TELA: Text to Layer-wise 3D Clothed Human Generation

Junting Dong[1], Qi Fang[2], Zehuan Huang[3], Xudong Xu[1],
Jingbo Wang[1], Sida Peng[4], and Bo Dai[1]

[1] Shanghai AI Laboratory [2] NetEase Games AI Lab
[3] Beihang University [4] Zhejiang University
http://jtdong.com/tela.github.io/

Abstract. This paper addresses the task of 3D clothed human generation from textural descriptions. Previous works usually encode the human body and clothes as a holistic model and generate the whole model in a single-stage optimization, which makes them struggle for clothing editing and meanwhile lose fine-grained control over the whole generation process. To solve this, we propose a layer-wise clothed human representation combined with a progressive optimization strategy, which produces clothing-disentangled 3D human models while providing control capacity for the generation process. The basic idea is progressively generating a minimal-clothed human body and layer-wise clothes. During clothing generation, a novel stratified compositional rendering method is proposed to fuse multi-layer human models, and a new loss function is utilized to help decouple the clothing model from the human body. The proposed method achieves high-quality disentanglement, which thereby provides an effective way for 3D garment generation. Extensive experiments demonstrate that our approach achieves state-of-the-art 3D clothed human generation while also supporting cloth editing applications such as virtual try-on.

1 Introduction

The generation of 3D clothed human is of great need in a variety of applications such as AR/VR, immersive telepresence, and virtual try-on. In these applications, the generation process is hoped to be highly controllable and detachable, where the human model and its set of clothes can be created independently, and simultaneously support free clothing replacement and transfer. This requires a faithful disentanglement of the human model and clothes. While manually generating such decomposed clothed human can be labor-intensive and time-consuming, auto-generation conditioned on inputs from complex capture systems (e.g., 3D scans [47] and multi-view videos [8, 13] is also unscalable and inaccessible. In real applications users often desire the clothed human can be created with simple inputs, such as textual descriptions.

Benefiting from the rapid development of Large Language Models [4, 39] and Diffusion Models [40, 43, 45], some works have begun to explore various

Fig. 1: Given textural descriptions (e.g., "a man wearing jeans, a denim shirt, and a windbreaker"), this paper aims to generate clothing-disentangled 3D human models progressively. Meanwhile, our approach enables high-quality 3D cloth generation and supports applications like cloth composition.

text-to-3D tasks. Recently, DreamFusion [37] proposes to leverage a pretrained text-to-image diffusion model to generate 3D objects in an optimization-based pipeline, under the guidance of Score Distillation Sampling (SDS). However, due to the lack of human prior, they still struggle for high-quality clothed human generation. To address this, some works [5, 17, 22] propose to combine the linear blending skinning algorithm [3] into the optimization-based pipeline, which significantly improves the quality of generated clothed humans. However, these methods usually represent the human model and its set of clothes as a holistic model and generate the whole model in a single-stage optimization. This not only makes clothing editing such as clothing replacement infeasible, but also results in the loss of control over the whole generation process (e.g., can not specify the order of inside and outside of clothes) and thereby limits their applicability.

In this paper, we propose TELA, a novel approach for the new task of clothing disentangled 3D human model generation from texts. The proposed approach introduces a layer-wise clothed human representation, where the human body and each clothing are represented with separate neural radiance fields (NeRFs) [28]. Then, a progressive optimization strategy is proposed to generate the minimal-clothed human body and layer-wise clothes sequentially, which effectively out-

puts a disentangled clothed human model and also provides the users sufficient control of the generation process, as shown in Figure 1.

However, achieving high-quality decoupled clothing generation is challenging. First, existing multi-layer representations [42, 53] commonly define all NeRF layers in the same space and then adopt the *Max* or *Mean* operation to composite multi-layer outputs, which may lead to penetration between adjacent layers as shown in the experiments. It is more reasonable to define each NeRF in its own layer space. However, how to divide areas of different layers is unknown. Second, unlike those disentangled reconstruction methods that take accurate part masks as inputs, we aim to achieve the disentanglement with only texts, which makes the task more difficult. Our experiments show that only using body-level SDS loss [5, 17, 22] cannot obtain high-quality cloth decoupling.

To address these challenges, we first propose a novel stratified composition method. During volume rendering, this method can automatically divide areas of different layers according to the transparency of each point along each ray. On the divided area of a specific layer, only the corresponding NeRF is utilized to produce the density and color, which avoids the interpenetration between adjacent layers. For cloth disentanglement, we propose novel dual SDS losses to simultaneously supervise the rendered clothed human image and clothes-only image, which introduces more regularization on the cloth model and encourages it to decouple from the human body.

Extensive experiments demonstrate that TELA effectively disentangles the human body and each cloth. Thanks to the disentangled modeling, our approach supports clothing editing applications such as virtual try-on, which is hard to achieve in prior methods. It is worth mentioning that, from another perspective, our paper presents a novel method for high-quality 3D clothing generation by simultaneously considering the underlying human body.

In summary, this work makes the following contributions:

- We present a novel framework for the new task of cloth disentangled 3D human generation from textual inputs, which also provides an effective way for 3D garment generation.
- We propose a novel decoupled cloth generation algorithm that introduces stratified composition method for multi-layer rendering and dual SDS losses for cloth disentanglement.
- Compared to the holistic modeling method, our approach achieves better clothed human generation quality while the disentangled modeling unleashes the potential of many downstream applications (e.g., virtual try-on).

