

# ATLAS: Decoupling Skeletal and Shape Parameters for Expressive Parametric Human Modeling (Supplementary Materials)

Jinhyung Park<sup>1,2</sup> Javier Romero<sup>1</sup> Shunsuke Saito<sup>1</sup> Fabian Prada<sup>1</sup> Takaaki Shiratori<sup>1</sup>  
 Yichen Xu<sup>1</sup> Federica Bogo<sup>1</sup> Shoou-I Yu<sup>1</sup> Kris Kitani<sup>1,2</sup> Rawal Khirodkar<sup>1</sup>  
<sup>1</sup>Meta <sup>2</sup>Carnegie Mellon University

## A. Supplementary Overview

In the supplementary video, we present video results of fitting ATLAS to high-fidelity 3D scans, demonstrate controllability of skeletal attributes for a dynamic sequence, and show results of fitting ATLAS to RGB videos in the wild.

In this supplementary document, we provide additional details on skeletal attributes, visualizations of the training data, implementation details, and qualitative results of ATLAS. The sections are organized as follows:

- Section B provides additional details regarding the 76 individually controllable skeletal attributes of ATLAS.
- Section C outlines the specific formulation of the Linear Blend Skinning (LBS) function.
- Section D visualizes some sample registrations from our Goliath dataset.
- Section E contains additional details regarding the training of ATLAS and the pose prior.
- Section F shows the skin weights before and after training.
- Section G includes visualizations of the first few external shape & internal skeleton latent components.
- Section H demonstrates the full expressiveness of ATLAS by visualizing generated subjects through random sampling of external shape, internal skeleton, body poses, hand poses, and facial expressions.
- Section I provides additional results on our single image to mesh prediction pipeline on in-the-wild images.

## B. Details on Controllable Skeletal Attributes

ATLAS defines 76 controllable skeletal attributes that modify different parts of the skeleton. As described in the main

Shape	Skeleton	3DBodyTex	Goliath-Test
✓	✗	6.47	4.76
✗	✓	3.17	2.67
✓	✓	<b>2.48</b>	<b>2.34</b>

Table 4. Mesh fitting error (mm) with shape and skeleton params.

paper, 15 of these attributes directly scale a local joint space (and those of its kinematic children). These consist of scales that affect the full-body, head, hands, feet, and individual fingers. The remaining 61 are bone length parameters that directly adjust each joint’s center location with respect to its kinematic parent. These include the spine, neck offset, neck length, shoulder width, upper arms, lower arms, hip location, upper legs, lower legs, and each bone in the finger for precise controllability. We visualize the skeletal attributes that affect major parts (excluding individual finger bone adjustments) in Figure 12.

Further, we demonstrate that the surface shape basis and the skeletal basis are both necessary and are complementary by evaluating mesh fitting with disentangled parameters in Table 4. Shape alone misses height and limb length variations, while skeleton alone overlooks soft tissue. Using both, like ATLAS, best captures diverse body shapes.

## C. Linear Blend Skinning Formulation

In this section, we provide the precise formulation for the LBS skinning function  $M$  used in Section 3.1 of the main paper. This transformation  $M$  to yield a scaled and posed vertex  $x_i$  is written as:

$$x_i = \sum_{j=1}^I \omega_{ij} \mathcal{T}_j(\bar{\theta}, \bar{t}, \theta, \sigma, t) \mathcal{T}_j(\bar{\theta}, \bar{t}, \vec{0}, \vec{0})^{-1} \tilde{x}_i \quad (1)$$

where  $\bar{\theta}$  and  $\bar{t}$  define the rest pose of the skeleton. These rest pose definitions of each joint’s rotation and offset with respect to its parent are necessary because unlike SMPL [1] where each joint’s coordinate system is root axis-aligned, our rotations are skeleton-aligned. The forward kinematic transformation  $\mathcal{T}_j$  is then defined by:

$$\mathcal{T}_j(\bar{\theta}, \bar{t}, \theta, \sigma, t) = \Pi_{a \in K(j)} \begin{bmatrix} 2^{\sigma(a)} R(\theta_a) R(\bar{\theta}_a) & t(a) t_e(a) + \bar{t}_l \\ 0 & 1 \end{bmatrix} \quad (2)$$

where  $K(j)$  are the kinematic tree parents of joint  $a$  in ascending order,  $\sigma(a)$ ,  $t(a)$ , and  $t_e(a)$  are zero if joint  $a$  lacks a corresponding skeleton modification. Thus,

