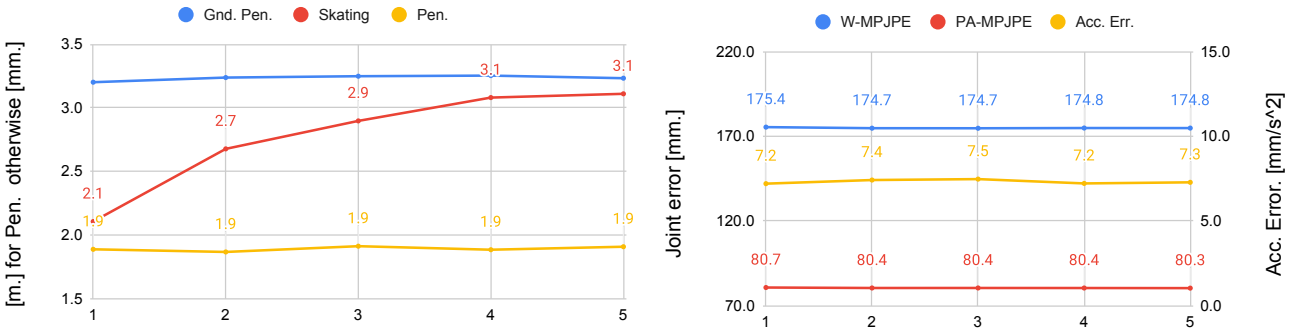
| Aerial | Name | Couple 1 | Couple 2 |
|---|---|---|---|
| $A_1$ | A-frame | ✓ | ✓ |
| $A_2$ | Around the back | ✓ | ✓ |
| $A_3$ | Coochie | ✓ | ✓ |
| $A_4$ | Frog classic | ✓ | ✓ |
| $A_5$ | Noser | ✓ | ✓ |
| $A_6$ | Toss out | ✓ | ✓ |
| $A_7$ | Cartwheel | ✓ | ✓ |
| $A_8$ | Back flip | ✓ | |
| $A_9$ | Big ben | ✓ | |
| $A_{10}$ | Chandelle | ✓ | |
| $A_{11}$ | Check the change | ✓ | |
| $A_{12}$ | Frog-turn | ✓ | |
| $A_{13}$ | Twisted toss | ✓ | |
| $A_{14}$ | Crunch-toast | | ✓ |
| $A_{15}$ | Frog-kick | | ✓ |
| $A_{16}$ | Ninja-kick | | ✓ |

Figure 3. **Dance sequences included in ExPI [6].** This shows the overall content of the dataset. In general, it contains highly dynamic motions where the dancers move fast and have their feet off the ground for a considerable amount of time.

## 3. Ablation Studies

In Fig. 4 and Fig. 5, we include detailed information where we show the effect of our *loop-N* component for the CHI3D and ExPI datasets, respectively. For similar graphs of Hi4D, please refer to the main text. By looking at the physics-based metrics, in Fig. 4a and Fig. 5a, we see that skating increases with $N_l$ similar to Hi4D, as discussed in Sec. 4.2 of main text. We also see that ground penetration is much less severe for ExPI than CHI3D. This is explained by the fact that in the ExPI, most of the frames contain at least one person with his/her feet off the ground.