2 Related works

Text-guided 3D content generation. Capitalizing on the significant advancements in Text-to-Image (T2I) generation models [41, 43, 44], numerous studies have delved into the domain of text-to-3D generation. Certain approaches [20, 30] advocate for the development of a text-conditioned 3D generative model using

paired 3D and text data. Nevertheless, the limited size of existing paired 3D-text datasets, compared to their image-text counterparts [48], constrains the generalization capability of these methods. Consequently, many works have investigated the use of pre-trained 2D generation models to generate 3D content without any 3D data. Early works propose to leverage the CLIP model [38] to optimize underlying 3D representations (e.g., neural radiance fields [18] or meshes [21]) by minimizing the embedding difference between rendered images and accompanying text descriptions. Owing to the limited capacity of the CLIP space, these endeavors often result in the generation of unrealistic 3D models. Recently, [24,37] propose the utilization of powerful text-to-image diffusion models for optimization with Score Distillation Sampling (SDS), leading to impressive results. Furthermore, [52] suggests a particle-based variational approach, namely variational score distillation, to enhance the quality of 3D generation. Despite their success in general object generation, these methods still face challenges in achieving a high-quality generation of clothed human models.

Text-guided 3D human generation. Building upon the foundation of 2D pre-trained generative models, several studies aim to enhance 3D human generation by incorporating a human prior. Avatar-CLIP [14] enhances the parametric human model SMPL [3] by integrating the neural radiance field as the 3D human representation, followed by optimization using the CLIP model. Likewise, leveraging the parametric human model as a human prior, some works [5] utilize pre-trained text-to-image diffusion models and Score Distillation Sampling for 3D human generation. DreamHuman [22] introduces a deformable and pose-conditioned NeRF model, achieving animatable 3D clothed human generation. Recently, to alleviate the Janus (multi-face) problem, DreamWaltz [17] introduces 3D-consistent score distillation sampling. The approach involves projecting the canonical 3D human skeleton to each view and adopting a 2D human skeleton-conditioned diffusion model [60] for view-aligned optimization, yielding impressive results. AvatarVerse [59] replaces the human skeleton condition with the densepose map. Moreover, TADA [23] replaces the NeRF model with a deformable SMPL-X [32] mesh and a texture map, seamlessly integrating the results into existing Computer Graphics workflows. HumanNorm [16] proposes to fine-tune the diffusion model to generate normal maps and achieve remarkable human geometry generation. GAvatar [57] introduces the Gaussian splatting representation for efficient and high-quality human rendering. While these methods demonstrate remarkable success in clothed human generation, they commonly represent the clothed human as a holistic model, neglecting the layer-wise nature and facing challenges in cloth editing.

Clothed human modeling. Early methods [1,2,26] commonly depend on parametric human meshes and incorporate additional vertex offsets for clothed human modeling. Benefiting from advancements in implicit functions [27,28,31], recent methods [7,10,11,33–35,46,47,49,51,54] have presented impressive clothed human reconstruction or generation from images and 3D scans. Typically, these approaches treat the human body and clothing as an integrated entity. To

disentangle the representation of the human body and clothing, some studies [6, 12, 36, 55, 56] propose a multi-layer human representation, achieving high-quality reconstruction from 3D scans, depth sensors, and even monocular videos. These techniques rely on a parametric human model [3] and then reconstruct separate layers for clothing. In addition, there are some works proposing the learning of 3D generative clothed models from 3D garment datasets. BCNet [19] proposes the learning of parametric mesh-based garment models, employing networks to predict parameters crucial for clothed human reconstruction. Furthermore, SMPLicit [9] employs implicit unsigned distance functions to represent the cloth model. Despite yielding promising results, the restricted availability of 3D garment data leads to the unrealism and limited diversity of these cloth models. Recently, some concurrent works [15, 58] also attempt to explore multi-layer human generation. HumanLiff [15] learns layer-wise human generation from 2D images. However, their cloth models are not disentangled with the human body, which limits cloth editing applications. [58] mainly focuses on disentangling the hair and head ornament from the upper body while this paper explores the multi-layer clothes and human body disentangled generation. [50] proposes generating disentangled clothing using offset mesh construction on the SMPL model. However, due to the topological constraints of SMPL, they are unable to generate loose garment types such as dresses and skirts. In contrast, TELA supports various garment types and can achieve more photorealistic rendering.

3 Methods

Given textual descriptions of a clothed person as input, TELA aims to generate the disentangled 3D human model, where each garment is decoupled from the human body. Figure 2 presents an overview of our approach. To disentangle each component, we propose a multi-layer clothed human model that is parameterized by multiple neural radiance fields (NeRFs) (Section 3.1) and introduce a novel transparency-based stratified compositional rendering approach (Section 3.2). Then, a layer-wise generation method is designed to produce each component based on the text prompts (Section 3.3).

