AvatarFusion: Zero-shot Generation of Clothing-Decoupled 3D Avatars Using 2D Diffusion

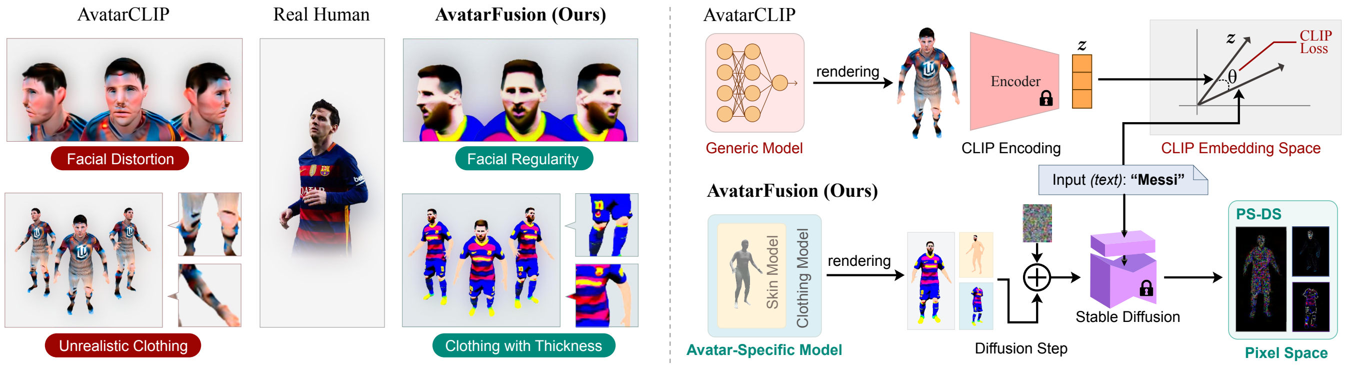
Shuo Huang
huangs22@mails.tsinghua.edu.cn
Department of Computer Science and Technology, Tsinghua University
Beijing, China

Zongxin Yang
yangzongxin@zju.edu.cn
ReLER, CCAI, Zhejiang University
Zhejiang, China

Liangting Li
llt21@mails.tsinghua.edu.cn
Department of Computer Science and Technology, Tsinghua University
Beijing, China

Yi Yang
yangyics@zju.edu.cn
ReLER, CCAI, Zhejiang University
Zhejiang, China

Jia Jia*
jjia@tsinghua.edu.cn
Department of Computer Science and Technology, Tsinghua University
Beijing National Research Center for Information Science and Technology
Beijing, China



(a) Facial Distortion and Unrealistic Clothing

(b) Avatar-Specific Model and Pixel-level Supervision

Figure 1: Given a text prompt "Messi", our method (AvatarFusion) effectively alleviates (a) the problem of facial distortion and unrealistic clothing and generates more photo-realistic avatars. To achieve this, we introduce (b) an avatar-specific model and a pixel-level diffusion supervision (PS-DS) which separates the generation of skin and clothing for better realism.

ABSTRACT

Large-scale pre-trained vision-language models allow for the zero-shot text-based generation of 3D avatars. The previous state-of-the-art method utilized CLIP to supervise neural implicit models that reconstructed a human body mesh. However, this approach has two limitations. Firstly, the lack of avatar-specific models can cause facial distortion and unrealistic clothing in the generated avatars. Secondly, CLIP only provides optimization direction for the overall appearance, resulting in less impressive results. To address these limitations, we propose AvatarFusion, the first framework to use a latent diffusion model to provide pixel-level guidance for generating human-realistic avatars while simultaneously segmenting clothing from the avatar's body. AvatarFusion includes the first clothing-decoupled neural implicit avatar model that employs a novel Dual Volume Rendering strategy to render the decoupled skin and clothing sub-models in one space. We also introduce a novel optimization method, called Pixel-Semantics Difference-Sampling (PS-DS), which semantically separates the generation of body and clothes, and generates a variety of clothing styles. Moreover, we establish the first benchmark for zero-shot text-to-avatar generation. Our experimental results demonstrate that our framework outperforms previous approaches, with significant improvements observed in all metrics. Additionally, since our model is clothing-decoupled, we can exchange the clothes of avatars. Code are available on our project page https://hansenhuang0823.github.io/AvatarFusion.

*Corresponding author.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612022

CCS CONCEPTS

• **Computing methodologies → Computer vision problems**.

KEYWORDS

neural implicit models, zero-shot text-to-avatar generation, diffusion models

ACM Reference Format:

Shuo Huang, Zongxin Yang, Liangting Li, Yi Yang, and Jia Jia. 2023. AvatarFusion: Zero-shot Generation of Clothing-Decoupled 3D Avatars Using 2D Diffusion. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3581783.3612022

1 INTRODUCTION

Recently, the zero-shot text-to-avatar generation task has become feasible due to the emergence of large-scale vision-language models [34, 35, 37, 38]. This task utilizes these pre-trained models as supervision to create highly compelling avatars that represent celebrities or virtual novel characters by simply inputting text. Compared with the 3D GANs currently evolving for human body generation [2, 26, 54], zero-shot text-to-avatar generation does not require large amounts of richly-annotated human body datasets, extensive computing resources, or the training of difficult-to-converge 3D human body generations. Furthermore, it can generate more diverse avatars, making it a promising technology with significant potential for various applications. However, effectively integrating generalizable large-scale models with parametric human models to leverage multi-knowledge representations [52] remains an open problem for improving avatar generation.

The previous methods of generating avatars from text were limited to 2D representations [12, 13, 29, 47, 49, 53]. This was due to the lack of 3D datasets. However, recent advancements in differentiable rendering [14, 20, 22, 25] have made it possible to supervise a 3D model by using a vision-language model to supervise its rendered images. This has led to the development of zero-shot text-to-3D object methods [10, 15, 19, 24, 32, 44, 46]. Despite the progress in generating common 3D objects, generating 3D avatars still poses significant challenges. The flexible movements and complex body structure of 3D avatars make it difficult for these models to generate avatars that exhibit optimal performance. However, AvatarCLIP [9] was specifically designed to address these issues. It uses an implicit neural network to reconstruct a human body prior [16] and leverages CLIP (Contrastive Language-Image Pretraining) [34], a vision-language model, to optimize the reconstructed human body towards the text description. In this approach, CLIP encodes both the text and the rendered image to embeddings in a multi-modal feature space, and then minimizes their cosine distance.

Despite significant advancements made by AvatarCLIP in the zero-shot text-to-avatar task, there are still several limitations that require attention. 1) A lack of avatar-specific model. Current methods do not employ specialized models designed for avatars. Instead, they rely on generic 3D models, which leads to two major issues: facial distortion and unrealistic clothing, which greatly compromise the quality of the generated avatars as shown in Figure 1. Facial distortion arises from a poor combination of these models with human body priors, resulting in imprecise human body representations with shapeless faces. Unrealistic clothing is caused by the current models treating clothing and skin as mere textures on the body surface, without differentiating between clothing and body models, which results in a mixed texture and a lack of clothing thickness. 2) Limited generative power of CLIP. Although CLIP has successfully bridged the gap between natural language and avatar images, it has limited generative power as it is not a generative model. It only calculates an embedding of the overall appearance, which causes generated avatars to mismatch with the text in terms of details.

To address the first limitation, it is necessary to develop more sophisticated avatar models that can also effectively separate clothing and skin textures to simulate realistic clothing. One promising approach is the use of clothing-decoupled models, such as those proposed in [4, 7, 11, 50, 56]. However, these models do not use a complete neural implicit representation, which has been shown to be the most suitable 3D representation for receiving guidance from vision-language models [9, 19]. As for the second limitation, one possible solution is to use diffusion models' supervision [15, 32] to optimize the human body model, as diffusion models are pixel-level generative models. However, due to the lack of ground truth annotated with clothing segmentation labels, existing clothing-decoupled models cannot obtain an optimization direction that effectively distinguishes clothing from skin through current optimization methods.

