# JADE: Joint-aware Latent Diffusion for 3D Human Generative Modeling

Haorui Ji  Rong Wang  Tao Jun Lin  Hongdong Li
The Australian National University
{haorui.ji, rong.wang, taojun.lin, hongdong.li}@anu.edu.au

## Abstract

*Generative modeling of 3D human bodies have been studied extensively in computer vision. The core is to design a compact latent representation that is both expressive and semantically interpretable, yet existing approaches struggle to achieve both requirements. In this work, we introduce JADE, a generative framework that learns the variations of human shapes with fined-grained control. Our key insight is a joint-aware latent representation that decomposes human bodies into skeleton structures, modeled by joint positions, and local surface geometries, characterized by features attached to each joint. This disentangled latent space design enables geometric and semantic interpretation, facilitating users with flexible controllability. To generate coherent and plausible human shapes under our proposed decomposition, we also present a cascaded pipeline where two diffusions are employed to model the distribution of skeleton structures and local surface geometries respectively. Extensive experiments are conducted on public datasets, where we demonstrate the effectiveness of JADE framework in multiple tasks in terms of autoencoding reconstruction accuracy, editing controllability and generation quality compared with existing methods.*

## 1. Introduction

Generative modeling of 3D human bodies is key to many practical applications, such as multimedia, healthcare, virtual and augmented reality [6, 18, 44]. To better facilitates these applications, the human model should achieve two main goals: (i) it can accurately encode body shape details to generate high-quality samples without artifacts (ii) it enables join-level fine-grained manipulation, while producing realistic and plausible pose-dependent deformations. However, existing works mostly fail to meet both requirements. [2, 24, 33] adopt a statistical model with a linear parametric space to encode body shapes, which is not sufficiently expressive thus resulting in degraded. While recent learning-based methods [17, 49, 54] improves reconstruction quality, their latent encodings are not disentangled, making them hard to flexibly edit the human bodies.

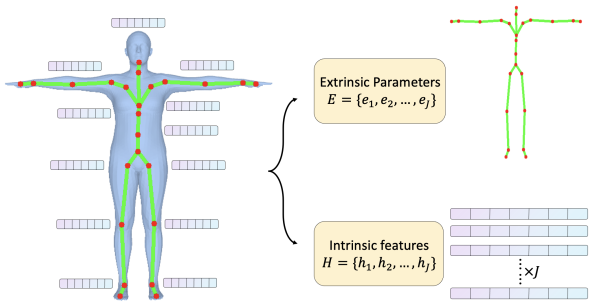



Figure 1. Overview of our joint-aware latent representation. We dispatch modeling of the overall human body into a sequence of *joint tokens*, where each token contains extrinsic parameters for encoding skeleton structures, and intrinsic features for modeling local surface geometries.

In this work, we introduce **J**oint-aware L**A**tent **D**iffusion for 3D Human G**E**nerative Modeling (**JADE**), a new method that satisfy both the expressiveness and controllability requirements in a principled way by building a diffusion-based generative model that operates on human body surface point cloud. Firstly, unlike previous works that often represent human bodies as a holistic latent code [24, 33], we introduce a joint-aware latent representation by dispatching body encoding into individual joints, so that the overall human model can be described as a sequence of *joint tokens*, akin to the tokenization process in natural language processing [53]. Each joint token are further split into extrinsic parameters and intrinsic features, where the former encodes body kinematic configurations and pose-dependent deformations, and the latter encodes canonical body shapes through local surface geometries, as shown in Figure 1. This disentangled joint-level representation naturally allows local sampling and editing human parts associated to each joint, thus are more flexible and controllable. Furthermore, to capture joint-wise correlation, we utilize a Transformer-based architecture [8] to fuse features across joints via attention, which enhances model expressiveness and ensures consistency between articulated joints.

We also develop a cascaded diffusion pipeline that allows for better modeling the latent distribution under our proposed factorization so that we can generate plausible human shapes through independently sampled tokens. Leveraging the fact that joint-wise tokens can be factorized into extrinsic-intrinsic pairs, our pipeline learns two diffusions successively, one generating extrinsic parameters to explicitly express any spatial and structural information, while the other producing the intrinsic features conditioned on the extrinsics to supplement local geometric details and improves the generation quality. This design allows each generated surface point to be informed of both global skeletal information as well as individual local structure style, enriching the synthesized geometry with useful structural information and allowing for fine-grained manipulations.

To summarize, our main contributions are as follows:

- We propose a joint-aware latent representation for generative 3D human modeling that encode both the skeleton information as well as local structure geometric details.
- We propose a two-phase cascaded diffusion pipeline to effectively model the distribution of disentangled latent space and generate plausible human shapes through independently sampled tokens.
- We showcase that our method not only allows for generating high-fidelity and diverse human shapes, but also enables multiple shape editing and variation applications, demonstrating its effectiveness.

## 2. Related Work

### 2.1. Human Body Modeling

Human body modeling refers to learning a statistical model by exploiting extensive collection of 3D body scans with or without hand articulation and facial expression. Classical parametric models like SMPL, SCAPE, etc. [2, 24, 33], factor full body deformations into identity-dependent and pose-dependent components, accompanied with a skinning algorithm that deforms surface vertices as a function of underlying skeleton change. Despite success in various applications, they still possess inherent drawbacks. For example, their expressivity is limited by the linearity of PCA subspace and struggle to represent highly nonlinear soft-tissue deformations. Moreover, fully separation of shape and pose parameters may not capture the interdependencies between body shape changes and specific poses, leading to less accurate representations in dynamic scenarios.