3.1 Multi-layer clothed human model

In contrast to previous methods [17, 22] that encode the holistic clothed human in a single model, we decompose the clothed human into a minimal-clothed human body and multiple clothes (e.g., shirt, coat, and pants). The geometry and appearance of clothes are represented as neural radiance fields F, which are implemented by MLP networks. To better represent human hair, we also employ NeRF as the representation of the minimal-clothed human body rather than the mesh [3]. Each component is represented as a single network. Specifically, for the component p, the model can be written as follows:

$$\sigma(\mathbf{x}), \mathbf{c}(\mathbf{x}) = F_p(h_p(\mathbf{x})), \tag{1}$$

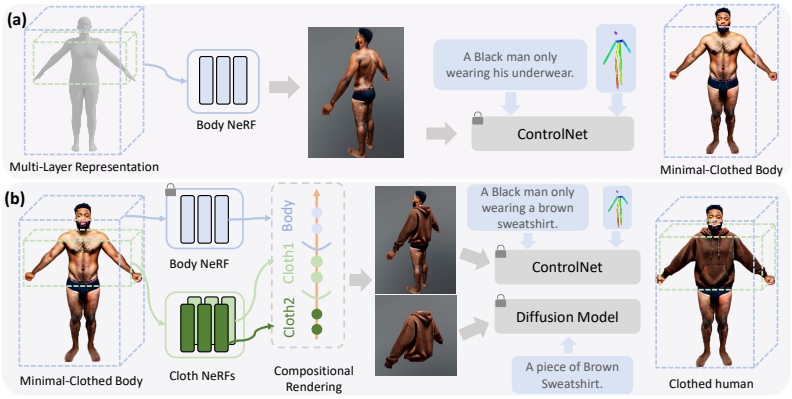

Fig. 2: Overview of TELA. (a) Minimal-clothed body is the first component to generate. To train the body NeRF, we render the image and the corresponding 2D human skeleton under a random viewpoint and then utilize the 2D human skeleton conditioned ControlNet [60] for SDS optimization. (b) Given the fixed body NeRF, we aim to progressively generate each cloth. For generating cloth p, we render an image of the human with cloth p and another image of cloth p only through the proposed transparency-based stratified compositional rendering. Then, the dual SDS losses are proposed to supervise these two images. For the cloth-only image, the original stable diffusion model is adopted for SDS optimization.

where $\sigma(\mathbf{x})$ and $\mathbf{c}(\mathbf{x})$ denote the predicted density and color of the sample point $\mathbf{x}$. $h_p(\mathbf{x})$ denotes the learnable hash encoding function [29].

In addition, similar to prior works, we define an 'A-pose' human mesh (i.e., SMPL) in the NeRF space to introduce the human prior. Specifically, we utilize the SMPL mesh to determine rough 3D boxes for different clothing, such as upper-body and lower-body clothes. Moreover, during the NeRF optimization, the 3D SMPL skeleton is projected to each viewpoint, and the resulting 2D human skeleton is incorporated with the ControlNet [60] for training (see Section 3.3 for details).

3.2 Transparency-based stratified composition

Based on the predicted density and color, novel view images can be synthesized through volume rendering:

$$\tilde{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma(\mathbf{x}_i)\delta_i))\mathbf{c}(\mathbf{x}_i), \text{and } T_i = \exp(-\sum_{j=1}^{i-1} \sigma(\mathbf{x}_j)\delta_j), \tag{2}$$

where $\tilde{C}(\mathbf{r})$ denotes the rendered color of ray $\mathbf{r}$, $\delta_i = ||\mathbf{x}_{i+1} - \mathbf{x}_i||_2$ is the distance between adjacent sampled points, and T_i is the accumulated transparency at sample point $\mathbf{x}_i$ along the ray. N is the number of sample points along each ray.

Due to the space overlap between different components (e.g., the human body and upper-body clothes), we propose a novel transparency-based stratified compositional rendering method to prevent the penetration between adjacent components. In particular, when training the component p, we leverage the inside NeRF networks $\{F_k|k = 1, ..., p-1\}$ that have been optimized before to predict the densities of sample points along each ray. For each sample point, we take the maximum density value $max\{\sigma_1, ..., \sigma_{p-1}\}$ as its final density. Then, we can calculate the transparency of sample points along each ray with equation (2). *Note that different from previous works [42, 53] using maximum operation to composite multi-layer outputs, we utilize the maximum density for transparency calculation and then divide areas for each layer.* Based on the transparency, we can select sample points whose transparency is larger than a predefined threshold th for training the NeRF network F_p. After training, the stratified composition is also utilized to render final clothed human models with multiple NeRFs. Please also refer to equation (7) for the composition rendering.

3.3 Training

With the multi-layer clothed human representation, we propose a progressive generation framework to obtain each component model sequentially. This framework not only ensures the disentanglement of each component but also provides enhanced flexibility and controllability in the process of clothed human generation. Specifically, we begin by generating a minimal-clothed human body and then progressively generate the clothing outward.

Score Distillation Sampling (SDS). Before introducing the component generation in detail, we first briefly describe the Score Distillation Sampling [37]. To leverage the supervision from the text descriptions, the SDS loss is proposed to utilize a pre-trained diffusion model $\boldsymbol{\epsilon}_\phi$ for training. Given a 3D model parameterized by $\boldsymbol{\theta}$ and its differentiable rendering image $\mathbf{u} = g(\boldsymbol{\theta})$, the SDS gradients of the 3D model parameters $\boldsymbol{\theta}$ can be written as follows:

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{SDS}}(\phi, \mathbf{u}) = \mathbb{E}_{t,\boldsymbol{\epsilon}} \left[w(t)(\boldsymbol{\epsilon}_\phi(\mathbf{u}_t; y, t) - \boldsymbol{\epsilon}) \frac{\partial \mathbf{u}}{\partial \boldsymbol{\theta}} \right], \tag{3}$$

where $w(t)$ denotes a weighting function depending on the timestep t, $\mathbf{u}_t$ the noised image, and y the input text prompt. $\boldsymbol{\epsilon}$ is the injected noise added to the rendered image $\mathbf{u}$.