To tackle these challenges, we present AvatarFusion, a pioneering method for generating clothing-decoupled 3D avatars from text prompts using diffusion models as supervision. Notably, to the best of our knowledge, our approach is the first to combine zero-shot 3D avatar generation and segmentation. In AvatarFusion, we present the first clothing-decoupled avatar model of complete neural implicit representations, along with a novel Dual Volume Rendering strategy for its rendering. Additionally, we propose a novel optimization loss called Pixel-Semantics Difference-Sampling (PS-DS) to optimize the model from diffusion models and separate the generation of clothing and skin. To elaborate on our proposed approach, we first utilize an off-the-shelf Signed Distance Function (SDF) field to create a SDF-based Avatar Model (SAvM) that accurately captures the intricate details of the human body prior, SMPL [16]. SAvM is also equipped with a deformation field that allows for the manipulation of the avatar's pose during training. Secondly, we introduce a clothing-decoupled model (CDM) that uses two SAvMs to represent the skin and clothing separately. Our proposed Dual Volume Rendering strategy is compatible with traditional volume rendering and provides a way to jointly render two implicit neural representations in the same space. Finally, our proposed Pixel-Semantics Difference-Sampling (PS-DS) aligns the **D**ifferences in **P**ixel space to the **D**ifference of text **S**emantics as a **S**ampling strategy of a latent diffusion model known as Stable Diffusion [37]. This allows the clothing model to generate only the clothing around the avatar's body without covering the skin.

Prior to our work, evaluation metrics for this field were limited to user studies, and no benchmark was available. To fill this gap, we propose a new benchmark called Famous-Character-50 (FC50). It comprises fifty varied text prompts featuring famous people or fictional characters of different cultures and genders. Our benchmark enables quantitative evaluation by comparing the generated avatars with the outcomes of mature 2D text-to-image models.

To assess the generated avatar images’ quality, we employ face recognition distance [8] and Fréchet Inception Distance (FID) score [40] as evaluation metrics. Our experimental results indicate that AvatarFusion outperforms baselines in both quantitative and qualitative evaluations on our benchmark. Specifically, AvatarFusion attains a lower face recognition distance of 14.04%, implying that the generated avatars are more faithful to the given text prompts. Furthermore, our method significantly enhances the FID score, indicating that AvatarFusion generates more realistic avatars. The model’s high performance in user studies also suggests that the generated avatars align better with the human perspective. Additionally, our avatar-specific model is clothing-decoupled, enabling us to exchange clothing between different characters.

Our contributions are as follows.

- We propose AvatarFusion, a novel framework for zero-shot 3D avatar generation that is the first to combine zero-shot clothing segmentation and avatar generation. AvatarFusion includes a clothing-decoupled neural implicit model, a Dual Volume Rendering strategy, and a PS-DS optimization method.
- We propose the use of diffusion-based optimization methods to enhance the visual details of the generated avatars, specifically by improving the distinction between clothing and skin.
- We also propose the first benchmark for the field, called Famous-Character-50 (FC50). Our experiments demonstrate the effectiveness of our proposed approach in comparison to state-of-the-art methods.

2 RELATED WORKS

Large-scale vision-language models. Recently, CLIP [34] has significantly advanced in bridging the gap between natural language and images by providing a multi-modal embedding space for various downstream tasks such as image generation [35], segmentation [3, 5, 48, 51], classification [34], and captioning [21, 42]. Text-conditioned diffusion models [23, 35, 37, 38] are another breakthrough. DALL-E 2 [35] uses CLIP’s embedding space to generate images of complex text prompts, while Imagen [38] employs a cascade of super-resolution models to improve generation efficiency. Stable Diffusion [37] takes a different approach by using a low-resolution latent space to generate images.

Zero-shot text-to-3D-objects generation. Thanks to the recent development of pre-trained vision-language models, several works have contributed to generating 3D objects in a zero-shot manner from text prompts. Text2mesh [19], AvatarCLIP [9], and NeRF-Art [44] utilize CLIP [34] loss as supervision to optimize mesh and implicit function representations. In addition to CLIP, diffusion models [35, 37, 38, 55] can also be used for text-based model supervision [15, 17, 18, 24, 32, 36, 46]. DreamFusion [32] employs Imagen [38] for Score Distillation Sampling (SDS) to supervise the generation of NeRF [20] models. Although diffusion-based methods generate more impressive content, they often fail to converge when generating human bodies and cannot control the generated results even when they do. Therefore, using diffusion models to supervise the generation of 3D avatars remains a challenge.

Clothing-decoupled models. Clothing-body separation was initially developed for simulating clothing in physics simulations [1, 30, 39, 43]. Recently, clothing-body separation models [4, 7, 11, 56] have emerged in avatar generation and reconstruction tasks. These tasks are more challenging because separating 3D clothing and body purely from 2D data is difficult and requires extensive pixel-level annotation of clothing and body data. SMPLicit [4] is a generative clothing-decoupled model that uses a neural implicit network to represent clothes outside the SMPL mesh. Neural implicit models can represent clothing with any topology, thus significantly enhancing the expressive power compared with previous complete mesh-based models. The task AvatarFusion faces is even more challenging as there is no ground truth.

3 PRELIMINARIES

Human body priors. SMPL (Skinned Multi-Person Linear) [16] is a statistical body model that represents the surface of the human body as a triangulated mesh $M(\beta, \theta)$ with vertices $\mathbf{v}_i$. The model uses shape parameters β and pose parameters θ to control the body’s shape and pose, respectively. To enable deformation of the mesh based on pose, SMPL follows a linear blending skinning function for vertex transformations as:

$$\mathbf{v}_i' = \sum_{k=1}^{K} \omega_{k,i} G_k'(\theta, \mathbf{J})\mathbf{v}_i, \tag{1}$$

where $\omega_{k,i}$ are the blend weights for the k-th joint, and $\mathbf{G}_k'(\theta, \mathbf{J})$ is the relative transformation matrix of the k-th joint dependent on the pose parameter θ and joint locations $\mathbf{J}$.

Neural implicit surfaces. Neural Implicit Surfaces (NeuS) [45] is a recently developed method for learning implicit representations of 3D surfaces from 2D images. The model uses the volume rendering [6, 20], a technique to project a 3D volume onto a 2D image plane, to train a neural network to reconstruct the 3D surface as Signed Distance Function (SDF). The volume rendering strategy estimates the pixel color $\hat{C}$ by shooting a ray $\mathbf{p}(t)$ from the camera and calculating properties of sampled points $\mathbf{p}(t_i)$ on the ray, using three primary equations:

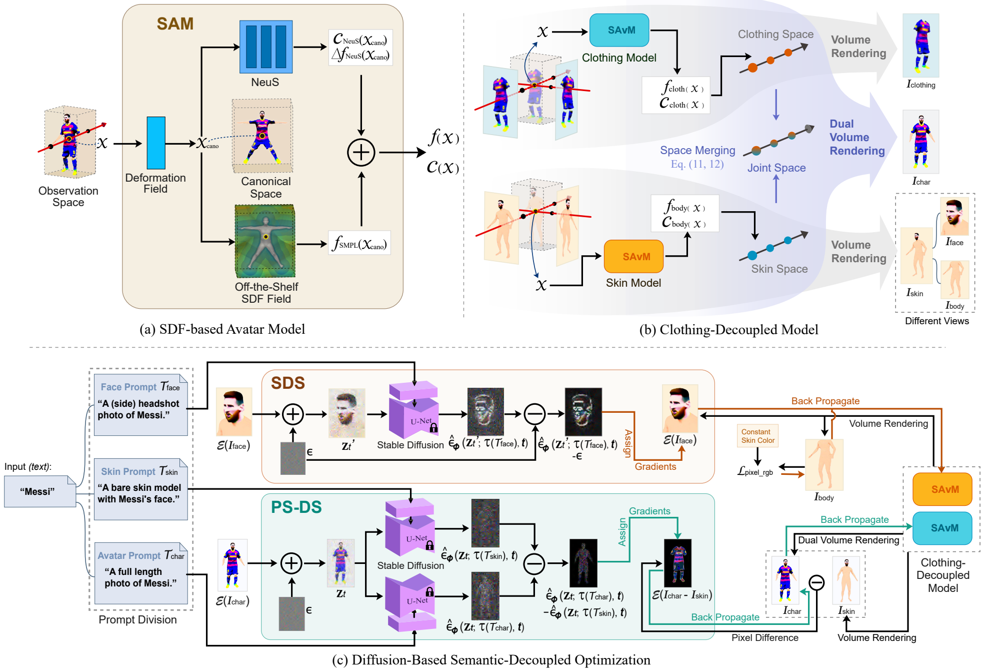
$$\alpha_i = 1 - \exp(-\int_{t_i}^{t_{i+1}} \rho(t) \mathrm{d}t), \tag{2}$$

$$T_i = \prod_{j=1}^{i-1} (1 - \alpha_j), \tag{3}$$

$$\hat{C} = \sum_{i=1}^{n} T_i \alpha_i \mathbf{c}_i. \tag{4}$$

Here α_i represents the discrete transparency of each point, which is calculated as the integral of the density function $\rho(t)$ between neighbour points. T_i represents the the accumulated transparency of all points behind it along the viewing ray. The pixel color $\hat{C}$ is then calculated based on α_i, T_i, and point color $\mathbf{c}_i$.

