Human Motion Generation
CSE 527 : Introduction to Computer Vision

Rishabh Tamhane
115939004

1 Objective

To develop an innovative approach for generating human body motion within the RGB space, by integrating the capabilities of 3D Gaussian Splatting (3DGS) with the advanced generative properties of a diffusion model. The project seeks to construct a comprehensive 3DGS framework tailored for human figures, enabling the synthesis of realistic human motions derived from either text-guided or audio-guided diffusion models.

2 Introduction

3DGS-Avatar [3] is a technique designed to create animatable clothed human avatars. It achieves this by leveraging millions of 3D Gaussians to represent the 3D structure and appearance of the clothed human body. These Gaussians encode information like color, density, and deformation, allowing for detailed and realistic rendering of the avatar during animation.
This approach offers significant advantages as it enables the creation of high-quality animations with short training times and real-time rendering capabilities.

In the human motion generation domain, many works have been proposed to learn a diffusion model that generate text-guided human motion, like MAS [1], MDM [5], PriorMDM [4], or music-guided human motion, like EDGE[6].

The **Motion Diffused Model (MDM)** and **Prior MDM** are diffusion models primarily utilized for their capability to directly translate textual descriptions, such as 'walk' or 'run', into realistic animations, demonstrating their proficiency in High-Quality Text-to-Motion conversion. In addition to this, they also excel in the Action-to-Motion Refinement domain, where they enhance existing animations (e.g., walking styles) by creating nuanced variations, thereby showcasing their versatile application in generating and refining motion sequences based on textual or basic action inputs.

The projects aims to build a :

- **Connector**: Connect the output of Prior MDM to the input prediction sequence of 3DGS-Avatar.
- **Fine Tuned Model**: Exploring Test Time Training (TTT) Methods to finetune the output of the predicted sequence.

Phase	Start Date	End Date
Preliminary Study of 3DGS-Avatar	12-Apr	16-Apr
Generation of Motion Sequences via Diffusion Models	17-Apr	24-Apr
Mid Term Project Report and Presentation		24-Apr
Integration	25-Apr	5-May
Optimization and Finetuning	6-May	13-May
Final Project Report and Presentation		15-May

Table 1: Project Timeline

3 Baseline Study and Application of the 3DGS-Avatar Technique

In this preliminary study, we comprehensively reviewed and executed the 3DGS-Avatar algorithm to establish a foundational understanding of the 3D Gaussian Splatting methodology applied to human body models. This involved selecting and analyzing data from two to three subjects in the ZJU-MoCap dataset[2].

Through this stage, we aimed to gain baseline knowledge regarding the input format of SMPL parameters and the loss functions employed for estimating the 3D RGB pose. Specifically, the following loss functions were examined and utilized:

- **RGB Loss** (L_{rgb}): Measures the difference between the predicted and actual RGB values.
- **Mask Loss** (L_{mask}): Measures the difference between the predicted and actual object masks.
- **Skinning Weight Regularization Loss** (L_{skin}): Regularizes the skinning weights.
- **As-Isometric-As-Possible Regularization Loss for Position** (L_{isopos}): Ensures the distances between neighboring 3D Gaussian centers remain similar after deformation.
- **As-Isometric-As-Possible Regularization Loss for Covariance** (L_{isocov}): Ensures the covariance matrices of neighboring 3D Gaussians maintain similar distances after deformation.

