HumanCoser: Layered 3D Human Generation via Semantic-Aware Diffusion Model

Yi Wang[1 2†], Jian Ma[1†], Ruizhi Shao[3], Qiao Feng[1], Yu-Kun Lai[4], Kun Li[1*]

[1]Tianjin University, China [2]Changzhou Institute of Technology, China
[3]Tsinghua University, China [4]Cardiff University, British

Figure 1: Our method can generate layered 3D humans guided by text prompts, which are physically-decoupled and structurally consistent. This allows our generated clothing to be reused, exchanging between digital avatars with different identities.

Abstract

This paper aims to generate physically-layered 3D humans from text prompts. Existing methods either generate 3D clothed humans as a whole or support only tight and simple clothing generation, which limits their applications to virtual try-on and part-level editing. To achieve physically-layered 3D human generation with reusable and complex clothing, we propose a novel layer-wise dressed human representation based on a physically-decoupled diffusion model. Specifically, to achieve layer-wise clothing generation, we propose a dual-representation decoupling framework for generating clothing decoupled from the human body, in conjunction with an innovative multi-layer fusion volume rendering method. To match the clothing with different body shapes, we propose an SMPL-driven implicit field deformation network that enables the free transfer and reuse of clothing. Extensive experiments demonstrate that our approach not only achieves state-of-the-art layered 3D human generation with complex clothing but also supports virtual try-on and layered human animation. More results and the code can be found on our project page at https://cic.tju.edu.cn/faculty/likun/projects/HumanCoser.

†Equal contribution.
*Corresponding author. e-mail: lik@tju.edu.cn

Index Terms: 3D Human Generation, Layered Clothing, Physical Decoupling, Human Animation.

1 Introduction

The generation of 3D humans with changeable clothing plays an important role in movies, games and AR/VR. Existing methods [4, 35, 38, 34, 29, 26] only produce a unified surface encompassing both the body and clothing, leading to a body-clothing coupling. This limits their ability to edit clothing and body separately, restricting detailed customization and accurate adjustments for virtual try-ons, animated character design, and personalized avatar creation. In this paper, we aim to generate high-fidelity layered 3D human which can be edited and exchanged for clothing via representation-decoupling, as shown in Fig. 1.

Recently, owing to the high-quality image synthesis capability of pre-trained diffusion models [30], methods [29, 26, 21] introduce a novel Score Distillation Sampling (SDS) strategy [30] to self-supervise the 3D human generation process. However, these methods ignore the diversity and self-occlusion of human shapes, which leads to inconsistencies in generated human structures. Furthermore, most data-driven 3D avatar generation methods [1, 37, 17, 39, 7, 11] generate 3D clothed humans in a coupled manner, and as a result, clothing cannot be exchanged between arbitrary bodies. Overall, the above methods fail to ensure structural consistency of the human body and lack the capability to generate and edit bodies and clothes in a layered and flexible manner.

This paper introduces HumanCoser, a novel framework based

on a physically-decoupled diffusion model. It aims to generate representation-decoupling and animatable 3D dressed humans with consistent body structure in a layer-wise manner, guided by text. To achieve accurate layer-wise clothing representation, we propose a dual-representation decoupling framework designed to generate clothing independent from the human body. This framework is complemented by an innovative multi-layer fusion volume rendering method. HumanCoser, thus, effectively generates multi-layer clothing consistent with the text prompts. Moreover, to ensure accurate geometric alignment between decoupled clothing and the body, we present a 3D implicit deformation field leveraging SMPL [47] as a clothing proxy for matching clothing with the body. Furthermore, to enhance details, we introduce a normal prediction network for smooth normals, combined with optimized spherical harmonic (SH) lighting. Hence, the proposed HumanCoser can generate reusable and intricate multi-layered dressed 3D humans that can be edited and changed separately as shown in Fig. 1.

Our main contributions are summarized as follows:

- We propose a layered 3D human generation framework with a multi-layer representation decoupling method. To our best knowledge, this is the first work that can make the 3D dressed human truly decoupled physically and support layered generation and editing for 3D dressed human. We also introduce a decoupled shape prior to generating structurally consistent 3D content.
- We propose a dual-representation decoupling strategy to improve the semantic consistency of generated clothing, combined with an innovative multi-layer fusion volumetric rendering approach. The strategy not only improves the semantic consistency of the clothing but is also generalizable to the enhancement of 3D semantics for other wearable outfits of humans.
- We propose a 3D implicit deformation method based on SMPL vertex prediction to achieve the geometric matching of human bodies and clothing in an implicit manner so that the clothing can be transferred between different human subjects.

2 Related Work

Text-guided 3D Content Generation. CLIP-Forge [33] and Dream-Field [13] optimize Neural Radiance Fields (NeRFs) to generate 3D shapes by aligning the embedding of the generated image with the text embedding in the space of the image-text model CLIP. CLIP-mesh [23] also uses CLIP to optimize meshes to represent 3D shapes. However, by directly generating images aligned with text in CLIP space, it is not possible to generate highly realistic images. Recently, diffusion modeling [24, 30, 32] has seen rapid growth due to its excellent performance in synthesizing high-quality images. DreamFusion [27] proposes Score Distillation Sampling (SDS) based on a pre-trained diffusion model [30] to optimize trainable NeRFs. Magic3D [20] uses a 2-stage training strategy to bootstrap 3D texture networks to optimize 3D content generation. Although the above diffusion-based 3D generation models have some 3D generation capability, the generation of 3D human is a challenge for the above methods due to the complexity of their shapes and the diversity of their poses.

Text-guided 3D Human Generation. AvatarCLIP [8] initializes the 3D human body shape via a VAE (Variational Autoencoder) encoder and then performs geometric shaping and texture generation guided by an image-text model [28]. However, since the method focuses on shaping localized structures, it lacks in the generation of global structures such as skirts, long hair and loose clothing. In addition, Latent-NeRF [21] and TADA [19] both utilize pre-trained text-to-image diffusion models for 3D avatar generation work. In particular, Latent-NeRF [21] employs a Sketch-Shape to constrain the generation of the diffusion model, but the results of this method lack details due to the lack of optimization of normals and illumination. TADA [19] is limited by the representation ability of the confined mesh, and thus cannot represent non-convex structures or transparent materials well. Neither of the above methods can generate 3D avatars with layer-wise bodies and clothing. Dreamhuman [16] produces animatable coupled avatar based on text and human posture. It combines 3D human prior to generate and re-pose the generated results, but it cannot arbitrarily adjust and replace the clothes of humans without retaining human identity. Avatarcraft [14] transforms text into a 3D avatar, using a diffusion model to stylize geometry and texture, while shape and pose are controlled by a parametric human model. Avatarcraft uses a bare neural human avatar as a template. Given a text prompts, Avatarcraft uses the diffusion model to guide the creation of the avatar by updating the template so that the geometry and texture are consistent with the text. Although Avatarcraft updates the avatar with new pose and shape parameters without training, the generated avatar hardly shows details such as loose clothing and fluffy hair. Dreamwaltz [12] generates 3D digital avatars from text prompts, leveraging prior knowledge of human body shapes and poses, and facilitating animation and interactive compositions between avatars, objects, and scenes. It learns the distribution of human animations through prior knowledge of human actions, enabling the generation of plausible human animations. However, Dreamwaltz's learnable human action deformation module lacks generalizability for generating multi-layered humans, thereby hindering capabilities such as dress-up and clothing editing. DreamAvatar [2] uses SMPL for shape guidance and introduces a dual-observation space design to optimize shape and pose jointly. It addresses the "Janus" problem and enhances facial details. However, it fails to fully consider human body occlusion information, and it also couples clothing and human body generation. Distinct from non-layered methods, HumanLiff [9] generates human body based on the diffusion model in a layer-by-layer manner. However, the features depend on the tri-plane features of the previous layer. This coupling of features among layers impedes the separate editing and reuse of each layer.