Diffusion-supervised training. DreamFusion [32] proposes a method for generating 3D models from text using a pre-trained diffusion model called Imagen [38] with a denoising function, denoted as $\hat{\epsilon}_\phi(\mathbf{z}_t; y, t)$, where $\mathbf{z}_t$ represents a noisy image at noise level t, and y is the text embedding. The optimization of the neural implicit model g begins by adding noise ϵ to the rendered image



(a) SDF-based Avatar Model

(b) Clothing-Decoupled Model

(c) Diffusion-Based Semantic-Decoupled Optimization

Figure 2: Overview of AvatarFusion. The upper left part shows (a) the SDF-Based Avatar Model (SAvM) which takes a point x as input, and output its SDF value and color value. The upper right part shows (b) the clothing-decoupled model, which takes two SAvMs representing skin and clothing, and merge the space to render avatars with clothes. The lower part shows (c) our diffusion-based optimization methods with PS-DS separating the clothing from skin semantically. For clarity, we omit the image encoder $\mathcal{E}$ in the figure.

$\mathbf{x} = g(\theta)$ at a specified noise level t, obtaining $\mathbf{z}_t$. The Score Distillation Sampling (SDS) technique is then employed to optimize the model towards the intended text meaning by calculating the difference between the predicted text-conditioned noise, $\hat{\epsilon}_\phi(\mathbf{z}_t; y, t)$, and the added noise ϵ, as follows:

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\phi, \mathbf{x} = g(\theta)) \triangleq \mathbb{E}_{t,\epsilon}[\omega(t)(\hat{\epsilon}_\phi(\mathbf{z}_t; y, t) - \epsilon)\frac{\partial \mathbf{x}}{\partial \theta}], \quad (5)$$

where $\omega(t)$ is a weighting function.

4 METHODOLOGY

4.1 Overview

In this study, we present AvatarFusion, a novel framework designed to generate photo-realistic avatars with separate clothing for the zero-shot text-to-avatar task. Figure 2 illustrates the pipeline of AvatarFusion. Our framework incorporates a Clothing-Decoupled neural implicit Model (CDM) which consists of two SDF-based Avatar Models (SAvM), one for skin and the other for clothing. To optimize the models, we utilize Diffusion-Based Supervisions.

We begin by introducing the SAvM in Section 4.2. The SAvM extends NeuS [45] by incorporating a deformation field and an off-the-shelf SDF field generated by an offline SDF generator. The deformation field transforms points in an observation space to points in the canonical space. Meanwhile, the SDF field provides a detailed human body prior, SMPL [16], enabling our model to render well-formed facial structures and detailed SMPL models before any optimization, which effectively mitigates facial distortion problems.

In Section 4.3, we provide detailed explanations of CDM. CDM is our full model that represents the skin and clothing of an avatar using two SAvMs. These SAvMs can produce skin and clothing images using SDF-based volume rendering [45]. To render avatars with clothing, we merge the space points from two neural implicit representations using Dual Volume Rendering.

In Section 4.4, we describe the process of optimizing the models using Stable Diffusion, a latent diffusion model [37]. Firstly, we divide the text description into three prompts: face, skin, and avatar prompt. The face prompt is used as the text condition of the SDS method [32] to optimize the facial part of the skin model. For the body part of the skin model, we use a selected skin color. The skin and avatar prompts are regarded as two text conditions for Pixel-Semantic Difference-Sampling (PS-DS) to generate clothes, as the overall avatar minus the bare skin leaves only the clothes.

4.2 SDF-Based Avatar-Specific Model

NeuS has the ability to produce high-precision surface reconstructions for static scenes. However, in order to better represent an avatar, we have extended NeuS to the SAvM by introducing a deformation field and an off-the-shelf SDF field.

Deformation field. To create a dynamic avatar, we require a deformation field that maps the observation space, where the avatar takes on arbitrary poses θ, to a canonical space, where the avatar is in a standard pose θ_{cano}. Following the approach of [27, 28, 31, 33], we align any given point $\mathbf{x}$ in the observation space to the nearest vertex $\mathbf{v}$ of SMPL mesh $M(\beta, \theta)$ based on Euclidean distance, and assign the blend weights of $\mathbf{v}$, denoted as ω_i, to $\mathbf{x}$. This allows $\mathbf{x}$ to rotate with SMPL joints and correspond to a point $\mathbf{x}_{\text{cano}}$ in the canonical space, following Equation 1. The deformation field equation is then defined as follows:

$$\mathbf{x}_{\text{cano}} = \sum_{k=1}^{K} \omega_k \mathbf{G}'_k(\theta, \mathbf{J})\mathbf{x}. \tag{6}$$

Off-the-shelf SDF field. The optimization process begins with the models representing SMPL prior, as a starting point [9, 19]. To encode the SMPL prior directly into the model, we use an offline SDF generator to create a 256×256×256 grid of SDF values that correspond to the SMPL mesh space at its canonical pose $M(\beta, \theta_{\text{cano}})$. Any point $\mathbf{x}_{\text{cano}}$ in the mesh space can be assigned an SDF value denoted as $f_{SMPL}(\mathbf{x}_{\text{cano}})$ using bilinear interpolation. The Multi-Layer Perceptrons (MLPs) of NeuS generate only a residual SDF value $\Delta f_{NeuS}(\mathbf{x}_{\text{cano}})$. This ensures that the model can render highly detailed SMPL even before any training has been conducted.
Summary Therefore, the SDF value $f(\mathbf{x})$ and color $\mathbf{c}(\mathbf{x})$ at point $\mathbf{x}$ for volume rendering Equation 4 is

$$\Delta f_{\text{NeuS}}(\mathbf{x}_{\text{cano}}), \mathbf{feat} = MLPs(\mathbf{x}_{\text{cano}}), \tag{7}$$

$$\mathbf{c}(\mathbf{x}) = \mathbf{c}_{\text{NeuS}}(\mathbf{x}_{\text{cano}}) = MLPs(\mathbf{x}_{\text{cano}}, \mathbf{feat}), \tag{8}$$

$$f(\mathbf{x}) = f_{\text{SMPL}}(\mathbf{x}_{\text{cano}}) + \Delta f_{\text{NeuS}}(\mathbf{x}_{\text{cano}}). \tag{9}$$

4.3 Clothing-Decoupled Model

Skin, clothing and joint spaces. Our goal is to create a more realistic clothing simulation by developing a clothing-decoupled neural implicit model that represents skin and clothes separately. To achieve this, we utilize two SAvMs, one for rendering the images of skin space I_{skin} and the other for rendering the clothing space I_{clothing}. However, both SAvMs must be integrated into a joint space to render the complete image I_{char}. One direct approach involves dividing the joint space into an inner space, where only the skin space points exists, and an outer space, where only the clothing space points exists, as proposed in [7].

$$\mathbf{x} = \begin{cases} \mathbf{x}_{\text{body}}, & f_{\text{body}}(\mathbf{x}) <= \delta \\ \mathbf{x}_{\text{cloth}}, & f_{\text{body}}(\mathbf{x}) > \delta \end{cases}, \tag{10}$$

where δ is a small positive number.

However, connecting the skin and clothing spaces directly to form the joint space may result in a biased skin color that differs from the skin color rendered in the skin space. This is because skin color is determined by integrating every point's color along the ray during volume rendering. When skin space points are neglected in the outer joint space, their skin color contribution may also be disregarded, leading to a biased skin color. Therefore, the skin space points must also exist in the outer joint space to maintain their distribution.

