BodyShapeGPT: SMPL Body Shape Manipulation with LLMs

Baldomero R. Árbol[0000−0002−8320−9752] and Dan Casas[0000−0002−3664−089X]

Universidad Rey Juan Carlos, Madrid, Spain

https://github.com/baldoarbol/BodyShapeGPT

Abstract. Generative AI models provide a wide range of tools capable of performing complex tasks in a fraction of the time it would take a human. Among these, Large Language Models (LLMs) stand out for their ability to generate diverse texts, from literary narratives to specialized responses in different fields of knowledge. This paper explores the use of fine-tuned LLMs to identify physical descriptions of people, and subsequently create accurate representations of avatars using the SMPL-X model by inferring shape parameters. We demonstrate that LLMs can be trained to understand and manipulate the shape space of SMPL, allowing the control of 3D human shapes through natural language. This approach promises to improve human-machine interaction and opens new avenues for customization and simulation in virtual environments.

Keywords: Large Language Model · SMPL · Fine-Tuning.

1 Introduction

The increasing integration of artificial intelligence in our daily lives challenges us to continually explore and expand its capabilities to enhance human-machine interaction. Large Language Models (LLMs) represent a revolution in how machines understand and generate human language, offering virtually unlimited possibilities in terms of content creation and response automation. This research aims to leverage the capabilities of LLMs to interpret complex descriptions and convert them into accurate visual representations, using the SMPL-X [2,19] model to generate three-dimensional avatars from textual descriptions. Current works [18,13,15] that utilize LLMs to modify SMPL parameters primarily focus on adjusting the pose rather than the shape. Additionally, while some methods [21,3] generate avatars from text, they do not provide explicit control of SMPL shape space. Our approach addresses this gap by enabling the control of SMPL shape parameters directly from textual descriptions, enabling the interactive manipulation and generation of human avatars using natural language.

2 Related Work

Avatars from Image. In the field of computer vision, generating 3D avatars from images is a widely explored research area [1,20,4,5,16]. SHAPY [6] uti-

lizes a deep learning architecture to analyze the photograph and reconstruct a detailed 3D model, capturing the nuances of human shape and appearance. POCO [12] focuses on achieving high-fidelity shape generation by leveraging convolutional neural networks to extract and translate image features into a 3D form. AG3D [10] synthesizes high-quality 3D humans from 2D image collections, capturing detailed appearances and geometries, including various clothing styles. AvatarBooth [21] introduces a novel method for generating high-quality personalized 3D avatars using casual face or body images, and it also supports generation and editing based on text. FitMe [17] presents a facial reflectance model and a differentiable rendering optimization pipeline that can capture detailed human avatars from one or multiple images, achieving photorealistic results and preserving identity. StyleAvatar3D [22] uses image and text diffusion models to generate high-quality stylized 3D avatars, while DreamAvatar [3] employs a combination of textual and shape guidance to produce controllable 3D human avatars with specific poses.

This paper differs in that it does not require a photograph as input. Instead, the only input to the network is a *textual description of the body shape* of a human subject. This allows for greater flexibility and application in fields such as storytelling and virtual character creation, where body descriptions are common and often preferred.

Animations from Text. In the realm of text-based generative networks, PoseGPT [18], MotionGPT [15] and ChatPose [13], among others [8,7], are notable for their ability to interpret text inputs and transform them into avatar poses or brief movement sequences. PoseGPT [18] leverages a transformer-based architecture to decode textual descriptions into precise body poses, which can then be used in various animation and virtual reality applications. MotionGPT [15] extends this capability by generating continuous motion sequences from text, allowing for dynamic and expressive avatar animations. ChatPose [13] further advances this field by integrating conversational context to generate more nuanced and contextually appropriate avatar movements.

Our work is based on a similar approach but with the goal of generating the avatar shape instead of the pose. By focusing on shape generation, this work complements pose generation methods [18,15,13], offering a solution for creating complete avatars from natural language.