Table 1: Comparison of 3D human generation methods, including layered generation, geometric complexity, clothing transfer and clothing reusability.

Method	Multilayer	Geometry (non-skin tight)	Clothing Transfer	Reusability
AvatarCLIP [8]	✗	✗	✗	✗
TADA [19]	✗	✗	✗	✗
Latent-NeRF [21]	✗	✓	✗	✗
HumanLiff [9]	✓	✓	✗	✗
Ours	✓	✓	✓	✓

In summary, existing methods either generate 3D dressed humans as a whole or support only tight and simple clothing generation, which limits their applications to virtual try-on and part-level editing. In contrast, our method can generate reusable and intricate multi-layered 3D dressed humans that can be edited and changed separately. It achieves realistic body and clothing generation by predicting normals and employing improved spherical harmonic lighting. Moreover, we ensure the semantic consistency of the generated clothing through an optimized dual-representation decoupling framework. The layered clothing can be seamlessly transferred between different shapes of human bodies using an implicit deformation network based on SMPL. We summarize the main differences between our work and related work in Tab. 1.

3 Method

3.1 Overview

The proposed HumanCoser is a two-stage method to generate realistic 3D humans with consistent body structures guided by text

in a layer-wise manner. The first stage (a) is to generate a minimized human body, and the second stage (b) performs decoupled generation of clothing and matches the clothing with the human body. Specifically, as Fig. 2 shown, stage (a) consists of a NeRF and ControlNet with SMPL skeleton conditions as inputs and generates a minimized human body in canonical space (Sec. 3.2). In addition, stage (b) consists of a dual-representation decoupling network (Sec. 3.3) and an implicit deformation network driven by SMPL (Sec. 3.4). In which, the dual-representation decoupling network generates dis-entangled clothing on the basis of the minimized human body combined with a multi-layer fusion rendering method. Finally, the decoupled clothing is matched with the human body through the deformation network mentioned above.

3.2 Canonical Body Generation

In order to obtain the minimized human body, we adopt ControlNet with SMPL as conditional input and generate human body in canonical space, as shown in Fig. 2. We first use NeRF as a representation of layered humans. The inner body and each layer of clothing are represented by a separate network as follows:

$$F_\theta(\gamma(\mathbf{x})) = (\sigma, c), \tag{1}$$

where $\gamma(.)$ is the frequency encoder. We render the scene using the volume rendering equation $F_\theta(.)$ [22]. σ and c denote the density and color predicted by each sampling point $\mathbf{x}$. The color of each layer is predicted as follows:

$$\begin{aligned} C(\mathbf{r}) &= \sum_i w_i c_i, \\ w_i &= \alpha_i \prod_{j<i} (1-\alpha_j), \end{aligned} \tag{2}$$

where $\alpha_i = 1 - \exp(-\sigma_i \|\mathbf{x}_i - \mathbf{x}_{i+1}\|)$ and $\|\mathbf{x}_i - \mathbf{x}_{i+1}\|$ is the interval between sample i and $i+1$. w_i is the weight of the i^{th} sampling point [22]. c_i and σ_i is the predicted color and density of the i^{th} sampling point [22].

Multi-Layer Fusion Rendering. In order to fuse the layered human body and clothing for rendering, we proposed a multi-layer composite rendering method based on the density and weight of each sampling point according to (Eq. (2)), which is defined as follows:

$$C'(\mathbf{r}) = \sum_{i=1}^{N} w_i^j c_i^j, w_i^j = \max\left(w_i^1 \cdots w_i^n\right), \tag{3}$$

where $C'(\mathbf{r})$ is the rendering formula (Eq. (2)), w_i^j is the weight with the highest density in the n-layer component, and c_i^j is the corresponding color. In addition, in order to make the generated surface normals smoother, we calculate the normal loss between the predicted normal $\mathbf{n}'$ and the surface normal $\mathbf{n}$:

$$\mathscr{L}_n = \sum_i w_i \left\|\mathbf{n}_i' - \mathbf{n}_i\right\|, \tag{4}$$

where w_i is the weight of the i^{th} sampling point, which follows the definition of Eq. (2). Moreover, in order to regularize the normal and reduce redundant semantically generated artefacts, a loss of regular constraints is added:

$$\mathscr{L}_n^{reg} = \sum_i w_i \left(1 - \sum \mathbf{n}_i' \cdot \mathbf{n}_i\right), \tag{5}$$

where w_i is the weight of the i^{th} sampling point. Definitions of $\mathbf{n}_i'$ and $\mathbf{n}_i$ refer to Eq. (4). Furthermore, we introduce SDS loss to optimize 3D models of the body and clothing:

$$\nabla_\theta \mathscr{L}_{\text{SDS}}(\phi, \mathbf{x}) = E_{t,\varepsilon}\left[\mathbf{w}(t)\left(\varepsilon_\phi(\mathbf{x}_t; y, t, \mathbf{c}) - \varepsilon\right)\frac{\partial \mathbf{x}}{\partial \theta}\right] \tag{6}$$

where $\mathbf{w}(t)$ represents a weight function dependent on the time step t, $\mathbf{x}$ is the noisy image, and y is the input text prompt. The noise injected by ε is added to the rendered image $\mathbf{x}$. To maintain the consistency of the human body structure, we input the SMPL skeleton $\mathbf{c}$ as the conditional image. Therefore, the overall loss of the canonical body generation is as follows:

$$\mathscr{L}_{body} = \lambda_{\text{SDS}}^{body} \mathscr{L}_{\text{SDS}}^{body} + \lambda_n \mathscr{L}_n + \lambda_n^{reg} \mathscr{L}_n^{reg}, \tag{7}$$

where $\mathscr{L}_{\text{SDS}}^{body} = \mathscr{L}_{\text{SDS}}\left(\mathbf{x}^b; y^b, \mathbf{c}^b\right)$, $\mathbf{x}^b$ is the supervised body image, y^b is the prompt of body, and $\mathbf{c}^b$ is the input condition of SMPL skeleton. $\lambda_{\text{SDS}}^{body}$, λ_n, λ_n^{reg} are the weights attributed to each loss. More details of Sec. 3.2 are provided in the supplementary material.