Dual Volume Rendering. To integrate the two neural implicit models in the outer joint space, an extended volume rendering technique called Dual Volume Rendering is proposed. Specifically, we update the properties of joint space points, α_i and $\mathbf{c}_i$ in Equation 4, with combined values from both models. To be concise, we provide the summarized space merging equations below and leave the detailed derivation process in the supplementary material. Please note that the derivation of the merged α_i is in line with the definition of volume rendering [6], while the merged $\mathbf{c}_i$ is a linear interpolation approximation as there is no physical meaning of adding the RGB values.

Space merging equations. Given the sampled points on a ray, denoted as $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i, \ldots\}$, we derive $f_{\text{body}}(\mathbf{x}_i)$, $f_{\text{cloth}}(\mathbf{x}_i)$, $\mathbf{c}_{\text{body}}(\mathbf{x}_i)$, and $\mathbf{c}_{\text{cloth}}(\mathbf{x}_i)$ from the SAvMs. Then, we calculate $\alpha_{\text{body}}(\mathbf{x}_i)$ and $\alpha_{\text{cloth}}(\mathbf{x}_i)$ based on the discretization of Equation 2 (refer to [45] for more details). Finally, Dual Volume Rendering procedure can be summarised as:

$$\alpha_i = \begin{cases} \alpha_{\text{body}}(\mathbf{x}_i), & f_{\text{body}}(\mathbf{x}_i) <= \delta \\ \alpha_{\text{body}}(\mathbf{x}_i) + \alpha_{\text{cloth}}(\mathbf{x}_i) & \\ -\alpha_{\text{body}}(\mathbf{x}_i) \cdot \alpha_{\text{cloth}}(\mathbf{x}_i), & f_{\text{body}}(\mathbf{x}_i) > \delta \end{cases}, \tag{11}$$

$$\mathbf{c}_i = \begin{cases} \mathbf{c}_{\text{body}}(\mathbf{x}_i), & f_{\text{body}}(\mathbf{x}_i) <= \delta \\ \frac{\alpha_{\text{body}}(\mathbf{x}_i)}{\alpha_{\text{body}}(\mathbf{x}_i)+\alpha_{\text{cloth}}(\mathbf{x}_i)}\mathbf{c}_{\text{body}}(\mathbf{x}_i) & \\ +\frac{\alpha_{\text{cloth}}(\mathbf{x}_i)}{\alpha_{\text{body}}(\mathbf{x}_i)+\alpha_{\text{cloth}}(\mathbf{x}_i)}\mathbf{c}_{\text{cloth}}(\mathbf{x}_i), & f_{\text{body}}(\mathbf{x}_i) > \delta \end{cases}. \tag{12}$$

Hereafter, we can proceed with the subsequent volume rendering calculations using Equation 3 and so forth.

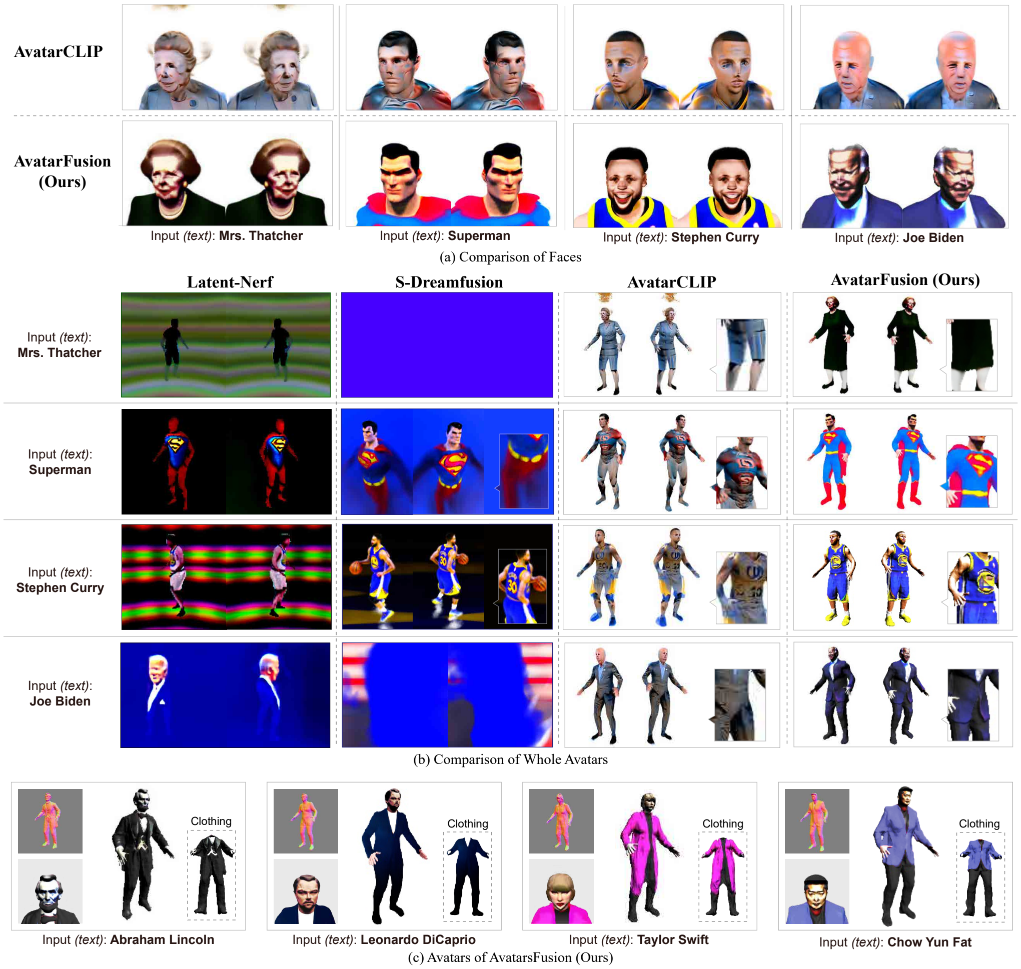
4.4 Diffusion-Based Semantic-Decoupled Optimization

Prompt division. In this sub-section, we present our optimization methods using Stable Diffusion for generating a complete avatar with decoupled skin and clothing components. To achieve this, we generate three prompts: the face prompt $\mathcal{T}_{\text{face}}$, the skin prompt $\mathcal{T}_{\text{skin}}$, and the character prompt $\mathcal{T}_{\text{char}}$, which correspond to the avatar's facial features, skin tone, and overall appearance, respectively.

We use $\mathcal{T}_{\text{face}}$ as the text condition for the SDS method [32] to optimize the facial part of skin model. We use a pre-selected color to optimize I_{body}, the body image rendered in the skin space, with pixel-wise loss $\mathcal{L}_{\text{pixel\_rgb}}$. $\mathcal{T}_{\text{skin}}$ and $\mathcal{T}_{\text{char}}$ are used to generate the clothing of avatar indirectly. This is because the diffusion model can only generate clothing along with skin, and directly applying the SDS method with a clothing prompt as a condition would result in a new layer of skin outside the original skin. To address this issue, we introduce the Pixel-Semantic Difference-Sampling (PS-DS) method.
PS-DS method. This method takes two text prompts as input and generates semantic differences between them. By using $\mathcal{T}_{\text{skin}}$ and $\mathcal{T}_{\text{char}}$ as input, we obtain clothing as their semantic difference.

