TeRA : Rethinking Text-guided Realistic 3D Avatar Generation

Yanwen Wang[1], Yiyu Zhuang[1], Jiawei Zhang[1], Li Wang[1], Yifei Zeng[1],
Xun Cao[1], Xinxin Zuo[2], Hao Zhu[1]
[1] Nanjing University [2] Concordia University

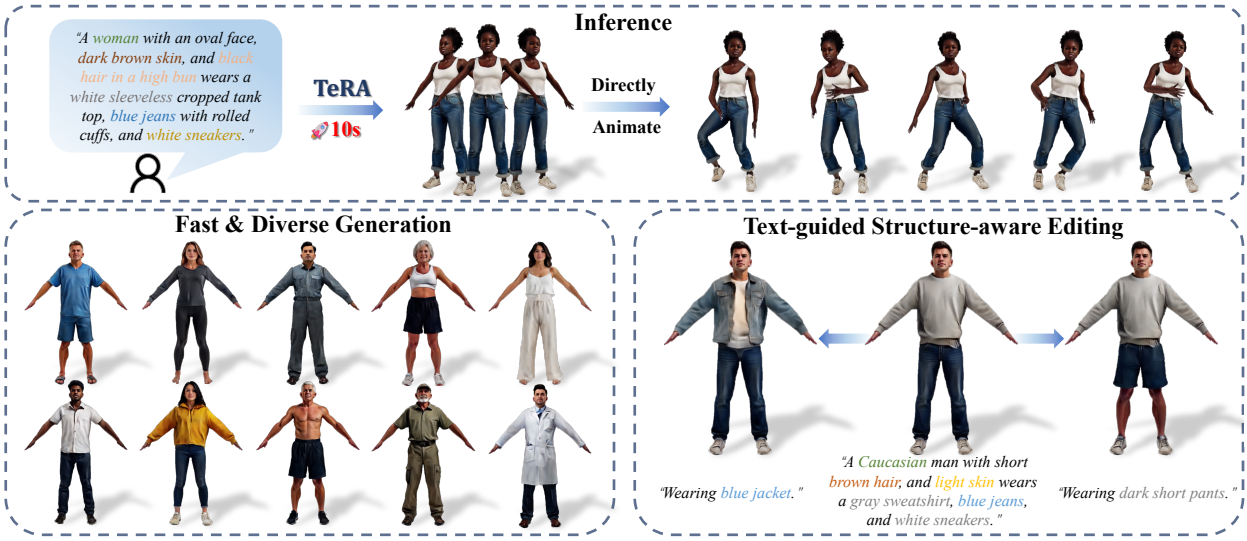

Figure 1. We propose TeRA , the first latent diffusion model specifically designed for text-guided 3D avatar generation. Leveraging a carefully curated dataset and designed network architecture, TeRA achieves superior inference speed, text-to-3D alignment, and visual quality. Our method naturally supports direct animation and enables customizable editing via a structure-aware editing technique.

Abstract

Efficient 3D avatar creation is a significant demand in the metaverse, film/game, AR/VR, etc. In this paper, we rethink text-to-avatar generative models by proposing TeRA , a more efficient and effective framework than the previous SDS-based models and general large 3D generative models. Our approach employs a two-stage training strategy for learning a native 3D avatar generative model. Initially, we distill a deencoder to derive a structured latent space from a large human reconstruction model. Subsequently, a text-controlled latent diffusion model is trained to generate photorealistic 3D human avatars within this latent space. TeRA enhances the model performance by eliminating slow iterative optimization and enables text-based partial customization through a structured 3D human representation. Experiments have proven our approach's superiority over previous text-to-avatar generative models in subjective and objective evaluation. The code and data will be publicly released upon publication.

1. Introduction

With the explosive growth of the metaverse, film/game production, and the AR/VR industry in recent years, the creation of rapid, convenient, and efficient 3D human avatars has emerged as a critical bottleneck. Conventional approaches to 3D human avatar reconstruction often involve intricate and time-consuming modeling techniques using a single camera [64, 77, 92, 93], camera arrays [3, 19, 23], or range sensors [80, 81]. On another line of research, large 3D generative models [75, 76, 86] have recently emerged as an efficient method for producing 3D models from image or text descriptions. Nevertheless, experimental results indicate that all existing 3D generative large models fail to produce plausible results for *photorealistic* 3D human avatars. Such defect is attributed to a profound style bias in the training samples, where there is an overabundance of designed cartoon-like character models and a dearth of high-precision realistic human models.