3 Methodology

3.1 Dataset Generation

Since there is no existing dataset that relates SMPL-X shape parameters to their corresponding verbal descriptions, a new dataset has been meticulously created to train the network for this task. The dataset generation process began by taking measurements of randomly sampled avatars. These measurements were then analyzed to understand their distribution within a random set of avatars with a normal distribution of body shapes.

In addition to these measurements, a custom labeling algorithm was developed to assign detailed verbal descriptions to each avatar. Our strategy took into account various aspects of the avatar physical attributes, ensuring that each verbal description was comprehensive and precise. For example, an avatar with a certain shoulder width and waist size might be described as having a "broad-shouldered and slim-waisted" appearance. By carefully correlating these measurements with descriptive language, the labeling process generated a significant volume of data.

After the initial creation of the dataset descriptions, an additional layer of variability was introduced by rephrasing the text using Llama-3 [11] to rewrite the same information with different wording and attribute order, enhancing the richness of the dataset.

The combination of detailed measurement analysis, precise verbal labeling, and varied rephrasing ensures that the neural network can learn the complex relationships between physical shapes and their descriptive counterparts, leading to more accurate and realistic avatar generation based on text input. Figure 1 shows two samples of our dataset.

```
{"description": "Person with an average height, tall
    neck, long arms, and broad shoulders.",
"shape_params": "[1.131, 1.928, -2.347, -0.793, 0.251, 0
    .58, 1.707, -2.888, -1.904, 2.772]"}
{"description": "Medium-built individual with a slender
    profile, featuring a mid-range stature, a petite
    torso, and limbs of moderate length.",
"shape_params": "[-1.016, -0.504, 0.948, 1.092, -0.514,
    -1.941, 0.415, 2.089, 0.509, 1.626]"}
```

Fig. 1: Example of dataset entries that relate SMPL-X shape parameters to their corresponding verbal descriptions

3.2   Fine-Tuning Process

Our model is trained by fine-tuning the popular Llama-3 [11] model to the specific task of interpreting verbal descriptions of avatars and generating the corresponding SMPL-X parameters. We use Low-Rank Adaptation (LoRA) [14] and quantization [9] to optimize the model for efficient training and execution on NVidia RTX-4090. The fine-tuning was carried out starting from LLaMA-3 8B, using the generated dataset consisting of 18,000 training cases and 2,000 evaluation cases. To train our model, we propose to extend the regular Cross Entropy loss used to train LLMs. Our key intuition is that the output of the model is not arbitrary text, but a string of characters that must encode a vector of 10 float values (i.e., the SMPL shape parameters). This observation allows us to formulate two new shape-specific loss terms that enable the model to focus

on shape parameters rather than arbitrary text. Our loss term is defined as

$$\mathcal{L} = \mathcal{L}_{\text{LLM}} + \mathcal{L}_{\text{shape}} + \mathcal{L}_{\text{measurements}} \tag{1}$$

where $\mathcal{L}_{\text{LLM}} = \text{CE}(\hat{Y}_t, Y_t)$ is the standard cross-entropy loss of tokens used to train LLMs, which ensures that the output is a correct string of digits. The term $\mathcal{L}_{\text{shape}} = |\hat{\beta} - \beta|$ is the L1 difference between shape parameters, weighting each beta according to its impact on the human parametric model. Finally, $\mathcal{L}_{\text{measurements}} = \text{CE}(f_{\text{SMPL}}(\hat{Y}_t), f_{\text{SMPL}}(Y_t))$, where f_{SMPL} is the SMPL function that evaluates a body mesh given Y shape parameters, is the cross-entropy loss of a set of body measurements (e.g., height, arm length, volume). This term is key to guarantee that the generated avatar falls within the expected range.