The PS-DS method is implemented as shown in Figure 2. First, we render two images I_{skin} and I_{char} and encode I_{char} to its latent image $\mathcal{E}(I_{\text{char}})$ using Stable Diffusion's encoder. Then, noise ϵ at level t is added to $\mathcal{E}(I_{\text{char}})$ to obtain $\mathbf{z}_t$, which serves as a noisy sample for Stable Diffusion. Next, we predict noise based on $\mathcal{T}_{\text{char}}$ and $\mathcal{T}_{\text{skin}}$, respectively. The difference between the predicted noises provides optimizing directions for the clothing model. We compute these directions using the equation below:

$$\nabla_\theta \mathcal{L}_{\text{PS-DS}}(\phi, \mathbf{x} = g(\theta)) \triangleq \mathbb{E}_{t,\epsilon}[\omega(t)(\hat{\epsilon}_\phi(\mathbf{z}_t; \tau(\mathcal{T}_{\text{char}}), t) \\ -\hat{\epsilon}_\phi(\mathbf{z}_t; \tau(\mathcal{T}_{\text{skin}}), t))\frac{\partial \mathbf{x}}{\partial \theta}], \tag{13}$$


AvatarCLIP
AvatarFusion (Ours)
Input (text): Mrs. Thatcher
Input (text): Superman
Input (text): Stephen Curry
Input (text): Joe Biden
Latent-Nerf
S-Dreamfusion
AvatarCLIP
AvatarFusion (Ours)
Input (text): Mrs. Thatcher
Input (text): Superman
Input (text): Stephen Curry
Input (text): Joe Biden
Clothing
Clothing
Clothing
Clothing
Input (text): Abraham Lincoln
Input (text): Leonardo DiCaprio
Input (text): Taylor Swift
Input (text): Chow Yun Fat

(a) Comparison of Faces

(b) Comparison of Whole Avatars

(c) Avatars of AvatarsFusion (Ours)

Figure 3: Qulitative Comparison with baselines for face and body generation and more results of AvatarFusion. Latent-NeRF and Stable-DreamFusion can sometimes fail to generate content or produce distorted bodies. On the other hand, AvatarCLIP suffers from poor details. In comparison, our method can robustly generate avatars with vivid faces and realistic clothing. More examples are provided in the supplementary material.

where τ represents the text encoder of Stable Diffusion. For definitions of other variables, please refer to Equation 5. We assign this result as the gradients to the latent image $\mathcal{E}(I_{\text{char}} - I_{\text{skin}})$ and backpropagate it.

In addition to SDS and PS-DS method, we also use SDF loss to control the overall shape of the model and the pixel entropy loss to assist PS-DS method in segmenting the clothing. By calculating the entropy of the proportion of skin and clothing model in the color of each ray as a loss function, one side can dominate the overall color, which can better distinguish whether there is clothing coverage.

5 EXPERIMENTS

5.1 Implementation Details

We represented the SAvMs of the body and clothing using two-layer MLPs with a hidden size of 128. Our latent diffusion model is the open-source Stable Diffusion model version 1.5. We trained the whole avatar in a standing pose as in AvatarCLIP [9]. We trained the skin model for 25000 iterations and then used PS-DS to train the clothing model for 25000 iterations with $\delta = 1e-5$, followed by an additional 25000 iterations with $\delta = 1e-3$. More details are provided in the supplementary material.

5.2 Benchmark

The evaluation of zero-shot text-to-avatar generation task is challenging due to the absence of a unified dataset of selected characters and evaluation metrics beyond user studies. To address this, we developed a benchmark called Famous-Character-50 (FC50) that includes 50 descriptions of famous real and fictional characters, and the corresponding images generated using the Stable Diffusion [37] checkpoints of version 2.1. FC50 provides a diverse representation of races and genders while covering a broad range of cultures, as shown in Figure 4. This diversity guarantees a comprehensive evaluation that accurately reflects the diversity of human appearance.

Figure 4: The constructed dataset of the benchmark.

Table 1: Comparison with baselines and ablation studies. For all the metrics, the lower (↓), the better.

Model	FRD	Face-FID	Body-FID
Latent-NeRF [18]	0.8956	105.87	114.73
S-DreamFusion [41]	(-)	(-)	119.59
AvatarCLIP [9]	0.6907	79.08	96.65
Ours	**0.5937**	**71.73**	**83.21**
Full model	0.6007	82.13	92.16
32 * 32 SDF Reso	0.7581	104.69	95.42
64 * 64 SDF Reso	0.6754	87.94	96.85
w/o diffusion	0.7023	96.08	104.32
w/o speartaion	0.6018	84.72	95.98

Evaluation metrics. Since there is no ground truth available for the zero-shot text-to-avatar task, we propose comparing the similarity between 3D model rendering images and mature 2D image generation results to evaluate the models. To measure this similarity, we employ a face recognition model [8] to capture and measure the distance between the generated face and the corresponding face in the image dataset, which we refer to as the Face Recognition Distance (FRD). A smaller distance indicates that the generated avatar is closer to the corresponding face in terms of identity recognition, which suggests that the avatar better captures the character features described in the text prompt. Additionally, we compare the FID score between the generated avatars and image dataset, evaluating the quality of the generated avatars in terms of fidelity to the ground truth images. We compare the faces (Face-FID) and the whole body (Body-FID) separately.

5.3 Comparing with Baselines

We compared AvatarFusion with three baseline methods: DreamFusion [32], Latent-NeRF [18], and AvatarCLIP [9]. DreamFusion is a generative model that converts text to 3D objects using Imagen, a diffusion model [38]. As we were unable to obtain Imagen, we used an unofficial implementation called Stable-DreamFusion (S-DreamFusion) [41], which is based on Stable Diffusion [37]. Latent-NeRF modifies the generation space of S-DreamFusion from RGB space to the latent space of Stable Diffusion and uses a loss function to align the 3D implicit representation to a mesh prior representing basic shapes. We used SMPL [16] prior for avatar generation in this method. AvatarCLIP is specifically designed for avatar generation, and its robustness and better combination with the SMPL prior make it a strong competitor. For AvatarCLIP, we used the official text prompt "A 3D rendering of character's name in Unreal Engine," while for methods utilizing diffusion models, we used the text prompt "A full-length photograph of character's name."

Table 2: Results of user studies from 19 users. All metrics are measured on a scale of 1 (worst) to 5 (best). AvatarFusion achieves highest rates on all aspects.

Model	Overall	Face	Texture	Text Cons
Latent-NeRF [18]	2.21	1.89	2.47	2.73
S-DreamFusion [41]	3.74	(-)	3.95	2.84
AvatarCLIP [9]	3.89	3.00	4.16	4.58
Ours	**4.32**	**4.68**	**4.42**	**4.79**

Quantitative comparison. Table 1 shows the evaluation results of our model against baselines. Our superior performance on all three evaluation metrics indicates that we are capable of generating more accurate and visually appealing avatars from textual descriptions. Specifically, our models achieve a face distance of 0.5937, which is 14.04% higher than the baseline, indicating our generated faces are more realistic and faithful to the text.

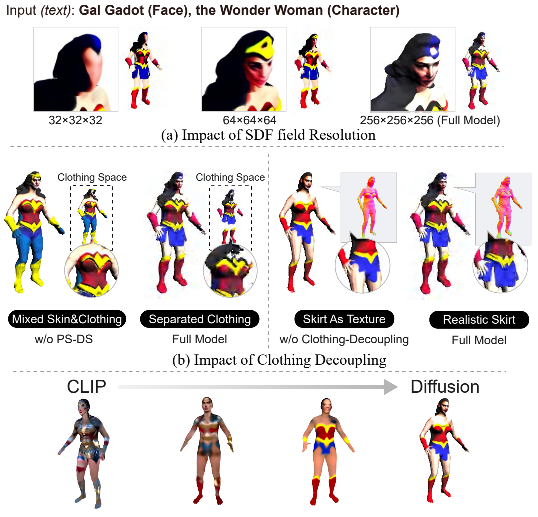
Qualitative comparison. The generation of complex human bodies using diffusion-based methods, such as Latent-NeRF and S-DreamFusion, is not always successful. S-DreamFusion failed to generate meaningful content for Mrs. Thatcher and Joe Biden, while Latent-NeRF initially converged but later lost all content. We present the best results before this issue occurred. While S-DreamFusion can generate avatars in some cases, it falls short in terms of facial and clothing details compared with AvatarFusion. Additionally, it can cause body distortions, such as generating three arms for Curry or not producing hands for Superman. These problems of diffusion-based method arise because of the models' ineffective combination with SMPL prior, which results in rendered images with weak correlations across different perspectives. Applying diffusion-based pixel-level supervision to different perspectives at this stage may lead to convergence failure or distortions.

Both AvatarCLIP and AvatarFusion can robustly generate complete avatars. However, the generic model of AvatarCLIP and the embedding-level supervision of CLIP cannot distinguish between clothing and skin, which results in characters with distorted faces, mixed clothing and skin textures, and clothing lacking thickness.

