MBTI: Masked Blending Transformers with Implicit Positional Encoding for Frame-rate Agnostic Motion Estimation

Supplementary Material

A. Implementation Details

A.1. Model Configuration

The high reference frame rate is fixed at $f_{\text{ref}} = 120$, while the input frame rate f_i and time interval length T are configurable by design. We fix the interval length $T = 0.5$ empirically. For all transformer encoders and decoders, we set the embedding dimension to $d_m = 512$ and use 8 heads for multi-head self-attention. The number of layers is set to 3, except for the trajectory refiner, which uses 2 layers.

A.2. Pretraining

The model is pretrained without the feature integrator F on the AMASS dataset [6], following the setup of WHAM [9]. The input frame rate f_i is randomly sampled from the set $\mathcal{F} = \{10, 15, 24, 25, 30, 45, 48, 50, 60, 75, 90, 120\}$. Given the input frame rate, 2D joints and angular velocity are interpolated using linear interpolation, following Eq. (9) and Eq. (10) from the main paper. The output 3D motion sequence is upsampled if the target frame rate is lower than $f_{\text{ref}} = 120$, ensuring consistency across training. The loss function weights are set to $\lambda_{2D} = 0.1$, $\lambda_{3D} = 0.4$, $\lambda_{\text{SMPL}} = 8.0$, $\lambda_V = 0.5$. The model is trained for 200 epochs with a batch size of 64, twice as long as WHAM, as learning to predict high reference frame rate from varying lower frame rate inputs is more challenging. We use the AdamW optimizer [5] with an initial learning rate of 5×10^{-4} and a weight decay of 0.05. To stabilize training, learning rate decays by a factor of 10^{-1} at epochs 120 and 160.

A.3. Finetuning

Using the pretrained model, we finetune on video datasets, including 3DPW [10], Human3.6M [2], MPI-INF-3DHP [7], and InstaVariety [3], and BEDLAM [1] to adapt the model to video inputs. To leverage diverse motion patterns, the AMASS dataset is also included in finetuning with zero image feature vectors. For AMASS, we follow the same pretraining protocol, randomly sampling f_i from $\mathcal{F}$ and interpolating input and output sequences. For video datasets, the input frame rate matches the original video frame rates: 30 FPS for 3DPW, 50 FPS for Human3.6M, 25 FPS for MPI-INF-3DHP, and 24–30 FPS for InstaVariety. Despite finetuning on fixed-frame-rate videos, the model generalizes well to variable-frame-rate inputs due to the inclusion of AMASS during training. The loss function weights are set to $\lambda_{2D} = 3.0$, $\lambda_{3D} = 6.0$, $\lambda_{\text{SMPL}} = 1.0$, $\lambda_V = 0.01$. The finetuning process spans 80 epochs with a batch size of 64, again double that of WHAM as in the pretraining stage. The learning rate for the feature integrator is set to 10^{-4}, while the rest of the model uses a learning rate of 10^{-5}, both with a weight decay of 0.05. Learning rates decay by 10^{-1} at epochs 40 and 60 for stabilization.

B. EMDB-FPS Dataset

To evaluate frame rate-agnostic human motion estimation, we constructed the EMDB-FPS dataset by augmenting the EMDB dataset [4]. We first augmented the videos to various frame rates, specifically frame rates $F = \{10, 20, 30, 60, 120\}$. Starting from the source video frame rate of 30 FPS, we generated lower frame rate videos (10 FPS and 20 FPS) via downsampling and higher frame rate videos (60 FPS and 120 FPS) via upsampling. Figure 1 illustrates the construction process for lower and higher frame rates from the original source video. For upsampling, we use a video frame interpolation method [8] to synthesize intermediate frames. The annotation labels, such as the 2D joint locations, 3D SMPL parameters, and camera trajectory, are also augmented to match the new frame rates. For the 2D joint locations, they are initially labeled via linear interpolation, and further fine-tuned using ViTPose detections [11]. For the 3D SMPL parameters, we use the original SMPL parameters from the EMDB dataset and interpolate them to match the new frame rates. As the interpolated motion does not represent the actual human motion, we further refine the 3D SMPL parameters using the ViTPose detections. Specifically, we reproject the 3D joint locations computed from the SMPL parameters to the 2D joint locations, and optimize the SMPL parameters to minimize the reprojection error with the VITPose detections. The camera trajectory is also interpolated to match the new frame rates, assuming a constant velocity model. This augmentation ensures a comprehensive evaluation of the model's robustness across a wide range of frame rates.