Minimal-clothed body generation. The first component to generate is the minimal-clothed body. Similar to previous work [17], we adopt the 2D human skeleton conditioned diffusion model to improve the multi-view consistency. Specifically, relying on the predefined 'A-pose' SMPL mesh, we can obtain a 3D human skeleton. Then, when the radiance field renders images, the corresponding 2D human skeletons can be obtained by projecting the 3D skeleton to the same viewpoints. Conditioned on the 2D human skeleton $\mathbf{s}$, the SDS loss can be written as follows:

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}^{\mathbf{s}}_{\text{SDS}}(\phi, \mathbf{u}) = \mathbb{E}_{t,\boldsymbol{\epsilon}} \left[w(t)(\boldsymbol{\epsilon}_\phi(\mathbf{u}_t; y_b, t, \mathbf{s}) - \boldsymbol{\epsilon}) \frac{\partial \mathbf{u}}{\partial \boldsymbol{\theta}} \right], \tag{4}$$

where y_b denotes the corresponding text prompt of the body.

In addition, we also introduce a regularization loss to prevent floating 'radiance clouds':

$$L_r = \lambda_1 \cdot BE(\mathbf{M}) + \lambda_2 \cdot \|\mathbf{M}\|_1, \tag{5}$$

where $\mathbf{M}$ denotes the rendered mask of the NeRF model. The first item binary entropy function $BE(\cdot)$ encourages the mask to be binarized and the second item introduces the sparsity regularization. λ_1 and λ_2 are constants.

Thus, the full training loss for the minimal-cloth body generation is as follows:

$$L_b = \mathcal{L}^{\mathbf{s}}_{\text{SDS}} + L_r. \tag{6}$$

Cloth generation. Given the minimal-clothed body, we aim to progressively generate each garment model. To generate the p-th piece of clothing, we propose the compositional rendering of the body NeRF F_b and the current cloth NeRF F_p. Based on the transparency-based stratified compositional rendering (section 3.2), we employ the cloth NeRF model F_p to predict the density σ_p and color $\mathbf{c}_p$ for those sample points with transparency larger than the threshold th along each ray. For the left sample points, we use the trained body NeRF model F_b to predict corresponding σ_b and $\mathbf{c}_b$. This compositional process can be formally written as follows:

$$\begin{aligned}\tilde{C}_{bp}(\mathbf{r}) = &\sum_{i=1}^{k} T_i(1-\exp(-\sigma_p(\mathbf{x}_i)\delta_i))\mathbf{c}_p(\mathbf{x}_i) + \\ &\sum_{i=k+1}^{N} T_i(1-\exp(-\sigma_b(\mathbf{x}_i)\delta_i))\mathbf{c}_b(\mathbf{x}_i),\end{aligned} \tag{7}$$

where k is the last point with transparency larger than th. Intuitively, the above compositional rendering produces an image $\mathbf{u}_{bp}$ of a person wearing the cloth p. Based on this image, we can train the cloth model F_p using a loss function similar to L_b. Here, we use a text prompt y_{bp} describing a person wearing the cloth p and a new mask $\mathbf{M}_{bp}$.

When only using the compositional rendering images for training, we could ultimately obtain a high-quality model of a person wearing clothes. However, the learned clothing model tends to be coupled with the body model, i.e., the resulting cloth model may include parts of the human body, such as the arms and legs (see Figure 7). To address this problem, we introduce an additional cloth-only SDS loss, which encourages the clothing model to only contain the clothing. Specifically, based on the stratified composition, we additionally render an image only using the cloth model as follows:

$$\tilde{C}_p(\mathbf{r}) = \sum_{i=1}^{k} T_i(1-\exp(-\sigma_p(\mathbf{x}_i)\delta_i))\mathbf{c}_p(\mathbf{x}_i). \tag{8}$$

Then, we utilize the vanilla SDS loss (3) to supervise the cloth-only image $\mathbf{u}_p$ with a text prompt y_p describing the cloth p. Furthermore, we also introduce the

regularization loss (5) for the cloth-only mask $\mathbf{M}_p$. Thus, the full loss for cloth generation is as follows:

$$L_p = \mathcal{L}^{\mathbf{s}}_{\text{SDS}}(\mathbf{u}_{bp}; y_{bp}) + L_r(\mathbf{M}_{bp}) + \mathcal{L}_{\text{SDS}}(\mathbf{u}_p; y_p) + L_r(\mathbf{M}_p). \tag{9}$$

Note that the parameters of the body model are fixed here.

4 Experiments

Implementation details. We train each NeRF model for $10k$ iterations with batchsize 1, which takes around 3 hours on a single A100 GPU. The NeRF models render images with a resolution of 256×256 for the first $5k$ iterations and 512x512 for the last $5k$ iterations. Given a text prompt that describes the clothed human, e.g., "a man wearing cloth A, B, ...", we can generate corresponding text prompts for optimization programmatically such as "a man only wearing underwear" for mini-clothed body generation, "a man only wearing cloth A" and "a piece of cloth A" for cloth A generation.