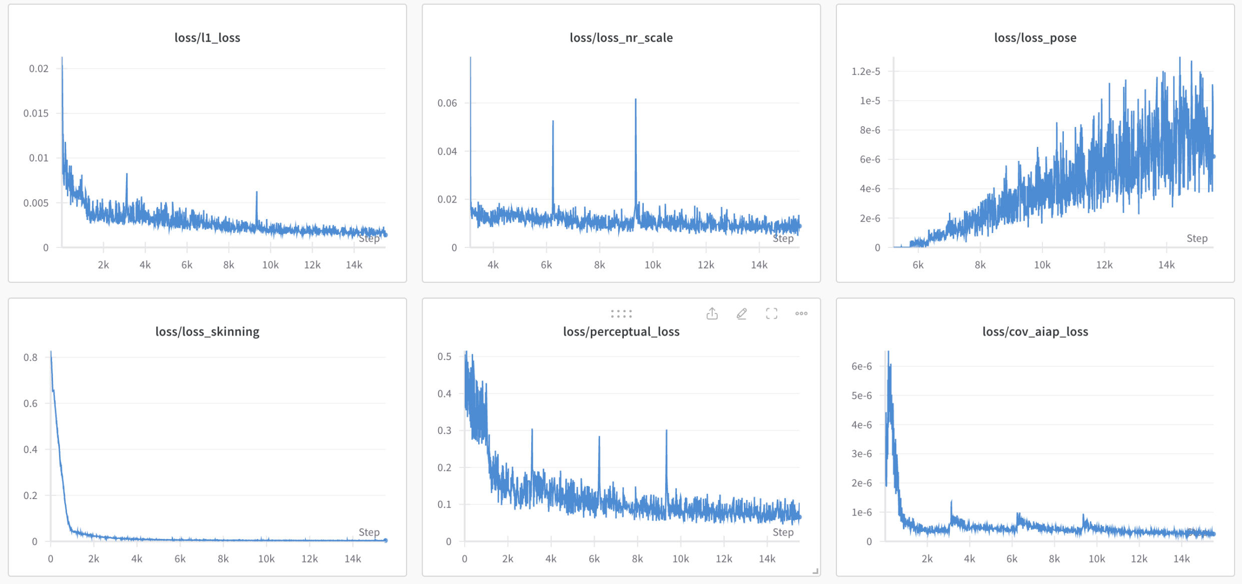


Figure 1: **Training Loss for Subject 392**

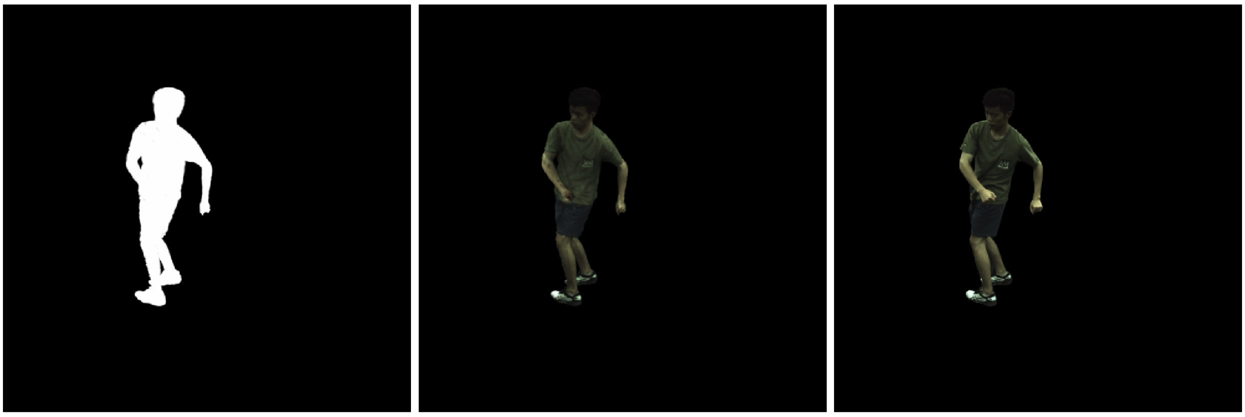

train_view_c01_f000275/render_opacity train_view_c01_f000275/render train_view_c01_f000275/ground_truth

Figure 2: **Training Instance for Subject 392**

4 Integration of Prior MDM and 3DG-Avatar

The output of the Prior MDM were SMPL Parameters in the 6D space. While, the expected input of the 3DGS-Avatar was in the 3D space. By using quarternions and accounting for the following points, we were able to establish the converter.

- **Normalization Check**: Added checks to ensure the norm of **u** and **v** is not zero before normalization.
- **Numerical Stability**: Clipped the trace_val to ensure it is within the valid range for np.arccos.
- **Small Angle Handling**: Added a check to set a default axis if the angle θ is too small or if $\sin(\theta)$ is zero to prevent division by zero.

