Magic Tailor: a Latent-Anchor based Diffusion Model for 3D Clothes Generation

Zibo Zhao[1,2*] Wen Liu[2] Xin Chen[2] Gang Yu[2] Shenghua Gao[1†]
[1]ShanghaiTech University [2]Tencent PCG, China

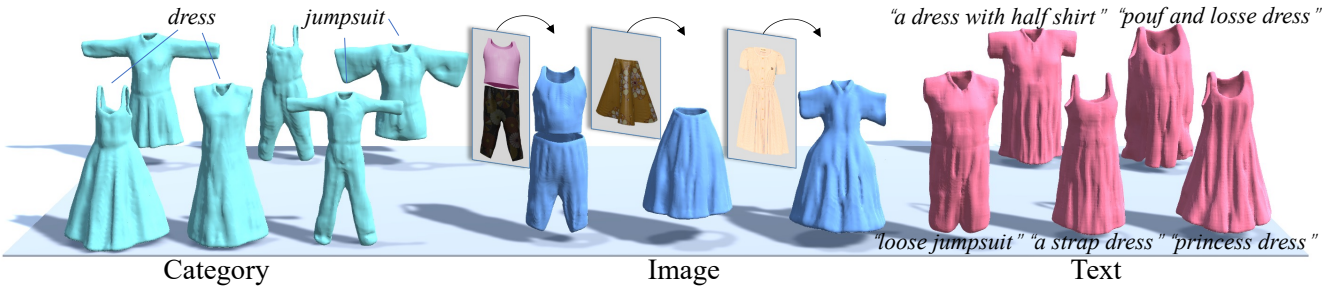

Figure 1. **An illustration of our motivation.** We propose the Magic Tailor, a latent-anchor diffusion model (LAD), which generates high-quality 3D clothes with multiple modalities of conditional inputs, including a specific category (e.g., dress), a conditioning image, and a textual prompt. The figure shows results generated by our LAD under different conditions and demonstrates that our model can produce elegant results that conform to the provided conditions.

Abstract

We study conditional 3D clothes generation to synthesize high-quality 3D clothes models that conform to various conditions, such as clothes categories, images, and texts. Traditional methods to generate 3D clothes depend on registering 3D clothes to human parametric models or predefined templates. However, this registration process inevitably compromises the fidelity and topology of clothes. Thus, we propose a topology-free and computation-friendly latent-anchor representation for 3D clothes to tackle this restriction. Specifically, we employ a Vector Quantised-Variational AutoEncoder (VQ-VAE) to encode each 3D clothes model into groups of latent anchors, and each latent anchor contains an anchor point and anchor embedding. Based on the latent-anchor representation, we introduce a novel two-level latent-anchor diffusion model (LAD) that first learns a probabilistic mapping function from various conditional inputs to anchor points. The anchor points and conditional inputs are used to generate the anchor embeddings. Then, anchor points and anchor embeddings are fed into the decoder of VQ-VAE for 3D clothes generation. Extensive experimental results demonstrate the effectiveness of LAD in producing 3D clothes models. The codes of our work will be released later to facilitate further research in this field.

1. Introduction

3D clothing generation [92] is promising for promoting the fashion industry, virtual try-on experiences, film production, and 3D gaming asset design. Condition-based 3D clothing generation offers an efficient and user-friendly method for generating 3D clothing shapes, accommodating a range of conditional inputs such as categories, images, and textual prompts.

Nevertheless, generating plausible 3D clothes conforming to conditional inputs is a formidable challenge, and the absence of an effective shape representation for 3D clothes and the limited capabilities of previous generative models on 3D garments constitute two primary obstacles for this task. Since 3D clothes are in diverse topology structures, conventional methods typically register each type of 3D clothes to a template [20, 38] or a human body parametric model [2, 56, 64, 65]. However, template-based registration inevitably compromises the fidelity of the original 3D clothes due to the simplification of mesh topology.

In recent years, the neural field [54, 62] has demonstrated its capability for 3D shape representation because its topology-free data structure, such as global latent code [14,

[1]*Work was partially done while Zibo Zhao was a Research Intern with Tencent PCG.
[2]†Corresponding author.

22, 23, 39, 60] and regular or irregular grid latent code [55, 98, 103, 106], can be processed by neural networks in an implicit functional manner. However, implicit neural representation for 3D clothing shapes is still unexplored. Inspired by the success of LION [101] and 3DILG [102] for object representation, we investigate the latent-anchor representation for 3D clothes in a topology-free and neural network-friendly manner in this paper. Specifically, we use a Vector Quantised-Variational AutoEncoder (VQ-VAE) [70, 90] to encode each 3D clothing into latent anchors consisting of latent points and latent embeddings that can reconstruct the original 3D clothing shape with high fidelity. Compared to the LION [101] approach, which employs a decoder to transform the latent points into dense point clouds and reconstruct the entity surface via an additional shape-as-points [63] network, our latent anchors contain more representative anchor points with richer feature descriptors of the local surface. Additionally, our decoder can directly reconstruct a high-fidelity 3D clothing shape from a fixed number of latent anchors, simplifying the decoding process while maintaining reconstruction quality. Thus, our representation is more suitable for generative models to learn the distribution of 3D clothing shapes.

Based on the latent-anchor representation for 3D clothes, we focus on learning the adequately capable generative models to map the multi-modality conditions to the distribution of the 3D clothing shape or its latent space. Previous approaches utilize the generative adversarial network (GAN) [5, 10, 46, 93] or the variational auto-encoder (VAE) [2, 4, 48] to model the distribution of the shape latent codes. However, these methods have limited capabilities in modeling various distributions, unlike generating diverse 3D clothes. Contemporaneous auto-regressive-based models [55, 98, 102] achieve surprising conditional 3D shape generation performance but suffer from error accumulations and linear time computational overhead in sampling.

Fortunately, the recent advancements in diffusion-based generative models [27] have showcased remarkable success in various domains such as image [69, 71, 72, 105], video [17, 28], audio [37], and motion [86, 97]. The diffusion models [15, 27] generate plausible results with enhanced diversity and stability during training without the need for adversarial loss, as opposed to VAEs [35] or GANs [19]. Additionally, they exhibit less error accumulation compared to auto-regressive models. Moreover, the current efficient sampling strategies, such as DDIM [82], DPM Solver [45], and stochastic sampler [34], have reduced the sampling steps to fewer than 50, making them faster than auto-regressive-based generative models. Inspired by the success of the diffusion model, this paper proposes a two-level Latent-Anchor Diffusion (LAD) model, which leverages the effective latent-anchor representation to synthesize plausible and diverse 3D clothing shapes that adhere to various conditional inputs. In particular, utilizing the latent-anchor representation for 3D clothing, the first diffusion model predicts the anchor points based on conditional inputs. Subsequently, the second diffusion model employs the coordinates and conditional inputs to predict the latent embedding of each latent anchor. Then, the anchor points and latent embeddings are fed into the VQ-VAE decoder to generate a complete 3D clothes surface.

We summarize the contributions of this paper as follows: 1) we propose a latent-anchor representation for generating 3D clothing shapes; 2) a two-level diffusion model to learn mapping functions from various conditions to the 3D clothing shape distribution; 3) extensive experiments demonstrate the effectiveness of our proposed framework for 3D clothing generation under various conditions.

2. Related Work

2.1. 3D Clothes Representation and Generation

In response to the fashion industry's growing demand for intelligent systems capable of generating high-quality 3D garments, researchers are exploring methods [1, 24, 94–96] to streamline the intricate design process, which typically involves multiple stages, such as sketching, fabric assembly, and pattern creation. Despite the assistance provided by existing tools like Optitex [58] and Marvelous Designer [52], designing clothing from scratch takes time and effort, even for skilled professionals. Deep learning-based approaches have recently emerged as promising solutions to this challenge. Early work laid the groundwork for achieving this objective, developing several 3D clothing datasets, such as MGN [3], TailorNet [61], SIZER [87], CAPE [48], Cloth3D [2], DeepFashion3D [108], and TightCap [9]. Most current methods [10, 13, 40, 49–51, 100] focus on the generative reconstruction of articulated humans, and only a few [11, 18, 67, 92] studies have directly targeted 3D clothing generation.

Prior studies have investigated two prevalent approaches for representing clothing in generation and reconstruction tasks. The first approach employs a template-based representation, while the second involves mapping 3D clothing mesh onto a parametric human body model, such as the Skinned Multi-Person Linear (SMPL) model [43]. Registration on the SMPL mesh has become a popular method for representing clothing in 3D generation [2], reconstruction [25, 29, 73, 74], animation [75–77, 88, 104]. In the Cloth3D [2], arbitrary 3D clothing models were simplified and registered onto an SMPL mesh. Subsequently, the researchers designed a generative model, the Conditional Variational Auto-Encoder (C-VAE), to synthesize the processed meshes. This C-VAE utilized graph convolutions to model mesh vertices more effectively. However, mesh distortion during simplification and registration is unavoidable,

resulting in the model learning from defective information in the dataset. Consequently, the generated clothing models may exhibit unrealistic features, such as unclear boundaries.

These approaches [8, 57, 85, 92] devise an algorithm or rule for mapping 3D coordinates to a 2D representation and subsequently train a generative model on the 2D UV coordinates or sewing pattern. One advantage of this representation is the ease of leveraging powerful 2D generative models. However, the 2D-3D mapping rule is limited to handling limited situations. It unavoidably comprises the topology structure, which leads to difficulty when models synthesize plausible results when dealing with clothing that exhibits diverse topologies. Our proposed latent-anchor diffusion model offers a more effective solution for the 3D clothing generation task than these alternative methods. By leveraging a topology-free latent-anchor representation, our model gets rid of handling topology structures and can generate cross-category results.

2.2. Generative Models in 3D

In recent years, generative models [6, 30, 59, 69, 72] have gained prominence for their generation quality. Researchers have primarily concentrated on two types of generative models: Variational Auto-Encoders (VAEs) and Generative Adversarial Networks (GANs) [31–33]. VAEs [89] generate 3D shapes by learning a low-dimensional latent representation of input shapes, enabling the creation of novel 3D shapes through sampling from the learned latent space. Various VAE-based models for 3D shape generation have been proposed, including 3D-VAE-GAN [93] and PointFlow [99]. Meanwhile, GANs employ a generator network to produce shapes resembling real 3D shapes, with a discriminator network distinguishing between generated and ground-truth shapes.

However, VAE-based methods are often limited with restricted generative ability and are unlikely to yield diverse shapes, while GAN-based methods are prone to unstable training. Recently, auto-regressive [16, 102] models with transformer-based architectures have demonstrated remarkable performance in conditional 3D shape generation, but they suffer from error accumulation and linear time computational overhead during sampling. On the other hand, several diffusion-based point cloud generation methods [47, 101, 107] reveal competitive performance. Nevertheless, these methods struggle to produce smooth surfaces by solely manipulating points due to the high degree of freedom in point coordinates. Additionally, these models necessitate operating on highly dense point clouds to capture fine-grained surface details, which is often infeasible.

To address these challenges, we propose the Latent-Anchor Diffusion (LAD) model, which combines the expressiveness of diffusion models with the flexibility of neural fields based on the latent-anchor representation. This approach enhances the generative model's ability to reconstruct high-quality surfaces, resulting in a powerful solution for generating the 3D clothes model.

3. Method

We aim to develop a framework that generates diverse 3D clothing based on various conditions, such as specific categories, images, and textual prompts. However, devising a universal representation for 3D clothing is challenging, as different clothing types may possess distinct structures, leading to significantly different mesh topologies. Furthermore, even within the same category, variations in surface details can result in differences in mesh vertices and faces.

To tackle this issue, we introduce the flexible latent-anchor representation, which circumvents to handle 3D clothing models with traditional representations (details in Section 3.1). Thus, our model avoids cumbersome mesh operators and reduces computational costs by learning the distribution of sparse yet representative latent anchors. Consequently, the model exhibits enhanced effectiveness in learning the probabilistic mapping from multi-modal guidance to the latent-anchor distribution (details in Section 3.2).

In particular, our proposed Latent-Anchor Diffusion (LAD) model consists of two primary modules, as outlined in Figure 2. The first module is a Vector Quantised-Variational AutoEncoder (VQ-VAE) that encodes each 3D clothing model into a set of latent anchors, comprising anchor latents and anchor embeddings. The second module is a two-level latent-anchor diffusion (LAD) model. The first-level diffusion model learns to predict the anchor point based on the conditional inputs, and a second-level one subsequently predicts the anchor embedding of each latent anchor based on the coordinates and conditional inputs, ultimately generating a 3D clothes model.

3.1. Latent-Anchor Representations for 3D Clothes

In more detail, our clothing VQ-VAE, denoted as $\mathcal{V}$, is composed of a clothing encoder $\mathcal{E}$, a clothing decoder $\mathcal{D}$ and a quantized codebook $\mathbb{Z}$. The encoder contains a pointnet-like module and a transformer-based extractor, aiming to extract anchor latent and continuous embeddings. The codebook $\mathbb{Z} = \{z_j \in \mathbb{R}^D\}_{j=1}^J$ stores J discrete embedding z_j to transform continuous embeddings into anchor embeddings. The decoder employs a transformer-based architecture and a multi-layer perceptron (MLP) based head to reconstruct the neural field of a 3D clothes model.