4   Results

4.1   Quantitative Evaluation

To validate our loss function, we carry out an ablation study for each of the terms. To this end, we evaluate the following models

- $\mathcal{L} = \mathcal{L}_{\text{LLM}}$: only considers the cross-entropy error of the generated text tokens.
- $\mathcal{L} = \mathcal{L}_{\text{LLM}} + \mathcal{L}_{\text{shape}}$: takes into account both the cross-entropy error of the generated text tokens and the error the shape parameters (i.e. β parameter of SMPL).
- $\mathcal{L} = \mathcal{L}_{\text{LLM}} + \mathcal{L}_{\text{shape}} + \mathcal{L}_{\text{measurements}}$: our full model, it considers the cross-entroy of the text tokens, the error in shape parameters, and and also adds the cross-entropy of a set of body measurements, including limb lengths, BMI, and height.

As we show in Figure 2, our proposed loss achieves lower error than other baselines. The error depicted in the graph represents the difference between the ranges in which the avatars are labeled and the desired target ranges. A lower error indicates that the model predictions are more accurate in matching the desired labeling ranges, leading to more precise and consistent avatar descriptions.

The significant reduction in the cross-entropy loss value for the various measurements represents a clear improvement in the model's results regarding the accuracy of body measurements within the desired ranges. Studying its reduction throughout the training process helps identify the impact of the applied improvements.

This accuracy can also be observed in the plots from Figure 3, which evaluates the accuracy of the distributions of our predictions. It is worth noting that the accuracy of baseline model, which considers fewer factors when calculating error, is much lower, tending to generate avatars with measurements very close to the average and making errors typical of only considering text tokens as the value to account for, such as representing "very low" values as "very high." Our model, on the other hand, has a distribution where the values fit much better with the ground truth and align with the descriptions within the expected range.
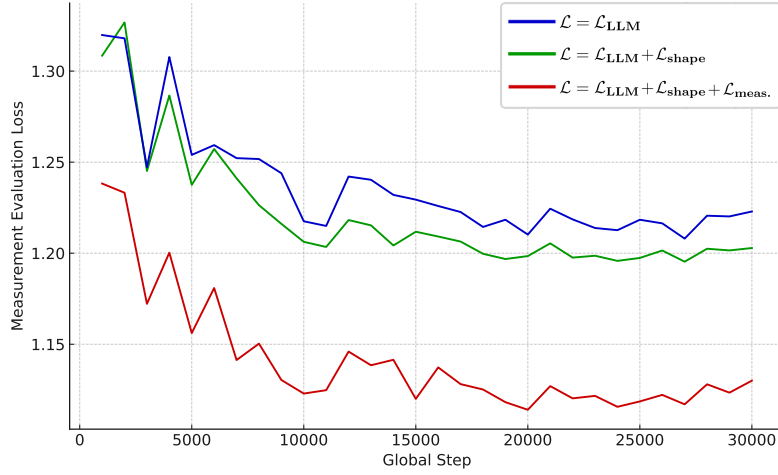
Fig. 2: Cross Entropy Loss of the validation set for each training iteration.

(a) $\mathcal{L} = \mathcal{L}_{\mathrm{LLM}}$

(b) $\mathcal{L} = \mathcal{L}_{\mathrm{LLM}} + \mathcal{L}_{\mathrm{shape}} + \mathcal{L}_{\mathrm{meas.}}$
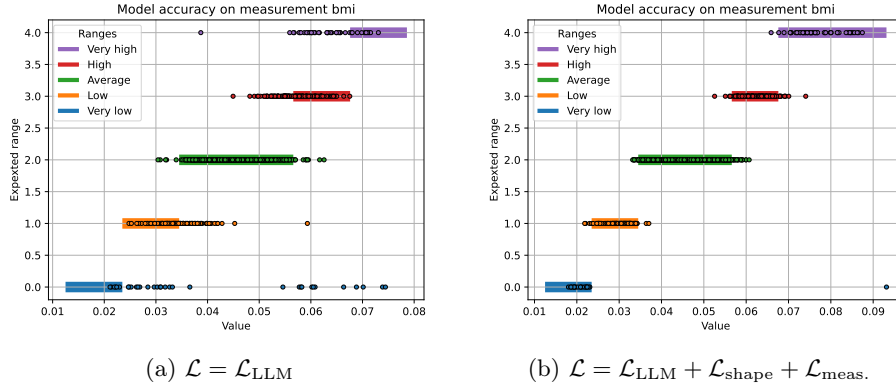
Fig. 3: Accuracy of baseline model (a) and our model (b), evaluated using the Body Mass Index (BMI) computed from the regressed avatars. The ground truth range for each category is shown as horizontal solid boxes. Using our model, most of test samples (i.e., color dots) fall into the correct range.