3.3 Dual-Representation Decoupling

In order to accurately obtain the shape of clothing, we introduce a dual-representation decoupling framework (DRD) to eliminate the parts that are inconsistent with the semantics of clothing.

Decoupling Clothing Representation. As illustrated in Fig. 2(b), the DRD model consists of a multi-layer component composition network and a clothing generation network. During the training of the clothing component at the N^{th} layer, we take the density of the sampling point with the largest weight in the first $N-1$ layers as the combination density of the first $(N-1)^{th}$ layers. This combined density is defined as follows:

$$\underset{\delta}{\text{argmax}}(w(\delta)), \delta \in \{\delta_1 \ldots \delta_{N-1}\}, \tag{8}$$

where $w(.)$ is defined in Eq. (2). The final combination weight $w(\delta)$ is then calculated based on the combined density δ. We use the following loss to constrain the density of the overlapping parts of the N^{th} clothing and the first $N-1$ components:

$$\begin{aligned} \mathscr{L}_{reg\_ds} &= \left\|w\left(\delta_c^N\right)\right\|_2, \\ &\left\{\delta_c^N \mid w\left(\delta_c^N\right) > \lambda \cup \delta_c^N < \delta_{bc}\right\}, \end{aligned} \tag{9}$$

where $w(.)$ is Eq. (2) which only considers the input of density, δ_c^N is the density of the N^{th} clothing component, δ_{bc} is the combination density of the first $N-1$ components, and λ is the defined threshold. Finally, as shown in Fig. 2(b), we perform a composite rendering of the N^{th} clothing component and the first $N-1$ components based on the Eq. (2) of multi-layer fusion rendering. The composite rendering is defined as follows:

$$\begin{aligned} C_{comp}(r) &= \sum_{x_i \in M_c} w_i c_c(x_i) + \sum_{x_j \in M_{bc}} w_j c_{bc}(x_j), \\ M_c &= \left\{x_i \mid w(x_i) < \lambda \cup \delta(x_i) > \delta(x_j)\right\}, \\ M_{bc} &= \bar{M}_c, \end{aligned} \tag{10}$$

where $C_{comp}(.)$ is the rendering formula Eq. (2), x is the sampling point, $w(x)$ and $\delta(x)$ is the weight and density of x, λ is the defined threshold, $c_c(x)$ and M_c is the predicted color and set of x for the N^{th} cloth component, $c_{bc}(x)$ and M_{bc} is the predicted color and set of x for the first $N-1$ components.

Dual SDS optimization. By using the above method, we obtain a rendered image composed of N components. However, direct utilization of this outcome for the training of clothing leads to issues with semantic inconsistency with clothing. As shown in Fig. 2(b), apart from utilizing SDS loss to supervise the composite rendering results, we also employ a single volume rendering combined with Diffusion model solely to supervise the clothing. After using the above decoupling strategy, we thus get clothing that is disentangled

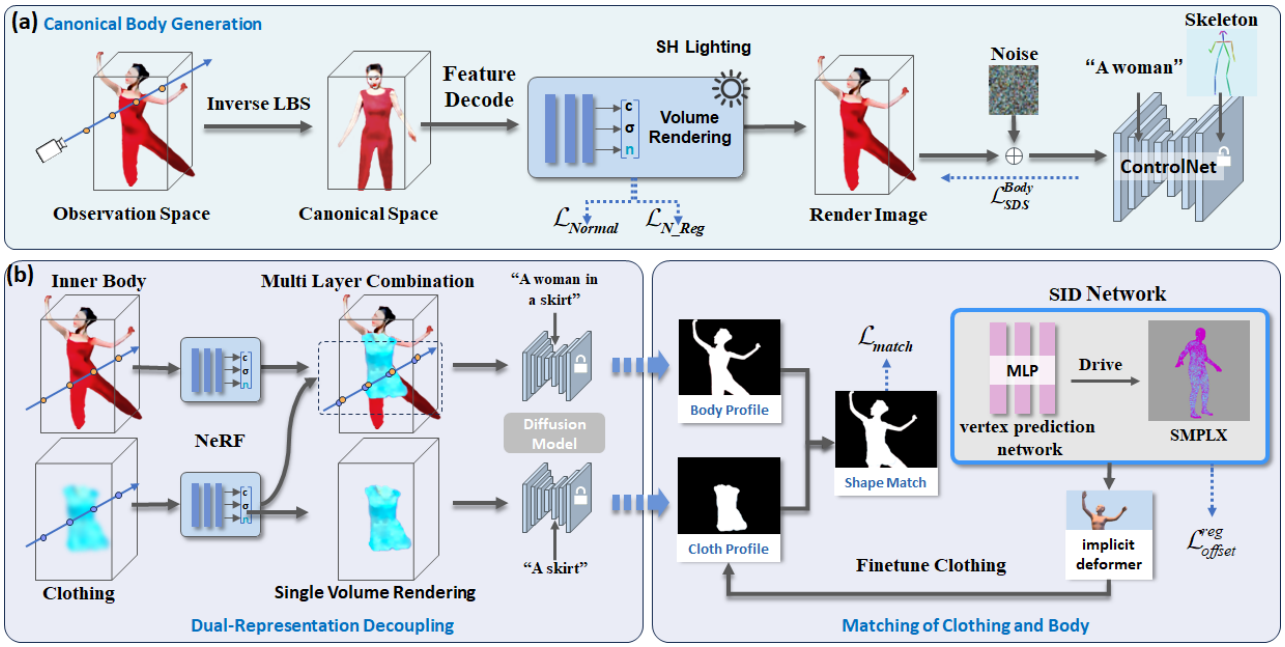

Figure 2: Illustration of our framework for generating the clothes and body of a dressed human in a layered manner. (a) shows the generation of the minimized body, and (b) shows the layered generation of clothing and the matching of clothing with the body.

with the body. Additionally, we introduce NeRF density regularization loss $\mathscr{L}_r(.)$ to eliminate floating clouds. The loss function of the decoupling generation stage for clothing is as follows:

$$\mathscr{L}_{clothing} = \lambda_{\text{SDS}}^{cloth}\mathscr{L}_{\text{SDS}}^{cloth} + \lambda_{\text{SDS}}^{comp}\mathscr{L}_{\text{SDS}}^{comp} + \lambda_{reg\_ds}\mathscr{L}_{reg\_ds} + \lambda_r\mathscr{L}_r, \tag{11}$$

where $\mathscr{L}_{\text{SDS}}^{cloth} = \mathscr{L}_{\text{SDS}}(\mathbf{x}^c; y^c)$, $\mathbf{x}^c$ is the supervised clothing image, y^c is the prompt of clothing. $\mathscr{L}_{\text{SDS}}^{comp} = \mathscr{L}_{\text{SDS}}(\mathbf{x}^{cp}; y^{cp})$, $\mathbf{x}^{cp}$ is the supervised composite image, y^{cp} is the prompt of composite image. $\lambda_{\text{SDS}}^{cloth}$, $\lambda_{\text{SDS}}^{comp}$, λ_{reg_ds}, λ_r are the weights attributed to each loss.