Given a clothes mesh with arbitrary vertices and faces, we randomly sample surface points $P_s \in \mathbb{R}^{N\times 3}$ on the mesh surface and apply Farthest Point Sampling (FPS) on the surface points P_s to produce **anchor points** $P_c \in \mathbb{R}^{M\times 3}$. For each anchor point $p_m \subset P_c$, we find its nearest $K-1$ points from surface points P_s via K-nearest

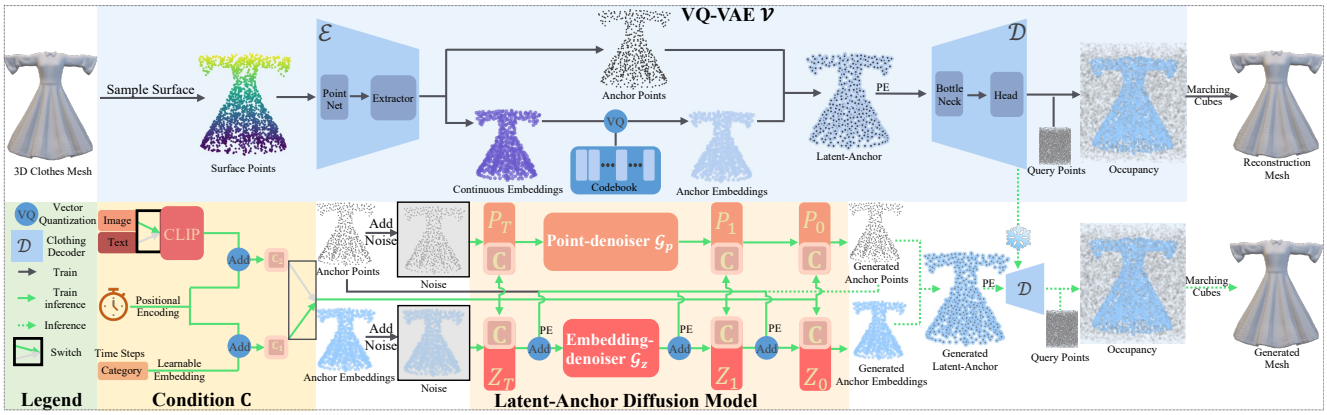

Figure 2. **Network Overview.** Our Latent-Anchor Diffusion (LAD) model comprises a VQ-VAE $\mathcal{V}$ (detailed in Section 3.1) and a two-level diffusion model $\mathcal{G}$ (described in Section 3.2). The model employs a two-stage training strategy. In the first stage, the focus is on learning the latent-anchor representation of 3D clothing meshes, wherein the encoder $\mathcal{E}$ encodes 3D clothing into anchor points and anchor embeddings, which decodes to the 3D clothing shape via the decoder $\mathcal{D}$. The second stage involves the two-level latent-anchor diffusion model learning distribution anchor point and anchor embedding. During the inference phase, the latent-anchor diffusion model predicts the latent anchors, efficiently reconstructed into the 3D clothes by the decoder.

neighbor and form a clustered patch with K points, including the anchor point p_m. Subsequently, a point-net-like [66] module consisting two-layer MLPs to extract a feature $f_m \in \mathbb{R}^C$, and ultimately produces the point-feature pair $\{p_m, f_m\}_{m=1}^M$. The extractor further extracts continuous embeddings $\tilde{Z} = \{\tilde{z}_m \in \mathbb{R}^D\}_{m=1}^M$. Performing an element-wise quantize operation $\mathcal{Q}(\cdot)$ on each $\tilde{z}_m$, we query the discrete **anchor embedding** $\hat{Z} = \{\hat{z}_m \in \mathbb{R}^D\}_{m=1}^M$ from the quantized codebook $\mathbb{Z}$:

$$\hat{Z} = \mathcal{Q}(\tilde{Z}) = \{\arg\min_{z_j \in \mathbb{Z}} \|\tilde{z}_m - z_j\|\}_{m=1}^M. \tag{1}$$

We define the **latent-anchor representation** of clothes by pairing anchor points P_c and the anchor embeddings $\hat{Z}$:

$$\{P_c, \hat{Z}\} = \{p_m \in \mathbb{R}^3, \hat{z}_m \in \mathbb{R}^D\}_{m=1}^M. \tag{2}$$

Each sub-pair $\{p_m, \hat{z}_m\}$ effectively represents local information for the 3D clothing, providing a compact and expressive representation for the modeling process.

After processing through the clothing decoder, $\mathcal{D}$, the latent anchor is converted into a weight indicating whether a query point x resides inside or outside the clothing. Specifically, we employ MLPs with a sigmoid activation function as the classifier to predict the result. During the inference, we sample all grid points within a volume as query points and predict their indicators based on the latent anchors. Finally, we use contouring methods, marching cubes [44] to obtain a 3D clothes mesh.

We optimize the clothing VQ-VAE $\mathcal{V}$ by a binary-cross-entropy loss $\mathcal{L}_{BCE}$, criticing the predicted and ground-truth, and a reconstruction regularization to maximize the representation capacity of the constructed latent anchor. Moreover, we train the model with two distinct types of reconstruction regularization. The first of these is the commitment loss, denoted as:

$$\mathcal{L}_{VQ} = \|\text{sg}[\hat{Z}] - \mathcal{E}(P_s)\|^2, \tag{3}$$

where sg[·] denotes a stop-gradient operation. The second type of reconstruction regularization is the Kullback-Leibler divergence loss, denoted as $\mathcal{L}_{KL}$. Training the model with both $\mathcal{L}_{BCE}$ and $\mathcal{L}_{KL}$ essentially reduces it to a Variational AutoEncoder (VAE) without a vector quantization operation. Table 3 shows that the VQ-VAE scheme demonstrates superior representation capabilities. Thus, we employ it in the subsequent generative process.

3.2. Conditional Latent-Anchor Diffusion Model

Diffusion Models [27] models a Markov noising process and learns the data distribution $p(x)$ through a sequence of denoising operations that convert Gaussian noise to a real signal. Inspired by the similarity between particles in a thermodynamic system [81] and points in a point cloud, prior works [47, 107] have introduced diffusion models for synthesizing point clouds. These generative models $\mathcal{G}_\theta(x_t, t)$ train to predict a denoised variant with input x_t, where x_t denotes a disturbed x_0 and $t = \{1, 2, ..., T\}$. The corresponding objective can be reduced to an $\mathcal{L}_2$ loss between the input and noise ϵ as:

$$\mathcal{L}_2 = \|\epsilon - \mathcal{G}_\theta(x_t, t)\|_2^2, \quad \epsilon \sim \mathcal{N}(0, 1), \tag{4}$$

where t uniformly samples from $\{1, 2, ..., T\}$.

Conditional Latent-Anchor Diffusion Models. We devise the diffusion model based on a time-conditional transformer architecture. By representing 3D clothing via the latent anchor, $\{P_c, \hat{Z}\}$, the diffusion model learns on the latent anchor's distribution, enhancing its generative capacity while reducing its computational cost.

Since the anchor point P_c represents explicit shape information, and the anchor embeddings $\hat{Z}$ are high-dimension features providing implicit shape information, a domain gap exists between the anchor points distribution and the anchor embeddings distribution. It leads to difficulties in training the generative model. Thus, we introduce our two-level Latent-Anchor Diffusion (LAD) Model, $\mathcal{G} = \{\mathcal{G}_p, \mathcal{G}_z\}$ to enable the diffusion model to reach optimal generative capabilities. In this model, the point-denoiser $\mathcal{G}_p$ focuses on learning the anchor point P_c exclusively, while embedding-denoiser $\mathcal{G}_z$ is responsible for generating anchor embeddings using P_c produced by $\mathcal{G}_p$.

Moreover, unconditional generative models [19, 83, 84] are far from users' requirements in many scenarios, as the generated content might not adhere to semantic guidance or even a rudimentary category condition. Thus, we devise the conditional latent-anchor diffusion model, with condition input denoted as $\mathbf{c}$. The conditions could be specific categories, an image, and a text.

Similar to previous methods [69, 86], both the point-denoiser $\mathcal{G}_p$ and the embedding-denoiser $\mathcal{G}_z$ predict the starting signal when given any noisy version and conditions. Specifically, we adopt the original objective 4 as follows:

$$\begin{aligned} \mathcal{L}_{\mathcal{G}_p} &= \|P_c - \mathcal{G}_p(P_t, c, t)\|_2^2, \\ \mathcal{L}_{\mathcal{G}_z} &= \|\hat{Z} - \mathcal{G}_z(\hat{Z}_t, P_c, c, t)\|_2^2, \end{aligned} \tag{5}$$

where P_t denotes a disturbed P_c, $\hat{Z}_t$ denotes a disturbed $\hat{Z}$ and $t = \{1, 2, ..., T\}$.

Following the classifier-free guidance (CFG) [26], we randomly set the conditions as empty set ϕ to the model with 10% probability in the training phase. In the inference stage, the model generates latent anchors $\{P_c, \hat{Z}\}$ with CFG with a guidance weight λ to balance generative diversity and quality. For example, when sampling anchor points P_c, the CFG process expresses as:

$$\mathcal{G}_p(P_t, c, t) = \mathcal{G}_p(P_t, \phi, t) + \lambda(\mathcal{G}_p(P_t, c, t) - \mathcal{G}_p(P_t, \phi, t).) \tag{6}$$

3.3. 3D Clothing Generation with Various Conditions

Category-Conditioned generation refers to generating 3D clothing within a specific category. By capitalizing on flexible transformer architectures, we prepend the conditional token to the beginning of the sequence in the LAD. In particular, the conditional token adds the time step and category embedding, which allows the model to incorporate specific category information during the denoising process.

Image- and text-Conditioned generation aims to generate 3D clothes that conforms to a target image $I \in \mathbb{R}^{H \times W \times 3}$. We employ the powerful pre-trained vision-language model CLIP [68] to extract discriminative conditional features to the LAD model's ability. In this context, we denote the CLIP image encoder as $\mathcal{E}_i$, which facilitates the conversion of cross-modal information into a vector. The condition token adds the time step embedding and the CLIP embedding. Moreover, the CLIP image encoder's training aligns with the CLIP text encoder's domain, suggesting that once a model has mapped a distribution to the CLIP feature domain, it can handle two modalities. During training, we employ the CLIP image encoder to facilitate accepting a prompt as input in subsequent experiments. Importantly, our experiments demonstrate that the CLIP image encoder enables the LAD model to generate 3D clothing based on a textual prompt. Ultimately, the decoder $\mathcal{D}$ reconstructs 3D clothes corresponding to the sampled latent anchors.

3.4. Training and Inference

The clothing VQ-VAE comprises two six-layer transformer encoders. One integrates the extractor in the clothing encoder with a point-net-like module. At the same time, the other incorporates into the bottleneck of the clothing decoder, followed by an eight-layer MLP for reconstructing the neural field of 3D clothing. Both diffusion models in the two-level latent-anchor diffusion model employ an eight-layer transformer encoder to execute the diffusion process. Specifically, seven learnable embeddings integrated into the model correspond to seven clothes categories, and we utilize a frozen CLIP (ViT-B/32) to extract conditional information from a given image or text. Our diffusion models train with $T = 1000$ noising steps and a cosine noise schedule. All code is based on PyTorch and tested on two GPUs: NVIDIA GeForce RTX2080Ti and NVIDIA TITAN RTX. More details can be found in the supplementary.

In the training phase, we learn the VQ-VAE and the two-level diffusion model. In the inference phase, given the conditional input, we first encode the input condition and feed it into the diffusion model to get the anchor points and anchor embeddings, which are fed into the decoder of VQ-VAE $\mathcal{D}$ to generate 3D clothes corresponding to the input condition.

4. Experiments

4.1. Datasets

We employ the Cloth3D dataset [2] to validate our model. This dataset offers an extensive collection of 3D clothing meshes stored as quadrilateral meshes, each with distinct topology and metadata such as a category, texture, and com-

patible body shape (expressed as SMPL parameters). 3D clothes items are classified into six categories: dress, jumpsuit, t-shirt, top, trousers, and skirt. The training set comprises 8634 meshes, while the testing set contains 1345. To better evaluate our model, we thicken the mesh. Since we aim to develop a model for various generation tasks, we further render the mesh into a 2D image with a resolution of 512^2 by Blender [12] and register the SMPL model with our processed meshes. The preprocessing details are present in the supplementary. We optimize the clothing VQ-VAE and generative latent-anchor diffusion (LAD) model on the training set and assess their performance on the test set.

4.2. Experimental Setup

Baselines. We employ three baselines in our study: DPC [47], PVD [107], and 3DILG [102]. It is important to note that DPC and PVD are designed for point-cloud generation tasks, where the point scale is too small to generate a mesh. Therefore, we modify them slightly to learn about the latent-anchor distribution, the same distribution learned by our LAD, and provide the same pre-trained clothing VQ-VAE to encode 3D clothes into latent anchors and reconstruct the final 3D clothing for a fair comparison. Additionally, we select 3DILG as another baseline to investigate the effectiveness of the auto-regressive scheme and diffusion process for fitting the latent-anchor distribution. Similarly, the auto-regressive model learns on the latent-anchor representation, and both models decode 3D clothing using the sample clothing VQ-VAE.

Metrics. Following previous works [5, 36, 99], we employ chamfer distance (CD) and earth mover's distance (EMD) for evaluation. As described in PVD [107], we calculate 1-nearest neighbor (1-NN) to assess generative quality in the category-conditioned generation task. Notably, a 1-NN score closer to 50 indicates better quality. We report the chamfer distance on image-conditioned generation results, where a lower value indicates superior quality.