Table 1 depicts the numerical accuracy of each of the models. Our model achieves the highest accuracy on all the measurements that we evaluate.

Table 1: Measurements accuracy using different models.

	$\mathcal{L}_{\text{LLM}}$	$\mathcal{L}_{\text{LLM}} + \mathcal{L}_{\text{shape}}$	$\mathcal{L}_{\text{LLM}} + \mathcal{L}_{\text{shape}} + \mathcal{L}_{\text{meas.}}$
height_average	99,7%	99,1%	**99,9%**
hip_thickness_average	95,4%	96,1%	**99,1%**
hip_thickness_average	94,8%	96,0%	**98,8%**
bmi_average	94,5%	95,8%	**98,4%**
waist_thickness_average	94,5%	94,9%	**98,0%**
. . .			
neck_length_very_high	33,3%	50,0%	**82,6%**
arms_relation_very_high	30,4%	41,6%	**81,0%**
leg_thickness_very_low	24,0%	41,6%	**79,3%**
shoulders_relation_very_low	21,7%	41,1%	**78,1%**
bmi_very_low	18,7%	40,9%	**76,7%**

4.2   Qualitative Evaluation

Figure 4 showcases some of the avatars generated with the proposed model. We demonstrate that our model is robust, and has the capability to generate accurate 3D avatars from diverse natural language descriptions, validating the effectiveness of the fine-tuning process and the customized dataset.

5   Conclusions

The paper successfully presented a neural network capable of generating 3D avatars from natural language descriptions. This was achieved through the meticulous creation of a comprehensive dataset, which involved taking measurements of randomly generated avatars, analyzing their distribution, and assigning detailed verbal descriptions. The training process employed fine-tuning techniques such as LoRA and quantization, enabling efficient model training within the constraints of available computational resources. The fine-tuning of large language models (LLMs) specifically for avatar generation has proven highly effective, achieving elevated levels of precision. The ability to generate an avatar from a textual description in just a few seconds opens up numerous practical applications, particularly in fields like storytelling, where dynamic and personalized character creation is highly valued.

6   Future Work

Future work could explore further optimization of the network, aiming to enhance performance and accuracy. This includes expanding the dataset to encompass more diverse descriptions, thus improving the model's ability to handle

"A tall person with very long legs."

"A tall person with very short legs."

"Short, pearl-shaped person."

"Towering, muscular figure."

"Big shoulders but small hips person."

"A petite individual that has long arms."

Fig. 4: Generated avatar with different prompts

a wider range of inputs. Additionally, integrating pose generation capabilities, potentially in combination with models like PoseGPT or MotionGPT, would enable the creation of complete avatars with both animation and form from a single phrase. This would significantly enhance the dynamism and expressiveness of the generated avatars.

Another important avenue for future research is to train the network to recognize and appropriately respond to descriptions indicating the subject's gender, whether masculine, feminine, or neutral. This involves teaching the network to understand the differences in SMPL-X shape parameters for different genders, ensuring that the generated avatars accurately reflect these distinctions. By addressing these areas, future developments can create more sophisticated and versatile avatar generation systems, broadening their applicability and improving their accuracy and realism.