User studies. To further evaluate the quality of our generated avatars, we conducted a user study comparing them with baseline methods. We recruited 19 volunteers and asked them to rate the methods based on (1) overall quality, (2) facial quality, (3) texture quality, and (4) consistency with the given text prompt. For each aspect, we randomly selected 12 generated results and asked the volunteers to score each example on a scale of 1 (worst) to 5 (best). The final results are presented in Table 2. Our AvatarFusion approach achieved the highest rank in all aspects, demonstrating its effectiveness in mitigating the defects of previous methods.

5.4 Ablation Studies

We conducted ablation studies on a random sample of fifteen characters from FC50 to gain insight into the contributions of each



(a) Impact of SDF field Resolution

(b) Impact of Clothing Decoupling

(c) Impact of Diffusion-Based Supervision

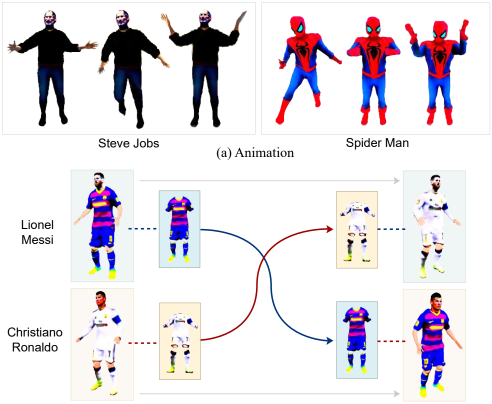
Figure 5: Ablation studies which demonstrate the following key findings: (a) High-resolution SDF fields contribute to well-formed faces. (b) The semantic separation of clothing from skin by PS-DS method allows for more realistic skirts than just texture alone. (c) The inclusion of Diffusion Supervision contributes to a more complete and detailed clothing.

component to the overall performance. The results of these studies are presented in Figure 5 and Table 1.

Impact of SDF field resolution. We examined the impact of using off-the-shelf SDF fields of different resolutions on our framework's performance. Our full model used a 256^3 SDF grid. Since there was no noticeable difference from the human perspective when using a 128^3 grid, we evaluated the results for SDF grids with resolutions of 32^3 and 64^3. The results showed that when we reduced the resolution, the avatar model at initial point lacked the necessary details, which ultimately led to facial distortion in the final output.

Impact of clothing decoupling. We investigated the impact of clothing decoupling on the results. First, we replaced PS-DS method with SDS method [32], which resulted in a loss of semantic decoupling between the Wonder Woman's skin and clothing in Figure 5. Then, we compared the results obtained with and without clothing decoupling and found that the decoupled model produced more expressive skirts than just a texture on the legs. Additionally, it achieved better performance in evaluation metrics. Decoupling clothing from skin enabled independent optimizing of clothing, loosening the constraints of the SMPL model, a body prior rather than a clothing prior.

Impact of diffusion-based supervision. We conducted an experiment to evaluate the impact of diffusion-based supervision by replacing it with CLIP [34]. In this step, we also removed clothing decoupling as it relies on diffusion supervision. The results demonstrated that CLIP-based optimization led to a blending of skin and clothing textures, whereas diffusion supervision resulted in complete clothing textures.



(a) Animation

(b) Exchanging Clothes

Figure 6: Results of (a) animation and (b) exchanging clothes.

5.5 Extra Abilities

As we align the generated avatars with the SMPL skeleton, they can be animated with SMPL pose sequences. And as our avatar is clothing-decoupled, we can exchange the clothes of avatars as shown in Figure 6.

6 DISCUSSION

Limitations. Similar to AvatarCLIP [9], our method currently cannot generate a reasonable backside because the vision-language models are not responsive to the prompt "the back of ...". Consequently, the backside of the avatar may resemble the front.

Ethical issues. The technology of generating realistic and clothing-decoupled avatars from text may raise concerns regarding ethics, privacy, and security. It is critical to continue conducting research and development in an ethical and responsible manner. Please note that in our avatar generation process, we only generate the inner model representing skin color and not the specific physique of an individual. We firmly oppose generating unethical outcomes.

7 CONCLUSIONS

In this study, we introduce AvatarFusion, a novel framework for zero-shot text-to-avatars generation. Our primary contribution is a clothing-decoupled neural implicit avatar model that employs a dual volume rendering strategy and a Pixel-Semantics Difference-Sampling process to separate the generation of body and clothing. Our experiments on the first benchmark in this field demonstrate that our approach surpasses state-of-the-art methods by a significant margin. We are the first to generate and segment 3D avatar models using a pre-trained vision-language model. Future research may focus on enhancing the backside of avatars or fitting loose clothing.

ACKNOWLEDGMENTS

This work is supported by the National Key R&D Program of China under Grant No. 2021QY1500, the State Key Program of the National Natural Science Foundation of China (NSFC) (No. 61831022), the Fundamental Research Funds for the Central Universities (No. 226-2022-00051).

REFERENCES

[1] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. 2020. CLOTH3D: clothed 3d humans. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, 344–359.

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

[3] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. 2023. Segment and track anything. *arXiv preprint arXiv:2305.06558* (2023).

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

[5] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. 2022. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11583–11592.

[6] Robert A Drebin, Loren Carpenter, and Pat Hanrahan. 1988. Volume rendering. *ACM Siggraph Computer Graphics* 22, 4 (1988), 65–74.

[7] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J Black, and Timo Bolkart. 2022. Capturing and animation of body and clothing from monocular video. *arXiv preprint arXiv:2210.01868* (2022).

[8] Geitgey. 2017. Face Recognition. https://github.com/ageitgey/face_recognition

[9] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. 2022. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535* (2022).

[10] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. 2022. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 867–876.

[11] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. 2020. Bcnet: Learning body and cloth shape from a single image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, 18–35.

[12] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. 2022. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–11.

[13] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. 2021. Tryongan: Body-aware try-on via layered interpolation. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–10.

[14] Tzu-Mao Li, Miika Aittala, Frédo Durand, and Jaakko Lehtinen. 2018. Differentiable monte carlo ray tracing through edge sampling. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 1–11.

[15] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. 2022. Magic3D: High-Resolution Text-to-3D Content Creation. *arXiv preprint arXiv:2211.10440* (2022).

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

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

[18] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. 2022. Latent-NeRF for Shape-Guided Generation of 3D Shapes and Textures. *arXiv preprint arXiv:2211.07600* (2022).

[19] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. 2022. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13492–13502.

[20] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

[21] Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734* (2021).

[22] Thomas Müller, Brian McWilliams, Fabrice Rousselle, Markus Gross, and Jan Novák. 2019. Neural importance sampling. *ACM Transactions on Graphics (ToG)* 38, 5 (2019), 1–19.

[23] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).

[24] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. 2022. Point-E: A System for Generating 3D Point Clouds from Complex Prompts. *arXiv preprint arXiv:2212.08751* (2022).

[25] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. 2020. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3504–3515.

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

[27] Xiao Pan, Zongxin Yang, Jianxin Ma, Chang Zhou, and Yi Yang. 2023. TransHuman: A Transformer-based Human Representation for Generalizable Neural Human Rendering. *Proceedings of the IEEE/CVF International conference on computer vision*.

[28] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. 2021. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5865–5874.

[29] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2337–2346.

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

[31] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. 2021. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14314–14323.

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

[33] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. 2021. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10318–10327.

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

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

[36] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. 2023. Texture: Text-guided texturing of 3d shapes. *arXiv preprint arXiv:2302.01721* (2023).

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

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

[39] Igor Santesteban, Miguel A Otaduy, and Dan Casas. 2019. Learning-based animation of clothing for virtual try-on. In *Computer Graphics Forum*, Vol. 38. Wiley Online Library, 355–366.

[40] Maximilian Seitzer. 2020. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid. Version 0.3.0.

[41] Jiaxiang Tang. 2022. Stable-dreamfusion: Text-to-3D with Stable-diffusion. https://github.com/ashawkey/stable-dreamfusion.

[42] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2021. Zero-shot image-to-text generation for visual-semantic arithmetic. *arXiv preprint arXiv:2111.14447* (2021).

[43] Raquel Vidaurre, Igor Santesteban, Elena Garces, and Dan Casas. 2020. Fully convolutional graph neural networks for parametric virtual try-on. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 145–156.

[44] Can Wang, Ruixiang Jiang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. 2022. NeRF-Art: Text-Driven Neural Radiance Fields Stylization. *arXiv preprint arXiv:2212.08070* (2022).