4.1 Comparisons with the holistic modeling method

Qualitative comparisons. To evaluate the quality of the generated clothed human, we compare TELA with the state-of-the-art holistic modeling method DreamWaltz [17] that introduces 3D consistent score distillation sampling and achieves impressive results. *Please note that while there exist other text-to-human generation approaches, this paper focuses on the cloth-disentangled 3D human generation. These approaches are orthogonal to our work, and our method can also be adapted to their frameworks.*

We present the qualitative results in Figure 3. DreamWaltz [17] generates clothed humans with blurry and distorted clothes. In contrast, thanks to the disentangled modeling, our method produces high-quality cloth details. Despite we separately model the human body and each cloth, the proposed approach achieves the natural composition of each component. In addition, the results of the second person (i.e., the Black woman) show that disentangled modeling can also help alleviate the multi-face problem. Moreover, we present more qualitative results of clothed humans and disentangled clothes in Figure 4.

Quantitative comparison. Quantitative evaluation of generated 3D human models is challenging. We adopt the Fréchet Inception Distance (FID) metric that compares the distribution of two image datasets. We compute the FID between the rendered images of generated models and the images produced by the stable diffusion model. As shown in Table 1, we achieve a lower FID score signifying our rendering aligns more closely with the high-quality 2D images from the SD model.

User study. We also conduct user studies to further assess the quality of generated clothed humans against the baseline [17]. We randomly select 30 text prompts and render generated clothed human models as rotating videos. Given

Fig. 3: Qualitative comparisons with the holistic modeling method [17]. (a) DreamWaltz [17], (b) Ours. Text prompts (from left to right): "A woman wearing a Brown Cycling Top and Brown Chiffon Skirt", "A Black woman wearing a blue turtleneck and blue Midi Skirt", "An Asian man wearing a Light Blue Varsity Jacket and Western Pants", "A Black man wearing a Brown Sweatshirt and jeans"

Table 1: Quantitative comparisons with the holistic modeling method.

Methods	DreamWaltz [17]	Ours
FID ↓	142.6	**125.9**

these videos, we asked 26 volunteers to assess the (1) overall clothed human quality, (2) cloth quality, and (3) consistency with text inputs, and selected the preferred results. The quantitative results are shown in Table 2, which demonstrates that the proposed method obtains much higher preference than the baseline over all three metrics.

Table 2: User preference study. Our method achieves consistently higher preference than the holistic modeling method DreamWaltz [17] in overall clothed human quality, cloth quality, and consistency with text inputs.

Comparison (Ours vs. DreamWaltz)	Preference (%)
Overall clothed human quality	71.46
Cloth quality	77.92
Consistency with text inputs	73.18

4.2 Qualitative comparisons of cloth generation

Due to the disentangled modeling of the human body and each cloth, our method also enables high-quality 3D cloth generation. To evaluate the quality of gener-

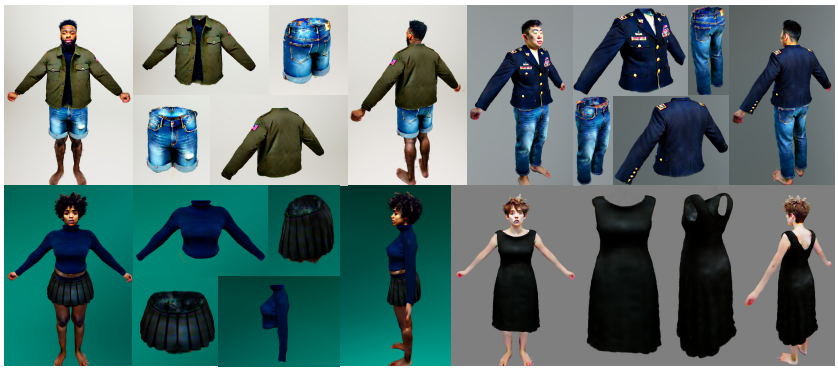

Fig. 4: Qualitative results of the proposed method. Text prompts: "A Black man wearing Khaki Outerwear and denim shorts", "An Asian man wearing a Navy Blue Military Jacket and jeans", "A Black woman wearing a blue turtleneck and Sporty Skirt", "A woman wearing Black Dress".

ated clothes, we compare it with the state-of-the-art text-to-3D methods: DreamFusion [37], Magic3D [24], and ProlificDreamer [52].

The qualitative results are shown in Figure 5, which presents that the proposed approach significantly outperforms the baselines. First, as shown in the first row (i.e., jeans), even with the view-dependent text augmentation technology [37], the baselines [24,37,52] still struggle for the multi-face problems for the cloth generation. In contrast, by composing with the human body, our method additionally introduces the 2D human skeleton conditioned diffusion model to enhance the multi-view consistency, which dramatically alleviates the multi-face problem. Second, the baselines often generate clothes coupled with the human body while our method enables high-quality disentanglement. Third, thanks to the stratified compositional rendering, as shown in the second row (i.e., sleeveless dress), our approach produces clothes with a hollow structure as real clothes, which is crucial for applications like clothing transfer.

4.3 Ablation studies

Transparency-based compositional rendering. As described in Section 3.1, to prevent the penetration between adjacent component NeRFs, Our method utilizes Transparency-based compositional rendering to compose multiple NeRFs. Here, we compare it to the baseline [42, 53] without this compositional rendering, i.e., directly using the maximum operation to fuse the densities of different models and then the corresponding colors are used for volume rendering. The qualitative results of generated clothed human models are presented in Figure 6. As shown in the first column, without the proposed compositional rendering, there is serious penetration between the human body and the dress. In addition, as shown in the second column, our compositional rendering can help align clothing models to the human body with the appropriate size.