3.4 Matching of Clothing and Body

In order to perform fine deformation of the clothing shape to fit the body, we introduce the SMPL-driven implicit deformation network (SID Net), as shown in Fig. 2(b). Furthermore, for precise clothing editing, we use SMPL-X [25] for our clothing shape proxy and add learnable vertex offsets o for each shape proxy. At the same time, we use the vertex prediction model $o = F_v(v)$ to predict the offset o of each vertex v of the SMPL shape proxy. The specific implementation of SMPL to drive the clothing to match the body is as follows:

Optimizing vertices. Given the body SMPL parameters (β, θ), the vertex offset $F_v : v \rightarrow o$ and the camera parameter ρ, we render a mesh proxy of the body as a binary mask image $\mathscr{R}_m\left(M_{body}(\beta,\theta,o),\rho\right) \rightarrow I_{smpl}^{body}$, where $\mathscr{R}_m$ is a differentiable raster renderer. At the same time, we render a meshes proxy of the clothing as binary mask images (where we use the SMPL model excluding vertices of the head, hands and feet) $\mathscr{R}_m\left(M_{cloth}(\beta,\theta),\rho\right) \rightarrow I_{smpl}^{cloth}$. Then, since the body masks should be within the region where the clothing proxy I_{mask}^{cloth} rendered by NeRF and I_{smpl}^{cloth} are merged, we perform the optimization using the following loss:

$$\mathscr{L}_{match} = \mathscr{L}_{huber}\left(I_{mask}^{cloth} + I_{smpl}^{cloth} - I_{smpl}^{body}\right), \tag{12}$$

where $\mathscr{L}_{huber}(.)$ [3] is a smoothed loss function. Also to smooth the predicted vertex offsets, we introduce a regularization loss for the vertex offset o:

$$\mathscr{L}_{offset}^{reg} = \|o\|_2, \tag{13}$$

where o contains the predicted vertex offsets for all vertices. Then, we update the vertex prediction model $F_v : v \rightarrow o_{opt}$ using the gradient of the $\mathscr{L}_{match}$ loss to obtain the optimized vertex offsets o_{opt} for optimizing the implicit geometry of the clothing. More details of Sec. 3.4 are provided in the supplementary material. The loss of clothing matching is delineated as follows:

$$\mathscr{L}_{matching} = \lambda_{match}\mathscr{L}_{match} + \lambda_{reg}\mathscr{L}_{offset}^{reg}, \tag{14}$$

where λ_{match}, λ_{reg} are the weights attributed to each loss. In conclusion, the overall loss of the decoupled generation and matching of bodies and clothing is as follows:

$$\mathscr{L}_{all} = \mathscr{L}_{body} + \mathscr{L}_{clothing} + \mathscr{L}_{matching}. \tag{15}$$

4 Experiments

In this section, we assess the efficacy of our proposed layered human generation framework. We commence by providing implementation details in Sec. 4.1, followed by generated results in Sec. 4.2. Subsequently, quantitative and qualitative comparisons between state-of-the-art methods and ours are presented in Sec. 4.3. To evaluate the effectiveness of proposed modules, ablation studies are discussed in Sec. 4.4. Finally, we showcase the applications of our method. Please refer to the demo case for some experimental results.

4.1 Implementation Details

Hyperparameters. We use ISM [18] to compute the SDS loss with normal CFG (7.5) for all stages. The warm-up period of ISM is 1,000 iterations. (1) *Canonical Body Generation:* The loss weights, $\lambda_{\text{SDS}}^{body}$, λ_n and λ_n^{reg}, are 1.0, 0.01, 0.05, respectively. The gradient scaling factor of ISM is 0.1. (2) *Dual-Representation Decoupling:* The loss function weights, $\lambda_{\text{SDS}}^{cloth}$, $\lambda_{\text{SDS}}^{comp}$, λ_{reg_ds} and λ_r, are 1.0, 1.0, 0.05, 2.0, respectively. The gradient scaling factor of ISM is

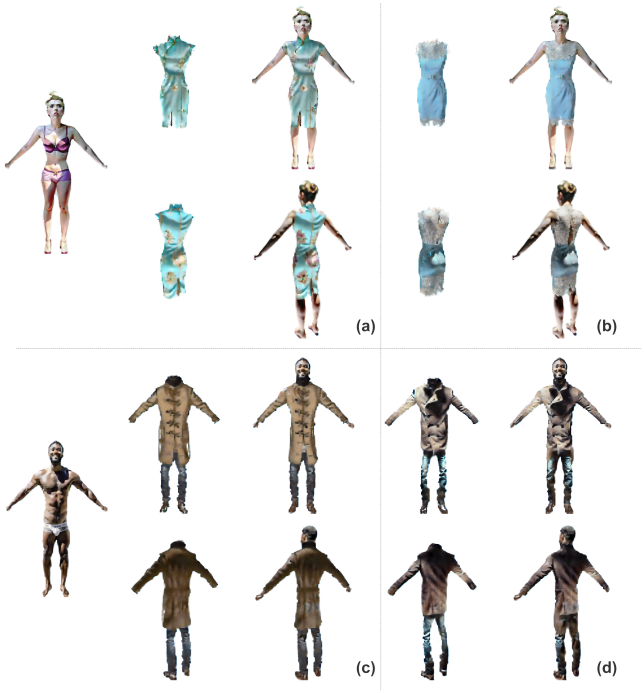

Figure 3: The decoupled generation of human body and clothing by our method. (a) clothing prompt: "A turquoise Cheongsam", (b) clothing prompt: "A deep-skyblue sleeveless sheath dress with lace trims", (c) clothing prompt: "A Duffle Coat and a baggy linen pants", (d) clothing prompt: "A Car Coat and a baggy jeans".

0.07. (3) *Matching of Clothing and Body:* The loss weights, λ_{match} and λ_{reg}, are 10.0, 1.0, respectively.

