

MotionPRO: Exploring the Role of Pressure in Human MoCap and Beyond

Supplementary Material

8. Dataset Details:

Implementation Details.

Each participant performs almost all the motion types. Each motion type is repeated two or three times. Each sequence represents a Sub-Motion Type in Fig. 3 and lasts about 10 minutes. Following Human3.6M, we split the dataset into training and test sets at a 5:1 ratio based on participants, ensuring that there is no overlap between training and test sets for any <Participant, Motion Type >pair.

Volunteers Details.

Gender: Our dataset consists of 70 individuals, comprising 29 females and 41 males, as shown in Fig. 9.

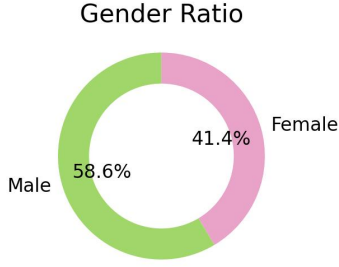


Figure 9. Gender Ratio of MotionPRO

Age: As shown in Fig. 10, our dataset encompasses individuals across a broad age range, spanning from 15 to 61 years, with an average age of 31.4 for women and 26.6 for men.

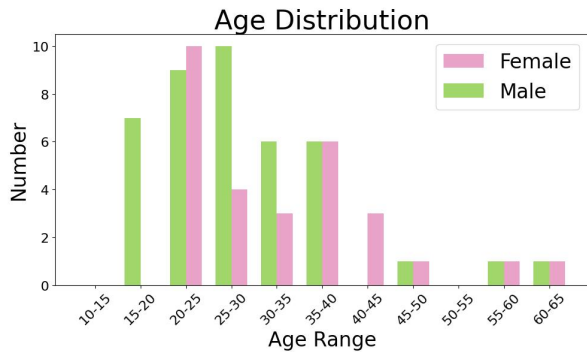


Figure 10. Age Distribution by Gender (5 years intervals)

Height: As shown in Fig. 11, our dataset includes individuals of varying heights, spanning from 157 cm to 185

cm, with an average height of 162.9 cm for women and 176.2 cm for men.

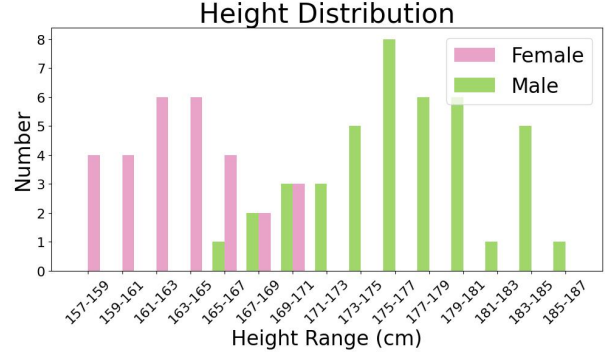


Figure 11. Height Distribution by Gender (5 cm intervals)

Weight: As shown in Fig. 12, our dataset includes individuals with a range of weights, spanning from 44.1 kg to 108 kg, with an average weight of 59.8 kg for women and 78.0 kg for men.

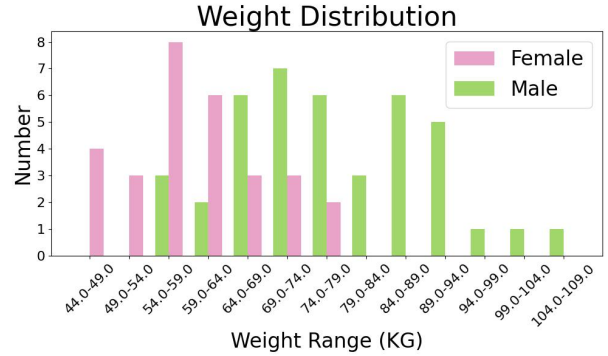


Figure 12. Weight Distribution by Gender (5 KG intervals)

Sensor Details.

Our system utilizes a multi-sensor setup for data acquisition:

- 4 Azure Kinect cameras [38] to capture high-quality RGB videos.
- 12 optical cameras (SWIFT 30) [34] to record raw marker data for precise motion tracking.
- 1 pressure mat, specially designed for our system, to measure whole-body pressure during various motions.

Motion Types.

The T-SNE [52] and UMAP [36] plot in Fig.13

and Fig.14 demonstrates that MotionPRO encompasses a wide range of motion types, nearly equivalent to the combined distribution of all currently available datasets (AMASS [35], MoYo [50], TIP [57], IC [33], SLP [30]). The figure on the left represents the T-SNE or UMAP distribution of the existing dataset, while the figure on the right illustrates the results of directly mapping MotionPRO based on the T-SNE or UMAP distribution observed on the left.

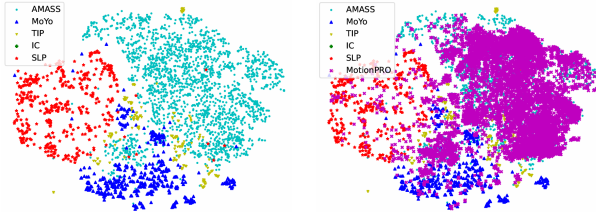


Figure 13. The distribution of poses in MotionPRO and existing MoCap datasets is visualized using T-SNE [52] dimensionality reduction.