For the creation of *photorealistic* 3D human avatars, the state-of-the-art approach leverages the score distillation sampling (SDS) strategy [57]. The core idea of SDS lies

in its utilization of pre-trained 2D diffusion models to steer the optimization process for generating 3D models, without the necessity for any 3D data for training. The 3D priors employed in SDS-based methods are derived from pre-trained 2D vision-language large models [57], rich in photorealistic human features. Nevertheless, the inherent absence of explicit 3D structures in 2D diffusion models poses a challenge in ensuring multi-view consistency. This leads to the suboptimal quality of 3D human avatars generated within the SDS framework. Furthermore, the iterative distillation procedure inherent in SDS methods often necessitates a substantial amount of time to complete the optimization process. The inherent limitations of the SDS route prevent it from achieving efficient and robust 3D avatar generation.

In this paper, we introduce TeRA , a feedforward text-to-avatar generative model tailored for the efficient, realistic, and editable creation of 3D humans driven by text. To tackle the challenge of insufficient 3D human data, we leverage a collaborative approach that integrates large vision-language and language models, providing highly accurate and detailed appearance descriptions for HuGe100K [97], an extensive 3D human dataset. Regarding network architecture and training, we utilize a two-stage feedforward prediction model. The first stage employs an autoencoder to extract a structured and readily generable latent space from a comprehensive human reconstruction model. Subsequently, the second stage trains a text-controlled latent diffusion model within this latent space, generating diverse and lifelike 3D human models. We noticed that directly connecting the diffusion model to the encoder results in significantly diminished performance, primarily due to posterior collapse. To address this issue, we propose a distillation module that improves the generated quality and reduces the training resources. Furthermore, by incorporating a structured human representation, we have enabled fine-grained editing of a partially customizable 3D avatar through text descriptions.

Our framework markedly improves the performance of the text-to-avatar model. Firstly, by embracing a single-pass prediction framework, TeRA obviates the necessity for the slow and cumbersome iterative optimization processes characteristic of SDS-based methods. Secondly, incorporating a structured 3D human representation enables text-based partial customization, significantly enhancing usability. Lastly, our model exhibits exceptional generation quality and text-model alignment, outperforming SDS-based approaches and general 3D generative large models. Comprehensive user studies and qualitative/quantitative experimental results substantiate our TeRA 's superior performance.

The contributions of this paper can be summarized as:

- We propose the pioneering text-to-3D avatar generative model built upon the latent diffusion model framework. In terms of speed, text-model alignment, and rendering quality, it surpasses previous state-of-the-art models that leverage scored distillation sampling.
- A distillation module that links the diffusion model to the VAE encoder has been introduced, serving as an essential component for generating high-quality avatar models.
- By introducing a structured 3D human representation, structure-aware editing is achieved for a partially customizable 3D avatar.

2. Related Work

2.1. 2D Diffusion-based Generative Model

Recent years have witnessed remarkable progress in vision-language technologies, driven by breakthroughs in cross-modal representation learning [58] and generative models [29, 45, 60, 62, 88]. These approaches, trained on massive-scale text-image datasets, demonstrate unprecedented capability in understanding and synthesizing visual content. Such advancements have propelled significant improvements in text-to-image generation systems [9, 59, 61, 63] and laid the foundation for text-to-video synthesis [11, 24, 48, 51]. With the large-scale data containing bilions of image-text pairs and video-text pairs, the diffusion model shows great understanding of general objects and enabling the synthesis of high-quality and diverse objects. Furthermore, many works have exploring the controlable generalization with addition condition, including camera motion [18, 25, 32, 38, 91] or others [6, 10, 85, 90].

2.2. Text-to-3D Generation

Recent text-to-3D generation methods can be broadly categorized into two main approaches: feedforward generation and optimization-based generation. **Feedforward generation** methods employ a variety of 3D representations, including point clouds [2, 65], voxel grids [74], meshes [17], implicit radiance fields [5, 13, 26, 95], and 3D Gaussian Splatting (3DGS)[39, 40, 96]. GAN-based approaches leverage conditional GANs[14, 49, 67, 73] to generate 3D assets, but they often struggle with limited diversity and suboptimal quality. Recently, diffusion-based methods for native 3D generation [68, 75, 76, 86] have shown promise by directly generating 3D shapes and textures from text prompts. However, these methods require high-quality, large-scale 3D asset datasets to achieve satisfactory results. **Optimization-based generation** methods adopt a per-prompt generation strategy by distilling 3D knowledge from rich priors learned in the 2D domain. For example, early approaches utilized CLIP guidance [58] to generate the multi-view information [36, 54, 55, 69]. More recent methods employ score distillation sampling (SDS) [57] to transfer the high-quality rendering capabilities of state-of-the-art text-to-image models [15, 16, 53, 57, 66, 70, 72, 79].

Despite their advancements, these methods are hindered by prolonged per-scene optimization times and often produce cartoonish or multi-face 3D outputs.