4.3. Results and Analysis

Characteristic Comparison We compare IG [92], Cloth3d [2], GGUSPI [80], PBM [100], SMPLicit [13], and NSM [11], in terms of characteristics with our method. As listed in Table 1, all baselines attempt to bypass operating mesh by leveraging explicit representations, such as registering clothes on human body or projecting onto human body-aligned sewing patterns and UV coordinates. However, these methods require generating the 3D clothes binding with a given human body. Moreover, some methods require heavy computation resources due to executing on dense point-cloud. This characteristic further prevents the extension of these methods, which runs counter to developing a versatile generative model. In contrast, our approaches represent 3D clothes in a flexible and topology-free manner. Besides, the latent-anchor representation has a lot of underlying applications related to 3D clothes, like the synthesis of texture 3D clothes or simulation of the clothes with motion sequences, due to its topology-free properties and plug-and-play characteristics.

Characteristics	Topology Free	Body Shape	Sewing Pattern	Image	Prompt
IG [92]	Limited	Required	Required	Limited	No
Cloth3d [2]	No	Required	No	No	No
GGUSPI [80]	Yes	Required	Required	Limited	No
PBM [100]	Limited	Required	No	Support	No
SMPLicit [13]	Limited	Required	No	No	No
NSM [11]	Yes	No	Required	Support	No
Ours	Yes	No	No	Support	Support

Table 1. **Characteristic comparisons.** Most baselines require generating 3D clothes with human bodies or expensive computation resources since they heavily rely on representing 3D clothes by body shapes, registered point clouds, or sewing patterns. In contrast, our proposed latent-anchor representation is topology-free and computation-friendly, which could handle up to three modality conditions, like specific category, image, and prompt.

	Dress		Jumpsuit		Tshirt		Top		Trousers		Skirt	
	CD	EMD	CD	EMD	CD	EMD	CD	EMD	CD	EMD	CD	EMD
DPC [47]	100.00	94.58	99.05	92.98	100.00	97.77	100.00	98.23	100.00	98.09	99.05	93.69
PVD [107]	93.83	97.22	92.05	96.67	91.50	92.83	98.96	97.97	96.52	95.45	98.68	100
3DILG [102]	85.08	84.21	**80.64**	84.55	85.27	**82.25**	83.87	84.55	90.47	**84.12**	**82.83**	**84.33**
Ours	**82.4**	**80.81**	87.09	**81.45**	**85.00**	82.57	**83.73**	**83.05**	**86.44**	87.70	93.54	88.18

Table 2. **Quantitative Comparison for Category-Conditioned Generation.** This table showcases a numerical comparison of the 1-NN accuracy between the LAD and baselines across each conditioning category, with the 1-NN accuracy indicating general shape quality. LAD substantially improves over the first two baselines while partially outperforming 3DILG.

Category-Conditioned Generation Comparison We present the quantitative results for each method in the category-conditioned generation task in Table 2. Our transformer-based two-level diffusion model achieves superior generative quality compared to the two diffusion paradigms, DPC [47] and PVD [107]. Furthermore, our LAD model outperforms, in most cases compared to the auto-regressive approach 3DILG [102]. We analyze the result based on the dataset's statistics. Compared to the 2037 dress meshes in the training set, there are only 468 skirt meshes. As the diffusion model necessitates substantial training data to learn the Markov process, training with limited data might negatively affect performance. Nevertheless, our LAD excels in most categories. The distinction between the qualitative results of our method and 3DILG [102] is evident, as illustrated in Figure 3. Although 3DILG generates clothing with smooth surfaces, it exhibits imperfections in local patches. Auto-regressive schemes predict subsequent latent anchors based on previous predictions, introducing uncertainty and resulting in error accumulation, which may cause noisy patches to appear near the end of the inference process or even earlier. Another piece

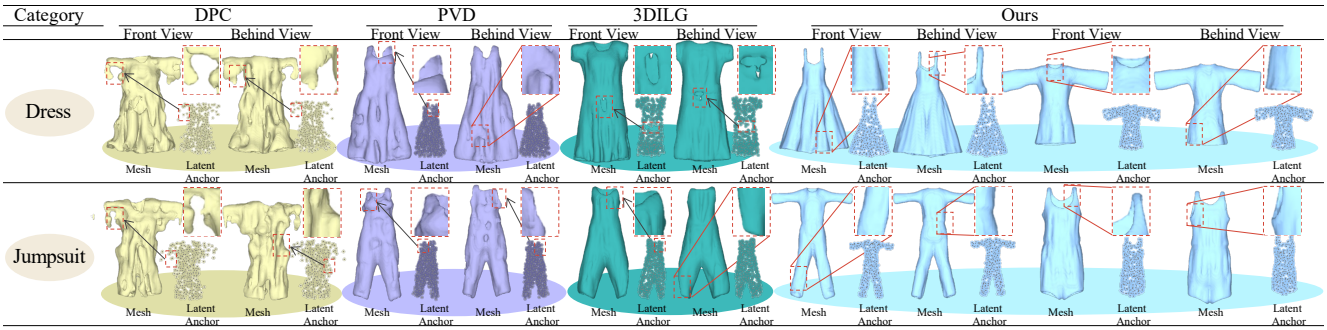

Figure 3. **Qualitative Comparison for Category-Conditioning Generation**. We present rendered images of the generated meshes for visual comparisons, displaying each mesh from the front and back views. Additionally, we showcase the latent anchor by its coordinates, which directly correspond to the distortion observed in the mesh. From left to right, we show results from DPC, PVD, 3DILG, and our LAD. The first two diffusion-based models generate only rough shapes, with the latent anchor associated with the defective areas on the mesh. 3DILG outperforms the previous methods, producing plausible shapes that are also evident in the latent anchor. However, it fails to create a smooth surface due to the non-uniform nature of the generated anchor (visualized in the magnifier). In contrast, our LAD generates elegant shapes with fine details, demonstrating superior performance in the comparison. Zoom in for more details.

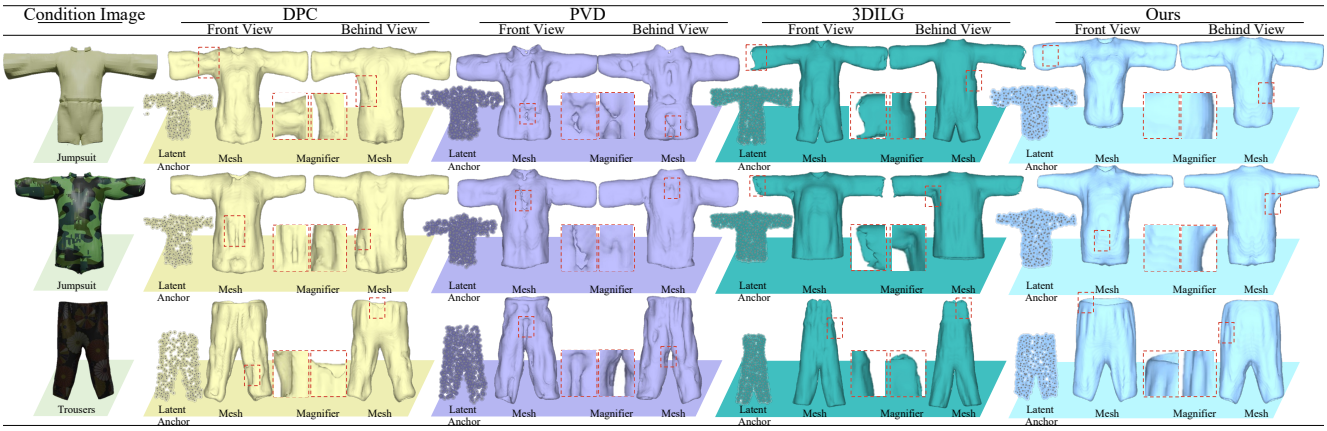

Figure 4. **Qualitative Comparison on Image-Conditioned Generation.** We select three cases to compare our results with other baselines. The condition images are displayed on the left. A notable difference from category-conditioned generation is that all baselines exhibit improved performance. However, the enhancement is minor, as DPC and PVD can only generate plausible shapes with fluctuating surfaces. At the same time, their unevenly generated latent anchors are also non-uniform (only the front view shown here). When guided by an image, 3DILG produces better results but exhibits a coarse boundary. In contrast, our method generates meshes with plausible shapes and smooth surfaces (visualized in the magnifier), outperforming the other approaches. Zoom in for more details.

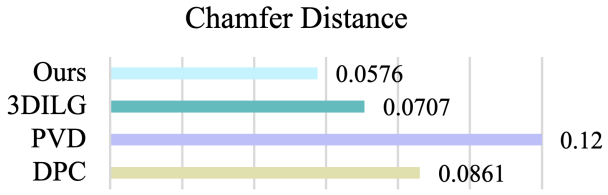

Figure 5. **Quantitative Comparison on Image-conditioned generation.** We employ chamfer distance to assess the results, and the value show that our method outperforms the others.

of evidence is the asymmetric latent anchors generated by 3DILG. In contrast, our LAD model predicts all latent anchors in a uniform and symmetric layout through an iterative denoising process. The other two diffusion-based methods [47, 107] generate meshes of inferior quality. A possible explanation is that these models do not account for low information density conditions during development. Consequently, when conditioned on the category embedding, the models struggle to fit the distribution of each category. As a result, only our LAD can produce 3D clothing with fine details.

Image-Conditioned Generation Comparison The

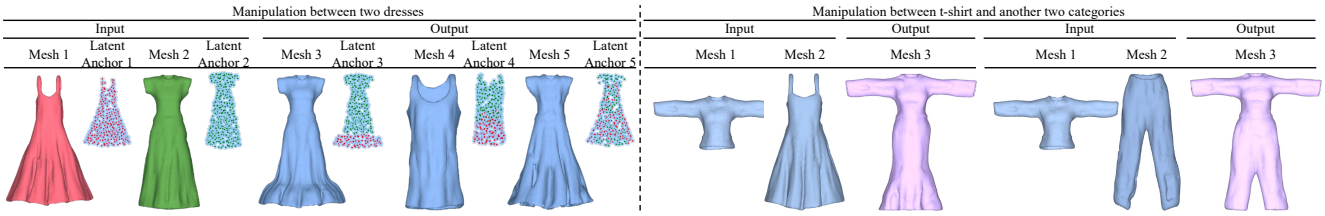

Figure 6. **Visualization for Editing 3D Clothes by Manipulating Latent Anchors.** The figure's left shows three ways to manipulate latent anchors for editing 3D clothes. The first and second columns under output indicate combining two latent anchors in sorted order, and the third column showcases combining randomly. The right figure shows the results of mixing clothes from different categories. The visual results demonstrate the flexibility of our latent-anchor representation.

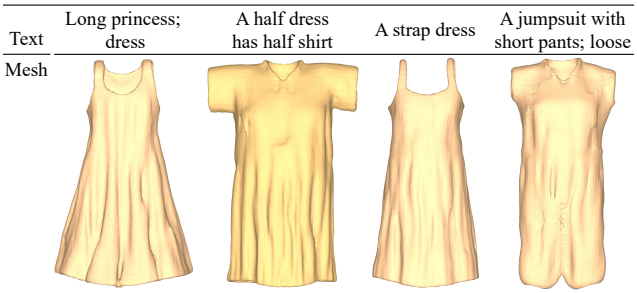

Figure 7. **Text-Conditioned Generation.** The figure presents the outcomes of the text-conditioned generation process. Based on the corresponding prompts, our model generates results that closely align with the semantic context. Zoom in for more details.

quantitative results in Table 5 indicate that our LAD model achieves the lowest chamfer distance and outperforms all other comparison methods. The qualitative comparison between LAD and baselines, as shown in Figure 4, supports this conclusion. When the condition type changes from low information density category embeddings to more informative image embeddings, the results of all baselines improve. However, upon closer examination, DPC and PVD can only generate plausible shapes, while 3DILG produces better results but still exhibits flaws in local patches. In contrast, our LAD generates accurate shapes with fine details preserved on the clothing surface.

Text-conditioned Generation. Benefiting from the CLIP, our LAD can directly employ a pre-trained text encoder. CLIP aligns its feature space between visual and textual inputs, enabling our LAD to generate 3D clothing using text input. We present visual results in Figure 7. These visual outcomes indicate that our model can produce results conform with the prompt input. However, due to the lack of text-3D clothes pair, we have to train our model with CLIP encoder as bridges, which limits the generative ability of our model. With the enlarging of 3D clothes database, these knotty problem will great waken.

Clothes Editing via Manipulating Latent Anchors.

	Commitment Loss (VQ-VAE)				KL-Divergence (VAE)			
	1024	512	256	128	1024	512	256	128
OverAll	89.08	**90.02**	85.95	76.00	89.76	88.90	85.57	78.72
Dress	86.19	**87.23**	82.17	69.64	86.77	85.81	81.91	72.65
Jumpsuit	89.98	**91.05**	87.72	79.44	90.53	90.09	87.27	81.57
Tshirt	90.20	**91.63**	88.41	80.02	90.85	90.68	87.84	82.61
Trousers	90.32	**90.66**	86.49	76.47	91.08	89.58	85.99	79.37
Top	90.38	**91.46**	87.41	79.43	**91.46**	90.61	87.32	81.80
Skirt	85.03	85.22	78.88	62.14	**85.86**	83.39	79.02	67.22

Table 3. **Ablation Study.** We perform an ablation study on the auto-encoder architecture utilized for latent-anchor representation, primarily investigating the number of latent anchors and the objective. The results span six categories, suggesting that a VQ-VAE with 512 latent anchors constitutes the optimal architecture.