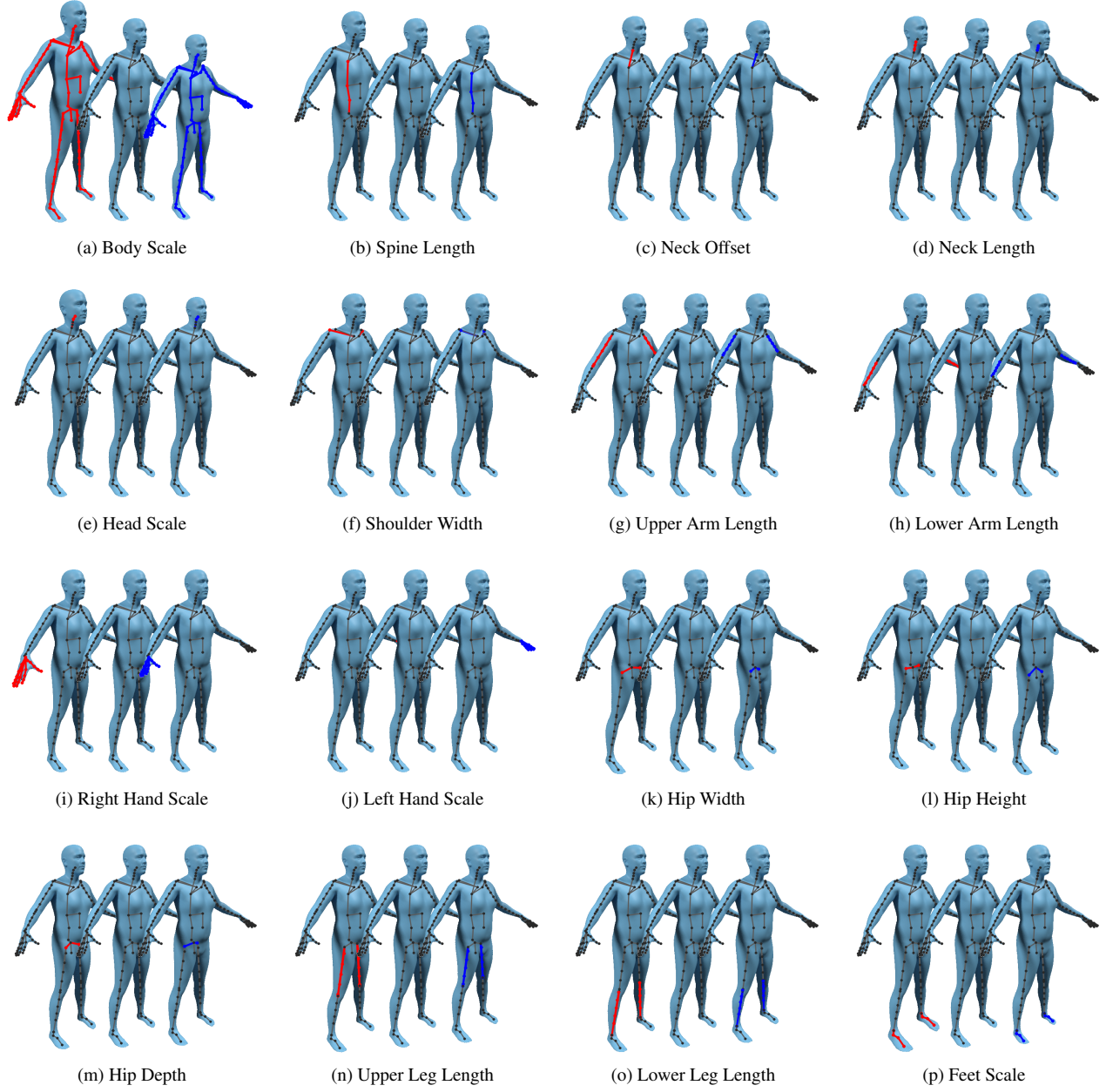


Figure 12. **Visualization of Body Skeletal Attributes.** For each skeletal attribute, we show three meshes - increasing the skeletal parameter, the base mesh, and decreasing the parameter. Bones affected by the changed parameter are colored red if they have increased in size, and blue if they have decreased. Each attribute either directly scales a local joint space, including those of its kinematic children, or adjusts joint translations relative to its own kinematic parent. For instance, Figure 12i shows an instance of the former, where the entirety of the right hand changes in size, while Figure 12f is an instance of the latter, where the shoulder joint center is moved, driving an increase or decrease in shoulder width.

$\mathcal{T}_j(\bar{\theta}, \bar{t}, \bar{0}, \bar{0}, \bar{0})^{-1}$  transforms from global to joint- $j$ 's local coordinates through kinematic tree traversal of an unposed, unscaled skeleton, while  $\mathcal{T}_j(\bar{\theta}, \bar{t}, \theta, \sigma, t)$  transforms from joint- $j$ 's local to global coordinates with skeleton posing and bone scale/length modifications.

#### D. Visualization of scans from the Goliath dataset

In Figure 14 we provide a sample of our Goliath dataset. To assemble a large and diverse set of scans to train our model, we capture 130 subjects in a diverse suite of poses including conversational settings, charades acting, and dy-



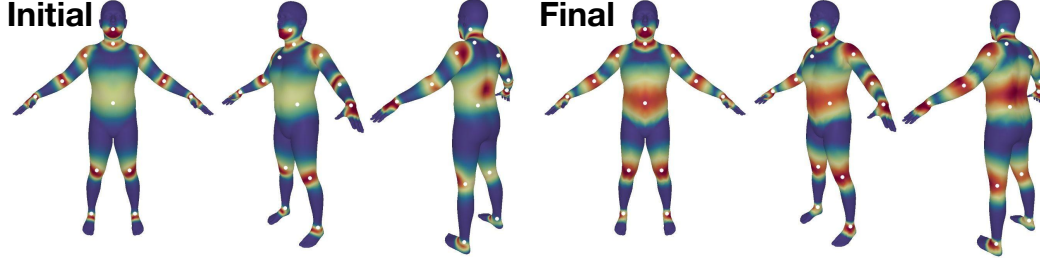


Figure 13. Skinning weights for the jaw, neck, upper arm, elbow, wrist, lower spine, knee, and ankle before and after optimization.

dynamic movements. The frames are captured using 240 high-resolution, synchronized cameras that yields meshes with approximately 1 million vertices. The scans are captured at 30-90 FPS, and we use furthest-point-sampling on pose to select an interesting and diverse set of 600k frames to train ATLAS.