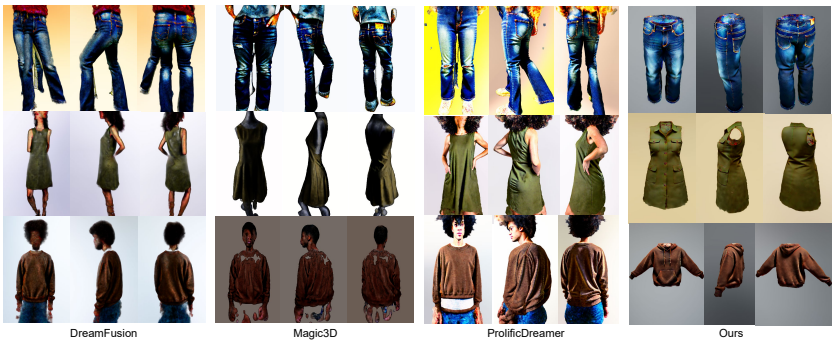

Fig. 5: Qualitative comparisons of cloth generation with the SOTA methods. Text prompts (from top to bottom): "a pair of jeans", "a piece of Khaki Sleeveless Dress", "a Brown Sweatshirt".

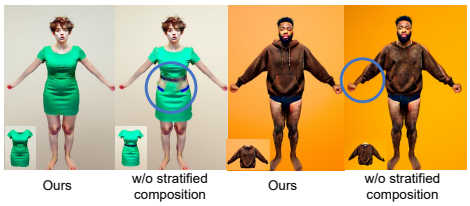

Fig. 6: Ablation studies for Transparency-based stratified compositional rendering.

Dual SDS losses. As described in Section 3.3, we propose the dual SDS losses for clothing optimization, i.e., additionally introducing cloth SDS loss. Here, we analyze the effect of the additional cloth SDS loss function. The qualitative results are presented in Figure 7. As the results show, when not using the cloth SDS loss, the generated clothing model will entangle with the human body. The reason is that the 2D human skeleton-conditioned SDS loss only supervises the compositional renderings of the clothed human to be consistent with the text input, which does not constrain the cloth model to only contain the clothes. In contrast, the cloth SDS loss provides additional constraints to the cloth model, which can remove the undesired artifacts. Moreover, as shown in section 4.2, with only cloth SDS loss, the clothing model can not align with the human body and the cloth quality will also degrade. Therefore, the proposed dual SDS losses are effective for high-quality cloth generation.

Regularization loss. As described in Section 3.3, we introduce a regularization loss to prevent floating 'radiance clouds'. We analyze the impact of this regularization loss, and the results are depicted in Figure 7. Removing this loss results in the emergence of floating artifacts around the generated body and clothing.

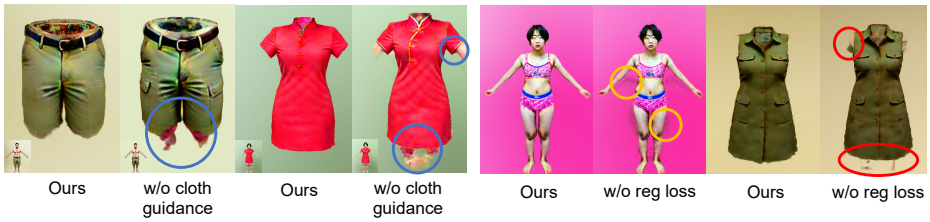

Fig. 7: Ablation studies for cloth guidance loss and regularization loss.

Fig. 8: Free composition of different clothes.

4.4 Applications

Thanks to the disentangled modeling, our method supports the free composition of different clothes on the same person. We present the free composition of three garments (charcoal gray jacket, coral T-shirt, and brown sweatshirt) and three pairs of pants (jeans, athletic pants, and khaki shorts) in Figure 8.

Additionally, our objective extends to facilitating clothing transfer among individuals with similar body shapes, a valuable pursuit given the clothing sizes similarity. While the NeRF model excels in representing various types of clothing, clothing transfer under NeRF representation poses challenges. [58] proposes to adjust the NeRF model in size manually. As shown in Figure 9, though they can avoid penetration, the resized clothes do not adapt well to the new person. To solve this, we introduce a clothing deformation field. Specifically, for each sample point $\mathbf{x}$ of the clothing model p, we predict its non-rigid deformation using an MLP network D_p and we leverage the same loss functions as cloth generation (9). The deformation field is trained for $2.5k$ steps. Note that only the parameters of the deformation field are trained here. Based on the deformation field, our method achieves better cloth transfer automatically.

Moreover, we can integrate a linear blend skinning module to generate animatable human model. Adopting a similar animation module to DreamWaltz, the proposed method can produce high-quality rendering under novel human poses. The results are presented in Figure 10.

5 Limitations

The proposed approach still has the following limitations. First, our generation takes hours for optimization, which is much slower than the text-to-image generation. Exploring a more efficient and generalizable model [25] for disentangled

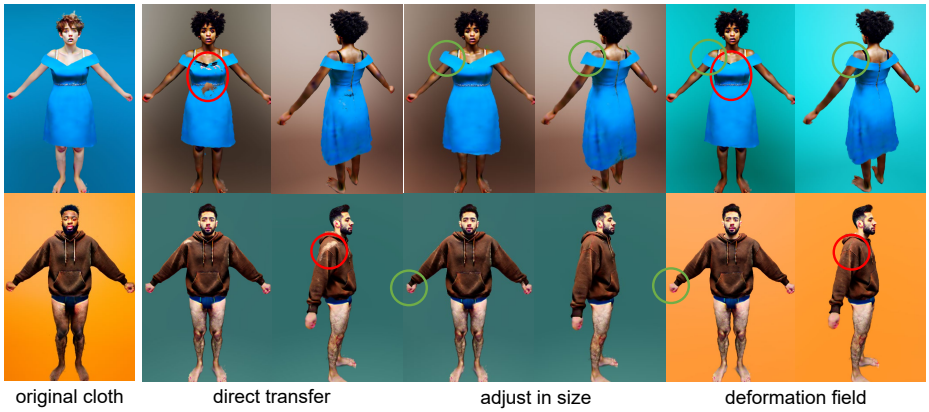

Fig. 9: Adjusting in size and deformation fields for cloth transfer.