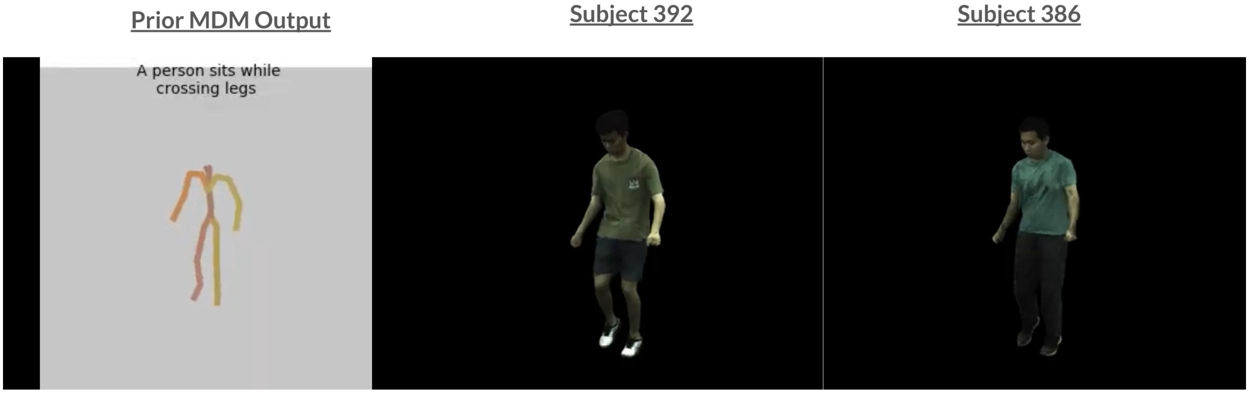


Figure 3: **Integration of Prior MDM and 3DGS-Avatar**

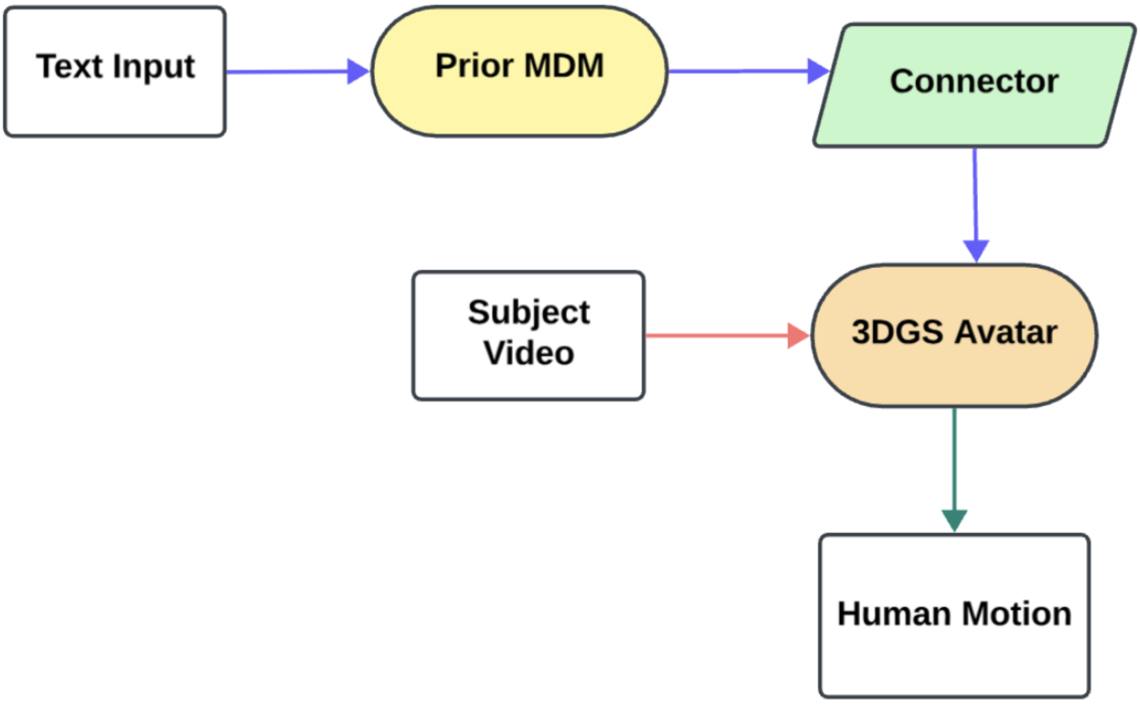


Figure 4: **Project Pipeline**

The preliminary results were good but they did not generalise very well on out of distribution poses.