Training Details. The overall framework is trained using Adam optimizer, with the *betas* of $[0.9, 0.99]$ and the learning rates of $5e-5$, $1e-3$ for the stage of decoupling dressed human and clothing-matching, respectively. The training of the body and clothing in the decoupling stage takes 12,000 and 8,000 iterations. Specifically, alternate training is used in clothing training, and the training ratio of the Nth layer to the combination of the first N layers is $1:6$. The training of clothing-matching requires 3,000 iterations. We use the training resolution of 512×512 with a batch size of 2 and the whole optimization process takes three hours on a single NVIDIA 4096 GPU. Further training details are available in the SupMat.

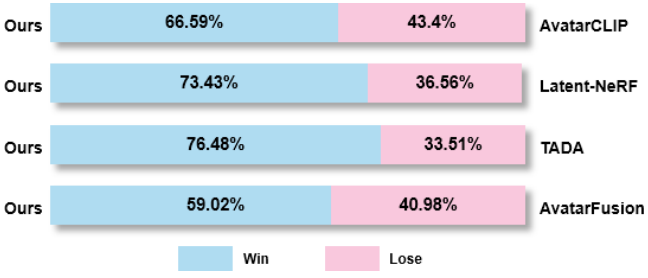

Figure 4: Quantitative results. Our method and methods [8, 21, 19, 10] are evaluated by using the method [15] to measure the visual quality of the generated 3D content, where higher scores are better.

4.2 Generated Results

We present physically-layered generated results in Fig. 3. When the same person is dressed in different clothes, our method generates 3D clothing that conform to the body shape. For example, when a woman wears a turquoise Cheongsam and then switches to a blue dress, our method generates her in the dress with a fitting body shape. In addition to all-in-one clothing, our method is capable of generating 3D humans in complex clothes, such as a man wearing a coat and pants or jeans. Notably, the generated clothes conform better to the body, including the waist position, suggesting that our physically-layered model not only accommodates various clothing changes but also ensures a better fit to the human body, resulting in a more natural appearance.

Table 2: Quantitative comparisons with non-layered and Layered methods.

Method	FID ↓	CLIP Score ↑
AvatarCLIP [8]	311.46	30.88
Latent-NeRF [21]	329.40	29.82
TADA [19]	392.61	25.39
AvatarFusion[10]	375.97	26.96
HumanLiff [9]	324.69	26.34
HumanCoser (Ours)	**298.54**	**31.61**

4.3 Comparison

We compare our approach with five SoTA methods. (1) AvatarCLIP [8] uses pre-trained vision-language CLIP model to guide NeuS [36] for 3D avatar generation; (2) TADA [19] creates 3D avatars from text by using hierarchical rendering with score distillation sampling; (3) Latent-NeRF [21] introduces sketch shape loss based on 3D shape guidance to supervise the training; (4) AvatarFusion [10] can generate avatars while simultaneously segmenting clothing from the avatar's body; (5) HumanLiff [9] firstly generates minimally clothed humans, represented by tri-plane features, in a canonical space and then progressively generates clothes in a layer-wise manner.

4.3.1 Quantitative Results

This section quantitatively compares the proposed method with [8, 19, 21, 10, 9]. Inspired by [15], we use user preference metrics to compare the generation quality to the SoTA methods [8, 19, 21, 10]. Fig. 4 demonstrates the superior performance of our method compared to [8, 19, 21, 10] in generation quality. Additionally, we calculate the FID [6] between the views rendered from the generated 3D humans and the images produced by Stable Diffusion [31]. As

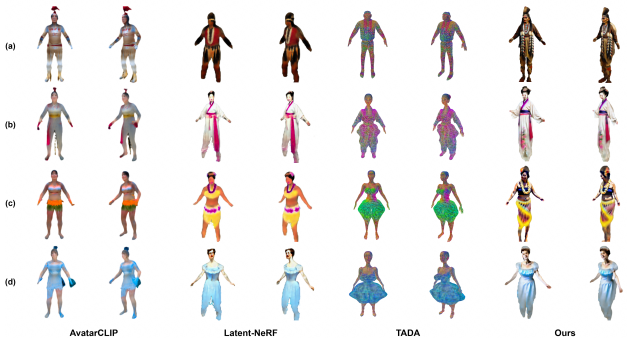

Figure 5: Qualitative comparison with coupled generation methods [8, 19, 4]. (a) prompt: "A north American Indian chief in full regalia", (b) prompt: "A Chinese lady wearing a gauzy hanfu", (c) prompt: "A Hawaiian woman wearing a hula skirt", (d) prompt: "A French woman wearing a light blue crinoline dress".

Table 3: User Study Results. We investigated user evaluations on geometric and texture quality, as well as consistency with text prompts.

Case	**AvatarCLIP** [8]			**TADA** [19]			**Latent-NeRF** [21]			**Ours**		
	Geometry	Texture	Text	Geometry	Texture	Text	Geometry	Texture	Text	Geometry	Texture	Text
case 1	2.41	2.82	2.88	2.58	2.74	2.34	3.18	3.22	4.04	3.72	4.26	4.29
case 2	3.85	2.79	2.85	2.24	2.48	2.72	4.06	2.86	3.29	4.53	4.51	4.16
case 3	3.05	3.07	2.33	2.47	2.31	2.42	3.58	2.27	3.82	4.61	3.79	4.63
case 4	2.57	2.51	3.27	2.78	2.43	2.51	3.12	3.76	2.89	3.08	3.74	4.53
case 5	3.24	2.46	2.74	2.54	2.10	2.03	3.41	3.54	3.24	4.66	3.64	3.79
case 6	2.59	2.41	3.95	3.02	2.61	2.49	3.14	3.57	3.17	4.16	3.93	4.89
case 7	2.37	2.60	2.58	2.67	1.83	2.17	3.70	3.43	3.95	4.68	4.13	4.23
case 8	2.55	3.11	2.08	2.57	2.26	1.90	2.97	3.65	3.71	4.32	3.73	4.64
case 9	2.88	2.93	3.08	2.51	2.10	2.33	3.87	3.81	3.22	4.47	4.40	4.11
case 10	3.79	2.40	2.64	1.92	2.74	2.29	2.47	2.79	2.97	4.37	3.97	4.36
Average	3.79	2.71	2.84	2.53	2.36	2.32	3.35	3.29	3.43	**4.26**	**4.01**	**4.37**

shown in the Tab. 2, our method achieves the lowest FID score, indicating the best generation quality. Furthermore, we adopt CLIP score [5] to measure the compatibility between the prompts with the rendered views of 3D humans. Tab. 2 shows our method achieves the highest CLIP score, indicating that the human model generated by our framework is more aligned with the prompt. Compared to layered-generation SoTA methods [10] and [9], our method not only achieves better generation quality, but also freely performs clothing transfer and generalizable animations.