## E. Training Details

### E.1. ATLAS Body Model Details

#### E.1.1. Vertex Resolutions

While ATLAS is natively trained at the highest resolution with 115,834 vertices, we define mappings to the 6890 and 10475 vertices of SMPL and SMPL-X. This enables transformations between ATLAS and SMPL/SMPL-X and allows ATLAS to operate with fewer vertices for improved efficiency.

#### E.1.2. Body Model Design

ATLAS leverages a joint structure designed by expert sculpting artists to ensure anatomical consistency. The joint locations adhere to the human bone structure, and in place of a standard single 3DoF rotation for major joints, ATLAS decomposes them into anatomically accurate sub-joints. For example, the shoulder includes a scapular joint, and the ankle is divided into subtalar and talocrural joints.

### E.2. ATLAS Training Details

ATLAS is trained end-to-end by sampling registrations with their corresponding rest-pose surface vertices, internal skeletal parameters, and full body pose. The surface vertices and skeletal parameters are input into their respective linear autoencoders [2], the pose is input into our sparse, non-linear pose correctives function, and the mesh is rigged with the reconstructed vertices, reconstructed skeletal parameters, pose correctives, and trainable skin weights.

We initialize autoencoders [2] using PCA of surface vertices and skeletal parameters from our multi-shape dataset. For each training iteration, we sample the number of components  $n \in [1, \max]$  and preserve only the first  $n$  features in the autoencoder latent bottleneck, zeroing out the remainder. This ordered dropout strategy maintains the component

importance hierarchy throughout optimization. We use 128 components for the shape and 16 components for the skeleton, as we found fitting error plateaued beyond these components.