Our latent-anchor representation enables the model to manipulate the generated mesh. As illustrated in the left of Figure 6, the model produces two groups of latent anchors, we pick partial latent anchors from each one and merge the picked as a new latent anchor, where the expressive latent-anchor representation ensures qualified results. Moreover, it also ensures cross-category manipulation, demonstrated in the right from Figure 6. Due to the decoder reconstructing the generated latent anchor without a claimed specific category, we can first mix latent anchors from two categories, use the target mesh category as a condition for generating a group of anchor features, and then send them to the decoder for generating the target 3D clothes mesh.

4.4. Ablation

We examine auto-encoder architectures 3.1 by investigating the impact of the anchor point's number and objectives. We assess each combination's performance through reconstruction tasks on the test set with the Intersection over Union (IoU) metric, where a higher IoU signifies a better performance. Based on Table3 results, we develop our clothing in the VQ-VAE approach and employ 512 latent anchors for subsequent generations.

5. Conclusion

We propose Magic Tailor for generating 3D clothes that accommodate various conditions, including clothes categories, images, and textual descriptions. Our framework employs a VQ-VAE module to encode diverse 3D clothing shapes within a compatible latent-anchor representation and learns a two-level diffusion model over the latent anchors' space, facilitating the efficient mapping of various conditions to the 3D clothing space. Comprehensive experiments on multi-modal conditioned 3D clothing generation tasks demonstrate the effectiveness of our proposed framework.

6. Limitations

This paper provides preliminary evidence of the feasibility of the latent-anchor representation, but there are still two directions worth exploring. The first aspect concerns the modeling of 3D clothes. This paper uses watertight mesh as a trade-off to investigate the approach better. A meaningful future research direction would be to combine this method with non-watertight mesh [7, 21, 41, 42, 53, 78, 79] for 3D clothes generation. Another direction is the generation of textured 3D clothes, which would enhance the practical value of the resulting models.

Magic Tailor: a Latent-Anchor based Diffusion Model for 3D Clothes Generation

Supplementary Material

In the supplementary, we describe the data preparation in section A and the Training details of Magic Tailor B. Furthermore, extensive visual results are illustrated in section C.

A. Data Preparation

We follow DualOctreeGNN [91] to convert the raw mesh from Cloth3D into watertight mesh to train the neural occupancy field and normalize all vertices inside $[-1, 1]$. We utilize the pysdf[3] to compute a groundtruth occupancy for the query point. Moreover, we pre-sample the surface points, query points, and labels of query points to accelerate training.

B. Training Details

While training the clothing VQ-VAE, we sampled surface points $P_s \in \mathbb{R}^{N\times 3}$ with $N = 2048$ as input and 2048 query points with their label as supervisions. The number of anchor points $P_c \in \mathbb{R}^{M\times 3}$ is $M = 512$, the codebook $\mathbb{Z} = \{z_j \in \mathbb{R}^D\}_{j=1}^{J}$ stores $J = 1024$ discrete embedding in dimension $D = 256$. The AdamW optimizer has a 1e-3 learning rate.

While training the two-level Latent-anchor-based Diffusion Model (LAD), we set eight transformer encoder layers with 512 latent dimensions for both point- and embedding-denoisers, and the sequence length is 512. Furthermore, each denoiser has an AdamW optimizer with a 1e-4 learning rate.

C. More Visual Results

We provide more visual results to demonstrate the quality of the generated clothes in the following.

C.1. Category-conditioned 3D Clothes Generation

Dress. Figure 8, 9, and 10 show the results conditioning on the category "Dress".

Jumpsuit. Figure 11, 12, and 13 show the results conditioning on the category "Jumpsuit".

Tshirt. Figure 14 shows the results conditioning on the category "Tshirt".

Trousers. Figure 15 shows the results conditioning on the category "Trousers".

Top. Figure 16 shows the results conditioning on the category "Top".

Skirt. Figure 17 shows the results conditioning on the category "Skirt".

[3]https://github.com/andreasBihlmaier/pysdf

C.2. Image-conditioned 3D clothes generation

Figure 18, 19, and 20 show the results conditioning on the image.

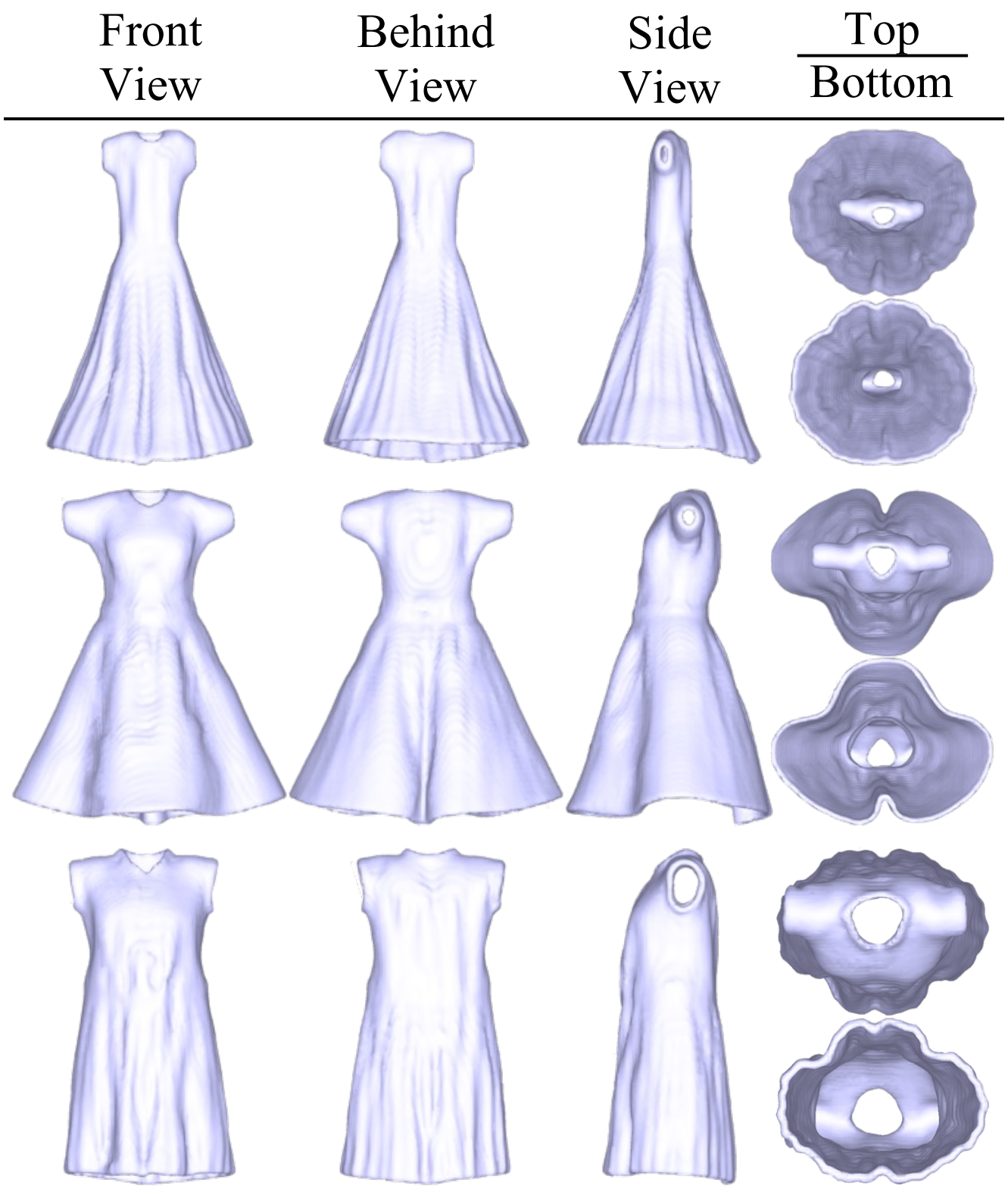

Figure 8. Conditional generation on category **Dress**.

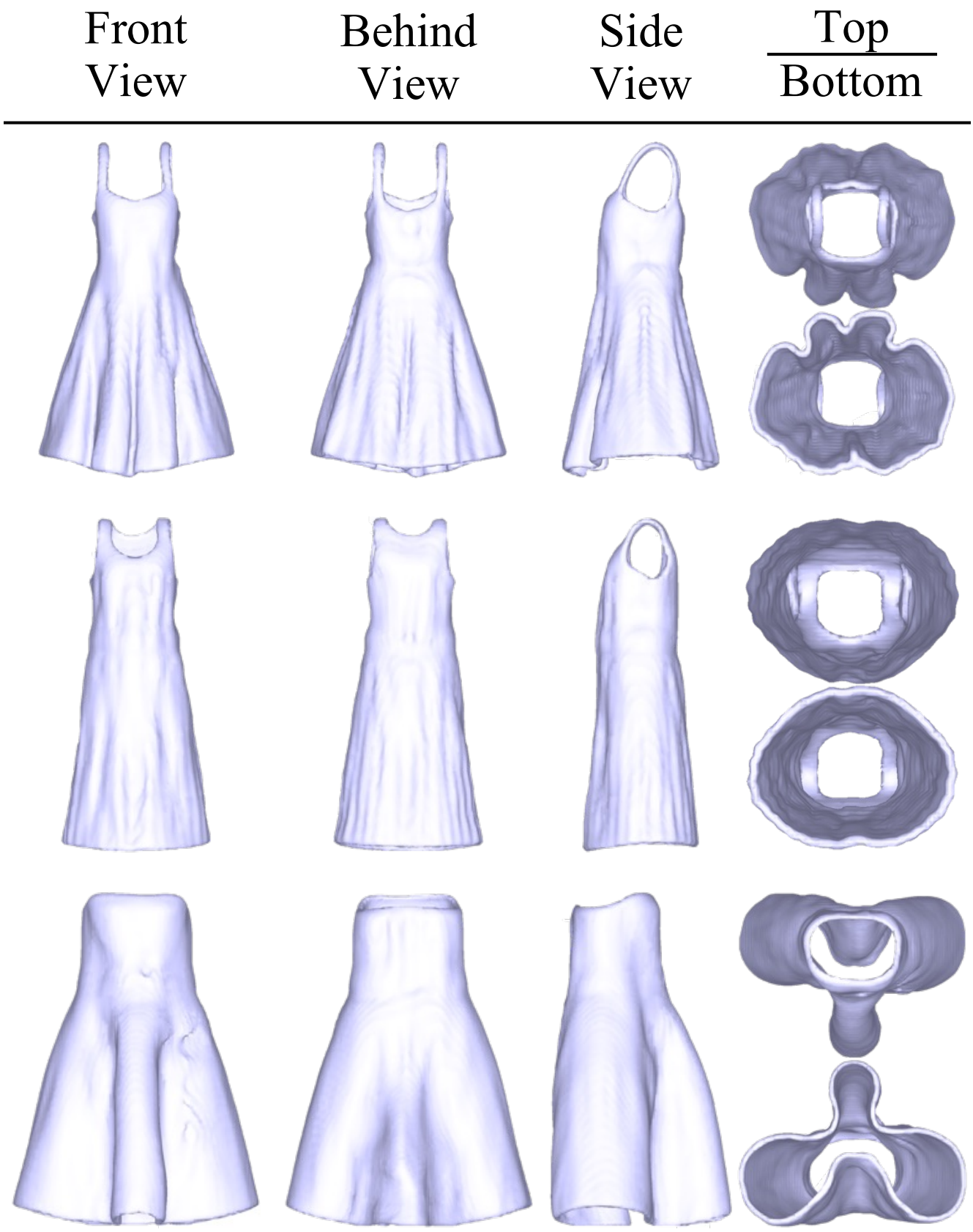

Figure 9. Conditional generation on category **Dress**.