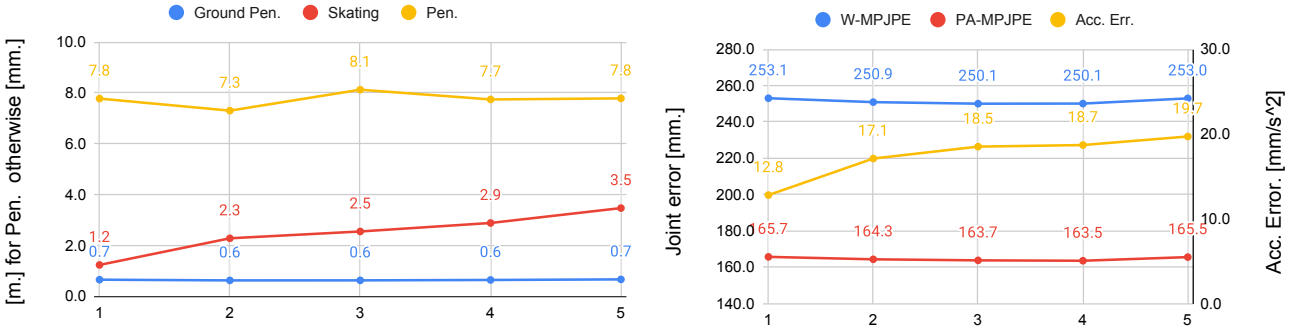
Looking at the pose metrics in Fig. 4b and Fig. 5b, we see




(a) Effect on physics-based metrics

(b) Effect on pose metrics

Figure 4. **Effect of $N_l$ for CHI3D [1] dataset.** We study the effect of different values of $N_l = \{1, ..., 5\}$ on both (a) physics and (b) pose metrics. We report Pen. in m, Gnd Pen. in mm, Skating in mm, W-MPJPE and PA-MPJPE in mm, Acc. Error in mm/s$^2$. We choose $N_l = 2$ for our experiments as it provides a good balance between physics and pose metrics.. Note that we scale Pen. metric by a factor of 1/10 to fit the graph.




(a) Effect on physics-based metrics

(b) Effect on pose metrics

Figure 5. **Effect of $N_l$. for ExPI [6] dataset.** The units in this figure are the same as for Fig. 4. Note that we scale Pen. metric by a factor of 1/10 to fit the graph.

that, in general, PA-MPJPE and W-MPJPE errors behave in a slightly different manner depending on the dataset. On the one hand, for CHI3D, when changing $N_l$ from 1 to 2, the pose error drops (both PA-MPJPE and W-MPJPE ) and then the values stay roughly the same when increasing $N_l$. On the other hand, we see that for ExPI the error drops until reaching its lowest value for $N_l = 4$ for both PA-MPJPE and W-MPJPE and then when $N_l = 5$ the error increases again. We hypothesise that this increase in error for $N_l = 5$ is due to the fact that it is hard for the policy to maintain a static reference pose when this pose is a valid static pose (*e.g.* sitting down, hugging) and not otherwise (*e.g.* a person doing a back-flip). Thus, the more iterations added to the loop (an increase in $N_l$) for a given pose in a highly dynamics motion will result in a degradation of this pose.

If we look at the acceleration metrics, we see that for ExPI (Fig. 5b), the acceleration error grows with $N_l$, this is due to the high dynamic nature of the motions.

## 4. Results Discussion

The discussion we present next is based on the main results table from the main text which we copy here as Tab. 1 for the ease of the reader.

**Difference in PA-MPJPE .** As it can be seen in Tab. 1, the pose metric PA-MPJPE is slightly higher for our system in comparison to SLAHMR [7] on the Hi4D and ExPI datasets. While we discussed this on the main text, we expand that discussion here. As mentioned on Sec. 2, there is a difference in the type of motion present in each dataset we use. We have seen that, for example, ExPI is composed of dance routines with aerial steps and it was built by recording professional dancers. As such, we expect that the data contained in ExPI is much more dynamic, *i.e.*, the movements are faster and can present uncommon poses, than CHI3D and Hi4D. Thus, it is harder for our physics-aware correction module to match the reference pose. For this reason, to better match the reference pose, we need to choose a bet-