ATLAS is trained by minimizing the loss:

$$\mathcal{L} = \mathcal{L}_{\text{data}} + \mathcal{L}_{\text{shape\_reg}} + \mathcal{L}_{\text{skele\_reg}} + \mathcal{L}_{\text{skin\_lapl}} + \mathcal{L}_{\text{pc\_lapl}} + \mathcal{L}_{\text{skin\_init}} + \mathcal{L}_{\text{pc\_act\_reg}}$$

where  $\mathcal{L}_{\text{data}}$  is the main data term minimizing vertex-to-vertex distance between the registration and the predicted mesh.  $\mathcal{L}_{\text{shape\_reg}}$  and  $\mathcal{L}_{\text{skele\_reg}}$  are L2 losses that regularize the intermediate latents of the surface vertex and skeleton attribute autoencoders.  $\mathcal{L}_{\text{skin\_lapl}}$  and  $\mathcal{L}_{\text{pc\_lapl}}$  regularize the skin weights and pose corrective blendshapes with a cotangent laplacian loss.  $\mathcal{L}_{\text{skin\_init}}$  regularizes the skin weights towards their artist-defined initialization through L2.  $\mathcal{L}_{\text{pc\_act\_reg}}$  imposes an L1 regularization loss on the pose corrective activation matrix, which is geodesic initialized, to encourage sparsity in vertex-joint correlations.

### E.3. Pose Prior Implementation Details

For our pose prior, we adopt a lightweight VAE architecture similar to that of SMPL-X [3]. The VAE has a 32 latent dimension, takes as input 6D continuous rotation vectors for the full body excluding hands, and is trained to reconstruct samples from our 600k multi-pose dataset. The model is trained for 40 epochs with a batch size of 512 and a learning rate of  $5e-3$ . We minimize three losses - the KL divergence loss, a reconstruction loss, and the angle difference loss between the input and output.

## F. Optimized Skin Weights

We initialize skin weights  $\Omega$  with artist-defined values and optimize them end-to-end during training. The weights before and after training are shown in Figure 13.

## G. Skeleton and Shape Latent Spaces

Our skeletal attribute definitions allow for direct controllability of individual aspects of the internal skeleton. Furthermore, for lower-dimensional keypoint fitting, scan registration, and skeleton modification, our skeleton latent space



Figure 14. **Sampled Visualizations of Our Multi-Pose Dataset.** We train ATLAS on a diverse set of 600k scans captured by a high-resolution scanner with 240 synchronized cameras.

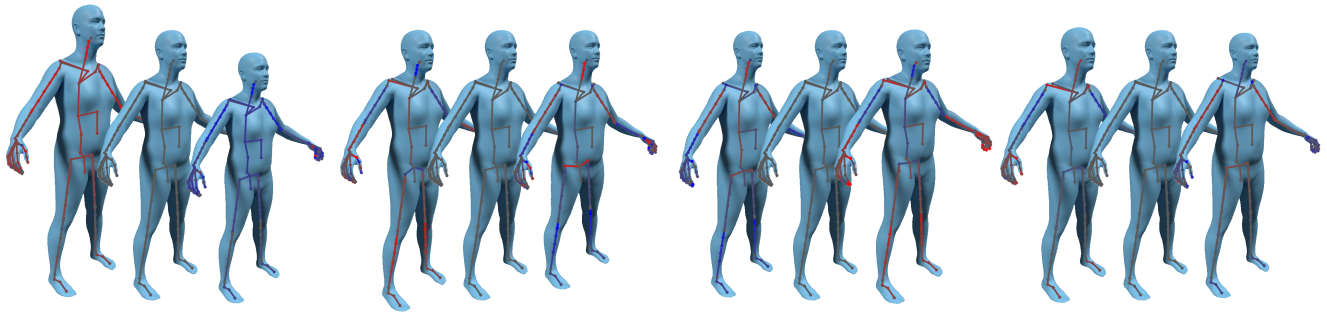


Figure 15. **Visualizations of the first four internal skeleton components.** For each component, we visualize changes in the mesh from decreasing and increasing the component. The skeleton is colored such that red indicates an increase in bone length while blue indicates a decrease. The skeletal components alone are sufficient to capture most human body variation. The first component is correlated with overall size of the subject, the second captures the neck and the hips, the third focuses on the shoulders and arms (decoupling upper and lower arm lengths), while the fourth captures length of the full arm.