Recent works [5, 17, 32, 45, 49, 54, 56] have started to adopt deep learning-based approaches on polygonal meshes of human bodies. By constructing different convolution-like operators for feature extraction on irregular meshes, they model articulations as vertex offsets field warped from template mesh in canonical space to deformed space. However, most of them still use single latent code to represent global shape variations. Not only does it tend to capture spurious long-range correlations and produce non-local deformation artifacts, but also makes it hard to conduct intuitive editing since the semantics of the latent is vague. Among the few who construct structure-level latents [17, 45], a common practice is to partition human bodies into anatomical parts, assign each surface vertex a hard segmentation label, and model the deformation of each vertex as dependent on the movement of corresponding part. We argue that this strategy tends to produce artifacts especially around joint positions where different parts connected with each other since vertex movement should have interdependencies with both local structures as well as full body articulation.

Our work combine the advantages of both parametric models and learning-based ones in a way by representing a human body as distribution of points on its surface. Such a formulation allows us to bypass the ambiguities in mesh topology modeling since the basic structure of human body remains consistent. Meanwhile, our factorized latent representation facilitates structure-level control, outperforming previous works that only generate a full shape with single latent code and addressing the limitations of part-based representations.

### 2.2. Denoising Diffusion Probabilistic Model

Denoising Diffusion Probabilistic Models (DDPMs) are a kind of generative models that learn the distribution of data samples in a given dataset through a sequence of forward and reverse processes [12, 41]. The forward process is defined as progressively injecting noise to the input with a variance schedule $\beta_1, \beta_2, \cdots, \beta_T$ until reaching an isotropic Gaussian distribution after sufficiently large $T$ steps:

$$\begin{aligned} q(\mathbf{x}_t|\mathbf{x}_{t-1}) &= \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \\ q(\mathbf{x}_{1:T}|\mathbf{x}_0) &= \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}). \end{aligned} \tag{1}$$

The reverse process learns a parameterized transition kernel $\theta$ that inverts the forward diffusion process. In this way, we can synthesize novel data that follow certain distributions by initializing from random noise and sampling from the kernel from $t = T$ back to $t = 0$ in an iterative fashion.

$$\begin{aligned} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) &= \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \mathbf{\Sigma}_\theta(\mathbf{x}_t, t)), \\ p_\theta(\mathbf{x}_{0:T}) &= p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \end{aligned} \tag{2}$$

Diffusion model's capability of modeling complex data distribution enables it to achieve remarkable performances in various research areas like 2D vision [7, 13, 34, 37, 41, 43], natural language processing [9, 10, 22]. Recently, the

applications are extended to 3D vision tasks [14, 23, 36, 47], revealing its potential in generative modeling.

Most relevant to ours are the works applying latent diffusion to 3D shape generation task, including LION [52], SLIDE [28], DiffFacto [30] and others [15, 19, 27]. Their success have brought us two key observations which highly inspire our work: (i) point clouds are naturally more suitable for diffusion models than other representations like meshes or voxels, because modeling the distribution of vertex positions is easier and more straightforward than modeling inherent interdependencies; (ii) to facilitate flexible manipulations, careful design of structure-aware latent representations, like segmented parts [11, 21, 30, 31] and hierarchies [28, 52], are needed. However, most existing works cannot be effectively applied to 3D human modeling since they primarily focus on the modeling of static rigid objects like cars, airplanes, chairs, etc, while human body is non-rigid deformable, whose characteristics make previous latent representations inapplicable. This motivates us to take advantage of the traditional parameterized human representations and design a joint-aware latent diffusion pipeline suitable for modeling the distribution of shape variations. To our best knowledge, we are the first to explore related concepts for 3D human modeling and also achieve good results.

# 3. Methods

## 3.1. Overview

Given the set of surface point clouds $\mathbf{X} \in \mathbb{R}^{N\times 3}$ of human meshes that consist of $N$ points with consistent connectivity, our goal is to learn a distribution $p(\mathbf{X})$ from which we can generate high-fidelity human bodies with fine-grained-level control. In the following sections, we first introduce the joint-aware latent representation design, which is core to JADE. Training is then performed in two stages - first, we implement a Transformer-based autoencoder to conduct representation learning through self-supervised setting; next, we build a cascaded diffusion pipeline on the factorized latent encodings to model their precise distributions and improve the generation quality.

## 3.2. Joint-aware Latent Representation

As raw point clouds are redundant to store and difficult to manipulate directly, we would like to encode them to a more compact latent representation that captures both low-level geometry and high-level semantic information. The primary objective that guide our design is that the latents should be structure-aware, i.e. each sub-structure of human body is encoded independently while their aggregation can still form a globally coherent shape.

To this end, we introduce a factorized representation of human shapes by partitioning them into $J$ different joints, as shown in Figure 1, whose number agrees with the cardinality according to anatomical splits, and utilize joint-wise embeddings as our learning goal. The intuition behind this comes from the concept of skinning in parametric human modeling [20, 24] which states that vertex positions on human bodies can be represented by a blend of multiple transformations associated with nearby joints. Therefore, we encode 3D human bodies into a set of joint-wise tokens $\mathbf{Z} = \{\mathbf{z}_i \in \mathbb{R}^{D_z}\}_{i=1}^{J}$, which can be further decomposed into two components. The first one is the skeleton structure $\mathbf{E} = \{\mathbf{e}_i \in \mathbb{R}^3\}_{i=1}^{J}$, referred to as extrinsics, characterized by joints positions and the second one are features attached to each joint $\mathbf{H} = \{\mathbf{h}_i \in \mathbb{R}^{D_h}\}_{i=1}^{J}$, identified as intrinsics, that model the interdependencies between joints and surface vertices to supplement detailed local surface geometry information. Our proposed shape factorization is then given as