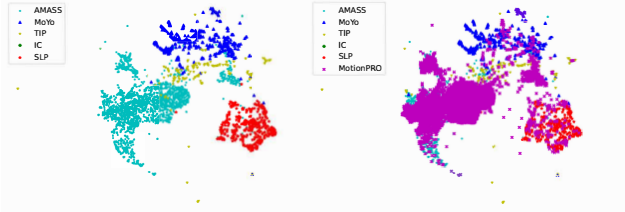


Figure 14. The distribution of poses in MotionPRO and existing MoCap datasets is visualized using UMAP [36] dimensionality reduction.

Motion Categories.

We define the six first-level categories as follows:

Daily: This category includes 172 common motions of daily life, such as basic postures, simple activities, and repetitive behaviors. These motions are characterized by natural, non-specialized patterns with high frequency, serving as a crucial baseline for developing human motions.

Robot: This category includes motions that simulate robotic or mechanized behaviors, characterized by mechanical patterns, fixed postures, high repetition, and predictability. Such data are essential for research on robotic motion simulation and human-robot interaction dynamics.

Flexibility Exercise: This category primarily includes motions involving large joint ranges of motion and the maintenance of slow, stable postures, such as leg stretches and splits.

Aerobic Exercise: This category comprises fitness activities defined by high-frequency, large-amplitude, full-body movements, typically associated with cardiovascular training.

Traditional Chinese Exercise: This category emphasizes movements characterized by fluidity, control, and balance, contrasting with high-intensity workouts and reflecting the characteristics of traditional Chinese fitness practices.

Ethics.

Volunteers in the MotionPRO dataset are well informed, and all participants have signed a Data Release Commitment Agreement, permitting the use of their data for research purposes.

9. Baseline Details:

Intuition of pose estimation from pressure

Through the spatial distribution and temporal changes of pressure, we verify that foot-to-floor pressure sensor readings can provide important discriminative prior information for pose estimation. Take standing and squatting as an example (shown in Fig.15), the CoP (Center of Pressure) is close to the heel and the toes exert almost no pressure on the ground when a person is standing. Conversely, when squatting, the CoP shifts closer to the forefoot and the toes generate pressure on the ground, helping to maintain balance. Additionally, the temporal relationship can provide more distinctive features. For example, when the posture transitions from standing to squatting, the body generates vertical acceleration, which leads to changes in both the total pressure value and the pressure distribution over time.

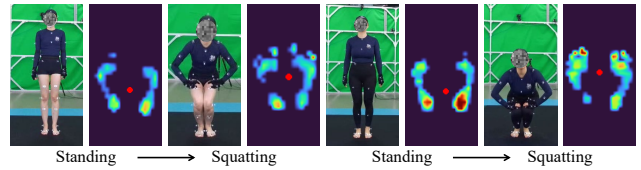


Figure 15. Comparison of pressure between standing and squatting.

Pressure network details.

When standing, the effective pressure area is small, requiring more fine-grained feature extraction. To address this, we reduce the size of the first convolution kernel in the pressure encoder, enabling us to capture more features within the limited pressure area. LSAM comprises two layers of bidirectional GRU and one layer of Self-Attention, with each layer incorporating a residual connection. The specific configuration of the network structure is determined through testing on toy examples.

Loss functions.

The loss of pose parameters \mathcal{L}_{pose} is the mean squared error between the predicted θ and ground-truth pose parameters $\tilde{\theta}$.

$$\mathcal{L}_{pose} = \|\theta - \tilde{\theta}\|_2^2 \quad (4)$$

The 3D joint loss, \mathcal{L}_{3d} , is the mean squared error between the predicted joints $\mathbf{J}(\boldsymbol{\theta}, \mathbf{T})$ and ground-truth whole-body joints $\tilde{\mathbf{J}}(\boldsymbol{\theta}, \mathbf{T})$, after performing pelvis alignment.

$$\mathcal{L}_{3d} = \|\mathbf{J}(\boldsymbol{\theta}, \mathbf{T}) - \tilde{\mathbf{J}}(\boldsymbol{\theta}, \mathbf{T})\|_2^2 \quad (5)$$

Global translation loss \mathcal{L}_{trans} is the mean squared error between predicted translation \mathbf{T} and ground truth translation $\tilde{\mathbf{T}}$.

$$\mathcal{L}_{trans} = \|\mathbf{T} - \tilde{\mathbf{T}}\|_2^2 \quad (6)$$

The ground contact loss, $\mathcal{L}_{contact}$, is the mean squared error between the predicted global whole-body in-contact joints $\mathbf{J}_C(\boldsymbol{\theta}, \mathbf{T})$ and the ground-truth global whole-body in-contact joints $\tilde{\mathbf{J}}_C(\boldsymbol{\theta}, \mathbf{T})$.