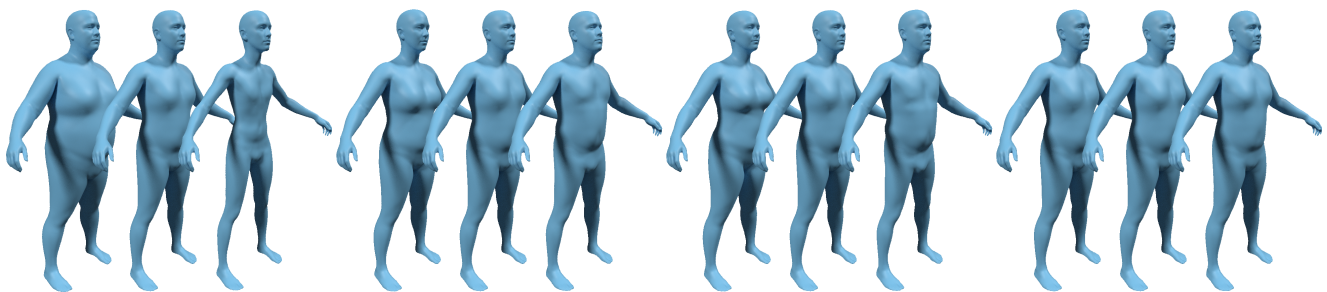


Figure 16. **Visualizations of the first four external surface components.** As most of the body variation (height, arm length, hand size, etc) are already captured by the skeleton, the surface components focus on soft tissue changes such as weight, neck width, arm thickness, and facial attributes. Note that we do not display the skeleton as it remains unchanged with variations in the surface vertices.

provides data-driven correlations between different aspects of the body. We visualize the first four components of the skeletal components in Figure 15. We find that the skeletal attributes themselves capture most of the variation in the human body, such as overall body size, shoulder width, arm length, etc. While the skeletal components focus on the internal structure of humans, the surface components, shown in Figure 16 instead focus on the external soft tissue changes. Our surface components are more subtle than the shape components of prior work, as previous methods entangle skeletal and surface attributes, forcing the same components to capture variations in both soft tissue attributes and internal skeleton.

## H. Latent Sampling of Shape, Skeleton, Pose, and Expressions

In this section, we further demonstrate the expressiveness of ATLAS by randomly sampling shaped, articulated human subjects in Figure 17. More specifically, we sample from our surface and skeletal latent spaces to model a random

identity, then sample from our pose prior and hand PCA space for full-body pose, and finally sample facial expressions. The resulting meshes are realistic, and they span a wide range of diverse human subjects in a variety of poses.

## **I. Additional Results on Mesh Prediction in the Wild**

We extend the results in Figure 11 of the main paper by providing additional results on in-the-wild images in Figure 18. Our fitting procedure complements ATLAS by yielding shape, scale, pose, and expression parameters from 2D RGB images in the wild. Of particular note is ATLAS’s ease at capturing undersized subjects such as children. By explicitly modeling the size of each skeletal part, ATLAS naturally predicts realistic shapes for children, accounting for their relatively larger heads compared to the rest of their body.





Figure 17. **Visualization of Random Latent Samples from ATLAS.** We randomly sample subject surface vertices and internal skeleton from the latent spaces, sample pose from our VAE pose prior and hand PCA space, and facial expressions from the FLAME space. ATLAS captures a wide breadth of realistic human shapes and articulates them into realistic poses.





Figure 18. **Additional Visualizations of Fitting ATLAS to Single Images.** Our fitting pipeline can capture a wide range of poses and shapes in addition to facial expressions.

## References

- [1] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. [1](#)
- [2] Andrew Ng et al. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19, 2011. [3](#)
- [3] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. [3](#)