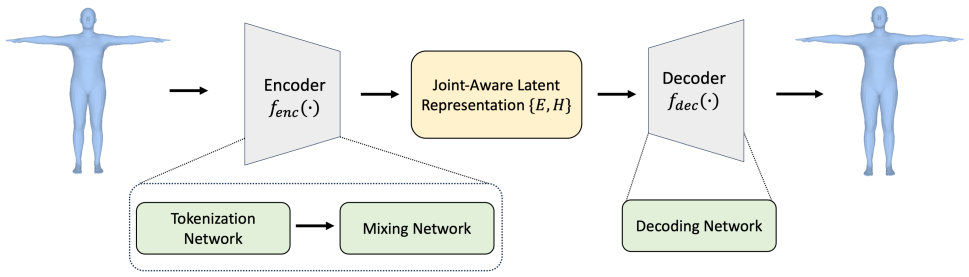
$$p(\mathbf{X}) = \prod_{i=1}^{J} p(\mathbf{z}_i) = \prod_{i=1}^{J} p(\mathbf{e}_i)p(\mathbf{h}_i|\mathbf{e}_i) \tag{3}$$

In contrast to common structure-aware latent designs that either decompose objects into predefined set of semantic parts [30], or use farthest point sampling (FPS) algorithm to sample sparse points as latent hierarchies [28, 52], our joint-aware representation has several advantages: (i) it addresses the critical problem that part segmentation labels are expensive to obtain, while also capable of adaptively decomposing human bodies into different semantically meaningful parts (ii) the extrinsic-intrinsic disentanglement makes our latent possess both geometric and semantic information, which facilitates flexible editing.

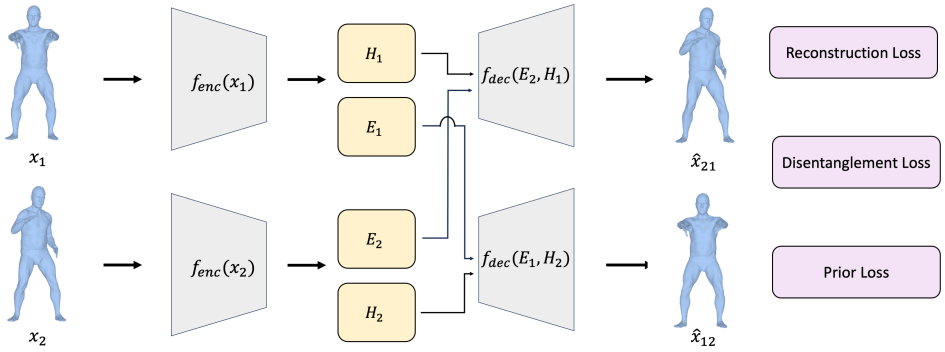
## 3.3. Autoencoder Architecture

In this section, we demonstrate how the aforementioned latent representation is learned in an autoencoding setup, whose architecture and training pipeline is illustrated in Figure 2. The whole architecture comprises of three parts. The first module, the Tokenization Network, transforms a 3D human point cloud to a sparse set of $J$ tokens. The second component, the Mixing Network, augments these tokens with contextual information from each other, which reinforce their capabilities to remain aware of both global skeletal structure and local geometric details, thus generating the joint-aware latent representation. The final building block, the Decoding Network, reconstructs these latents back to original input point cloud without any external supervision.

**Tokenization Network** Our first objective is to obtain a consistent joint-level decomposition of a given 3D human shape, which forms the basis of the joint-aware representation. Given the surface points of a 3D human body, we map them to a global latent using PointNet [38] and then split it into discrete set of tokens denoted as $\{\mathbf{f}_i \in \mathbb{R}^{D_z}\}_{i=1}^{J}$ with a simple MLP, where $D_z$ indicates the feature dimension.




(a) Overview of the autoencoder architecture.




(b) Joint-aware latent representation training pipeline.

Figure 2. A visual illustration of the autoencoder architecture that is used to train our joint-aware latent representation as well as its training pipeline. (a) The encoder $f_{enc}(\cdot)$, which consists of a tokenization network and a mixing network, maps a mesh human shape $x$ into the extrinsic parameters $E$ and intrinsic features $H$, and the decoder $f_{dec}(\cdot)$ aims to recover the original shape using the paired latents. (b) On top of typical reconstruction loss, we also employ a disentanglement loss to ensure the skeleton structure and the geometric details are independently preserved, as well as a prior loss to regularize a smooth latent space.

**Mixing Network** Given the outputs from tokenization module, we then aim to extract the extrinsic and intrinsic information out of each token to achieve semantic-geometric disentanglement. We first project them onto high dimension vectors and then sum up with a learnable positional embedding, which indicates the relative position of local area each token should attend to. The resulting features $\{\mathbf{z}_i^0\}_{i=1}^J$ are fed to a sequence of standard transformer encoder blocks which applies the self-attention mechanism to integrate information across all embeddings and obtain the output features $\{\mathbf{z}_i^L\}_{i=1}^J$ where the superscript indicates which encoder block we're extracting the feature from. On top of that, each token is projected to two sets of parameters: extrinsic parameters represented by joint location $\mathbf{e}_i \in \mathbb{R}^3$, and intrinsic surface geometry information $\mathbf{h}_i \in \mathbb{R}^{D_h}$.

$$\begin{aligned} \mathbf{e}_i &= \text{MLP}_e(\mathbf{z}_i^L) \\ \mathbf{h}_i &= \text{MLP}_h(\mathbf{z}_i^L) \end{aligned} \tag{4}$$

The combination of tokenization network and mixing network constitutes the encoder part of our framework $f_{enc}(\cdot) : \mathbb{R}^{N\times 3} \to \mathbb{R}^{J\times(3+D_h)}$.