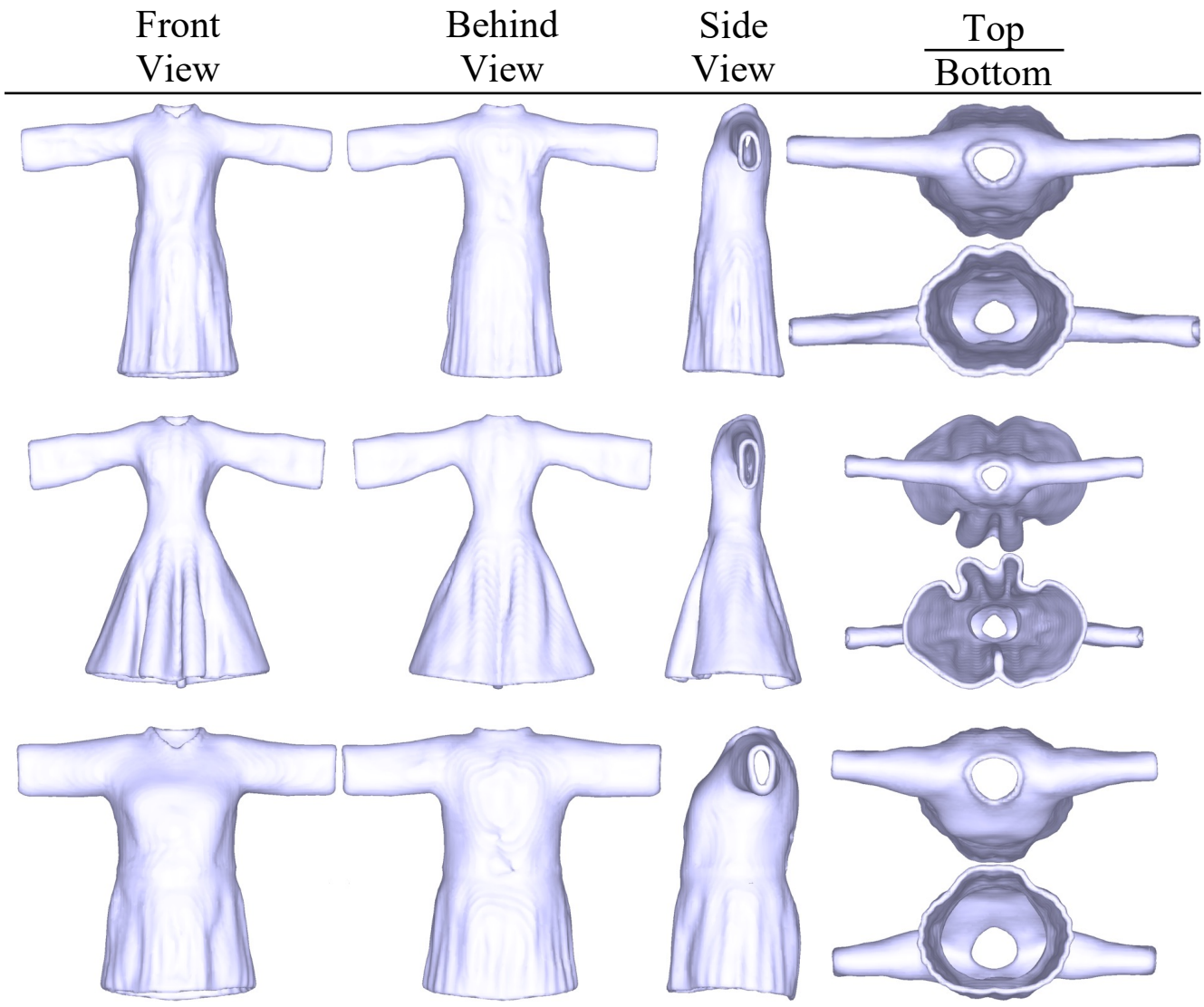

Figure 10. Conditional generation on category **Dress**.

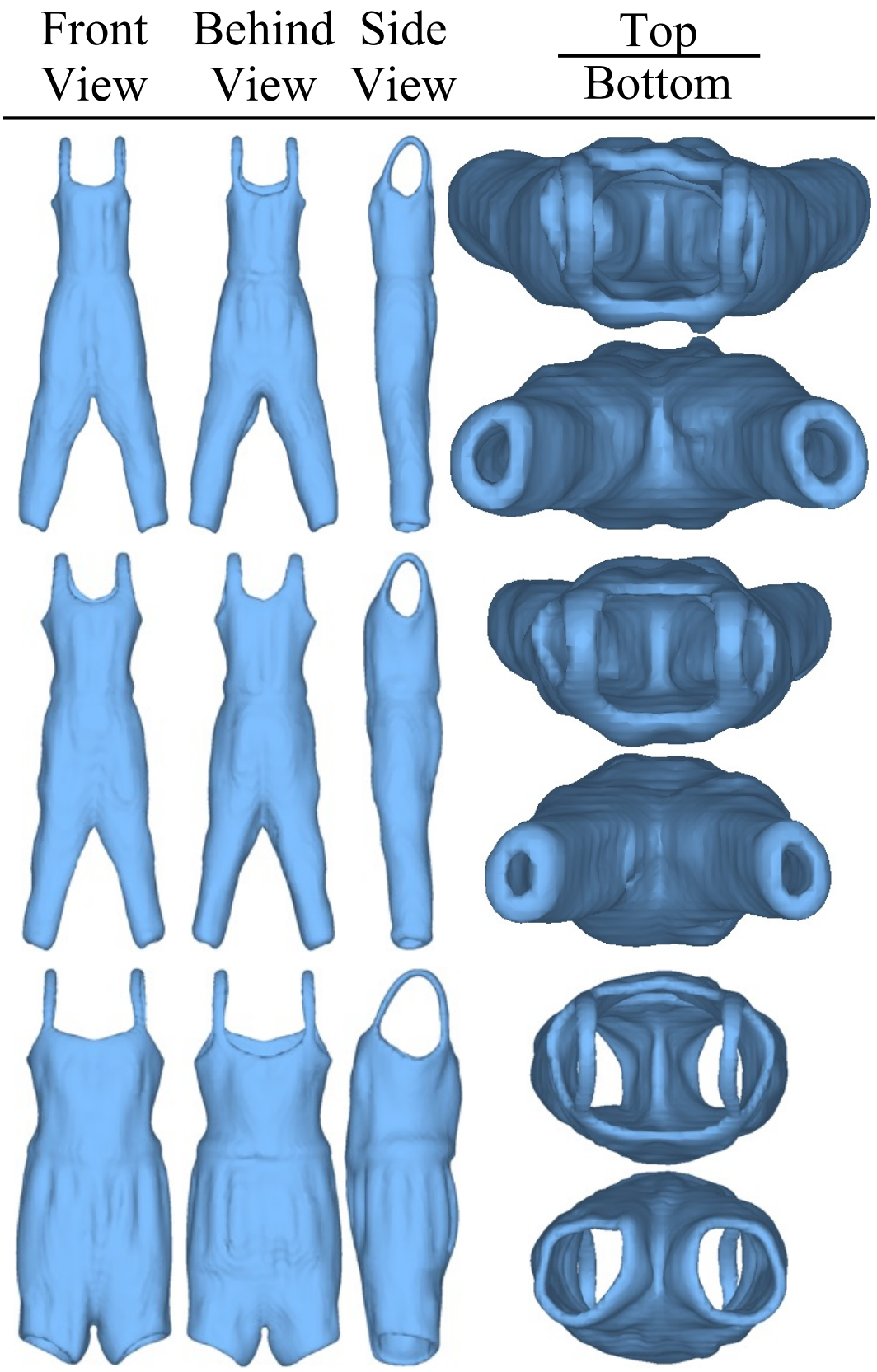

Figure 11. Conditional generation on category **Jumpsuit**.

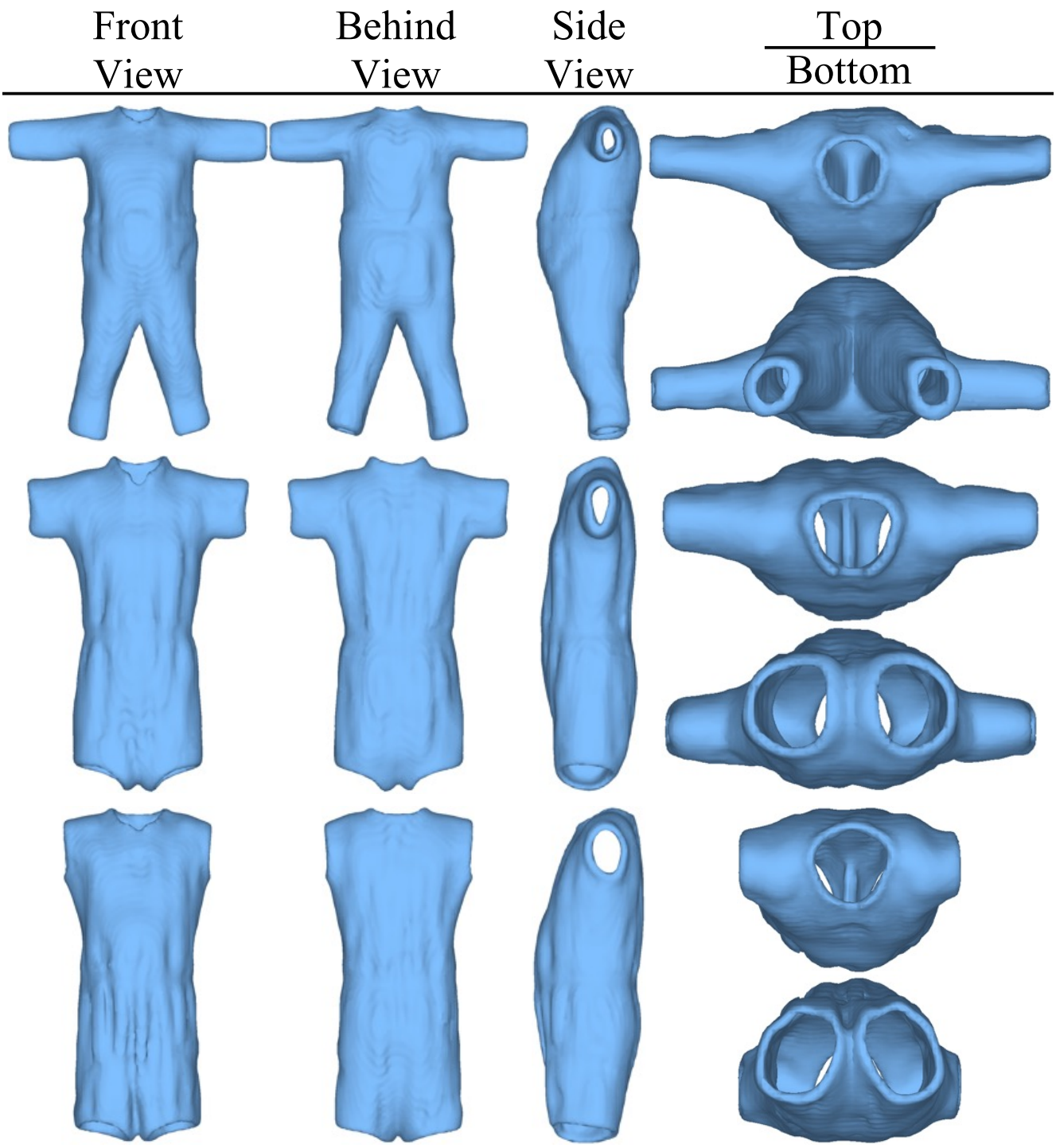

Figure 12. Conditional generation on category **Jumpsuit**.

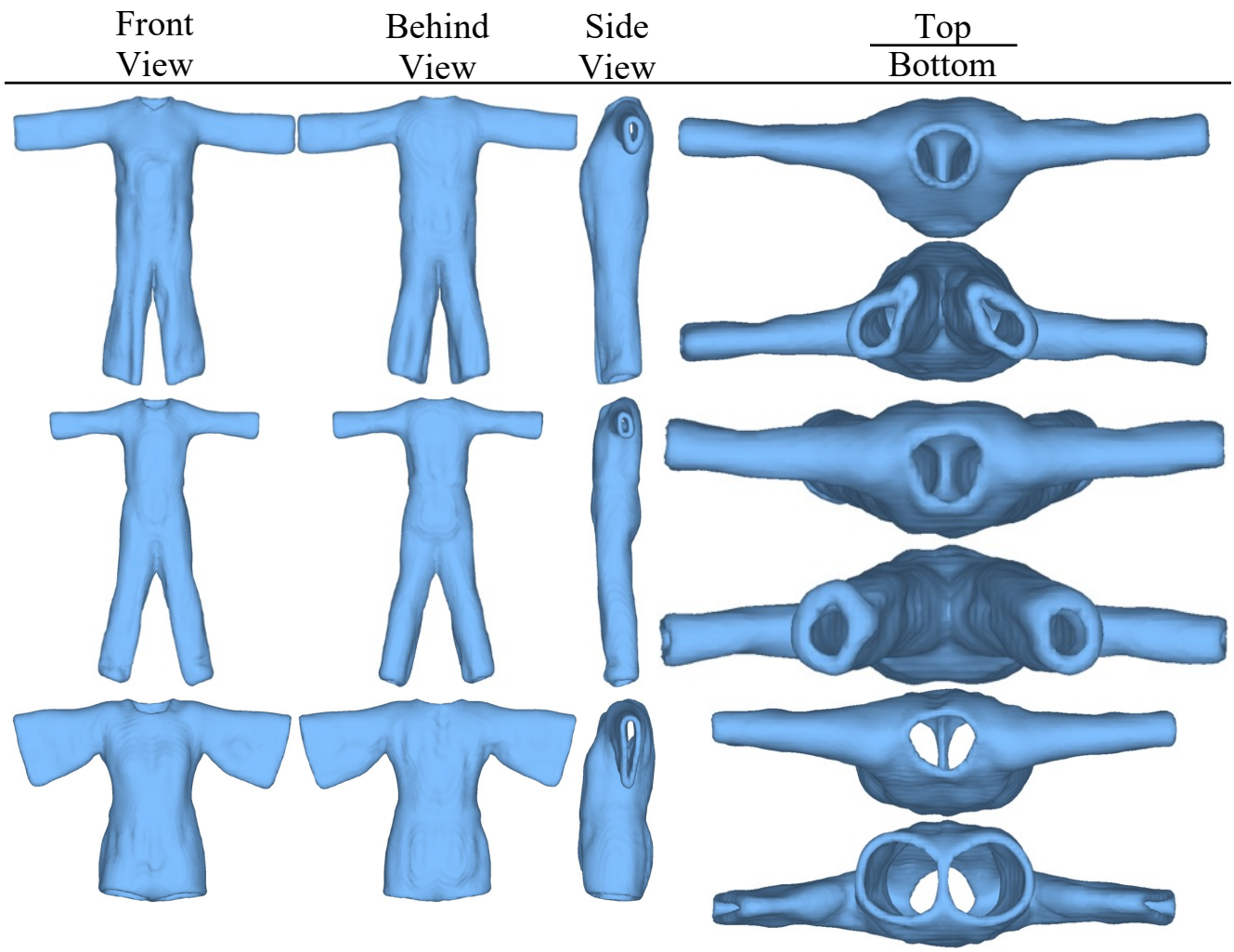

Figure 13. Conditional generation on category **Jumpsuit**.