[45] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. 2021. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689* (2021).

[46] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. 2022. Rodin: A Generative Model for Sculpting 3D Digital Avatars Using Diffusion. *arXiv preprint arXiv:2212.06135* (2022).

[47] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8798–8807.

[48] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. 2022. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

11686–11695.

[49] Shuchen Weng, Wenbo Li, Dawei Li, Hongxia Jin, and Boxin Shi. 2020. Misc: Multi-condition injection and spatially-adaptive compositing for conditional person image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7741–7749.

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

[51] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. 2021. A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. *arXiv preprint arXiv:2112.14757* (2021).

[52] Yi Yang, Yueting Zhuang, and Yunhe Pan. 2021. Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies. *Frontiers of Information Technology & Electronic Engineering* 22, 12 (2021), 1551–1558.

[53] Jae Shin Yoon, Lingjie Liu, Vladislav Golyanik, Kripasindhu Sarkar, Hyun Soo Park, and Christian Theobalt. 2021. Pose-guided human animation from a single image in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15039–15048.

[54] Jianfeng Zhang, Zihang Jiang, Dingdong Yang, Hongyi Xu, Yichun Shi, Guoxian Song, Zhongcong Xu, Xinchao Wang, and Jiashi Feng. 2023. Avatargen: a 3d generative model for animatable human avatars. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*. Springer, 668–685.

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

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

A SUPPLEMENTARY MATERIAL

A.1 Details of Methodology

A.1.1 Details of Dual Volume Rendering. Section 4.3 presents the equations used to merge the skin and clothing spaces in the clothing-decoupled model. In this section, we provide the derivation of these equations.

Initially, to compute the color for each pixel, we shoot a ray $\mathbf{p}(t)$ and sample points on it, denoted as $\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots\}$, where $\mathbf{x}_i = \mathbf{p}(t_i)$. Next, we compute the SDF values of each point from the skin and clothing models, represented as $f_{\text{body}}(\mathbf{x}_i)$, $f_{\text{cloth}}(\mathbf{x}_i)$, and the corresponding RGB color vectors $\mathbf{c}_{\text{body}}(\mathbf{x}_i)$, $\mathbf{c}_{\text{cloth}}(\mathbf{x}_i)$, respectively. Subsequently, we compute the discrete opacity values based on the SDF-based volume rendering technique [45], which can be expressed as:

$$\alpha_{\text{body}}(\mathbf{x}_i) = \max\{\frac{\Phi_s(f_{\text{body}}(\mathbf{x}_i)) - \Phi_s(f_{\text{body}}(\mathbf{x}_{i+1}))}{\Phi_s(f_{\text{body}}(\mathbf{x}_i))}, 0\}, \tag{14}$$

$$\alpha_{\text{cloth}}(\mathbf{x}_i) = \max\{\frac{\Phi_s(f_{\text{cloth}}(\mathbf{x}_i)) - \Phi_s(f_{\text{cloth}}(\mathbf{x}_{i+1}))}{\Phi_s(f_{\text{cloth}}(\mathbf{x}_i))}, 0\}, \tag{15}$$

where $\Phi_s(x) = (1 + e^{-sx})^{-1}$. This equation is a discretization of $\alpha_i = 1 - \exp(-\int_{t_i}^{t_{i+1}} \rho(t)\mathrm{d}t)$.

For the inner space, where only skin space points are present, we use the characteristics obtained from the skin model as the properties of the joint space points. This can be denoted as:

$$\alpha_i = \alpha_{\text{body}}(\mathbf{x}_i), \quad f_{\text{body}}(\mathbf{x}_i) <= \delta, \tag{16}$$

$$\mathbf{c}_i = \mathbf{c}_{\text{body}}(\mathbf{x}_i), \quad f_{\text{body}}(\mathbf{x}_i) <= \delta, \tag{17}$$

where the inner space refers to the space where the distance from the skin is less than δ.

For the outer space, where both skin space points and clothing space points exist, we need to merge the points from both spaces. To accomplish this, we begin by exploring the equations of volume rendering before discretization. It is worth noting that the opaque density ρ is proportional to the gas density in the primitive physical equations of volume rendering, as reported in [6]. Based on this relationship, we can approximate that the volume density ρ of the volume rendering functions can be added in the same way as gas densities are added. This can be expressed as:

$$\rho(t) = \rho_{\text{body}}(t) + \rho_{\text{cloth}}(t), \tag{18}$$

where $\rho_{\text{body}}(t)$ and $\rho_{\text{cloth}}(t)$ represent the densities of the skin and clothing spaces, respectively, at time t on the emitted ray. Then, the discrete opacity value should be calculated as follow:

$$\begin{aligned}
\alpha_i =& 1 - \exp(-\int_{t_i}^{t_{i+1}} \rho(t)\mathrm{d}t) \\
=& 1 - \exp(-\int_{t_i}^{t_{i+1}} (\rho_{\text{body}}(t) + \rho_{\text{cloth}}(t))\mathrm{d}t) \\
=& (1 - \exp(-\int_{t_i}^{t_{i+1}} \rho_{\text{body}}(t)\mathrm{d}t)) + (1 - \exp(-\int_{t_i}^{t_{i+1}} \rho_{\text{cloth}}(t)\mathrm{d}t)) \\
&- (1 - \exp(-\int_{t_i}^{t_{i+1}} \rho_{\text{body}}(t)\mathrm{d}t)) \cdot (1 - \exp(-\int_{t_i}^{t_{i+1}} \rho_{\text{cloth}}(t)\mathrm{d}t)) \\
=& \alpha_{\text{body}}(\mathbf{x}_i) + \alpha_{\text{cloth}}(\mathbf{x}_i) - \alpha_{\text{body}}(\mathbf{x}_i) \cdot \alpha_{\text{cloth}}(\mathbf{x}_i), \quad f_{\text{body}}(\mathbf{x}_i) > \delta.
\end{aligned} \tag{19}$$

For the color item $\mathbf{c}_i$, because there is no physical meaning of adding the RGB values, we make a simple approximation that

$$\begin{aligned}
\mathbf{c}_i = \frac{\alpha_{\text{body}}(\mathbf{x}_i)}{\alpha_{\text{body}}(\mathbf{x}_i) + \alpha_{\text{cloth}}(\mathbf{x}_i)}\mathbf{c}_{\text{body}}(\mathbf{x}_i) + \frac{\alpha_{\text{cloth}}(\mathbf{x}_i)}{\alpha_{\text{body}}(\mathbf{x}_i) + \alpha_{\text{cloth}}(\mathbf{x}_i)} \\
\cdot \mathbf{c}_{\text{cloth}}(\mathbf{x}_i), \quad f_{\text{body}}(\mathbf{x}_i) > \delta.
\end{aligned} \tag{20}$$

A.1.2 Details of Diffusion-Based Semantic-Decoupled Optimization. Section 4.4 provides a comprehensive overview of our optimization methods, with a particular emphasis on the proposed PS-DS method. This section delves into the specifics of additional losses used in our framework.

To optimize the skin model, we utilize the SDS method [32, 41] along with a binary cross-entropy mask loss $\mathcal{L}_{\text{mask}}$, which penalizes the difference between the silhouettes of the rendered avatar and the SMPL [16] mesh. The total loss is formulated as:

$$\mathcal{L}_{\text{skin}} = \lambda_1 \mathcal{L}_{\text{SDS}} + \lambda_2 \mathcal{L}_{\text{mask}}, \tag{21}$$

where λ_1 and λ_2 are hyperparameters, and $\mathcal{L}_{\text{SDS}}$ is the loss obtained from the SDS method.

For optimizing the clothing model, we employ the PS-DS method, along with an SDF loss and a pixel entropy loss. The SDF loss is used to control the clothing SDF values of each point close to its skin (body) SDF values and is formulated as:

$$\mathcal{L}_{\text{SDF}} = \mathcal{L}_{\text{MSE}}(f_{\text{cloth}}(\mathbf{x}_i), f_{\text{body}}(\mathbf{x}_i)), \tag{22}$$

where $\mathcal{L}_{\text{MSE}}$ is the mean squared error loss function, $f_{\text{cloth}}(\mathbf{x}_i)$ and $f_{\text{body}}(\mathbf{x}_i)$ are the SDF values of the clothing and skin models, respectively. Here we fix the parameters of the skin model, making $f_{\text{body}}(\mathbf{x}_i)$ a detached value.