5 Fine Tuning 3DGS-Avatar

5.1 Predicted Data Retraining

The current model does not perform well on out-of-distribution poses. Due to the lack of ground truth data, we checkpoint the model during the initial training phase. We then use the predicted outputs as pseudo-ground truth, in conjunction with the initial training data, to continue training the model. However,

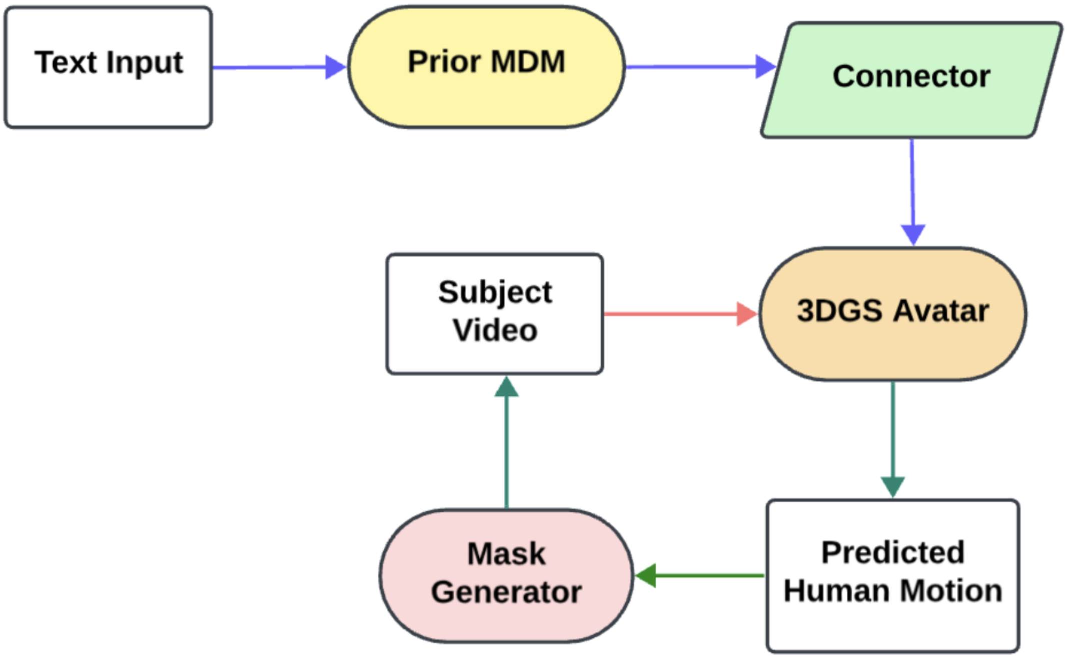


Figure 5: **Predicted Data Retraining Pipeline**

the inaccuracies of the mask generator, compared to actual annotated data, hinder the model's ability to generalize effectively to out-of-distribution poses.

The finetuned model does not perform as good as the actual model and tends to overfit on the actual training data and picks up on the inaccuracies of the predicted data used for training.

5.2 Test Time Training with Self Supervision

Since we do not have access to the ground truth for test data and the pseudo-ground truth may not be reliable for training, it is crucial to enhance the model's ability to generalize to out-of-distribution data. By leveraging TTT, we aim to improve the model's performance in scenarios where the test data differs significantly from the training data. Here's how we apply TTT to our image classification and human motion prediction model:

5.2.1 Training Phase

- **Main Task**: Our primary task is to predict human motion from images.
- **Auxiliary Task**: Additionally, we train the model to predict the orientation of an image. We rotate images by 0, 90, 180, or 270 degrees and ask the model to recognize the correct rotation.
- **Model Training**: During training, the model learns to perform both the main task (human motion prediction) and the auxiliary task (orientation prediction). This helps the model learn better feature representations of the data, enhancing its robustness to distribution shifts.

5.2.2 Test-Time Adaptation

- **Encountering New Data**: When the model encounters new test data, which might be different from the training data (e.g., due to variations in human poses, lighting conditions, or backgrounds), it adapts using the auxiliary task.
- **Self-Supervised Updates**: For each batch of test images, we first rotate them and ask the model to predict these rotations. The model adjusts its parameters based on this auxiliary task, which helps it understand the new test data better.
- **Making Predictions**: After these adjustments, the model makes predictions on the main task (human motion prediction) with updated parameters, leading to improved performance.