**Decoding Network** Our decoding network $f_{dec}(\cdot) : \mathbb{R}^{J\times(3+D_h)} \to \mathbb{R}^{N\times 3}$ also utilizes a Transformer-based architecture, takes the latent representations as input and reconstructs them back to the original shape. We separate extrinsic parameters $\mathbf{E} \in \mathbb{R}^{J\times 3}$ from intrinsic features $\mathbf{H} \in \mathbb{R}^{J\times D_h}$ in the latents and treat skeleton structures as an external condition for better guidance during the decoding process. The conditioning strategy is implemented by simply concatenating the two components before feeding in the network. After the input concatenation and summing up with the shared learnable positional embedding used in the mixing network, we can both anchor the absolute spatial location that corresponding token should attend to and preserve their relative order in the input sequence. The output features obtained after passing through transformer blocks are connected with a MLP head and convert them to a complete point cloud object, which concludes the autoencoding pipeline.

#### 3.3.1 Training Losses

To successfully complete the training pipeline and obtain the joint-aware latent representations, we introduce the loss function as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \mathcal{L}_{dis} + \mathcal{L}_{prior} \tag{5}$$

**Reconstruction loss**. In order to reconstruct the original human mesh as accurately as possible, we adopt the geometric reconstruction loss in both vertex-level and joint-level

$$\begin{aligned} \mathcal{L}_{rec} &= \mathcal{L}_{verts} + \lambda_j \mathcal{L}_{joints} \\ &= \|\mathbf{X} - \hat{\mathbf{X}}\| + \lambda_j \|\mathbf{J} - \hat{\mathbf{J}}\| \end{aligned} \tag{6}$$

where $\mathbf{X}$, $\mathbf{J}$ are the surface vertices and joints locations of the original mesh and $\hat{\mathbf{X}}$, $\hat{\mathbf{J}}$ are the ones from the reconstructed mesh. Recall that extrinsic parameters $\mathbf{E}$ are simply the joint positions so that $\hat{\mathbf{J}} = \mathbf{E}$. These losses force the reconstructed mesh to be close to the original one so that the latents can capture as much information as possible. In addition, the joint supervision loss also enables tokens to aggregate geometric information around specific joints, which to some extent facilitates the joint-aware design.

**Disentanglement loss**. After applying the reconstruction loss, our framework is already capable of accurate reconstruction, but its latent space is still entangled as geometric information might flow into semantic features. To decouple extrinsic and intrinsic components from the latent representation, we apply a disentanglement loss, as introduced in [56], during training:

$$\begin{aligned} \mathcal{L}_{dis} &= \lambda_c \mathcal{L}_{cross} \\ \mathcal{L}_{cross} &= \|\hat{\mathbf{X}}_{12} - \mathbf{X}_1\| + \|\hat{\mathbf{X}}_{21} - \mathbf{X}_2\| \\ \hat{\mathbf{X}}_{ij} &= f_{dec}(\mathbf{E}_i, \mathbf{H}_j) \end{aligned} \tag{7}$$

More specifically, we sample a mesh tuple $\{\mathbf{X}_1, \mathbf{X}_2\}$ from the same subject but with different poses during every training iteration. The encoder takes mesh $\mathbf{X}_i, i = 1, 2$ as input and outputs corresponding extrinsic and intrinsic vectors $(\mathbf{E}_i, \mathbf{H}_i), i = 1, 2$. The cross consistency loss is computed by swapping the intrinsics features of two deformations of the same subject to reconstruct each other.

**Prior loss**. We further add a weighted Kullback–Leibler divergence loss between the distribution of intrinsic features $p(\mathbf{H})$ and the standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$\mathcal{L}_{prior} = \lambda_{kl} \left[ D_{KL}(p(\mathbf{H}) || \mathcal{N}(\mathbf{0}, \mathbf{I})) \right] \tag{8}$$

This regularization term aims to encourage the latent space to be simple and smooth, so that we can perform generation and interpolation through sampling data points in this space.

### 3.4. Latent Cascaded Diffusion

In principle, after finish training the autoencoder, we could use the standard Gaussian's priors to sample latent representations and generate new human body shapes. However, such simple priors may not accurately model the distribution of latent space and thus produce low-quality generated samples, whose phenomenon is also known as prior hole problem [46, 50, 51]. This motivates us to resort to training a highly expressive DDPM on the latent space to model its precise distribution.

To properly handle the diffusion of latent representations, we utilize the characteristics of extrinsic-intrinsic disentanglement and apply a cascaded diffusion scheme according the shape factorization in Equation 3. The architectures of two diffusion models are illustrated in Figure 3. The first DDPM learns the distribution of extrinsics $p(\mathbf{E})$ with a Transformer-based noise prediction network $\boldsymbol{\epsilon}_\theta$ [16], providing coarse-level skeletal structure information of the human body. The second one uses a DiT-based network $\boldsymbol{\epsilon}_\psi$ [34] to learn the conditional distribution of intrinsics $p(\mathbf{H}|\mathbf{E})$ conditioned on the skeleton structure $\mathbf{E}$ to capture finer geometric details. Specifically, the encoded timestamp $\gamma(t)$ and extrinsic embedding $\phi(\mathbf{E})$ are concatenated and fed into the adaptive normalization layers such as AdaLN [35] so that it can learn a scale and translation factor from the condition signals to adaptively modulate the network output. Both the noise prediction networks $\boldsymbol{\epsilon}_\theta$ and $\boldsymbol{\epsilon}_\phi$ are trained with the same variational bound loss as follows:

$$\begin{aligned} \mathcal{L}_{\mathbf{E}}(\theta) &= \mathbb{E}_{t,\mathbf{E},\boldsymbol{\epsilon}} \left[ \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{E}_t, t)\|^2 \right] \\ \mathcal{L}_{\mathbf{H}}(\psi) &= \mathbb{E}_{t,\mathbf{H},\boldsymbol{\phi}} \left[ \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\phi(\mathbf{H}_t, t, \mathbf{E})\|^2 \right] \end{aligned} \tag{9}$$

where the subscription $t$ indicates the attributes results after $t$-step forward process of adding Gaussian noise.

## 4. Experiments

In this section we evaluate the performance of JADE in various human-centric applications, including human shape representation, editing and generation. We first elaborate the dataset configurations, evaluation metrics and implementation details of our approach, and then demonstrate the advantages of JADE for each application compared with prior works. Additionally, we conduct thorough ablation studies to delve into the key elements contributing to our model's performance.

### 4.1. Experiment Setups

#### 4.1.1 Datasets

**DFAUST** [4] captures 14 different body motion sequences (e.g., hips, running, and jumping) for each of the 10 human subjects and register a reference template mesh with same

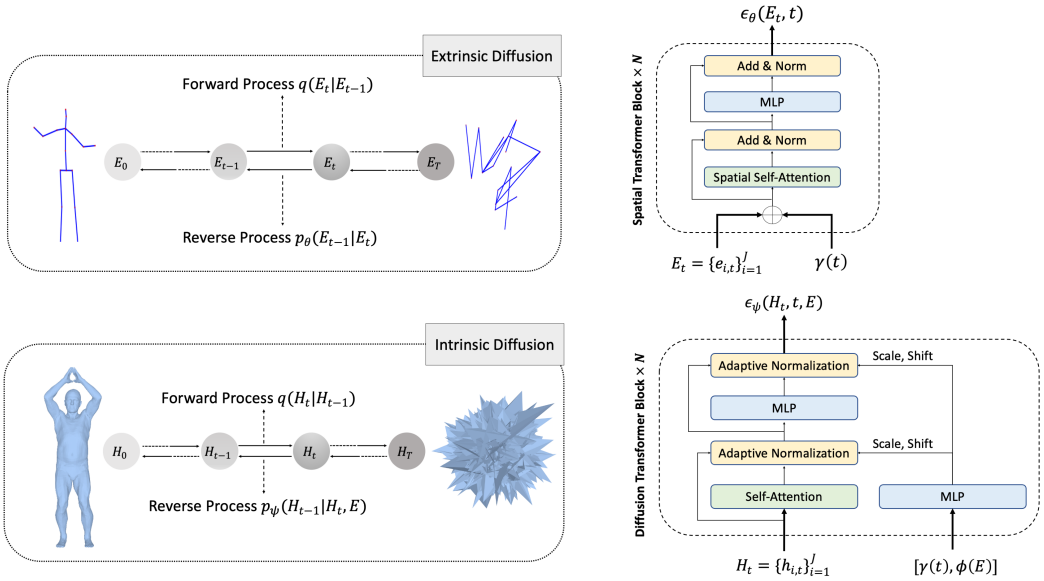



Figure 3. A visual illustration of the diffusion pipeline, where two cascaded diffusions are presented, one for extrinsic parameters $\mathbf{E} = \{\mathbf{e}_i\}_{i=1}^J$ and the other for intrinsic features $\mathbf{H} = \{\mathbf{h}_i\}_{i=1}^J$. In the first phase, we use a time-conditioned spatial transformer to handle diffusion on set data, and in the second phase, we utilize a DiT to handle more complex conditioning, where the concatenation of encoded timestamp $\gamma(t)$ and extrinsics outputs from phase one $\phi(\mathbf{E})$ is fed to the adaptive normalization layers to modulate the diffusion process.

topology to these raw scans. We split off two dynamic performances, i.e.,one-leg jump and chicken wings to conduct testing, which results in a training set with 29,005 samples and a testing set with 3,919 samples.

**SPRING** [55] provides a comprehensive collection of 3D meshes with a rough A-pose registered from the CAESAR dataset [40] using a non-rigid deformation algorithm. We preprocess the SPRING dataset in advance to make its data format compatible with the more commonly-used SMPL model so that the mesh connectivity can be consistent. For the subsequent experiments, the SPRING dataset is randomly split into 2743 training and 305 test meshes.

**AMASS** [29] aggregates motion capture data from various sources, transforming it into a unified format with 3D human body meshes using the SMPL model. It provides detailed 3D motion data, including SMPL parameters, joint angles, and shape coefficients. In our experiments, we adhere to the same train-test partition as previous works [26, 33, 48] but evenly extract one-tenth of the original dataset for the convenience of training, resulting in roughly two million data samples.

#### 4.1.2 Evaluation Metrics

To comprehensively evaluate our framework across various tasks, we adopt task-specific metrics accordingly. For representation ability evaluation, we simply utilize Mean Per Vertex Position Error (MPVPE), which is calculated as the average Euclidean distance between corresponding vertices in the ground truth and its reconstruction as metrics. For generation quality evaluation, both the diversity and fidelity of generated human bodies needs to be considered. To assess generation diversity, we utilize Average Pairwise Distance (APD) [1], defined as mean joint distance between all pairs of samples, as metrics, and for realism evaluation, we employ Self-Intersection rates (SI) [26], which is defined as the average percentage of self-intersecting faces in a batch of 3D meshes.

