Incorporating Physics Principles for Precise Human Motion Prediction
** Supplementary Materials **

Yufei Zhang[1], Jeffrey O. Kephart[2], Qiang Ji[1]
[1]Rensselaer Polytechnic Institute, [2]IBM Research
{zhangy76, jiq}@rpi.edu, kephart@us.ibm.com

In this supplementary material, we provide additional details of the experiment protocols in Section 1, and illustration of the fusion weights in Section 2.

1. Additional Details of Experiment Protocols

Model Architecture. PhysMoP includes different components to extract features, predict unknown physical parameters, and predict data-driven estimates and fusion weightis. Details of the model architecture are illustrated in Figure 1.

Datasets. For AMASS, its training subset consists of ACCAD [4], MPI-Limits [2], CMU [1], EyesJapanDataset [8], KIT [7], EKUT [9], TotalCapture [9], and TCDhandMocap [6]. When computing the error metrics, we evaluate on the same 256 samples and 22 body joints as [5] for Human3.6M. We evaluate on the same 18 body joints as used by [5] for both AMASS and 3DPW.

Implementation. Our implementation is on PyTorch. The frame rate of the motion data is 25 for Human3.6M, and 30 for AMASS and 3DPW, respectively. For data preprocessing, we apply a Gaussian filter to smooth the computed angles and the body translation to reduce the impact of unrelated noise introduced during data collection. We set a batch size of 64 for training the data-driven and physics-based models, and a batch size of 196 for training the fusion model. We utilize joint angles in SMPL format. For Human3.6M, we obtain SMPL data from [10] since the original dataset does not include SMPL data. As for AMASS and 3DPW, we directly use the provided SMPL data.

2. Illustration of Fusion Weights

To provide deep insights towards the fusion process, we present examples of the fusion weights in Figure 2. The examples are testing sequences from Human3.6M. The fusion weights indicate the importance of the data-driven estimates at a certain time stamps. As shown, the weights are

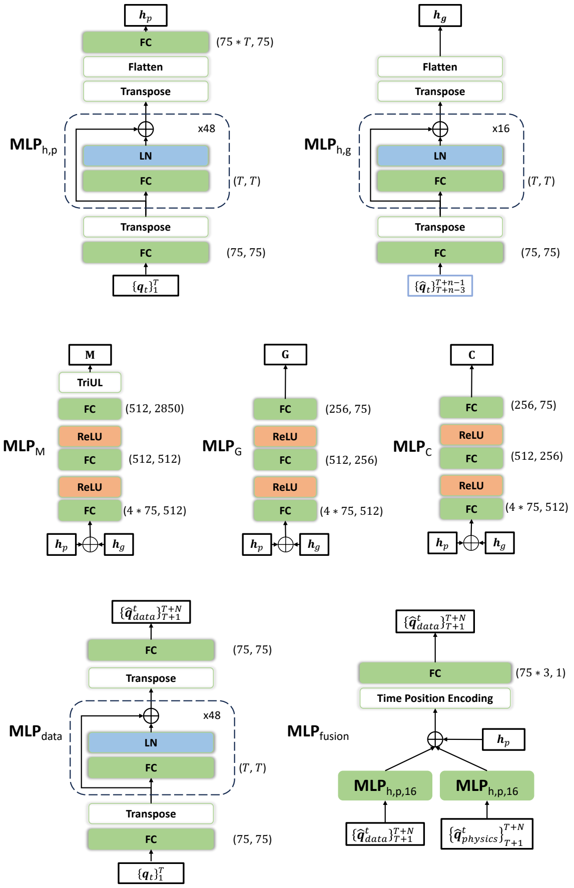

Figure 1. **Model architecture.** We illustrate the neural network architecture of the main components of PhysMoP. “FC”, “LN”, “ReLU” stands for the fully connected layer, layer normalization [3], and activation layer, respectively. The dimension of each fully connected layer is shown on its right side. “Transpose” exchanges the spatial and temporal dimension. “TriUL” generates a symmetric matrix from a vector prediction. “$\text{MLP}_{h,p,16}$” represents the framework that has the same architecture as “$\text{MLP}_{h,p}$” but is different in the number of blocks of the MLP (16 v.s. 48).

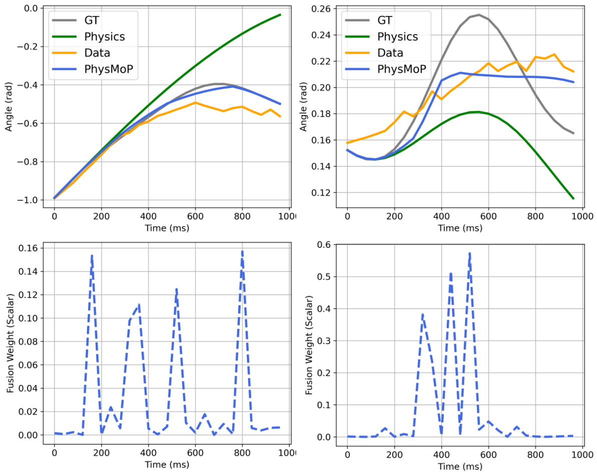

Figure 2. **Joint angles estimated by different models (top) and the corresponding fusion weights (bottom).**

smaller at shorter future time stamps as the physics-based model is advantageous in accurately capture the short-term physical movements. Moreover, the fusion weights become larger and occasionally exhibit oscillations around 500ms or longer time horizons. The reason is that we perform the fusion in an iterative manner which fully leverages the power of the physics-based model and makes rectification only when the estimates tend to diverge. Meanwhile, it is worth noting that values of the fusion weights are small if the estimates generated by the physics-based model are closer to the ground truth (as seen in the left column of Figure 2), indicating a reduced dependency on the data-driven model.

References

[1] NSF Grant 0196217. CMU graphics lab motion capture database. 1

[2] Ijaz Akhter and Michael J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2015*, June 2015. 1

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 1

[4] Advanced Computing Center for the Arts and Design. Accad mocap dataset. 1

[5] Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Back to mlp: A simple baseline for human motion prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4809–4819, 2023. 1

[6] Ludovic Hoyet, Kenneth Ryall, Rachel McDonnell, and Carol O'Sullivan. Sleight of hand: perception of finger motion from reduced marker sets. In *Proceedings of the ACM SIGGRAPH symposium on interactive 3D graphics and games*, pages 79–86, 2012. 1

[7] Franziska Krebs, Andre Meixner, Isabel Patzer, and Tamim Asfour. The kit bimanual manipulation dataset. In *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, pages 499–506, 2021. 1

[8] Eyes JAPAN Co Ltd. Eyes japan mocap dataset. 1

[9] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. Unifying representations and large-scale whole-body motion databases for studying human motion. *IEEE Transactions on Robotics*, 32(4):796–809, 2016. 1

[10] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Neuralannot: Neural annotator for 3d human mesh training sets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2299–2307, 2022. 1