References

[1] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023. 1

[2] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments.

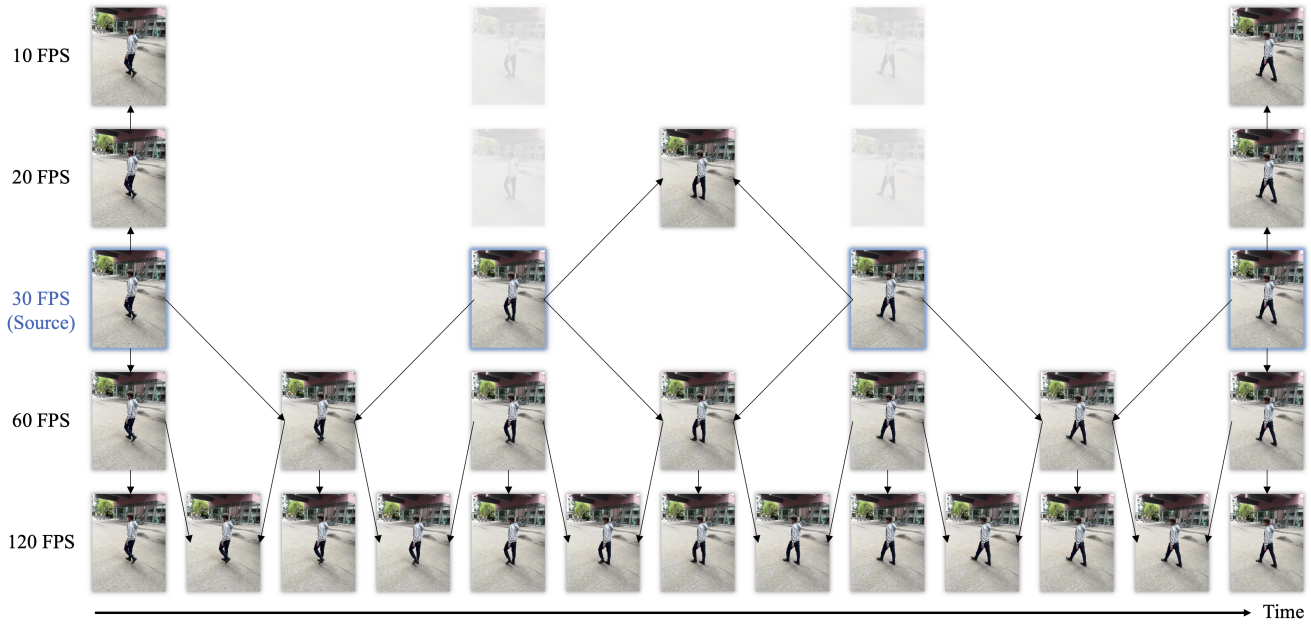

Figure 1. The construction process of multiple frame rate videos in the EMDB-FPS dataset.

IEEE transactions on pattern analysis and machine intelligence, 36(7):1325–1339, 2013. 1

[3] 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*, pages 5614–5623, 2019. 1

[4] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14632–14643, 2023. 1

[5] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 1

[6] 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*, pages 5442–5451, 2019. 1

[7] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017. 1

[8] Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion. In *European Conference on Computer Vision*, pages 250–266. Springer, 2022. 1

[9] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2070–2080, 2024. 1

[10] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018. 1

[11] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 1