To further improve our clothing model, we introduce the pixel entropy loss that considers the proportion of pixel colors from the skin and clothing models. The purpose of this loss is to ensure that one component dominates the pixel color, which allows us to control the thickness of the clothing.

To incorporate the entropy loss into our approach, we begin by computing the contributions of the skin model, denoted as p_{body}, and the clothing model, denoted as p_{cloth}, to each pixel color. Specifically, in Dual Volume Rendering, we update the α_i and $\mathbf{c}i$ parameters of the volume rendering equation $\hat{C} = \sum_{i=1}^{n} T_i \alpha_i \mathbf{c}_i$ using the merge space properties. This allows us to separate the contributions of the two models as follows:

$$T_i \alpha_i \mathbf{c}_i = w_{\text{body},i} \mathbf{c}_{\text{body}}(\mathbf{x}_i) + w_{\text{cloth},i} \mathbf{c}_{\text{cloth}}(\mathbf{x}_i), \tag{23}$$

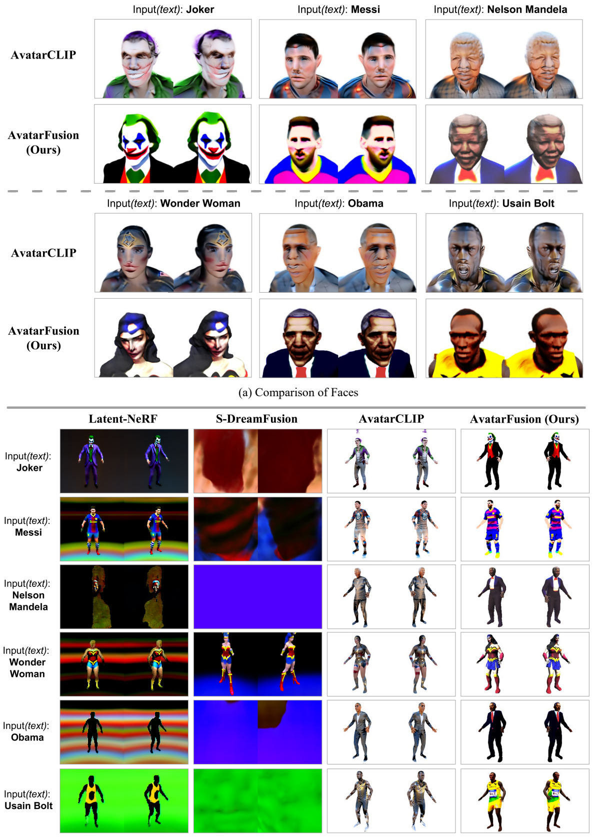
where $w_{\text{body},i}$ and $w_{\text{cloth},i}$ represent the calculated color weights. Subsequently, we can derive p_{body} and p_{cloth} as follows:

$$p_{\text{body}} = \frac{\sum_{i=1}^{n} w_{\text{body},i}}{\sum_{i=1}^{n} w_{\text{body},i} + \sum_{i=1}^{n} w_{\text{cloth},i}}, \tag{24}$$

$$p_{\text{cloth}} = \frac{\sum_{i=1}^{n} w_{\text{cloth},i}}{\sum_{i=1}^{n} w_{\text{body},i} + \sum_{i=1}^{n} w_{\text{cloth},i}}. \tag{25}$$

After obtaining the contributions, we can compute the entropy of these proportions using the following equation:

$$\mathcal{L}_{\text{entropy}} = -p_{\text{body}} \log(p_{\text{body}}) - p_{\text{cloth}} \log(p_{\text{cloth}}). \tag{26}$$



(a) Comparison of Faces

(b) Comparison of Whole Avatars

Figure 7: Comparison with Baselines.

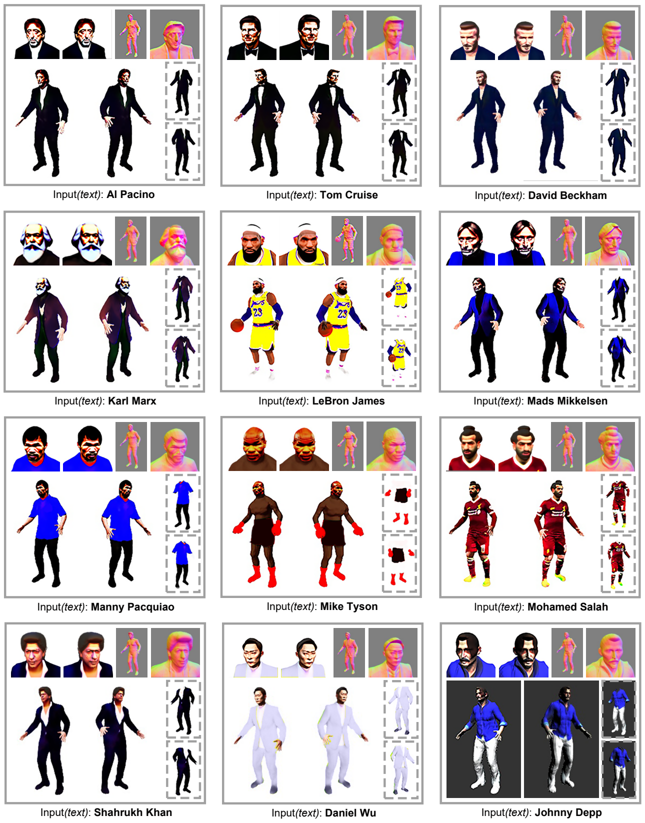

Input(text): Al Pacino
Input(text): Tom Cruise
Input(text): David Beckham
Input(text): Karl Marx
Input(text): LeBron James
Input(text): Mads Mikkelsen
Input(text): Manny Pacquiao
Input(text): Mike Tyson
Input(text): Mohamed Salah
Input(text): Shahrukh Khan
Input(text): Daniel Wu
Input(text): Johnny Depp

Figure 8: More Results of AvatarFusion.

The total loss for optimizing the clothing model is given by:

$$\mathcal{L}_{\text{clothing}} = \lambda_3\mathcal{L}_{\text{PS-DS}} + \lambda_4\mathcal{L}_{\text{SDF}} + \lambda_5\mathcal{L}_{\text{entropy}}, \tag{27}$$

where λ_3, λ_4, and λ_5 are hyperparameters, and $\mathcal{L}_{\text{PS-DS}}$ is the loss obtained from the PS-DS method.

A.2 Implementation Details

For volume rendering during training, we sample 32 points on each ray and set the scaling factor in Φ_s, to a fixed value of $s = e^7$. This fixed value ensures that the residual SDF value does not dominate the overall proportion. For the skin model optimization, we set the hyperparameters $\lambda_1 = 100$ and $\lambda_2 = 800$ to balance the contributions of the SDS loss and the mask loss. We optimize the skin model for a total of 25000 iterations. To update the model parameters, we use a learning rate of $1e-4$. Subsequently, we focus on optimizing the clothing model. We set the hyperparameters $\lambda_3 = 100$, $\lambda_4 = 300$, $\lambda_5 = 0$, and $\delta = 1e-4$. This optimization process for the clothing model also runs for 25000 iterations. The learning rate for this stage is set to $5e-4$. To further refine the generated clothing and filter out any noise, we perform an additional optimization stage. We set the hyperparameters as $\lambda_3 = 100$, $\lambda_4 = 300$, $\lambda_5 = 5000$, and $\delta = 2e-3$. The purpose of this stage is to facilitate the separation of clothing from body. The learning rate for this stage is set to $5e-4$. Similar to the previous stages, we conduct this optimization for 25000 iterations. Our model training is carried out on an NVIDIA Tesla V100 32GB GPU.

A.3 More Results

We present additional comparison results with baselines in Figure 7, detailed results of AvatarFusion in Figure 8. Due to page limitations, we have provided additional results on the project page and an internet image repository. We present more ablation results on https://postimg.cc/vDxRTnpj and https://postimg.cc/zyB5wFpk. For further comparison, we include results with Stable Diffusion 2D on https://postimg.cc/rRqpKXWj and backside comparison results on https://postimg.cc/c6HyVFSq.