Furthermore, we perform a user study comparing the human generation results of our method with those of other state-of-the-art methods [8, 19, 21]. We generate 3D human for different methods based on 10 text prompts. Fifty volunteers (including 26 males and 24 females, aged between 18 and 50 years) were invited to rank the methods in terms of (1) geometric quality, (2) appearance quality, and (3) consistency with the text prompts. Volunteers score each comparative indicator for each method from 1 (worst) to 5 (best). The final evaluation results are provided in Tab. 3. Our method achieves optimal scores across all three metrics, indicating superior generative quality for geometry and texture based on text inputs.

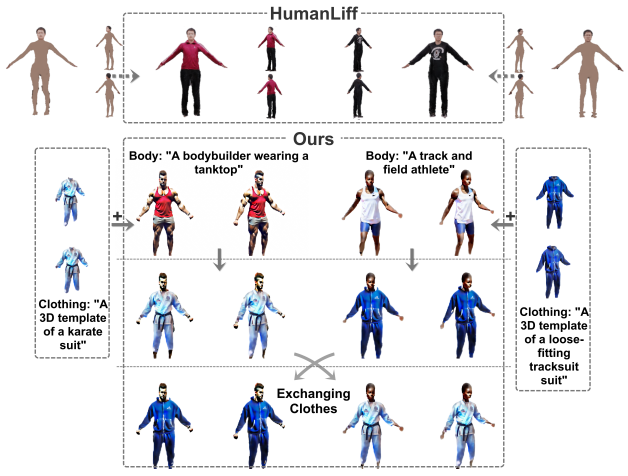

Figure 6: Qualitative comparison with the layered method [9].

4.3.2 Qualitative Results

Fig. 5 qualitatively compares to text-guided 3D generation methods [8, 21, 19]. Considering that [8, 21, 19] are based on coupled generation, we provide a coupled generation model for comparison. We render the model as multiple views for comparison. As shown in Fig. 5, although AvatarCLIP [8] generates view-consistent human bodies, it demonstrates limitations in effectively modeling

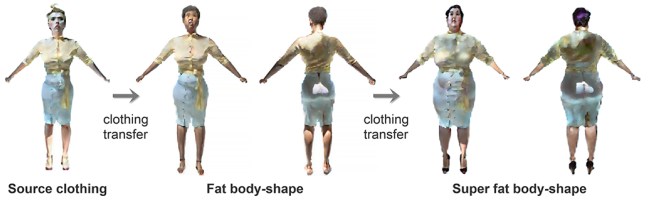

Figure 7: Editing results for adaptive matching of clothing to different body shapes.

global structures, such as skirts and long hair. Latent-NeRF [21] exhibits a limitation in its capacity to finely generate both geometry and texture. TADA [19] accuracy depends on the density of the mesh, and the discrete representation affects its geometric appearance. So, [8, 21, 19] exhibit deficiencies, either in the representation of geometric details or in the portrayal of fine textures. In contrast, our method produces humans characterized by enhanced geometric details, including loose clothing and diverse long hair, along with finer textures.

In addition, Fig. 6 illustrates the comparison of the layered 3D human generation approaches [9]. Since AvatarFusion [10] is not capable of multi-layer generation, we use HumanLiff [9][1] for the comparison of layered generation. HumanLiff [9] stands out as the most akin work to our method, employing a layer-by-layer generation approach. However, it lacks the capability to change clothes, as illustrated in the top row. HumanLiff generates a clothed human body by relying on a minimally-clothed human body. Instead, our method demonstrates the ability to generate the body and clothing independently, as depicted in the second row in Fig. 6. Subsequently, it engages in the matching of clothing and body, showcased in the third row. Finally, our method excels in the process of changing and reusing clothing, as illustrated in the last row in Fig. 6.

It's important to highlight that our method not only facilitates the transfer and matching of clothing across bodies of varying shapes but also enables the generation of multi-layer clothing using multi-layer fusion volume rendering. Fig. 7 shows that the clothing can adaptively match different shapes of body by our method including even extreme body shapes, i.e. the "super fat woman". Fig. 8 shows that a lady is wearing two-layer clothes, i.e. a dress as well as an outer clothing. Two distinct views showcase the harmony and naturalness achieved by our method in multi-layer clothing.

[1]HumanLiff currently does not provide the official implementation, and hence we compare with the visual results presented in [9]

Figure 8: The effectiveness of multi-layer decoupled clothing.

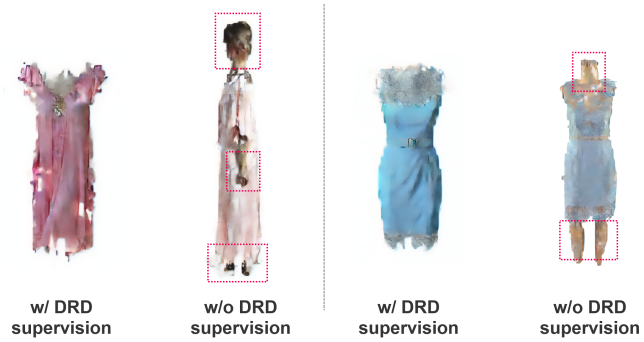

Figure 9: Ablation study on the effectiveness of the dual-representation decoupling framework.

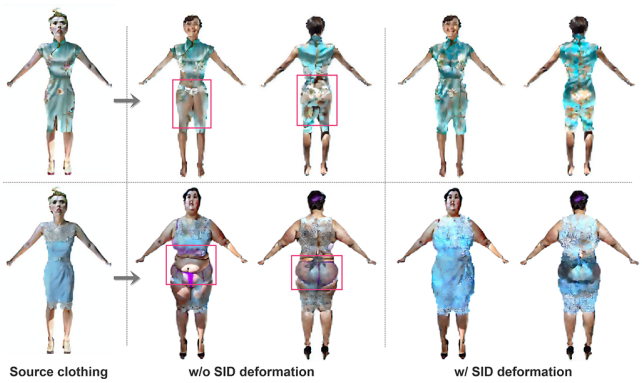

Figure 10: Ablation study on the implicitly deformed modules.

4.4 Ablation Study

Effectiveness of Dual-Representation Decoupling Framework. To assess the effectiveness of the dual-representation decoupling framework (DRD), we investigate the impact of employing dual SDS losses on clothing generation, as depicted in Fig. 9. Our findings indicate only utilizing a single SDS loss alongside a single volumetric rendering fails to accurately decouple the clothing from the human body and may result in incorrect clothing shapes as shown in the red box in Fig. 9. This is attributed to the single SDS loss supervising clothing generation, leading to the production of redundant non-clothing parts. However, by incorporating additional SDS to supervise the combined results of the human body and the clothing, we observe a significant improvement. This augmentation enables the elimination of redundant non-clothing parts and maintains semantic consistency with the clothing. Consequently, the proposed dual-representation decoupling framework validates its efficacy in generating intricate and semantically consistent clothing.