References

1. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Detailed Human Avatars from Monocular Video. In: International Conference on 3D Vision (3DV). pp. 98–109 (2018)
2. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In: European Conference on Computer Vision (ECCV) (2016)
3. Cao, Y., Cao, Y.P., Han, K., Shan, Y., Wong, K.Y.K.: DreamAvatar: Text-and-Shape Guided 3D Human Avatar Generation via Diffusion Models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 958–968 (2024)
4. Casas, D., Comino-Trinidad, M.: SMPLitex: A Generative Model and Dataset for 3D Human Texture Estimation from Single Image. In: British Machine Vision Conference (BMVC) (2023)
5. Casas, D., Volino, M., Collomosse, J., Hilton, A.: 4D Video Textures for Interactive Character Appearance. Computer Graphics Forum (Proceedings of Eurographics) **33**(2), 371–380 (2014)
6. Choutas, V., Müller, L., Huang, C.H.P., Tang, S., Tzionas, D., Black, M.J.: Accurate 3D Body Shape Regression using Metric and Semantic Attributes. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2022)
7. Dabral, R., Mughal, M.H., Golyanik, V., Theobalt, C.: MoFusion: A Framework for Denoising-Diffusion-based Motion Synthesis. In: Computer Vision and Pattern Recognition (CVPR) (2023)
8. Delmas, Ginger and Weinzaepfel, Philippe and Lucas, Thomas and Moreno-Noguer, Francesc and Rogez, Grégory: PoseScript: 3D Human Poses from Natural Language. In: ECCV (2022)
9. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: QLoRA: Efficient Fine-tuning of Quantized LLMs (2023), `https://arxiv.org/abs/2305.14314`
10. Dong, Z., Xu Chen, J.Y., J.Black, M., Hilliges, O., Geiger, A.: AG3D: Learning to Generate 3D Avatars from 2D Image Collections. In: International Conference on Computer Vision (ICCV) (2023)
11. Dubey, A., Jauhri, A., Pandey, A., et al., A.K.: The Llama 3 Herd of Models (2024), `https://arxiv.org/abs/2407.21783`
12. Dwivedi, S.K., Schmid, C., Yi, H., Black, M.J., Tzionas, D.: POCO: 3D Pose and Shape Estimation using Confidence. In: International Conference on 3D Vision (3DV) (2024)
13. Feng, Y., Lin, J., Dwivedi, S.K., Sun, Y., Patel, P., Black, M.J.: ChatPose: Chatting about 3D Human Pose. In: CVPR (2024)
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models. In: International Conference on Learning Representations (ICML) (2022), `https://openreview.net/forum?id=nZeVKeeFYf9`
15. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: MotionGPT: Human Motion as a Foreign Language. Advances in Neural Information Processing Systems (NeurIPS) **36** (2024)
16. Jiang, T., Chen, X., Song, J., Hilliges, O.: InstantAvatar: Learning Avatars from Monocular Video in 60 Seconds (June 2023)
17. Lattas, A., Moschoglou, S., Ploumpis, S., Gecer, B., Deng, J., Zafeiriou, S.: FitMe: Deep Photorealistic 3D Morphable Model Avatars. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8629–8640 (2023)

18. Lucas, T., Baradel, F., Weinzaepfel, P., Rogez, G.: PoseGPT: Quantization-based 3D Human Motion Generation and Forecasting (2022), https://arxiv.org/abs/2210.10542

19. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 10975–10985 (2019)

20. Yang, X., Chen, X., Gao, D., Wang, S., Han, X., Wang, B.: HAVE-FUN: Human Avatar Reconstruction from Few-Shot Unconstrained Images (2024), https://arxiv.org/abs/2311.15672

21. Zeng, Y., Lu, Y., Ji, X., Yao, Y., Zhu, H., Cao, X.: AvatarBooth: High-Quality and Customizable 3D Human Avatar Generation (2023), https://arxiv.org/abs/2306.09864

22. Zhang, C., Chen, Y., Fu, Y., Zhou, Z., YU, G., Wang, B., Fu, B., Chen, T., Lin, G., Shen, C.: Styleavatar3d: Leveraging image-text diffusion models for high-fidelity 3d avatar generation (2023)