HUMOS: Human Motion Model Conditioned on Body Shape
Supplemental Material

Anonymous ECCV 2024 Submission

Paper ID #2441

In the supplementary materials, we provide a detailed derivation of the ZMP and the dynamic stability term (Sec. 1), analyze the effect of body shape on motion (Sec. 2), provide additional qualitative results (Sec. 3), ablations for the latent embedding losses (Sec. 4), a discussion on AMASS shape diversity (Sec. 5), and finally additional implementation details (Sec. 6).

Video. Our research focuses on humans in motion with diverse body shapes and sizes, making motion a critical aspect of our results. Given the difficulty of conveying motion quality through a static document, we strongly recommend that readers view the provided supplemental video for an in-depth overview of our methodology and findings.

1 Detailed derivation for the Zero Moment Point (ZMP) and the dynamic stability term

Before we compute the ZMP, we first compute the body Center of Mass (CoM) by adapting the CoM formulation of Tripathi *et al.* [4] to dynamic humans. For every sequence, we use their body part segmentation and the differentiable *"close-translate-fill"* [4] to compute per-part volumes $\mathcal{V}^{P_i}$ by splitting the mesh in the first frame into 10 parts. Using the per-part volumes, the CoM is calculated for time instance t, as a volume weighted-average of $N_U = 6890$ mesh vertex points.

$$\mathcal{G}_t = \frac{\sum_{i=1}^{N_U} \mathcal{V}^{P_{v_i}} v_{i_t}}{\sum_{i=1}^{N_U} \mathcal{V}^{P_{v_i}}}, \tag{1}$$

The acceleration of the CoM, $a_{\mathcal{G}}$, is obtained using the central difference as,

$$a_{\mathcal{G}_t} = \frac{\mathcal{G}_{t+1} - 2\mathcal{G}_t + \mathcal{G}_{t-1}}{\Delta t^2} \tag{2}$$

With $a_{\mathcal{G}_t}$, the force of inertia, $\mathcal{F}^{gi}$, is computed as

$$\mathcal{F}^{gi} = mg - ma_{\mathcal{G}} \tag{3}$$

where m is the body mass. The moment around the projected CoM, $\mathcal{C}_m$, is

$$\mathcal{M}^{gi}_{\mathcal{C}} = \overrightarrow{\mathcal{C}_m\mathcal{G}} \times mg - \overrightarrow{\mathcal{C}_m\mathcal{G}} \times ma_{\mathcal{G}} - \dot{\mathcal{H}}_{\mathcal{G}} \tag{4}$$

where $\overrightarrow{\mathcal{C}_m\mathcal{G}}$ is the vector joining the projected CoM, $\mathcal{C}_m$ with the actual CoM, $\mathcal{G}$ and $\dot{\mathcal{H}}_\mathcal{G}$ is the rate of change of angular momentum at the CoM. For $\dot{\mathcal{H}}_\mathcal{G}$, we equally distribute the total m to point masses at the vertices of the body mesh proportional to the volume of the body part they are part of. The per-vertex mass and acceleration is

$$m_{v_i} = \frac{\mathcal{V}^{P_{v_i}}}{\sum_{i=1}^{N_U} \mathcal{V}^{P_{v_i}}} m, \quad a_{v_i} = \frac{v_{i_{t+1}} - 2v_{i_t} + v_{i_{t-1}}}{\Delta t^2} \tag{5}$$

And $\dot{\mathcal{H}}_\mathcal{G}$ is

$$\dot{\mathcal{H}}_\mathcal{G} = \sum_{i=1}^{N_U} \overrightarrow{v_i\mathcal{G}} \times m_{v_i} a_{v_i} \tag{6}$$

Finally, the ZMP is computed in closed-form as

$$\mathcal{Z} = \mathcal{C}_m - \frac{n \times \mathcal{M}^{gi}_{\mathcal{C}_m}}{\mathcal{F}^{gi} \cdot n} \tag{7}$$