Fig. 10: Animation results of generated human models.

human generation could significantly enhance its applicability. Second, while our method could support animatable human generation by introducing the linear blend skinning module like the previous work [17, 22], it is more interesting to explore clothed human animation with disentangled components (i.e., body and clothes) rather than a holistic model, which is left as future work. Third, the current human body and clothes are represented as NeRFs. The textured mesh representation may be further introduced like [24], which can be seamlessly integrated into existing computer graphics systems.

6 Conclusion

In summary, this paper presents a novel approach for clothed disentangled 3D human generation from text inputs. Different from previous methods, we propose a multi-layer clothed human representation and progressively generate each component. For superior disentanglement, we first introduce the transparency-based stratified compositional rendering, facilitating the separation of adjacent layers. Then, we propose novel dual SDS losses to help the clothing model decouple from the human body. Thanks to the effective disentanglement, our method enables high-quality 3D garment generation. Experiments show that our approach not only achieves better clothed human generation but also enables clothing editing applications such as virtual try-on.

Acknowledgement

This work is supported by Shanghai Artificial Intelligence Laboratory.

References

1. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Video based reconstruction of 3d people models. In: CVPR (2018)
2. Bhatnagar, B.L., Tiwari, G., Theobalt, C., Pons-Moll, G.: Multi-garment net: Learning to dress 3d people from images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5420–5430 (2019)
3. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. pp. 561–578. Springer (2016)
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
5. Cao, Y., Cao, Y.P., Han, K., Shan, Y., Wong, K.Y.K.: DreamAvatar: Text-and-Shape Guided 3D Human Avatar Generation via Diffusion Models. arXiv preprint arXiv:2304.00916 (2023)
6. Chen, X., Pang, A., Yang, W., Wang, P., Xu, L., Yu, J.: Tightcap: 3d human shape capture with clothing tightness field. ACM Transactions on Graphics (TOG) **41**(1), 1–17 (2021)
7. Chen, X., Jiang, T., Song, J., Yang, J., Black, M.J., Geiger, A., Hilliges, O.: gdna: Towards generative detailed neural avatars. arXiv (2022)
8. Collet, A., Chuang, M., Sweeney, P., Gillett, D., Evseev, D., Calabrese, D., Hoppe, H., Kirk, A., Sullivan, S.: High-quality streamable free-viewpoint video. TOG (2015)
9. Corona, E., Pumarola, A., Alenyà, G., Pons-Moll, G., Moreno-Noguer, F.: Smplicit: Topology-aware generative model for clothed people. In: CVPR (2021)
10. Dong, J., Fang, Q., Guo, Y., Peng, S., Shuai, Q., Zhou, X., Bao, H.: Totalselfscan: Learning full-body avatars from self-portrait videos of faces, hands, and bodies. Advances in Neural Information Processing Systems **35**, 13654–13667 (2022)
11. Dong, J., Fang, Q., Yang, T., Shuai, Q., Qiao, C., Peng, S.: ivs-net: Learning human view synthesis from internet videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22942–22951 (2023)
12. Feng, Y., Yang, J., Pollefeys, M., Black, M.J., Bolkart, T.: Capturing and animation of body and clothing from monocular video. SIGGRAPH Asia 2022 Conference Papers (2022)
13. Guo, K., Lincoln, P., Davidson, P., Busch, J., Yu, X., Whalen, M., Harvey, G., Orts-Escolano, S., Pandey, R., Dourgarian, J., et al.: The relightables: Volumetric performance capture of humans with realistic relighting. TOG (2019)
14. Hong, F., Zhang, M., Pan, L., Cai, Z., Yang, L., Liu, Z.: AvatarCLIP: Zero-Shot Text-Driven Generation and Animation of 3D Avatars. ACM Transactions on Graphics (TOG) **41**(4), 1–19 (2022)
15. Hu, S., Hong, F., Hu, T., Pan, L., Mei, H., Xiao, W., Yang, L., Liu, Z.: Humanliff: Layer-wise 3d human generation with diffusion model. arXiv preprint arXiv:2308.09712 (2023)