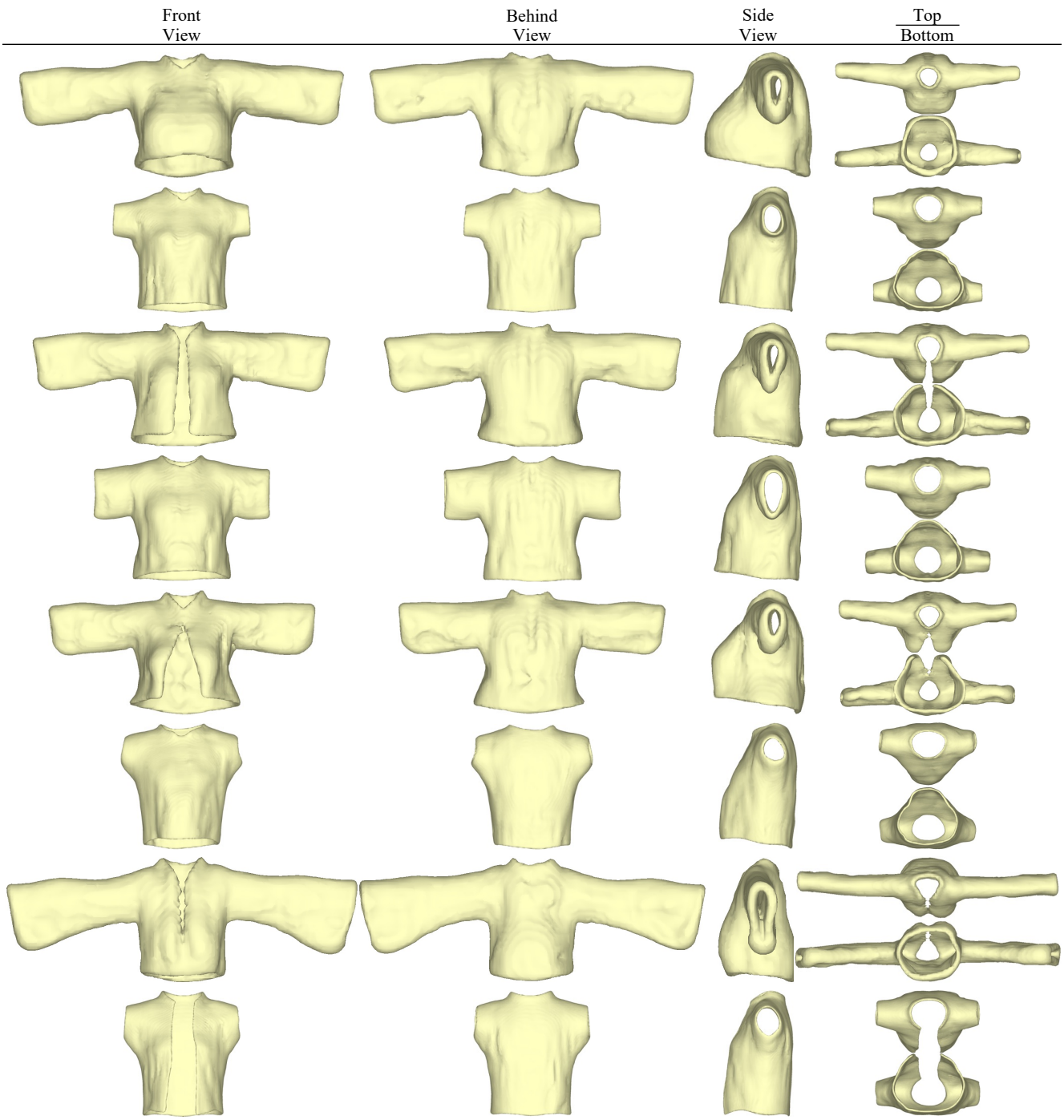

Figure 14. Conditional generation on category **Tshirt**.

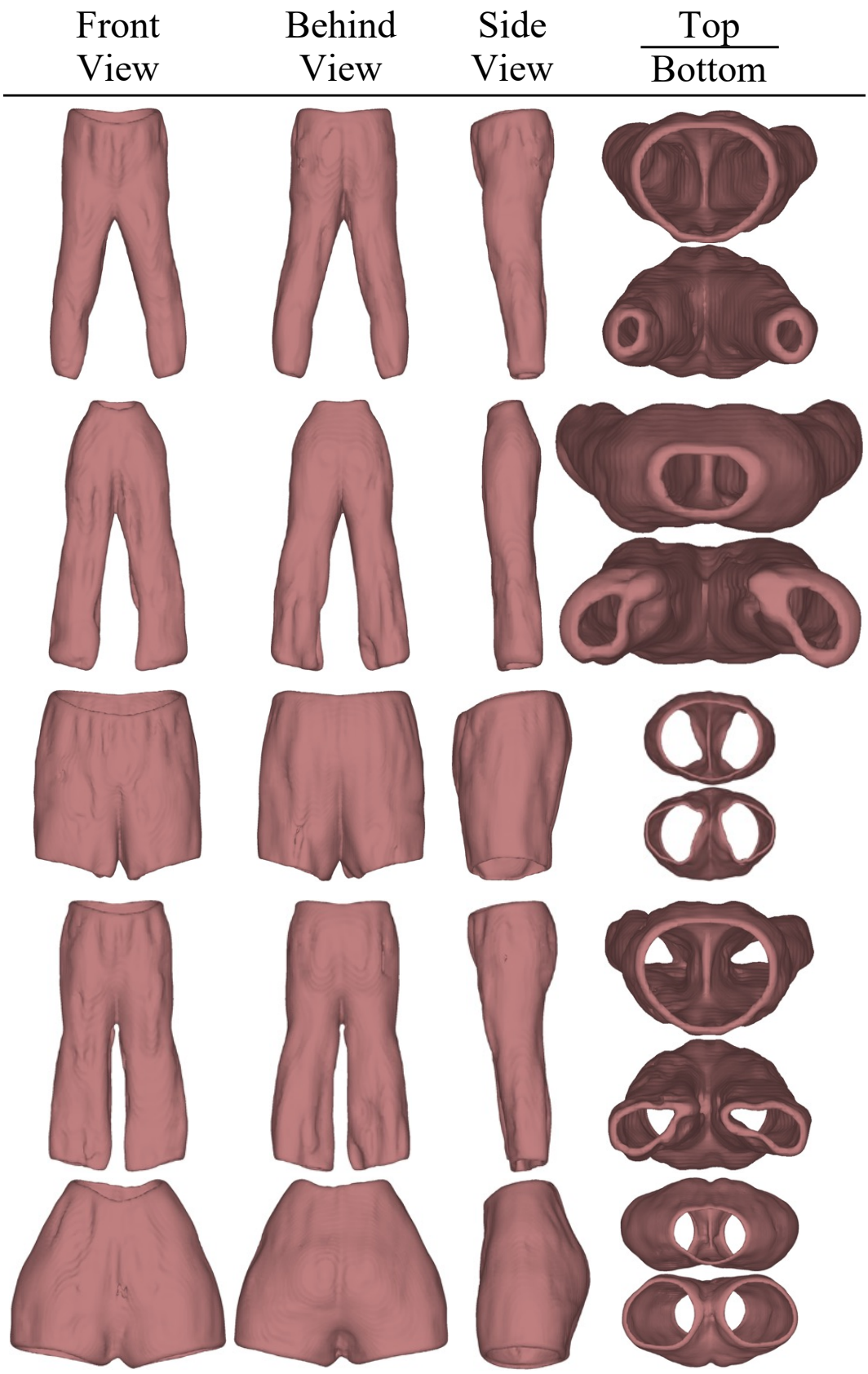

Figure 15. Conditional generation on category **Trousers**.

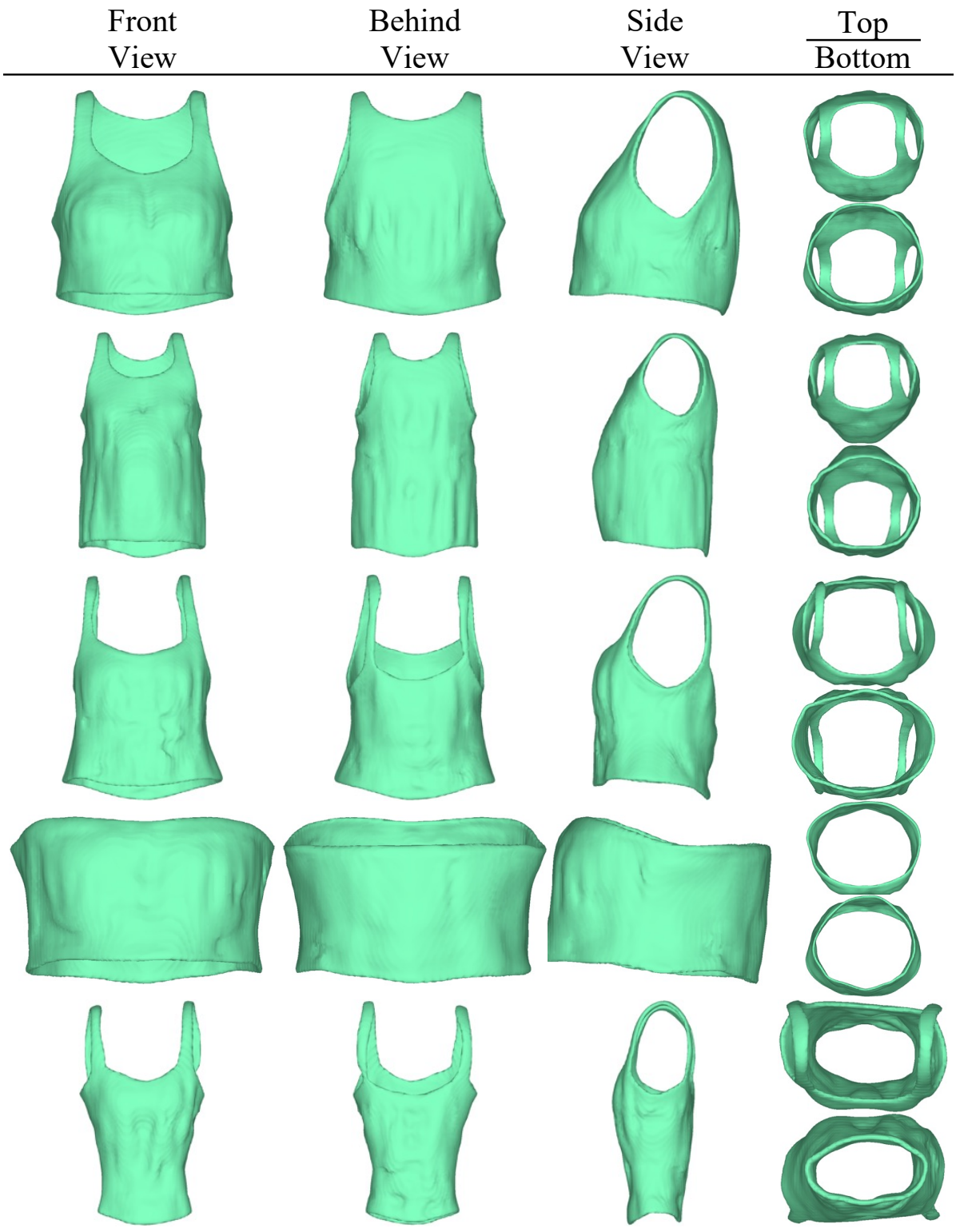

Figure 16. Conditional generation on category **Top**.

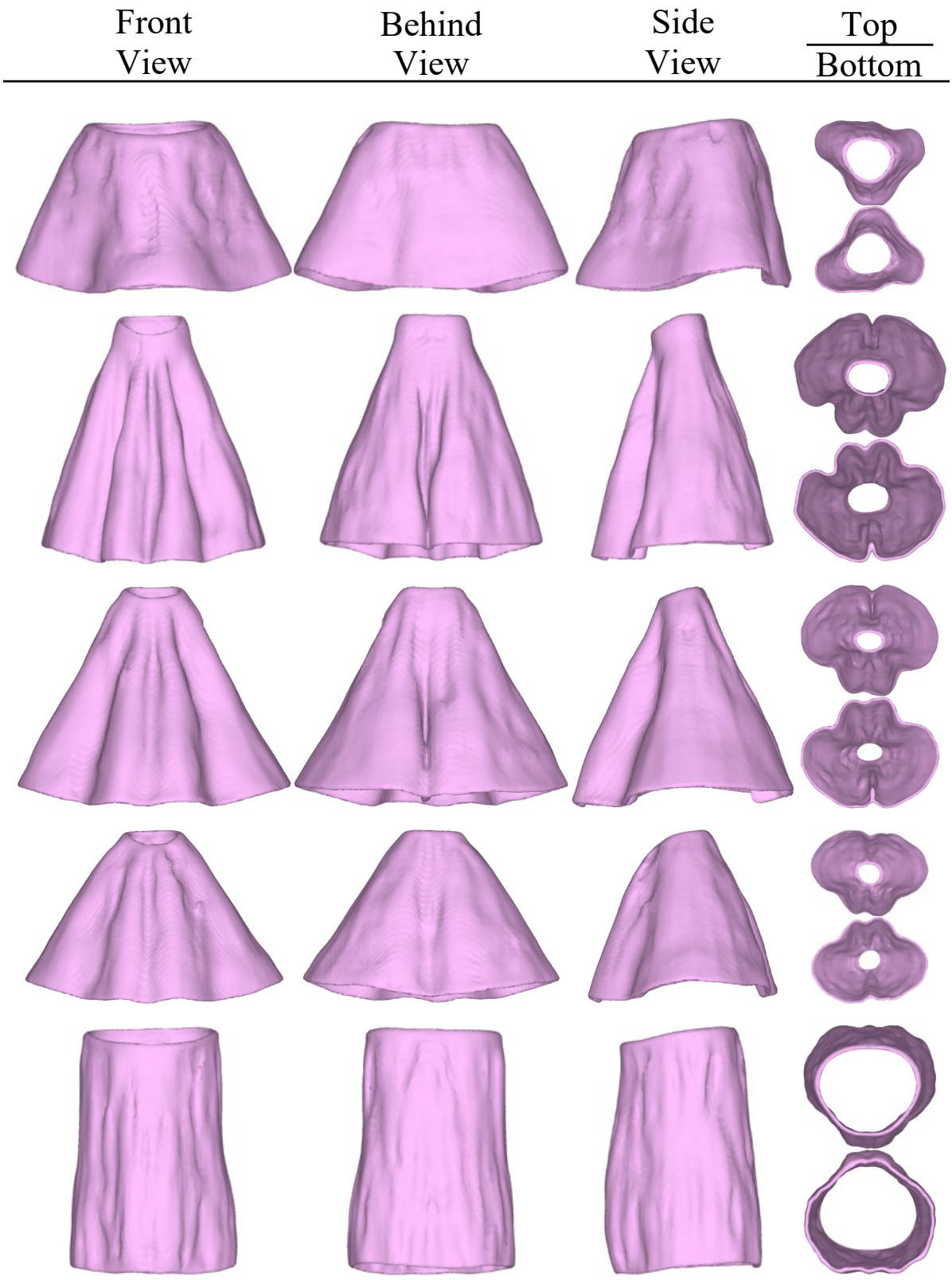

Figure 17. Conditional generation on category **Skirt**.