For CoP computation, we follow Tripathi *et al.* and uniformly sample the body mesh into $N_p = 20000$ uniformly sampled surface points. We, then, use their heuristic pressure field to compute per-point, p_i, pressure as

$$\rho_i = \begin{cases} 1 - \alpha h(p_i) & \text{if } h(p_i) < 0, \\ e^{-\gamma h(p_i)} & \text{if } h(p_i) \geq 0, \end{cases} \tag{8}$$

where $\alpha = 100$ and $\gamma = 10$ are scalar hyperparameters set empirically. The CoP is computed as,

$$\mathcal{C}_p = \frac{\sum_{i=1}^{N_p} \rho_i p_i}{\sum_{i=1}^{N_p} \rho_i}. \tag{9}$$

With the ZMP and the CoP, known the dynamic stability loss is defined as,

$$\mathcal{L}_{\text{dyn}} = \rho(\|\mathcal{C}_P - \mathcal{Z}\|_2) \tag{10}$$

where ρ is the Geman-McClure penalty function [1].

2 Effect of body shape on motion

In Fig. 1 (left), we assess the diversity of HUMOS generated motions across 100 β parameters obtained by interpolating between a short male and a tall male body. We report the maximum right knee joint angle ($|\theta|$) for the same walk sequence shown in the Sup. Mat. (SM) video (05:34). The graph illustrates that taller people bend their knees less for the same walking motion, indicating body parameters affect movement. Similarly, in Fig. 1 (center), we plot the right hand joint velocity across six different identities in the same walk sequence [SM video (05:34)]. The joint velocities differ across subjects in corresponding frames implying diversity induced by body shape variation. In Fig. 1 (right), we qualitatively show the same frame of the jumping jack sequence [SM video (05:47)] where the different arm positions indicate motion diversity.

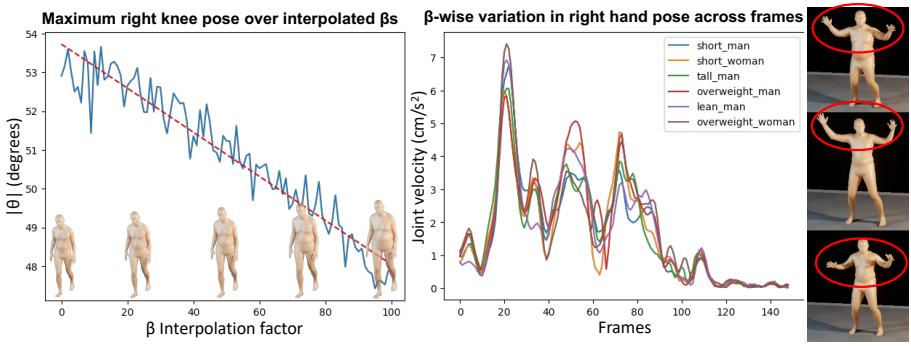

Fig. 1: Effect of body shape across (left) interpolated β parameters, (center) 150 frames for 6 different identities and (right) different identities for the same jumping jack frame 🔍 **Zoom in**.

3 Additional Qualitative Results

We include additional comparisons with baselines in Fig. 2. For video results, we recommend watching the **supplementary video**.

4 Additional Ablations

In Tab. 1, we conduct additional ablations to analyze the effect of latent embedding losses, $\mathcal{L}_E$ and $\mathcal{L}_{KL}$. We take the HUMOS model and successively remove the two loss terms individually. On ablating $\mathcal{L}_E$ during training, we observe a small improvement in ground penetration and float. However, the skate and dynamic stability metrics worsen. While the effect of $\mathcal{L}_E$ is minimal in terms of metrics, we empirically note faster and stable convergence when using it during training. $\mathcal{L}_{KL}$ also results in a slight improvement in physics metrics at the cost of dynamic stability. A significant advantage, however, of using $\mathcal{L}_{KL}$ is that it adds structure to the shape-agnostic latent space, making realistic motion generation easier.