Effectiveness of Implicitly Deformed Modules. To adaptively match the decoupled clothing to different body shapes, we introduce the SMPL-driven implicit field deformation network (SID Net). As seen from the red boxes in Fig. 10, the decoupled clothing is directly matched to different body shapes, which leads to the issue of interpenetration between the clothing and the body, and the clothing does not fit tightly and naturally to the body. Our SID Net can optimize the SMPL proxy model of the clothing to deform the implicit field of the clothing to match the body by calculating the shape deviation loss between the clothing and the body. As can be seen from columns 4,5 of Fig. 10, arbitrarily decoupled clothing can be freely and accurately matched with bodies of different shapes, even including extreme shapes of the human body, such as a super-fat or a very thin person. Our SID Net is validated to efficiently perform adaptive clothing-body matching via the above visualization results.

Effectiveness of Optimizable Spherical Harmonic (SH) Lighting. As detailed in Sec. 3.2, to mitigate the problem of color oversaturation stemming from SDS loss in the diffusion model, we introduced an optimizable SH lighting component to modulate the color of the sample point. As depicted in the red box in Fig. 11

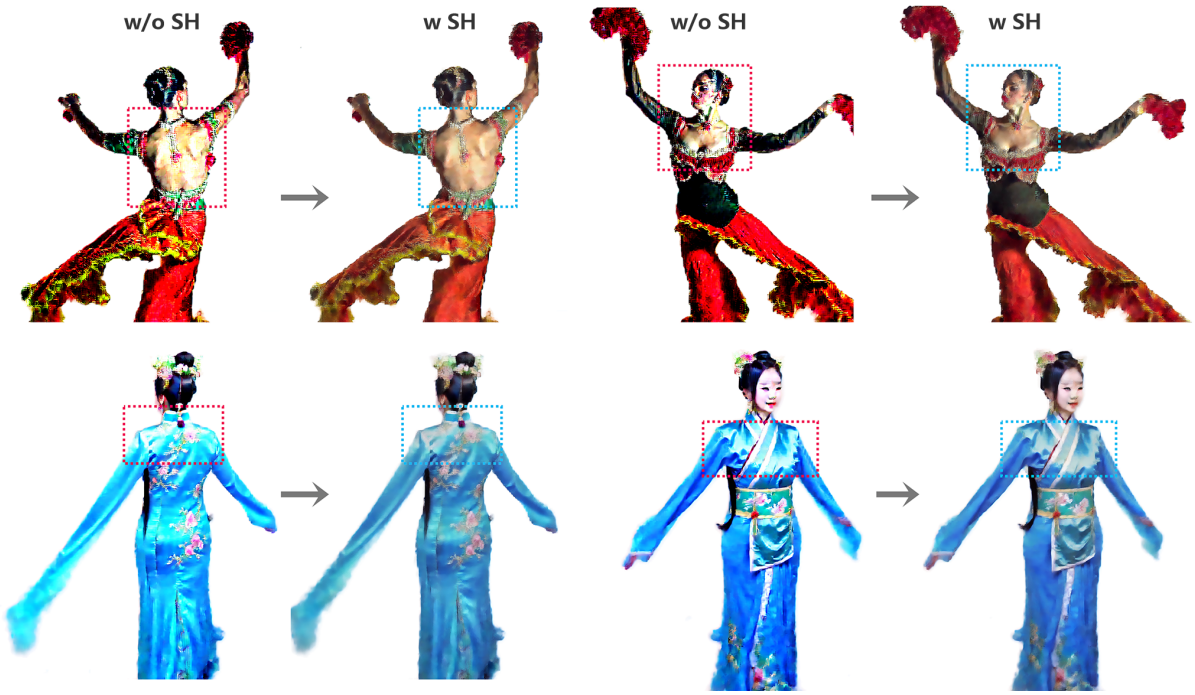

Figure 11: Ablation study on the effectiveness of spherical harmonic (SH) lighting.

without incorporating SH lighting, the color of the 3D dressed human exhibits oversaturation and lacks smoothness in surface rendering. Contrastingly, the blue box in Fig. 11 illustrates that integrating SH lighting enables the human model to achieve the correct coloration and a smoother visual effect. This enhancement not only addresses the issue of oversaturation but also contributes to improving the overall realism and visual fidelity of the rendered human models. The addition of SH lighting introduces subtle variations in color and shading, resulting in a more natural appearance that better aligns with real-world lighting conditions. Hence, this approach enhances the quality and believability of the generated results, providing more accurate representations of dressed human subjects.

4.5 Application

Thanks to our capability of generating layered 3D humans, our method also has the ability to transfer clothing across people and enable skeleton-driven layered human animation.

Clothing Transfer. Fig. 12 evaluates the effectiveness of our model in clothing transfer by exchanging avatars' clothes (left/right). In this case, the layered avatars are generated based on different SMPL shapes θ with the same pose β. We transfer the clothing layer of avatar left to the body layer of avatar right and vice versa: $(\text{cloth}^{\text{avatar left}} \rightarrow \text{body}^{\text{avatar right}}, \text{cloth}^{\text{avatar right}} \rightarrow \text{body}^{\text{avatar left}})$. Fig. 12 illustrates our model excels in adaptively shaping a match between the body and clothing layers, facilitating the transfer of the same clothing layer across different identity-based body layers.

Generalizable Poses and Animations. Fig. 13 demonstrates the effectiveness of SMPL skeleton-driven layered human animation by applying complex animations and poses to the body and clothing layers. We learn a generalizable density-weighted network by sampling the pose of the SMPL from the pre-trained VPoser model as conditional inputs to the ControlNet. This refines the SMPL-based pose deformations and supports SMPL-driven animations and complex poses without additional training.

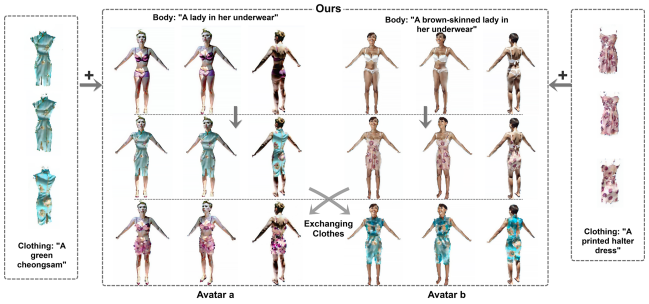

Figure 12: The effectiveness of clothing transfers.

Figure 13: The effectiveness of Pose-Driven Generation.