#### 4.1.3 Implementation Details

As mentioned in Section 3.1, training of JADE is performed in two stages. For the latent representation training, we set the number of joints $J = 24$, the feature dimension $D_z = D_h = 128$, and the autoencoder is trained using AdamW [25] optimizer with batch size 256, learning rate $10^{-3}$. For the diffusion process, we follow conventional DDPM training strategy, where the maximum iteration is set as $T = 1000$, the timestamp $t$ is uniformly sampled from $[1, T]$, and the variances of added noises are configured to linearly increase from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$. We also record the exponential moving average of the DDPM parameters along the training trajectory and the ratio is set to 0.9999. All 3D human shapes are normalized to pelvis-related coordinates. Our experiments are performed on one NVIDIA 3090 GPU, and the training takes about 2 days to complete. For all baselines, we train them using their released codebase and follow the default setting.

| Mode | Method | MPVPE ↓ DFAUST | SPRING |
|---|---|---|---|
| Par. | SMPL [24] | 24.67 | - |
| | SMPL-H [42] | 21.38 | - |
| | SMPL-X [33] | 18.71 | - |
| Lrn. | COMA [39] | 30.83 | 45.53 |
| | Neural3DMM [5] | 16.46 | 15.54 |
| | UnsupShapePose [56] | 9.61 | <u>13.65</u> |
| | SemanticHuman [45] | <u>5.70</u> | 18.76 |
| | **JADE** | **5.47** | **12.85** |

Table 1. Quantitative reconstruction results on DFAUST and SPRING datasets. *Par.* stands for parametric model and *Lrn.* means learning-based methods. *MPVPE* is the mean position error per vertex. − : not supported for this dataset. The best and second-best results are highlighted in bold and underline formats.

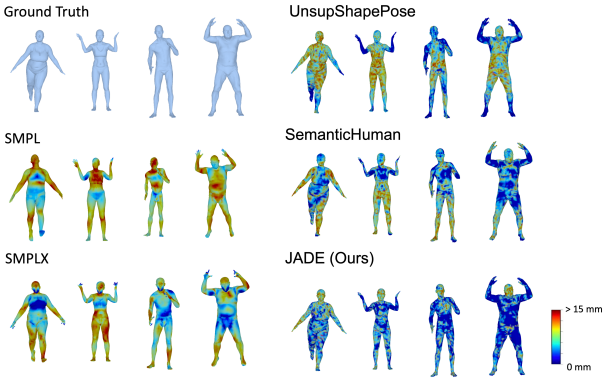



Figure 4. Qualitative visualization results on DFAUST dataset, showing the color coding of the MPVPE error of the reconstructions produced by our JADE framework and baseline methods[24, 33, 45, 56]. The error maps show that our method has better reconstruction accuracy.

## 4.2. Human Shape Reconstruction

In this section, we validate the representation ability of our approach on DFAUST and SPRING datasets. We compare the results against various kinds of methods, including classical parametric models [3, 24, 33], spectral-based approach [39], spiral-based method [5], and disentangled representations [45, 56]. As shown in Table 1, our approach outperforms the methods in all categories with high reconstruction precision, demonstrating the effectiveness of the joint-aware latent representation. Unlike mesh-based representations that use thorough topological information during training, our approach is point cloud-based, which only utilize vertex positions to model the surface geometry. Figure 4 also visualizes some qualitative results and their error maps so that we can observe that JADE exhibits better reconstruction accuracy especially for complex geometric details.

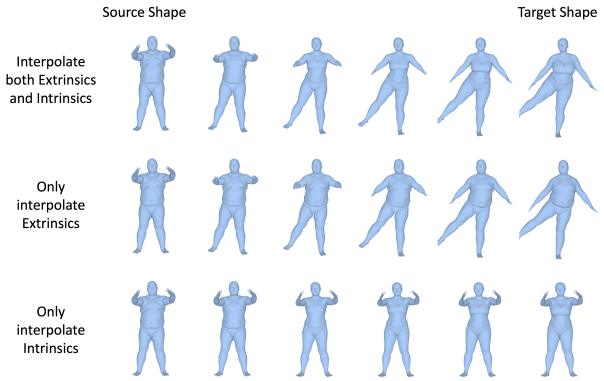



Figure 5. Qualitative example of transferring a given human character to a new identity with different body posture through interpolation, indicating that JADE can edit both the extrinsics and intrinsics components in the latent space independently to achieve desired human body movements.

| Sample source | APD ↑ | SI ↓ |
|---|---|---|
| AMASS [29] | 15.44 | 0.79 |
| GMM [3] | <u>16.28</u> | 1.54 |
| VPoser [33] | 10.75 | 1.51 |
| PoseNDF [48] | **18.75** | 1.97 |
| DPoser [26] | 14.28 | <u>1.21</u> |
| **JADE** | 14.95 | **1.05** |

Table 2. Quantitative generation metrics on AMASS dataset.

## 4.3. Human Shape Editing

In this section, we demonstrate that the joint-aware latent representation has great flexibility to perform human shape editing through transfer experiments. The top row of Figure 5 shows an example of shape transfer, who aims to transform from a given human character to a new identity with different body postures. JADE can complete this task through interpolation. Thanks to the disentangled latent design, we can interpolate both the extrinsic parameters and intrinsic features correspond to each joint to achieve desired human body movements. We also show the influence of extrinsic-intrinsic pair in the latent space decomposition by only interpolating each component respectively. From the second and third row of Figure 5, we can see that extrinsics control the overall skeleton structure and posture of the human body while the intrinsics control the local geometry.