Table 1: Ablation study for latent embedding losses, $\mathcal{L}_E$ and $\mathcal{L}_{KL}$

Method	Penetrate (cm) ↓	Float (cm) ↓	Skate (%) ↓	Dyn. Stability (%)↑	BoS Dist (cm) ↓
HUMOS	1.23	1.04	7.37	71.9	14.62
HUMOS - $\mathcal{L}_E$	1.20	0.98	9.3	71.0	15.01
HUMOS - $\mathcal{L}_{KL}$	1.14	0.93	6.96	71.05	15.21

5 AMASS Shape Statistics

For training and evaluation, we use the AMASS dataset [3]. AMASS is a comprehensive collection of human motion data, unifying various optical marker-based motion capture datasets. This dataset stands out due to its extensive volume, containing over 50 hours of motion data from 480 unique subjects, encompassing more than 11,000 distinct motions. Among the 480 unique subjects, we have 274 male and 206 female subjects. To understand the diversity of body shapes included in AMASS, in Fig. 3, we plot the mean and standard deviation of each principal component for the AMASS beta parameters. Following prior work, we use the first 10 shape principal components to represent body shape.

6 Additional Implementation Details.

Data processing. AMASS captures diverse human motions performed by real participants. Therefore, motions in AMASS start at arbitrary locations and facing directions. AMASS also includes motions where the person is supported by objects such as chairs, stairs or raised platforms. Only the human is captured in such sequences, and given the lack of a supporting object, these motions are physically implausible. These sequence, along with arbitrary start locations and facing directions, add unnecessary ambiguity and make the raw AMASS data unsuitable for neural network training. To prevent this, we process the raw AMASS data by removing all sequences where the lowest vertex in at least 5 frames is higher than 0.25m from the ground. Next, as described in the main text, we canonicalize all sequences to start at the origin with the same facing direction. To augment our training data, we mirror the pose parameters and global root translation from left-to-right and vice-versa, effectively doubling the training data. Figure 4 shows the effect of each step in our data processing pipeline.

Motion representation. The SMPL body model parameterizes the human body into body pose, shape and global root translation. The SMPL body pose is represented as parent-relative rotations in the axis-angle format. For our motion representation, we follow NeMF [2] and convert the parent-relative joint rotations to global root-relative rotations in 6d format [5]. This helps with convergence and produces better performance than using the SMPL parameters directly. We also experiment with using deltas in joint rotations and global translation in our motion representation. We empirically observe worse performance in this setting due to the propagation of errors in the integration step when recovering the joint rotations and translation from the predicted deltas.

6.1 Perceptual Study

We show the layout for our perceptual study in Fig. 5. We randomly sample sequence generations from our methods and baselines and every video is rated by 25 participants on Amazon Mechanical Turk. We ensure quality in ratings by adding two ground-truth videos and two catch-trial videos per worker with

extreme ground penetrations or floating sequences. Additionally, every participant is shown 5 *warming-up* sequences at the start of their annotation task which we discard. This allows the participant to get a sense of the task before they can reliably rate the generated motions. We report average ratings across all participants who qualify the quality checks.

To test statistical significance, we performed one-way ANOVA tests, yielding a significant p-value of $1.5 \times e^{-10}$. Tukey's HSD statistical test indicates that our method has statistically significant differences with TEMOS-Rokoko (mean diff $= 0.386$, $p < 0.001$) and TEMOS-Rokoko-G (mean diff $= 0.447$, $p < 0.001$).

Fig. 2: Additional qualitative comparison of shape-conditioned motion generation. Each row represents generations across different methods for a unique body shape and gender. The difference in quality between methods is particularly evident in their interaction with the ground. **Zoom in**.

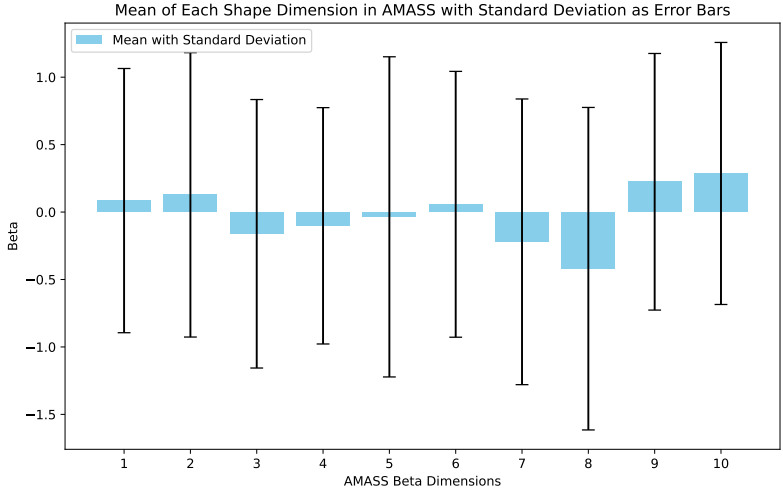

Fig. 3: Mean and standard deviation of the first 10 betas parameters in AMASS. This represents the diversity in body shapes.