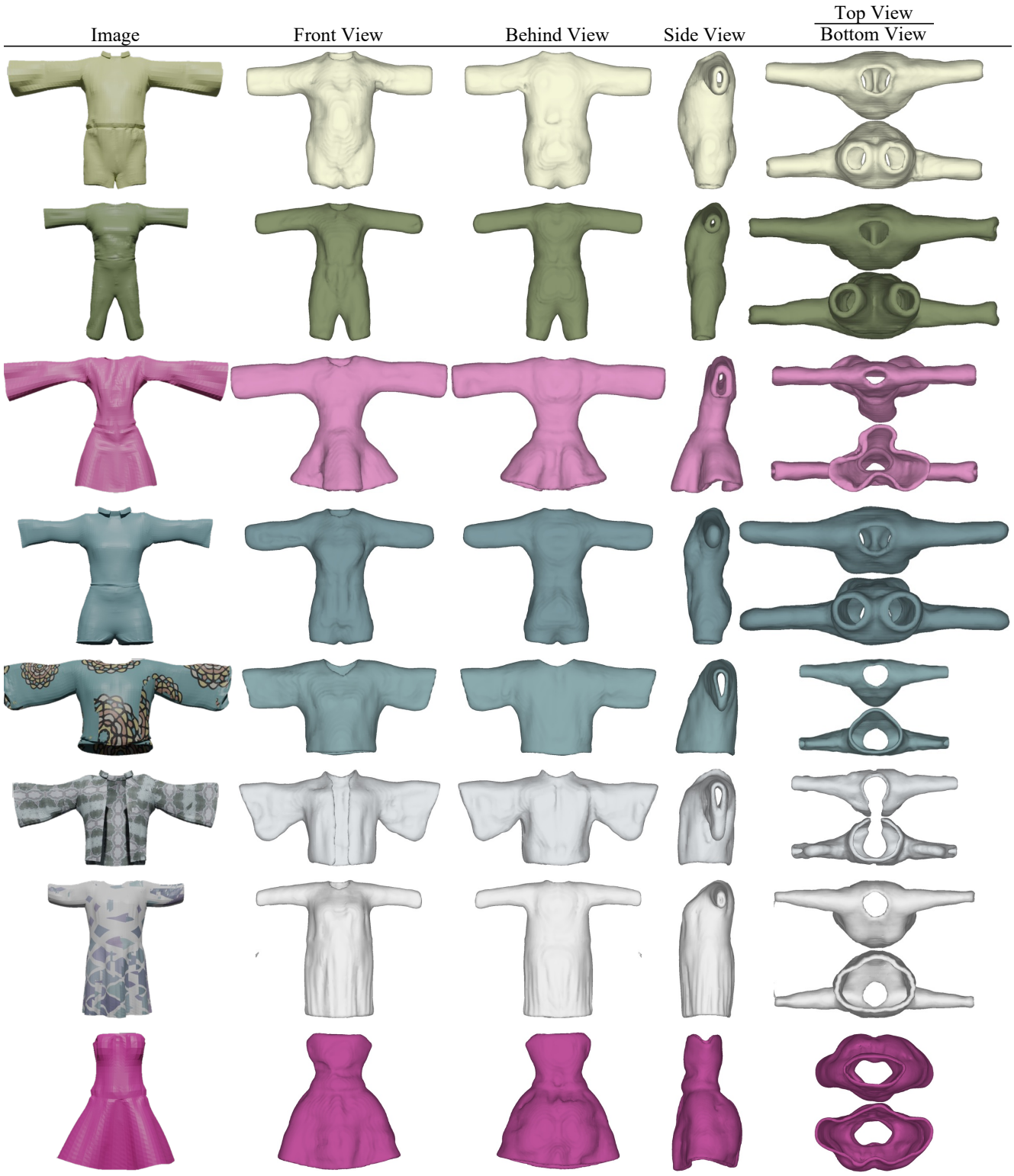

Figure 18. **Images** conditioning generation.

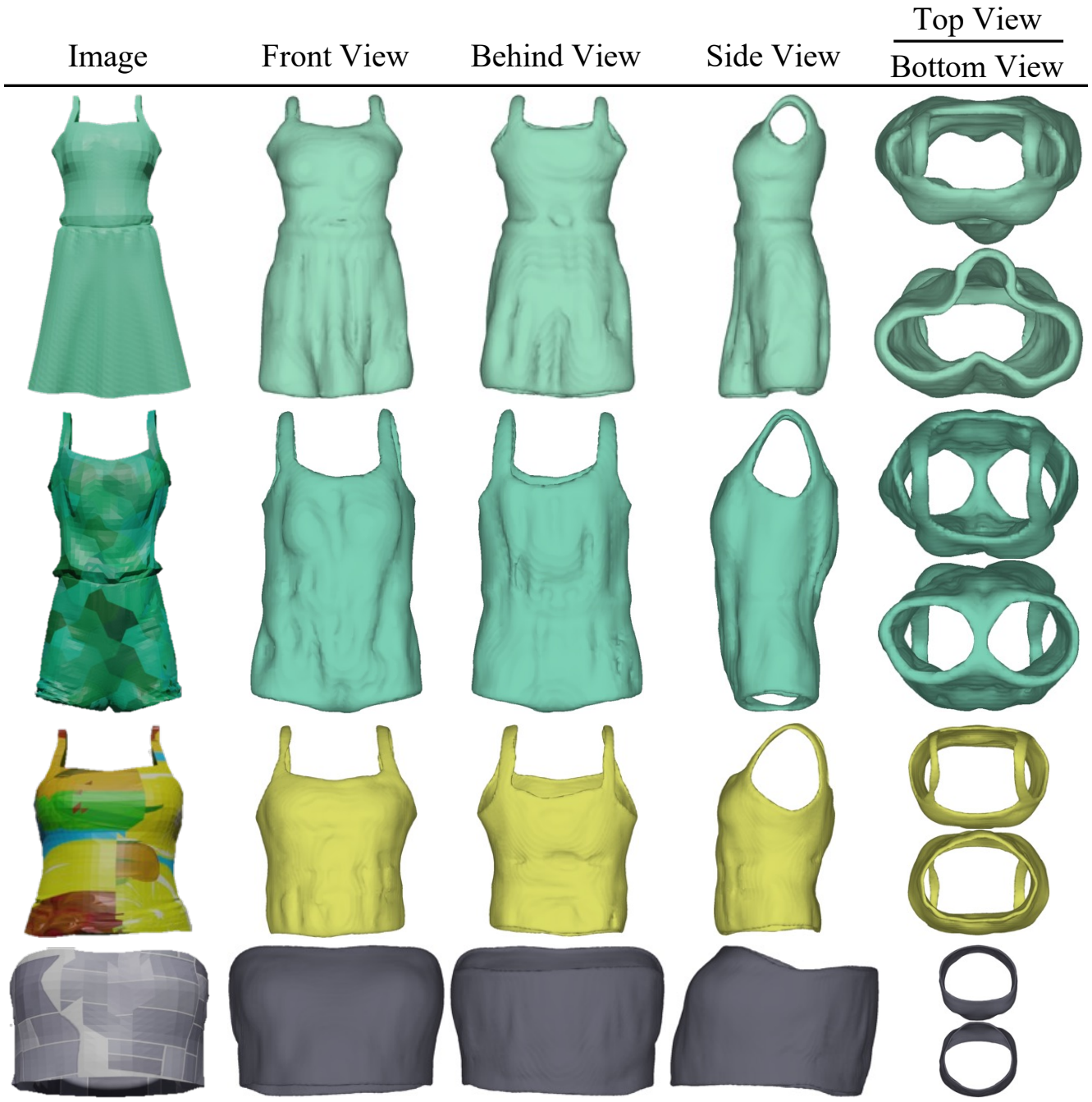

Figure 19. **Images** conditioning generation.

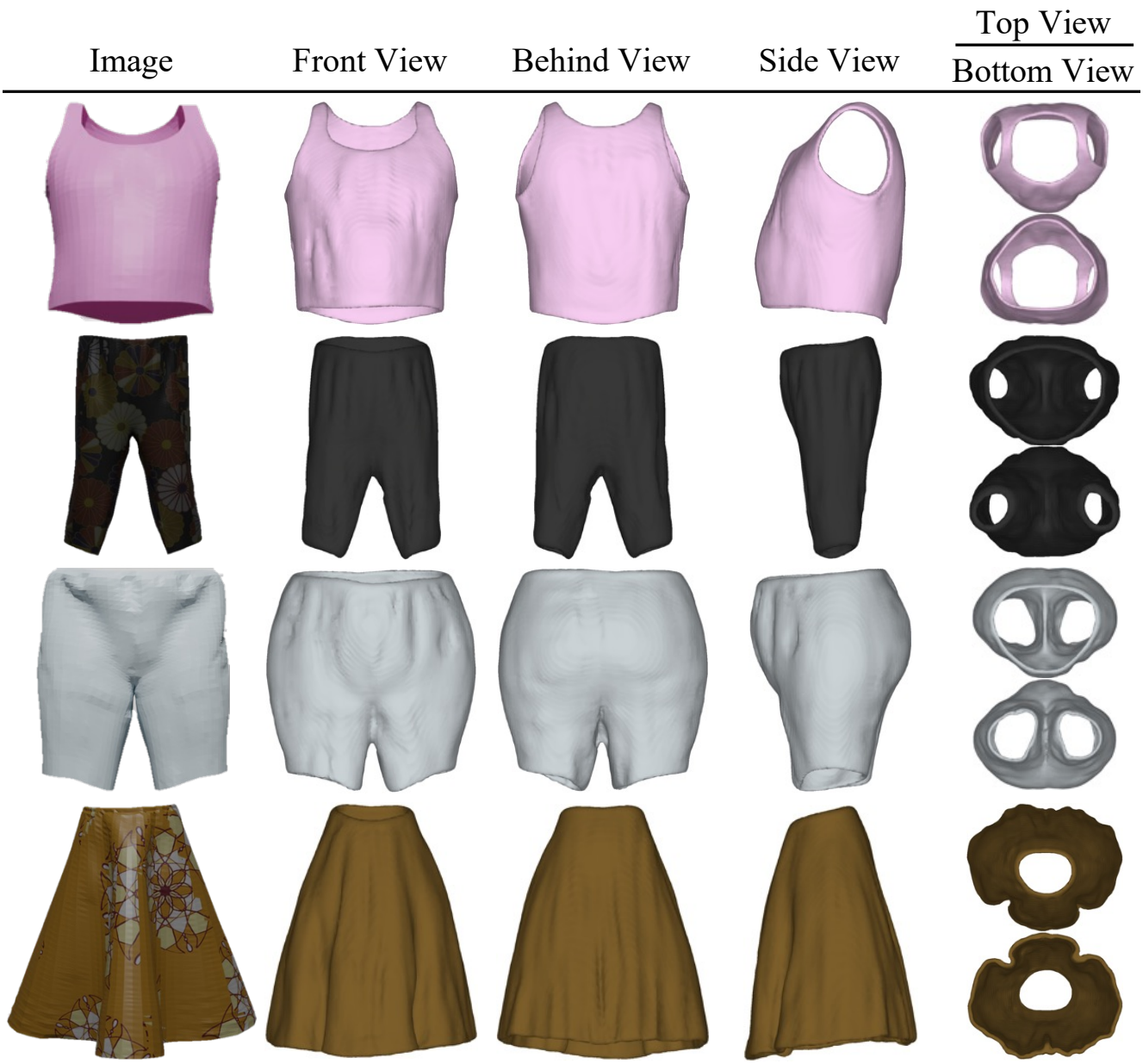

Figure 20. **Images** conditioning generation.