$$\mathcal{L}_{contact} = \|\mathbf{J}_C(\boldsymbol{\theta}, \mathbf{T}) - \tilde{\mathbf{J}}_C(\boldsymbol{\theta}, \mathbf{T})\|_2^2 \quad (7)$$

\mathcal{L}_{2d} is the mean squared error of orthographic projection $\mathcal{O}(\cdot)$ in the camera direction between the predicted joints and ground truth joints.

$$\mathcal{L}_{2d} = \|\mathcal{O}(\mathbf{J}(\boldsymbol{\theta}, \mathbf{T})) - \mathcal{O}(\tilde{\mathbf{J}}(\boldsymbol{\theta}, \mathbf{T}))\|_2^2, \quad (8)$$

Implement Details.

When driving virtual humans or robots in a 3D environment, their shapes typically remain constant over time. These shapes are often specifically designed and can differ significantly from those of human motion providers. Therefore, human body shape estimation is not our focus. In both **Pose and Trajectory Estimation using Only Pressure** experiment and **Pose and Trajectory Estimation by Fusing Pressure and RGB** experiment, we do not utilize FRAPPE to estimate body shape. Instead, we pre-calculate a more reasonable and representative shape based on the actual human body dimensions and maintain it fixed throughout training and evaluation. Similarly, the shapes for other comparison methods are also set to a consistent shape to ensure fairness in evaluation. FRAPPE outputs the SMPL pose and translation parameters $\boldsymbol{\theta}, \mathbf{T}$.

FRAPPE takes 20 frames of consecutive RGB and pressure images as input. The RGB images used in our method are captured from a frontal view monocular camera, providing a direct perspective for motion analysis. Notably, in the image branch, the encoder parameters are kept frozen during training. This ensures that the model focuses on learning the fusion of pressure and RGB features rather than re-learning image-specific features. At the same time, we also ensure fairness in comparison with other methods on the MotionPRO dataset, that is, our RGB image encoder, like other methods, is not trained on the MotionPRO dataset. We use AdamW optimizer with an initial learning rate of $5e^{-5}$ on 4 RTX 4090D GPUs.

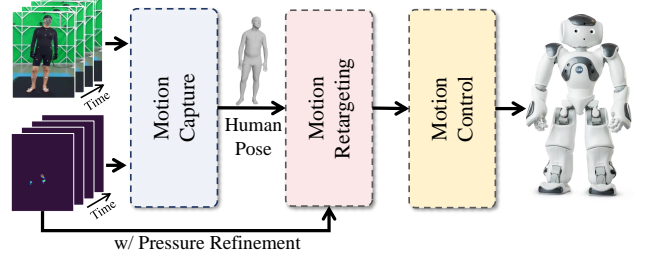


Figure 16. Framework of the robot demonstration system.

10. Robot Actuation Details:

We use the estimated human pose to actuate the robot. Our robot demonstration system is shown in Fig. 16. Specifically, we first extract human skeletal joint points from the SMPL model, which is estimated in motion capture module. The human joint points are then retargeted to corresponding target joint points that the robot can execute, involving coordinate transformation, scaling, Center of Mass (CoM) tracking, and other related processes. Finally, in the robot motion control module, we provide the retargeted pose to the robot controller for inverse kinematics optimization and whole-body control. For further details, refer to [26, 32, 40, 42].

Through the analysis of our framework, we argue that the performance of the robot’s action depends not only on the motion capture module but also on the other modules. Therefore, we investigate further optimization of the motion retargeting modules through the use of pressure data. Specifically, as the CoM distribution of the estimated human model does not perfectly align with the real pressure data, we refine the joint points using the pressure data, following [32], to ensure that the body CoM offset aligns with the pressure offset. Moreover, pressure data provides highly accurate information on human body contact, which can be used as a reference for controlling the robot’s support mode. We apply this approach to CLIFF and FRAPPE and corresponding results are shown in the main text.

We now clarify why CLIFF method performs better than ours in completeness, as discussed in the main text. For challenging actions that the robot cannot perform in the dataset, such as jumping, lying down, and the plank pose, etc, our method leads to the robot falling when imitating due to the higher accuracy of our estimated poses. In contrast, CLIFF’s less accurate poses allow the robot to remain standing and continue demonstrating the next action. In addition, it should be mentioned that the MPJPE-H metric primarily measures the difference between the estimated human pose and the ground truth. As we use the human pose captured by the optical system as the ground truth, resulting in a value of 0 for the optical MPJPE-H in Tab. 6 of the main text.

11. Future Work

Our dataset offers valuable opportunities for future research, particularly to examine the relationship between contact duration within the Base of Support (BoS), the distance between the Center of Mass (CoM) and the Center of Pressure (CoP), and demographic factors such as age, weight, and height. In addition, it supports applications in health monitoring and sports training. A key next step is to infer pressure information from visual input, which would expand its applicability by reducing the reliance on specialized sensors. Our dataset provides essential support for the advancement of these research directions.