16. Huang, X., Shao, R., Zhang, Q., Zhang, H., Feng, Y., Liu, Y., Wang, Q.: Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. arXiv preprint arXiv:2310.01406 (2023)
17. Huang, Y., Wang, J., Zeng, A., Cao, H., Qi, X., Shi, Y., Zha, Z.J., Zhang, L.: Dreamwaltz: Make a scene with complex 3d animatable avatars. NeurIPS (2023)
18. Jain, A., Mildenhall, B., Barron, J.T., Abbeel, P., Poole, B.: Zero-shot text-guided object generation with dream fields (2022)
19. Jiang, B., Zhang, J., Hong, Y., Luo, J., Liu, L., Bao, H.: Bcnet: Learning body and cloth shape from a single image. In: European Conference on Computer Vision. Springer (2020)
20. Jun, H., Nichol, A.: Shap-e: Generating conditional 3d implicit functions. arXiv preprint arXiv:2305.02463 (2023)
21. Khalid, N.M., Xie, T., Belilovsky, E., Tiberiu, P.: CLIP-Mesh: Generating textured meshes from text using pretrained image-text models. SIGGRAPH Asia 2022 Conference Papers (2022)
22. Kolotouros, N., Alldieck, T., Zanfir, A., Bazavan, E.G., Fieraru, M., Sminchisescu, C.: Dreamhuman: Animatable 3d avatars from text. arXiv preprint arXiv:2306.09329 (2023)
23. Liao, T., Yi, H., Xiu, Y., Tang, J., Huang, Y., Thies, J., Black, M.J.: Tada! text to animatable digital avatars. arXiv preprint arXiv:2308.10899 (2023)
24. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. arXiv preprint arXiv:2211.10440 (2022)
25. Liu, R., Wu, R., Hoorick, B.V., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object (2023)
26. Ma, Q., Yang, J., Ranjan, A., Pujades, S., Pons-Moll, G., Tang, S., Black, M.J.: Learning to dress 3d people in generative clothing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6469–6478 (2020)
27. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4460–4470 (2019)
28. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
29. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. TOG (2022)
30. Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., Chen, M.: Point-e: A system for generating 3d point clouds from complex prompts. arXiv preprint arXiv:2212.08751 (2022)
31. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 165–174 (2019)
32. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: 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. pp. 10975–10985 (2019)
33. Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Zhou, X., Bao, H.: Animatable neural radiance fields for modeling dynamic human bodies. In: ICCV (2021)

34. Peng, S., Xu, Z., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Bao, H., Zhou, X.: Animatable implicit neural representations for creating realistic avatars from videos. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
35. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: CVPR (2021)
36. Pons-Moll, G., Pujades, S., Hu, S., Black, M.J.: Clothcap: seamless 4d clothing capture and retargeting. ACM Trans. Graph. **36**, 73:1–73:15 (2017)
37. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. ICLR (2023)
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. ICML (2021)
39. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020), http://jmlr.org/papers/v21/20-074.html
40. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
41. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv preprint arXiv:2204.06125 (2022)
42. Ranade, S., Lassner, C., Li, K., Haene, C., Chen, S.C., Bazin, J.C., Bouaziz, S.: Ssdnerf: Semantic soft decomposition of neural radiance fields. arXiv preprint arXiv:2212.03406 (2022)
43. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis With Latent Diffusion Models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
44. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., et al.: Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. arXiv preprint arXiv:2205.11487 (2022)
45. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement (2021). https://doi.org/10.48550/ARXIV.2104.07636, https://arxiv.org/abs/2104.07636
46. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: ICCV (2019)
47. Saito, S., Yang, J., Ma, Q., Black, M.J.: Scanimate: Weakly supervised learning of skinned clothed avatar networks. In: CVPR (2021)
48. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: LAION-5B: An open large-scale dataset for training next generation image-text models. arXiv preprint arXiv:2210.08402 (2022)
49. Tiwari, G., Sarafianos, N., Tung, T., Pons-Moll, G.: Neural-gif: Neural generalized implicit functions for animating people in clothing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11708–11718 (2021)
50. Wang, J., Liu, Y., Dou, Z., Yu, Z., Liang, Y., Li, X., Wang, W., Xie, R., Song, L.: Disentangled clothed avatar generation from text descriptions. arXiv preprint arXiv:2312.05295 (2023)

51. Wang, S., Mihajlovic, M., Ma, Q., Geiger, A., Tang, S.: Metaavatar: Learning animatable clothed human models from few depth images. In: Advances in Neural Information Processing Systems (2021)
52. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. arXiv preprint arXiv:2305.16213 (2023)
53. Wu, J., Li, S., Ji, S., Wang, Y., Xiong, R., Liao, Y.: Dorec: Decomposed object reconstruction utilizing 2d self-supervised features. arXiv preprint arXiv:2310.11092 (2023)
54. Xiu, Y., Yang, J., Tzionas, D., Black, M.J.: Icon: Implicit clothed humans obtained from normals. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13286–13296. IEEE (2022)
55. Yu, T., Zheng, Z., Guo, K., Zhao, J., Dai, Q., Li, H., Pons-Moll, G., Liu, Y.: Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 7287–7296 (2018)
56. Yu, T., Zheng, Z., Zhong, Y., Zhao, J., Dai, Q., Pons-Moll, G., Liu, Y.: Simulcap : Single-view human performance capture with cloth simulation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5499–5509 (2019)
57. Yuan, Y., Li, X., Huang, Y., De Mello, S., Nagano, K., Kautz, J., Iqbal, U.: Gavatar: Animatable 3d gaussian avatars with implicit mesh learning. arXiv preprint arXiv:2312.11461 (2023)
58. Zhang, H., Feng, Y., Kulits, P., Wen, Y., Thies, J., Black, M.J.: Text-guided generation and editing of compositional 3d avatars. arXiv preprint arXiv:2309.07125 (2023)
59. Zhang, H., Chen, B., Yang, H., Qu, L., Wang, X., Chen, L., Long, C., Zhu, F., Du, D., Zheng, M.: Avatarverse: High-quality & stable 3d avatar creation from text and pose. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7124–7132 (2024)
60. Zhang, L., Agrawala, M.: Adding Conditional Control to Text-to-Image Diffusion Models. arXiv preprint arXiv:2302.05543 (2023)