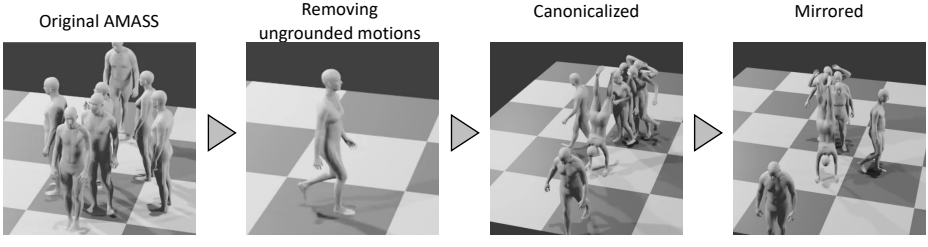

Fig. 4: We process the raw data from AMASS by 1) removing unsupported physically implausible motions *e.g.* walking up the stairs 2) canonicalizing all motions to start facing the same direction at origin and 3) mirroring the pose and root translation to augment data

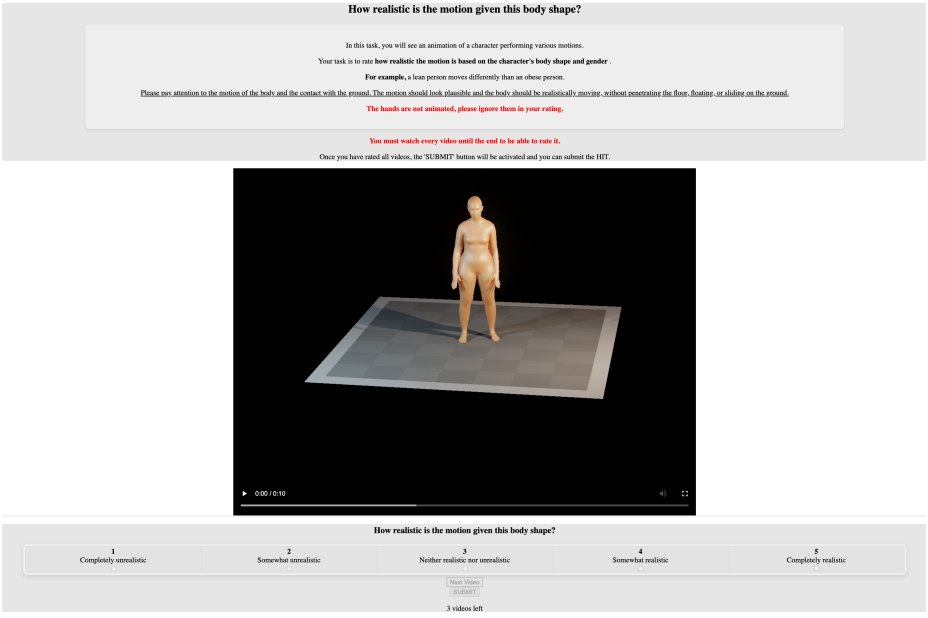

Fig. 5: Layout of the perceptual study.

References

1. Geman, S.: Statistical methods for tomographic image restoration. Bull. Internat. Statist. Inst. **52**, 5–21 (1987) 2
2. He, C., Saito, J., Zachary, J., Rushmeier, H.E., Zhou, Y.: NeMF: Neural motion fields for kinematic animation. In: NeurIPS (2022) 4
3. Mahmood, N., Ghorbani, N., F. Troje, N., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: International Conference on Computer Vision (ICCV). pp. 5441–5450 (2019) 4
4. Tripathi, S., Müller, L., Huang, C.H.P., Omid, T., Black, M.J., Tzionas, D.: 3D human pose estimation via intuitive physics. In: Computer Vision and Pattern Recognition (CVPR). pp. 4713–4725 (2023), https://ipman.is.tue.mpg.de 1
5. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: CVPR. pp. 5745–5753. Computer Vision Foundation / IEEE (2019) 4