| CHI3D | | | | | | |
|---|---|---|---|---|---|---|
| Method | Pen.↓ | Gnd Pen.↓ | Skating↓ | Acc. Error↓ | W-MPJPE↓ | PA-MPJPE (joint)↓ |
| SLAHMR [7] | 139.3 | 4.4 | **1.0** | **6.5** | 177.1 | 83.5 |
| EmbPose-MP [3] | 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 | | | | | | |
| Method | Pen.↓ | Gnd Pen.↓ | Skating↓ | Acc. Error↓ | W-MPJPE↓ | PA-MPJPE (joint)↓ |
| SLAHMR [7] | 367.3 | 12.2 | 4.9 | **6.9** | 121.6 | **69.1** |
| EmbPose-MP [3] | **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 | | | | | | |
| Method | Pen.↓ | Gnd Pen.↓ | Skating↓ | Acc. Error↓ | W-MPJPE↓ | PA-MPJPE (joint)↓ |
| SLAHMR [7] | 567.3 | 18.6 | 5.4 | **8.2** | 263.3 | **159.1** |
| EmbPose-MP [3] | 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 [1], Hi4D [8], and ExPI [6] datasets. Pose metrics are W-MPJPE and PA-MPJPE (joint) in mm.

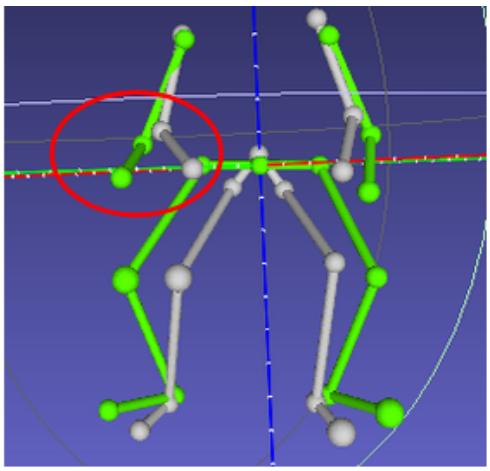


Figure 6. **Effect of scale ambiguity in the joint estimation.** Here we present the ground truth skeleton (green) and the estimated skeleton for P1 (gray) from the same frame as shown in Fig. 7. We see that the right hand deviates considerably from its real position. As explained in the text, this is caused by the depth misestimated in the initialization when combined with the physics-aware correction module.

ter value for $N_l$. In the particular case of ExPI dataset, if we look at Fig. 5 ignoring the rest of the metrics, we see that if $N_l = 4$ the PA-MPJPE (Fig. 5b) error drops from 165.7 ($N_l = 1$) and 164.3 ($N_l = 2$) to 163.5 ($N_l = 4$). For CHI3D (Fig. 4b), we also observe that if we increase $N_l$ the error decreases, however, in this case because the dataset presents less dynamic motions than ExPI, this change is much lower. Aside from this, there are two more subtle reasons why PA-MPJPE is slightly higher for our system in these datasets (CHI3D and ExPI): (i) scale ambiguity and (ii) propagation of initial incorrect poses. These causes are generally independent but, in some cases, (i) can cause (ii). We can explain this with the example in Fig. 7. Here we see a case where the initial estimates do not capture the correct depth for each person (reason (i)), specifically, the man (P1) is estimated closer to the camera than the woman (P2). This type of poses affect the physically corrected results more than in the case of purely kinematic estimates (reason (ii)). This is due to the fact that in the former the bodies cannot penetrate each other thus affecting the pose of the hand. This effect can be appreciated in more detail in Fig. 6. This does not happen with kinematic estimates. Therefore, the PA-MPJPE metric is affected and thus slightly higher compared to the initial estimates.

In general, (ii) can happen even when depth ambiguity is not the root cause of an initial inaccurate estimate but there are causes, for example, erroneous 2D keypoint estimates. In this case, only reason (ii) applies.