## 4.4. Human Shape Generation

We use AMASS dataset to train JADE solely for unconditional human shape generation and compare it against various human body generation methods such as classical

| | Intrinsics Dimension | Condition Mechanism | Loss function $\mathcal{L}_{joint}$ | Loss function $\mathcal{L}_{dis}$ | MPVPE ↓ DFAUST | MPVPE ↓ SPRING |
|---|---|---|---|---|---|---|
| Full Architecture | 128 | concat | ✓ | ✓ | 5.47 | 12.85 |
| Variant 1 | 16 | concat | ✓ | ✓ | 7.85 | 19.01 |
| | 32 | concat | ✓ | ✓ | 6.37 | 15.12 |
| | 64 | concat | ✓ | ✓ | 5.83 | 13.55 |
| | 256 | concat | ✓ | ✓ | 6.29 | 14.80 |
| | 512 | concat | ✓ | ✓ | 8.11 | 19.04 |
| Variant 2 | 128 | add | ✓ | ✓ | 5.61 | 13.43 |
| | 128 | cross-attention | ✓ | ✓ | 9.89 | 24.33 |
| Variant 3 | 128 | concat | | ✓ | 6.01 | 14.12 |
| | 128 | concat | ✓ | | 6.64 | - |

Table 3. Ablation studies on the intrinsics dimension, condition mechanism and the impact of different loss functions. We show the experiment results under reconstruction setting and report the performance of MPVPE on DFAUST and SPRING dataset.

GMM [3], VPoser [33], Pose-NDF [48], DPoser [26]. The quantitative and qualitative results are summarized in Table 2. Following the strategy of previous works, we randomly sample 500 different human shapes and evaluate their diversity and realism through APD and SI metrics. Compared with ground truth from AMASS dataset, Pose-NDF and classicl GMM exhibit both high APD and SI, suggesting more divergence from training samples but resulting in more unrealistic human shapes. On the contrary, VPoser and DPoser shows lower scores in both metrics, indicating they prefer realism over diversity. JADE proves to be a good trade-off between these two criteria, both numerically and visually.

### 4.5. Ablation Studies

To validate our architecture design and training settings, we compare the final model of the latent representation with several types of variants and evaluate their performances under the reconstruction experiment setting. The results of the ablation studies are summarized in Table 3 and we'll further discuss the details below.

**Effect of intrinsics feature dimension.** The first type of variants investigate how the dimension of intrinsic features affect the representation ability. We can see that as the feature dimension increases, the reconstruction error decreases at first, reaching a minimum at 128, and then increases. This reveals that enlarging the latent dimension in some extent contributes to the representation ability, but will introduce redundancy, similar to overfitting phenomena, and hamper its generalization expressivity once exceeds a threshold.

**Effect of conditioning mechanism.** The second type of variants examine multiple conditioning methods, including concatenation, addition, and cross-attention, to inject skeleton structure information in the decoding process. For all experiments, we first project the skeleton structure onto a high-dimensional embedding, after which these embeddings are fused into the decoding network. The best conditioning mechanism is concatenating the skeleton embedding and intrinsics latent feature, which provides a simple and effective way to guide the latent decoding.

**Effect of loss functions.** The third type of variants explore the influence of loss functions, where the network architecture remains the same, but we omit the joint supervision loss $\mathcal{L}_{joint}$ in Equation 6 and disentangled loss $\mathcal{L}_{cross}$ in Equation 7 from the training objective. Although these reduced variants achieve slightly worse reconstruction performances, the characteristics of latent disentanglement, structure-level control and the quality of editable human shapes degrades significantly, which assure the importance of our losses design.

## 5. Conclusion

In this work, we introduce JADE, a generative framework trained on surface point clouds for 3D human modeling. Our key insight is a joint-aware latent representation that decomposes human bodies into skeleton structures and local surface geometries. The disentangled design enables geometric and semantic interpretation, facilitating users with flexible controllability. We also present a cascaded pipeline to generate coherent and plausible human shapes under our proposed decomposition. Extensive experiments conducted on public datasets demonstrates the effectiveness of JADE framework in multiple downstream tasks. Currently, JADE is limited on fixed topology human shape modeling and can not directly generate textured shapes. To bypass the underlying mesh topology restriction, a promising extension would be to incorporate neural implicit representation to generate more smooth shapes. In addition, combining with image-based training and differentiable rendering techniques to also synthesize textures would also be a potential future research direction.

## References


[1] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5223–5232, 2020. 6

[2] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005. 1, 2

[3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016. 7, 8

[4] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J Black. Dynamic faust: Registering human bodies in motion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6233–6242, 2017. 5

[5] Giorgos Bouritsas, Sergiy Bokhnyak, Stylianos Ploumpis, Michael Bronstein, and Stefanos Zafeiriou. Neural 3d morphable models: Spiral convolutional networks for 3d shape representation learning and generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7213–7222, 2019. 2, 7

[6] Lewis Bridgeman, Marco Volino, Jean-Yves Guillemaut, and Adrian Hilton. Multi-person 3d pose estimation and tracking in sports. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 1

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[9] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022. 2

[10] Zhengfu He, Tianxiang Sun, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. Diffusionbert: Improving generative masked language models with diffusion models. *arXiv preprint arXiv:2211.15029*, 2022. 2

[11] Amir Hertz, Or Perel, Raja Giryes, Olga Sorkine-Hornung, and Daniel Cohen-Or. Spaghetti: Editing implicit shapes through part aware generation. *ACM Transactions on Graphics (TOG)*, 41(4):1–20, 2022. 3