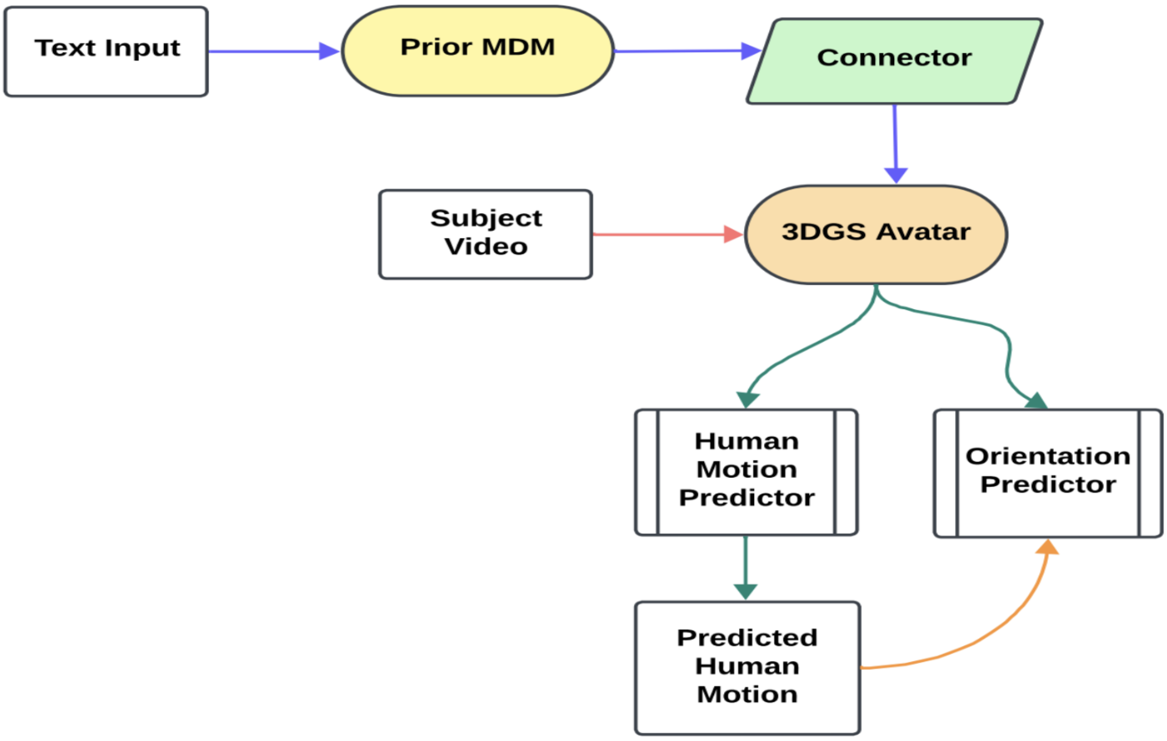


Figure 6: **Test Time Training with Self Supervision**

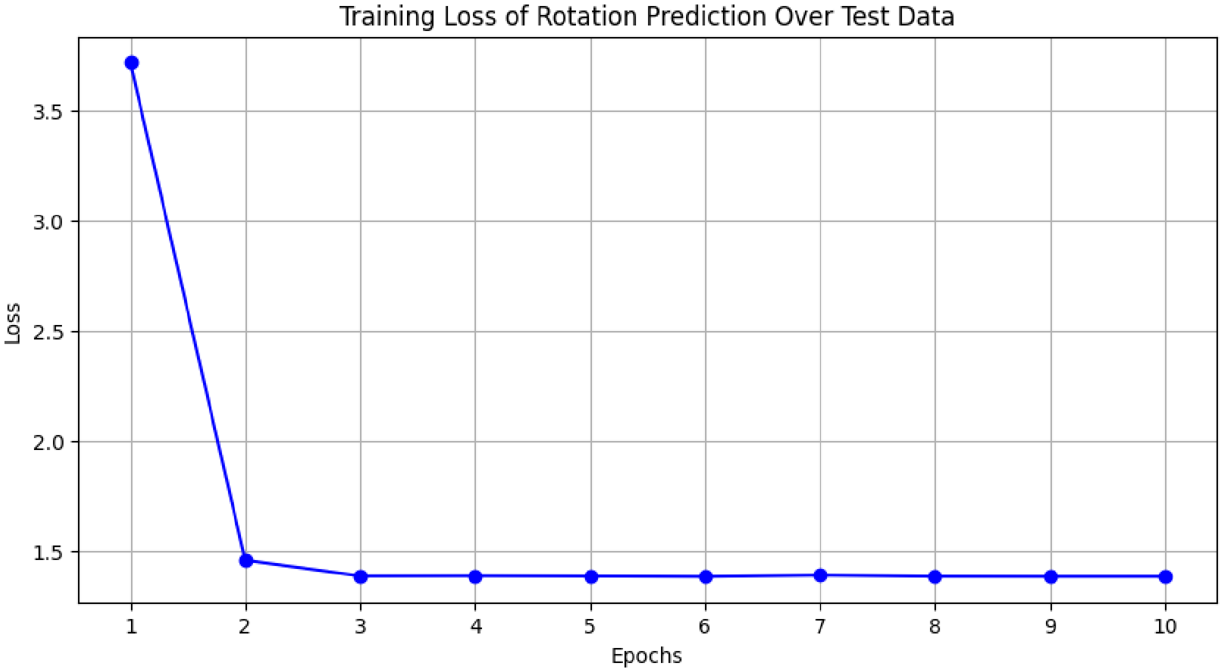
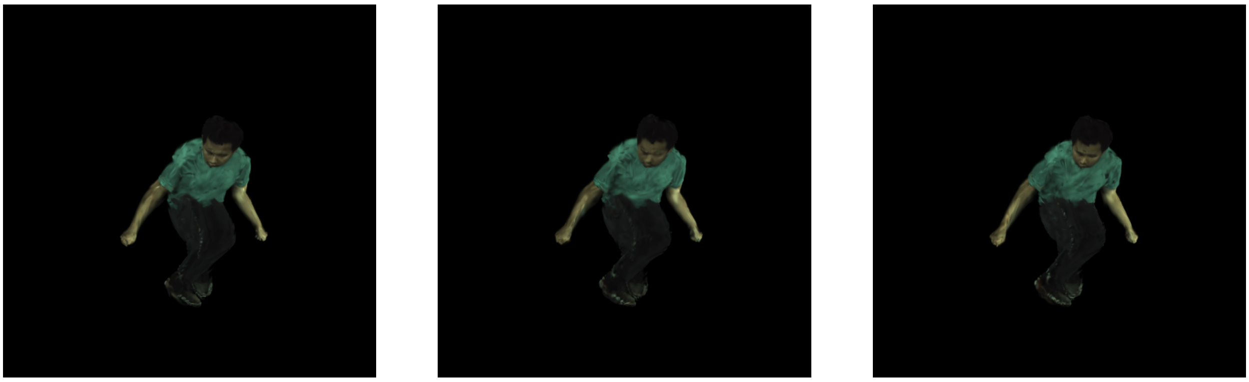


Figure 7: **Rotation Prediction Loss Curve during Test Time Training**

5.2.3 Results

We observe that there is not much difference in the 3 models. However, on closer observation, the model fine-tuned with the predicted data has more irregular facial features than the other 2 models.

On training the Rotation Predictor on test data we see a big drop in the loss from the 1st epoch to the 2nd epoch indicating that there is a large amount of data that is yet to be seen by the model. However, there is not much difference in the predicted images and the predicted images without fine-tuning.



(a) Connected Model without Fine-tuning

(b) Model Fine-tuned with Pseudo Ground Truth

(c) Test Time Training with Self Supervision

Figure 8: Qualitive Comparison of Models

5.2.4 Future Work

The current model is optimized for a relatively small task of rotation prediction. It is observed that the cross-entropy loss of the rotation predictor is significantly lower, almost 30 times, compared to the total loss. We should train on a bigger task with more similarity with the actual task such that the ground truth reality of which is easily available.

The discrepancy is primarily attributed to the black color of the hands and the body. To address this issue, future work should focus on integrating the rotation prediction task as a component of the network that measures the RGB loss. By incorporating the rotation orientation predictor within the RGB prediction framework, any changes in the rotation orientation would directly influence the RGB prediction, potentially leading to more accurate and robust results.

References

[1] Roy Kapon, Guy Tevet, Daniel Cohen-Or, and Amit H. Bermano. Mas: Multi-view ancestral sampling for 3d motion generation using 2d diffusion, 2023.

[2] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021.

[3] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. 2024.

[4] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023.

[5] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023.

[6] Jonathan Tseng, Rodrigo Castellon, and C. Karen Liu. Edge: Editable dance generation from music, 2022.