References

[1] Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabian Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. Driving-signal aware full-body avatars. *ACM Transactions on Graphics (TOG)*, 40(4):1–17, 2021. 2

[2] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. Cloth3d: clothed 3d humans. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 344–359. Springer, 2020. 1, 2, 5, 6

[3] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5420–5430, 2019. 2

[4] Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. Generative and discriminative voxel modeling with convolutional neural networks. *arXiv preprint arXiv:1608.04236*, 2016. 2

[5] Ruojin Cai, Guandao Yang, Hadar Averbuch-Elor, Zekun Hao, Serge Belongie, Noah Snavely, and Bharath Hariharan. Learning gradient fields for shape generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 6

[6] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 3

[7] Weikai Chen, Cheng Lin, Weiyang Li, and Bo Yang. 3psdf: Three-pole signed distance function for learning surfaces with arbitrary topologies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18522–18531, 2022. 9

[8] Xiaowu Chen, Bin Zhou, Feixiang Lu, Lin Wang, Lang Bi, and Ping Tan. Garment modeling with a depth camera. *ACM Transactions on Graphics (TOG)*, 34(6):1–12, 2015. 3

[9] Xin Chen, Anqi Pang, Wei Yang, Peihao Wang, Lan Xu, and Jingyi Yu. Tightcap: 3d human shape capture with clothing tightness field. *ACM Transactions on Graphics (TOG)*, 41(1):1–17, 2021. 2

[10] Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J Black, Andreas Geiger, and Otmar Hilliges. gdna: Towards generative detailed neural avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20427–20437, 2022. 2

[11] Xipeng Chen, Guangrun Wang, Dizhong Zhu, Xiaodan Liang, Philip HS Torr, and Liang Lin. Structure-preserving 3d garment modeling with neural sewing machines. *arXiv preprint arXiv:2211.06701*, 2022. 2, 6

[12] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 6

[13] Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11875–11885, 2021. 2, 6

[14] Luca De Luigi, Ren Li, Benoît Guillard, Mathieu Salzmann, and Pascal Fua. Drapenet: Generating garments and draping them with self-supervision. *arXiv preprint arXiv:2211.11277*, 2022. 1

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

[16] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3

[17] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 2

[18] Daiheng Gao, Xu Chen, Xindi Zhang, Qi Wang, Ke Sun, Bang Zhang, Liefeng Bo, and Qixing Huang. Cloth2tex: A customized cloth texture generation pipeline for 3d virtual try-on. *arXiv preprint arXiv:2308.04288*, 2023. 2

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2, 5

[20] Peng Guan, Loretta Reiss, David A Hirshberg, Alexander Weiss, and Michael J Black. Drape: Dressing any person. *ACM Transactions on Graphics (ToG)*, 31(4):1–10, 2012. 1

[21] Benoit Guillard, Federico Stella, and Pascal Fua. Meshudf: Fast and differentiable meshing of unsigned distance field networks. In *European Conference on Computer Vision*, pages 576–592. Springer, 2022. 9

[22] Erhan Gundogdu, Victor Constantin, Amrollah Seifoddini, Minh Dang, Mathieu Salzmann, and Pascal Fua. Garnet: A two-stream network for fast and accurate 3d cloth draping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8739–8748, 2019. 2

[23] Erhan Gundogdu, Victor Constantin, Shaifali Parashar, Amrollah Seifoddini, Minh Dang, Mathieu Salzmann, and Pascal Fua. Garnet++: Improving fast and accurate static 3d cloth draping by curvature loss. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):181–195, 2020. 2

[24] Oshri Halimi, Fabian Prada, Tuur Stuyck, Donglai Xiang, Timur Bagautdinov, He Wen, Ron Kimmel, Takaaki Shiratori, Chenglei Wu, and Yaser Sheikh. Garment avatars: Realistic cloth driving using pattern registration. *arXiv preprint arXiv:2206.03373*, 2022. 2

[25] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *Advances in Neural Information Processing Systems*, 33:9276–9287, 2020. 2

[26] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 4

[28] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[29] Fangzhou Hong, Liang Pan, Zhongang Cai, and Ziwei Liu. Garment4d: Garment reconstruction from point cloud sequences. *Advances in Neural Information Processing Systems*, 34:27940–27951, 2021. 2

[30] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134, 2023. 3

[31] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3

[32] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[33] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 3

[34] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022. 2

[35] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[36] Roman Klokov, Edmond Boyer, and Jakob Verbeek. Discrete point flow networks for efficient point cloud generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 694–710. Springer, 2020. 6

[37] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*. 2

[38] Maria Korosteleva and Sung-Hee Lee. Generating datasets of 3d garments with sewing patterns. *arXiv preprint arXiv:2109.05633*, 2021. 1

[39] Ren Li, Benoît Guillard, Edoardo Remelli, and Pascal Fua. Dig: Draping implicit garment over the human body. In *Proceedings of the Asian Conference on Computer Vision*, pages 2780–2795, 2022. 2

[40] Siyou Lin, Hongwen Zhang, Zerong Zheng, Ruizhi Shao, and Yebin Liu. Learning implicit templates for point-based clothed human modeling. In *European Conference on Computer Vision*, pages 210–228. Springer, 2022. 2

[41] Yu-Tao Liu, Li Wang, Jie Yang, Weikai Chen, Xiaoxu Meng, Bo Yang, and Lin Gao. Neudf: Leaning neural unsigned distance fields with volume rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 237–247, 2023. 9

[42] Zhen Liu, Yao Feng, Yuliang Xiu, Weiyang Liu, Liam Paull, Michael J Black, and Bernhard Schölkopf. Ghost on the shell: An expressive representation of general 3d shapes. *arXiv preprint arXiv:2310.15168*, 2023. 9

[43] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 2

[44] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998. 4

[45] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems*. 2

[46] Andrew Luo, Tianqin Li, Wen-Hao Zhang, and Tai Sing Lee. Surfgen: Adversarial 3d shape synthesis with explicit surface discriminators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16238–16248, 2021. 2

[47] 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, 4, 6, 7

[48] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J Black. Learning to dress 3d people in generative clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6469–6478, 2020. 2

[49] Qianli Ma, Shunsuke Saito, Jinlong Yang, Siyu Tang, and Michael J Black. Scale: Modeling clothed humans with a surface codec of articulated local elements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16082–16093, 2021. 2

[50] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J Black. The power of points for modeling humans in clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10974–10984, 2021.

[51] Qianli Ma, Jinlong Yang, Michael J Black, and Siyu Tang. Neural point-based shape modeling of humans in challenging clothing. *arXiv preprint arXiv:2209.06814*, 2022. 2

[52] Marvelous. Marvelous designer. https://www.marvelousdesigner.com/, 2021. 2

[53] Xiaoxu Meng, Weikai Chen, and Bo Yang. Neat: Learning neural implicit surfaces with arbitrary topologies from multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–258, 2023. 9

[54] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In

Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4460–4470, 2019. 1

[55] Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, and Shubham Tulsiani. Autosdf: Shape priors for 3d completion, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 306–315, 2022. 2

[56] Gyeongsik Moon, Hyeongjin Nam, Takaaki Shiratori, and Kyoung Mu Lee. 3d clothed human reconstruction in the wild. In *European conference on computer vision*, pages 184–200. Springer, 2022. 1

[57] Rahul Narain, Armin Samii, and James F O'brien. Adaptive anisotropic remeshing for cloth simulation. *ACM transactions on graphics (TOG)*, 31(6):1–10, 2012. 3

[58] optitex. Optitext fashion design software. https://optitex.com/, 2021. 2

[59] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13503–13513, 2022. 3

[60] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2

[61] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7365–7375, 2020. 2

[62] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 523–540. Springer, 2020. 1

[63] Songyou Peng, Chiyu Jiang, Yiyi Liao, Michael Niemeyer, Marc Pollefeys, and Andreas Geiger. Shape as points: A differentiable poisson solver. *Advances in Neural Information Processing Systems*, 34:13032–13044, 2021. 2

[64] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (ToG)*, 36(4):1–15, 2017. 1

[65] Albert Pumarola, Jordi Sanchez-Riera, Gary Choi, Alberto Sanfeliu, and Francesc Moreno-Noguer. 3dpeople: Modeling the geometry of dressed humans. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2242–2251, 2019. 1

[66] 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. 4

[67] Lingteng Qiu, Guanying Chen, Jiapeng Zhou, Mutian Xu, Junle Wang, and Xiaoguang Han. Rec-mv: Reconstructing 3d dynamic cloth from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4637–4646, 2023. 2

[68] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5

[69] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 3, 5

[70] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 2

[71] 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

[72] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2, 3

[73] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019. 2

[74] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 2

[75] Igor Santesteban, Nils Thuerey, Miguel A Otaduy, and Dan Casas. Self-supervised collision handling via generative 3d garment models for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11763–11773, 2021. 2

[76] Igor Santesteban, Miguel A Otaduy, and Dan Casas. Snug: Self-supervised neural dynamic garments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8140–8150, 2022.

[77] Yidi Shao, Chen Change Loy, and Bo Dai. Towards multi-layered 3d garments animation. *arXiv preprint arXiv:2305.10418*, 2023. 2

[78] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34:6087–6101, 2021. 9

[79] Tianchang Shen, Jacob Munkberg, Jon Hasselgren, Kangxue Yin, Zian Wang, Wenzheng Chen, Zan Gojcic,

Sanja Fidler, Nicholas Sharp, and Jun Gao. Flexible isosurface extraction for gradient-based mesh optimization. *ACM Transactions on Graphics (TOG)*, 42(4):1–16, 2023. 9

[80] Yu Shen, Junbang Liang, and Ming C Lin. Gan-based garment generation using sewing pattern images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 225–247. Springer, 2020. 6

[81] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 4

[82] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*. 2

[83] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 5

[84] 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. 5

[85] Zhaoqi Su, Tao Yu, Yangang Wang, and Yebin Liu. Deepcloth: Neural garment representation for shape and style editing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1581–1593, 2022. 3

[86] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 2, 5

[87] Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. Sizer: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 1–18. Springer, 2020. 2

[88] Lokender Tiwari and Brojeshwar Bhowmick. Garsim: Particle based neural garment simulator. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4472–4481, 2023. 2

[89] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020. 3

[90] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[91] Peng-Shuai Wang, Yang Liu, and Xin Tong. Dual octree graph networks for learning adaptive volumetric shape representations. *ACM Transactions on Graphics (TOG)*, 41(4): 1–15, 2022. 10

[92] Tuanfeng Y Wang, Duygu Ceylan, Jovan Popovic, and Niloy J Mitra. Learning a shared shape space for multimodal garment design. *arXiv preprint arXiv:1806.11335*, 2018. 1, 2, 3, 6

[93] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016. 2, 3

[94] Donglai Xiang, Fabian Prada, Timur Bagautdinov, Weipeng Xu, Yuan Dong, He Wen, Jessica Hodgins, and Chenglei Wu. Modeling clothing as a separate layer for an animatable human avatar. *ACM Transactions on Graphics (TOG)*, 40 (6):1–15, 2021. 2

[95] Donglai Xiang, Fabian Andres Prada, Timur Bagautdinov, Weipeng Xu, Yuan Dong, He Wen, Jessica Hodgins, and Chenglei Wu. Explicit clothing modeling for an animatable full-body avatar. *arXiv preprint arXiv:2106.14879*, 2, 2021.

[96] Donglai Xiang, Timur Bagautdinov, Tuur Stuyck, Fabian Prada, Javier Romero, Weipeng Xu, Shunsuke Saito, Jingfan Guo, Breannan Smith, Takaaki Shiratori, et al. Dressing avatars: Deep photorealistic appearance for physically simulated clothing. *ACM Transactions on Graphics (TOG)*, 41 (6):1–15, 2022. 2

[97] Chen Xin, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[98] Xingguang Yan, Liqiang Lin, Niloy J Mitra, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Shapeformer: Transformer-based shape completion via sparse representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6239–6249, 2022. 2

[99] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4541–4550, 2019. 3, 6

[100] Ilya Zakharkin, Kirill Mazur, Artur Grigorev, and Victor Lempitsky. Point-based modeling of human clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14718–14727, 2021. 2, 6

[101] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. *arXiv preprint arXiv:2210.06978*, 2022. 2, 3

[102] Biao Zhang, Matthias Nießner, and Peter Wonka. 3dilg: Irregular latent grids for 3d generative modeling. *Advances in Neural Information Processing Systems*, 35:21871–21885, 2022. 2, 3, 6

[103] Biao Zhang, Jiapeng Tang, Matthias Nießner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Trans. Graph.*, 42(4), 2023. 2

[104] Meng Zhang, Duygu Ceylan, and Niloy J Mitra. Motion guided deep dynamic 3d garments. *ACM Transactions on Graphics (TOG)*, 41(6):1–12, 2022. 2

[105] Xujie Zhang, Binbin Yang, Michael C Kampffmeyer, Wenqing Zhang, Shiyue Zhang, Guansong Lu, Liang Lin, Hang Xu, and Xiaodan Liang. Diffcloth: Diffusion based garment synthesis and manipulation via structural cross-modal

semantic alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23154–23163, 2023. 2

[106] Fang Zhao, Wenhao Wang, Shengcai Liao, and Ling Shao. Learning anchored unsigned distance functions with gradient direction alignment for single-view garment reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12674–12683, 2021. 2

[107] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5826–5835, 2021. 3, 4, 6, 7

[108] Heming Zhu, Yu Cao, Hang Jin, Weikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoguang Han. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 512–530. Springer, 2020. 2