**Improvement in poses** As mentioned in the main text (Sec. 4.1), one scenario where the estimated poses improve, after forcing the motion to be physically compliant, is when a person has both of their 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. In contrast, for kinematic estimates, this constraint usually does not exist and if it does, it is imposed only as a soft constraint through a loss in an opti-

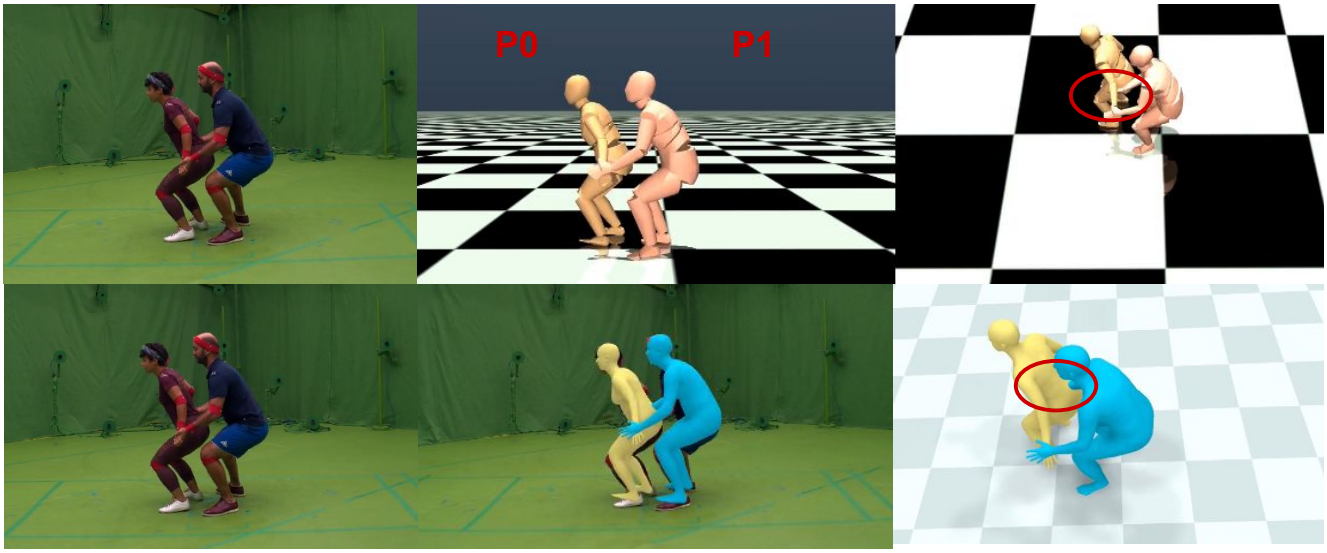



Figure 7. **Example of scale ambiguity in the physics correction.** Here, P0 should be at the same depth than P1, this inaccurate depth estimate that comes from the initialization (bottom row) causes the pose metric PA-MPJPE to increase for the physically corrected pose (top row). Note that in the rightmost column, while in the physically corrected pose, the right hand of P1 is displaced as it cannot penetrate P0's body, this does not happen in the kinematic pose. Thus in this case, the kinematic pose has a lower PA-MPJPE error.