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[13] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022. 2

[14] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 3

[15] Tao Hu, Fangzhou Hong, and Ziwei Liu. Structldm: Structured latent diffusion for 3d human generation. In *European Conference on Computer Vision*, pages 363–381. Springer, 2024. 3

[16] Haorui Ji, Hui Deng, Yuchao Dai, and Hongdong Li. Unsupervised 3d pose estimation with non-rigid structure-from-motion modeling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3314–3323, 2024. 5

[17] Boyi Jiang, Juyong Zhang, Jianfei Cai, and Jianmin Zheng. Disentangled human body embedding based on deep hierarchical neural network. *IEEE transactions on visualization and computer graphics*, 26(8):2560–2575, 2020. 1, 2

[18] Zhongyu Jiang, Haorui Ji, Samuel Menaker, and Jenq-Neng Hwang. Golfpose: Golf swing analyses with a monocular camera based human pose estimation. In *2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2022. 1

[19] Juil Koo, Seungwoo Yoo, Minh Hieu Nguyen, and Minhyuk Sung. Salad: Part-level latent diffusion for 3d shape generation and manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14441–14451, 2023. 3

[20] Peizhuo Li, Kfir Aberman, Rana Hanocka, Libin Liu, Olga Sorkine-Hornung, and Baoquan Chen. Learning skeletal articulations with neural blend shapes. *ACM Transactions on Graphics (TOG)*, 40(4):1–15, 2021. 3

[21] Shidi Li, Miaomiao Liu, and Christian Walder. Editvae: Unsupervised parts-aware controllable 3d point cloud shape generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1386–1394, 2022. 3

[22] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022. 2

[23] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 3

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 1, 2, 3, 7

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[26] Junzhe Lu, Jing Lin, Hongkun Dou, Yulun Zhang, Yue Deng, and Haoqian Wang. Dposer: Diffusion model as robust 3d

human pose prior. *arXiv preprint arXiv:2312.05541*, 2023. 6, 7, 8

[27] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021. 3

[28] Zhaoyang Lyu, Jinyi Wang, Yuwei An, Ya Zhang, Dahua Lin, and Bo Dai. Controllable mesh generation through sparse latent point diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 271–280, 2023. 3

[29] 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. 6, 7

[30] George Kiyohiro Nakayama, Mikaela Angelina Uy, Jiahui Huang, Shi-Min Hu, Ke Li, and Leonidas Guibas. Difffacto: Controllable part-based 3d point cloud generation with cross diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14257–14267, 2023. 3

[31] Charlie Nash and Christopher KI Williams. The shape variational autoencoder: A deep generative model of part-segmented 3d objects. In *Computer Graphics Forum*, pages 1–12. Wiley Online Library, 2017. 3

[32] Ahmed AA Osman, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Supr: A sparse unified part-based human representation. In *European Conference on Computer Vision*, pages 568–585. Springer, 2022. 2

[33] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 1, 2, 6, 7, 8

[34] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2, 5

[35] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 5

[36] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3

[37] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10619–10629, 2022. 2

[38] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 3

[39] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3d faces using convolutional mesh autoencoders. In *Proceedings of the European conference on computer vision (ECCV)*, pages 704–720, 2018. 7

[40] Kathleen M Robinette, Hans Daanen, and Eric Paquet. The caesar project: a 3-d surface anthropometry survey. In *Second international conference on 3-D digital imaging and modeling (cat. No. PR00062)*, pages 380–386. IEEE, 1999. 6

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[42] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 7

[43] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[44] Jan Stenum, Kendra M Cherry-Allen, Connor O Pyles, Rachel D Reetzke, Michael F Vignos, and Ryan T Roemmich. Applications of pose estimation in human health and performance across the lifespan. *Sensors*, 21(21):7315, 2021. 1

[45] Xiaokun Sun, Qiao Feng, Xiongzheng Li, Jinsong Zhang, Yu-Kun Lai, Jingyu Yang, and Kun Li. Learning semantic-aware disentangled representation for flexible 3d human body editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16985–16994, 2023. 2, 7

[46] Hiroshi Takahashi, Tomoharu Iwata, Yuki Yamanaka, Masanori Yamada, and Satoshi Yagi. Variational autoencoder with implicit optimal priors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5066–5073, 2019. 5

[47] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22819–22829, 2023. 3

[48] Garvita Tiwari, Dimitrije Antić, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision*, pages 572–589. Springer, 2022. 6, 7, 8

[49] Edgar Tretschk, Ayush Tewari, Michael Zollhöfer, Vladislav Golyanik, and Christian Theobalt. Demea: Deep mesh autoencoders for non-rigidly deforming objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 601–617. Springer, 2020. 1, 2

[50] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020. 5

[51] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based

generative modeling in latent space. *Advances in neural information processing systems*, 34:11287–11302, 2021. 5

[52] Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, Karsten Kreis, et al. Lion: Latent point diffusion models for 3d shape generation. *Advances in Neural Information Processing Systems*, 35:10021–10039, 2022. 3

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[54] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6184–6193, 2020. 1, 2

[55] Yipin Yang, Yao Yu, Yu Zhou, Sidan Du, James Davis, and Ruigang Yang. Semantic parametric reshaping of human body models. In *2014 2nd International Conference on 3D Vision*, pages 41–48. IEEE, 2014. 6

[56] Keyang Zhou, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Unsupervised shape and pose disentanglement for 3d meshes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 341–357. Springer, 2020. 2, 5, 7