5 Conclusion and Limitations

Conclusion. This paper introduces a layer-wise dressed human generation framework built upon a physically-decoupled diffu-

sion model. Central to our approach are the concepts of a dual-representation decoupling framework and a novel multi-layer fusion volumetric rendering technique. Building upon this decoupled representation, we achieve multi-layer 3D human wearing loose-fitting clothing while the existing coupled methods struggle to achieve layered dressed human. Additionally, unlike other methods that fail to arbitrarily change and exchange clothing, we introduce an implicit deformation module, guided by the SMPL model, which allows clothing to adaptively match different body shapes. Experimental results showcase that our method outperforms state-of-the-art approaches by generating high-quality multi-layered 3D humans wearing complex clothing and arbitrarily switching clothing across various body shapes.

Limitations. Given the absence of a uniform parametric clothing template, the assessment of matching loss to the body cannot be conducted through differentiable rendering employing a uniform 3D proxy tailored to the generated clothing. Consequently, we opt for a 3D implicit deformation field based on SMPL-X [25] to optimize the alignment between bodies and clothing. While our method enables the fitting of the clothing to various body shapes, it may yield unnatural matching outcomes when the shapes of the body and clothing differ significantly. In future, we will employ more accurate deformation proxies combined with object collision detection to optimize the matching of clothing and body bidirectionally in order to achieve better quality of layered generation.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62122058 and 62171317), and the Science Fund for Distinguished Young Scholars of Tianjin (No. 22JCJQJC00040).

References

[1] Y. Cao, Y.-P. Cao, K. Han, Y. Shan, and K.-Y. K. Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. *arXiv preprint arXiv:2304.00916*, 2023. 1

[2] Y. Cao, Y.-P. Cao, K. Han, Y. Shan, and K.-Y. K. Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. *arXiv preprint arXiv:2304.00916*, 2023. 2

[3] H. C. Carver, A. O'TOOLE, and T. RAIFORD. *The annals of mathematical statistics*. Edwards Bros., 1930. 4

[4] G. Gkioxari, J. Malik, and J. Johnson. Mesh r-cnn. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9785–9795, 2019. 1, 5

[5] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 6

[6] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[7] F. Hong, Z. Chen, Y. Lan, L. Pan, and Z. Liu. Eva3d: Compositional 3d human generation from 2d image collections. *arXiv preprint arXiv:2210.04888*, 2022. 1

[8] F. Hong, M. Zhang, L. Pan, Z. Cai, L. Yang, and Z. Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*, 2022. 2, 5, 6

[9] S. Hu, F. Hong, T. Hu, L. Pan, H. Mei, W. Xiao, L. Yang, and Z. Liu. Humanliff: Layer-wise 3d human generation with diffusion model. *arXiv preprint arXiv:2308.09712*, 2023. 2, 5, 6

[10] S. Huang, Z. Yang, L. Li, Y. Yang, and J. Jia. Avatarfusion: Zero-shot generation of clothing-decoupled 3d avatars using 2d diffusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 5734–5745, 2023. 5, 6

[11] Y. Huang, J. Wang, A. Zeng, H. Cao, X. Qi, Y. Shi, Z.-J. Zha, and L. Zhang. Dreamwaltz: Make a scene with complex 3d animatable avatars. *arXiv preprint arXiv:2305.12529*, 2023. 1

[12] Y. Huang, J. Wang, A. Zeng, H. Cao, X. Qi, Y. Shi, Z.-J. Zha, and L. Zhang. Dreamwaltz: Make a scene with complex 3d animatable avatars. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[13] A. Jain, B. Mildenhall, J. T. Barron, P. Abbeel, and B. Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 867–876, 2022. 2

[14] R. Jiang, C. Wang, J. Zhang, M. Chai, M. He, D. Chen, and J. Liao. Avatarcraft: Transforming text into neural human avatars with parameterized shape and pose control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14371–14382, 2023. 2

[15] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569*, 2023. 5

[16] N. Kolotouros, T. Alldieck, A. Zanfir, E. Bazavan, M. Fieraru, and C. Sminchisescu. Dreamhuman: Animatable 3d avatars from text. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[17] N. Kolotouros, T. Alldieck, A. Zanfir, E. G. Bazavan, M. Fieraru, and C. Sminchisescu. Dreamhuman: Animatable 3d avatars from text. *arXiv preprint arXiv:2306.09329*, 2023. 1

[18] Y. Liang, X. Yang, J. Lin, H. Li, X. Xu, and Y. Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. *arXiv preprint arXiv:2311.11284*, 2023. 4

[19] T. Liao, H. Yi, Y. Xiu, J. Tang, Y. Huang, J. Thies, and M. J. Black. Tada! text to animatable digital avatars. *arXiv preprint arXiv:2308.10899*, 2023. 2, 5, 6

[20] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 300–309, June 2023. 2

[21] G. Metzer, E. Richardson, O. Patashnik, R. Giryes, and D. Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12663–12673, 2023. 1, 2, 5, 6

[22] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[23] N. Mohammad Khalid, T. Xie, E. Belilovsky, and T. Popa. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 conference papers*, pp. 1–8, 2022. 2

[24] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[25] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10975–10985, 2019. 4, 9

[26] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1

[27] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2

[28] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021. 2

[29] A. Raj, S. Kaza, B. Poole, M. Niemeyer, N. Ruiz, B. Mildenhall, S. Zada, K. Aberman, M. Rubinstein, J. Barron, et al. Dreambooth3d: Subject-driven text-to-3d generation. *arXiv preprint arXiv:2303.13508*, 2023. 1

[30] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. 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. 1, 2

[31] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. 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. 5

[32] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding, 2022. *URL https://arxiv. org/abs/2205.11487*, 4. 2

[33] A. Sanghi, H. Chu, J. G. Lambourne, Y. Wang, C.-Y. Cheng, M. Fumero, and K. R. Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18603–18613, 2022. 2

[34] G. Te, X. Li, X. Li, J. Wang, W. Hu, and Y. Lu. Neural capture of animatable 3d human from monocular video. In *European Conference on Computer Vision*, pp. 275–291. Springer, 2022. 1

[35] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 52–67, 2018. 1

[36] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 5

[37] Z. Weng, Z. Wang, and S. Yeung. Zeroavatar: Zero-shot 3d avatar generation from a single image. *arXiv preprint arXiv:2305.16411*, 2023. 1

[38] Y. Xiu, J. Yang, X. Cao, D. Tzionas, and M. J. Black. Econ: Explicit clothed humans obtained from normals. *arXiv preprint arXiv:2212.07422*, 2022. 1

[39] J. Zhang, Z. Jiang, D. Yang, H. Xu, Y. Shi, G. Song, Z. Xu, X. Wang, and J. Feng. Avatargen: a 3d generative model for animatable human avatars. In *European Conference on Computer Vision*, pp. 668–685. Springer, 2022. 1