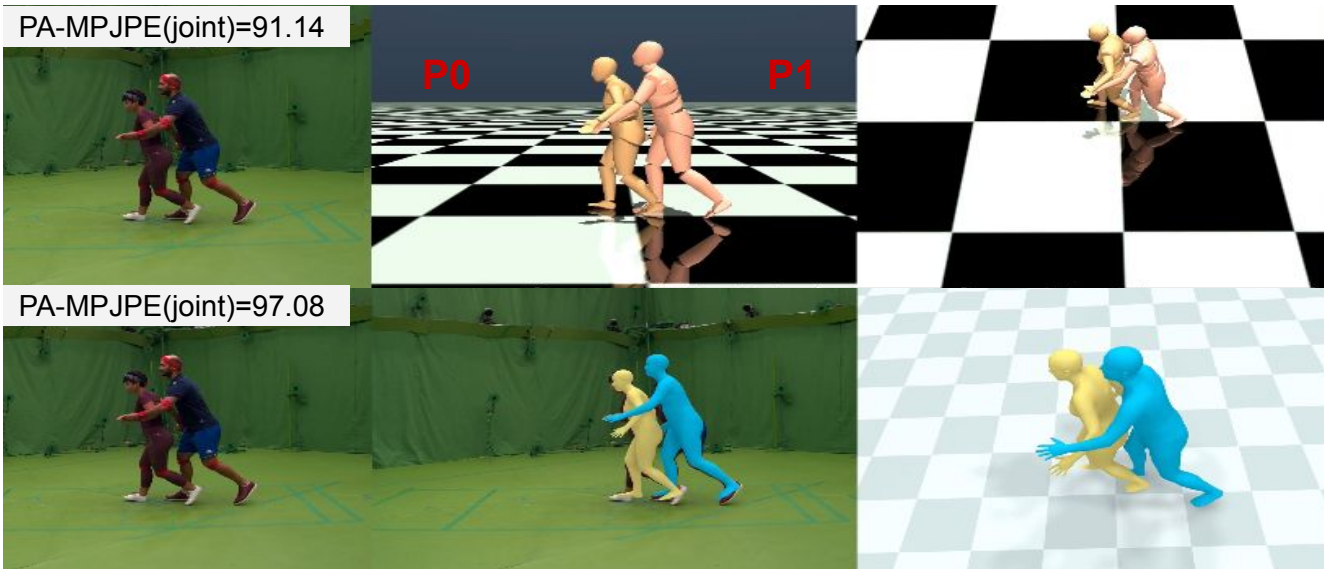



Figure 8. **Example of improved poses.** We show a example where poses are improved *w.r.t*. the initial poses when using our physics-aware correction module as measured by the PA-MPJPE metric. See the text and Fig. 9 for detailed analysis.

mization setup. Hence, kinematic estimates can penetrate the ground leading to less accurate poses. We illustrate this in Fig. 8, Fig. 9, and Fig. 10. In Fig. 8, we see the results for one frame of a motion sequence in ExPI dataset. Here, we present both the initial kinematic estimate (bottom row) and the physically corrected estimate (top row). Along with these results, we show the PA-MPJPE error for each case, where this value is higher for the kinematic estimate. What happens here is that in the kinematic case, the feet penetrate the ground leading to less accurate poses. This is better appreciated in Fig. 9, where we show the kinematic estimates in the simulation environment (rightmost image) together with superposed skeletons of P0 that correspond to the ground truth pose (green), kinematic pose (purple), and the physically corrected pose (light gray). Here, especially for the legs and the feet, the physically corrected pose is closer to the ground truth. On more example of these effect is shown in Fig. 10.

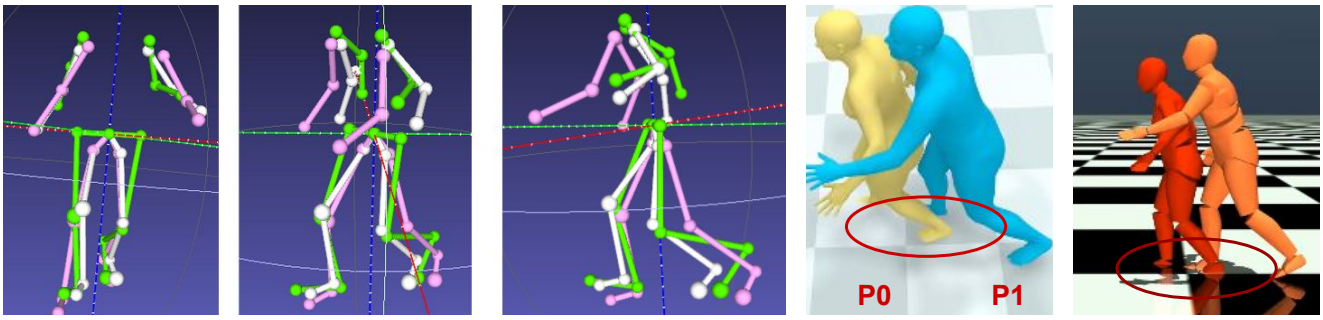



Figure 9. **Joints visualization for pose improvement.** Here, we show a detailed analysis of the joints corresponding to P0. The first three images show the ground truth pose for (green), kinematic pose (purple), and the physically corrected pose (light gray), respectively. The fourth image shows the kinematic estimates from a specific camera view. The last image shows the same kinematic poses in the simulation space without applying the physics constraints, note that here we see how the feet penetrate the ground.

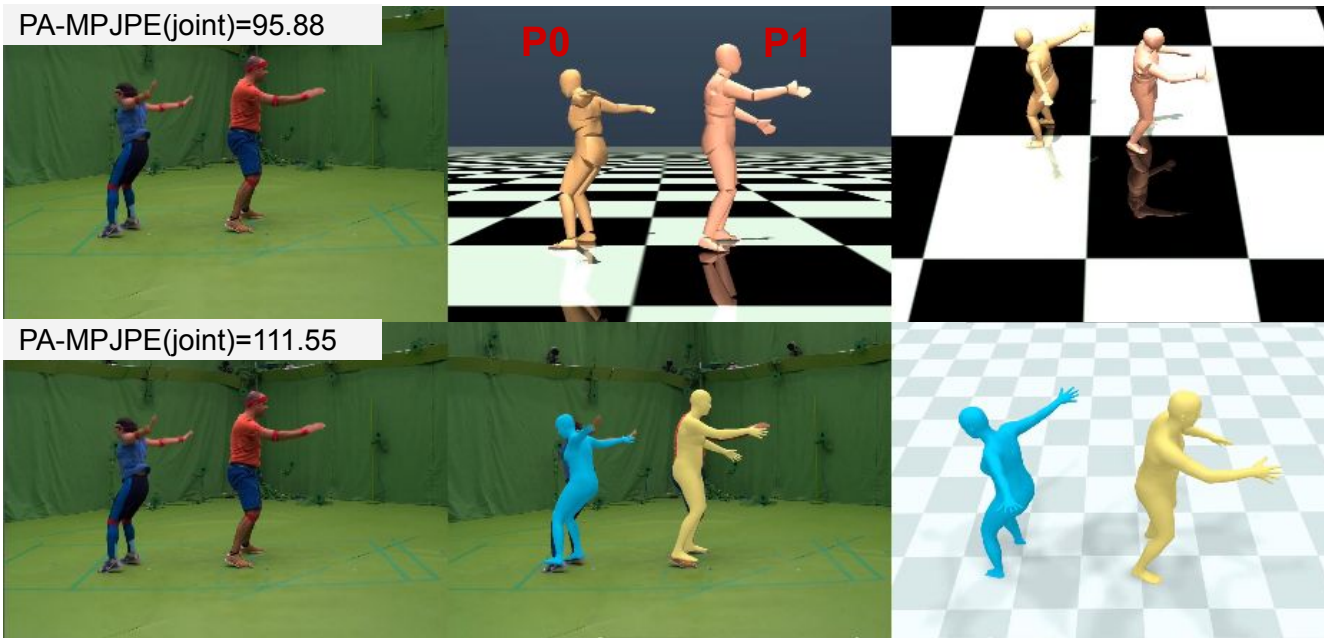



Figure 10. **Example of pose improvement.** Here we show another example where adding physics constraints with a simulation, using our physics-aware correction module, improves the estimated poses. Note how the PA-MPJPE metric improves in the physics-based estimates.

## References


[1] 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. 1, 2, 3, 4

[2] 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

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

[4] 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. 1

[5] 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. 1

[6] 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. 1, 2, 3, 4

[7] 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, 3, 4

[8] 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. 2, 4