2.3. Text-to-3D Avatar Generation

By incorporating human prior such as SMPL [50] and SMPL-X [56], the 3D human generation literature has emerged as a distinct subfield within text-to-3D research [12, 20, 30, 37, 46, 71, 82, 84, 89, 94]. AvatarCLIP [31] combines CLIP guidance with SMPL templates to generate 3D avatars. DreamWaltz [35] introduces 3D-aware skeleton conditioning and occlusion-aware SDS to mitigate the Janus (multi-face) problem. DreamHuman [41] utilizes imGHUM [4] to encode pose- and shape-conditioned signed distance fields, enhancing neutral human reconstruction. HumanNorm [34] enhances geometric details by fine-tuning a text-to-depth/normal diffusion models to provide explicit structural constraints. AvatarVerse [83] fine-tunes the ControlNet [85] branch with DensePose [22] as an SDS source for multi-view generation. HumanGaussian [47] integrates skeleton and depth maps to regulate the 3DGS embedded on the SMPL-X template, enabling efficient rendering. TADA [44] applies displacement maps to the SMPL-X shape and texture UV map to represent 3D avatars, optimizing them through a hierarchical rendering approach with SDS. However, these optimization-based methods often suffer from significant drawbacks, including long optimization times—sometimes requiring several hours per scene—and the generation of unrealistic results, such as cartoon-like appearances and oversaturation. Building on the insights from native 3D generation models, we propose a feedforward generation pipeline for 3D human avatars. This approach significantly improves both the generated results' efficiency and realism.

3. Method

This section introduces TeRA, an efficient, realistic, and editable text-guided 3D human generation model. The text-to-avatar data creation is outlined in Sec. 3.1, followed by 3D avatar representation in Sec.3.2. Subsequently, we elaborate on the network architecture and training strategy, encompassing a two-stage latent compression approach detailed in Sec. 3.3, and a structured latent diffusion model presented in Sec. 3.4. Lastly, Sec. 3.5 discusses structured-aware editing, which allows for fine-grained customization of a generated 3D human avatar.

3.1. Text-to-Avatar Dataset

To train a 3D Avatar generation model, the primary challenge lies in establishing a large-scale and diverse text-to-avatar dataset. The HuGe100K [97] dataset, comprising 100k photorealistic multi-view 3D human models, effectively fulfills the requirements for extensive and varied data. However, it suffers from a lack of text annotations. To address this, we enhance the HuGe100K dataset by incorporating semantic annotations, creating a comprehensive large-scale text-to-avatar dataset.

Early annotation of 3D objects often relies on vision-language models such as BLIP [43] or CLIP [58]. However, these models struggle to generate detailed and accurately aligned descriptions, which limits the diversity and precision of text-to-3D generative models. Recently, large vision-language models(VL model)[1, 8, 78] have demonstrated excellent performance in image understanding tasks. Therefore, following [21], we adopt a collaborative annotation approach using a large vision-language model Qwen2.5-VL[8] and a large language model Qwen2.5[78] to annotate multi-viewpoint human image data.

As illustrated in Fig. 2 (a), we first input front/back/right/left views of a human into the Qwen2.5-VL, obtaining comprehensive raw descriptions of various body parts through carefully designed prompts, including facial features, upper and lower clothing, shoes, and more. Subsequently, these raw descriptions are processed by the Qwen2.5 to extract essential information, producing a concise and precise description of approximately 60 words. Finally, Qwen2.5 further refines and condenses the content into a succinct phrase-based description of about 20 words. Our trained text-guided 3D human generation model demonstrates high textual consistency thanks to the meticulous and accurate text annotations. Additional details regarding the data annotation process can be found in *the supplementary material*.

3.2. 3D Human Representation

After creating the dataset, our next step is establishing text-to-avatar generative models. The first issue is deciding how to represent 3D human avatars. In this paper, we follow prior works [87, 97] to represent 3D human avatars with UV-structured 3D Gaussians. We will begin by providing preliminary knowledge about SMPL-X [56] and 3D Gaussian Splatting (3DGS) [39], followed by an introduction to the concrete representation settings we employ.

SMPL-X. is a deformable 3D parametric human model with excellent driving performance and decoupled shape and pose control, currently widely applied in human-driven reconstruction and generation tasks. SMPL-X generates a 3D human mesh using shape parameter β, pose parameter θ, and expression parameter ψ. The generated mesh consists of 10,475 vertices and 54 joints. The deformed human mesh $M(\beta, \theta, \psi)$ is derived from the mesh $T(\beta, \theta, \psi)$ in the canonical space through linear blend skinning (LBS). The process is formulated as:

$$M(\beta, \theta, \psi) = \text{LBS}(T(\beta, \theta, \psi), J(\beta), \theta, \psi, W) \quad (1)$$

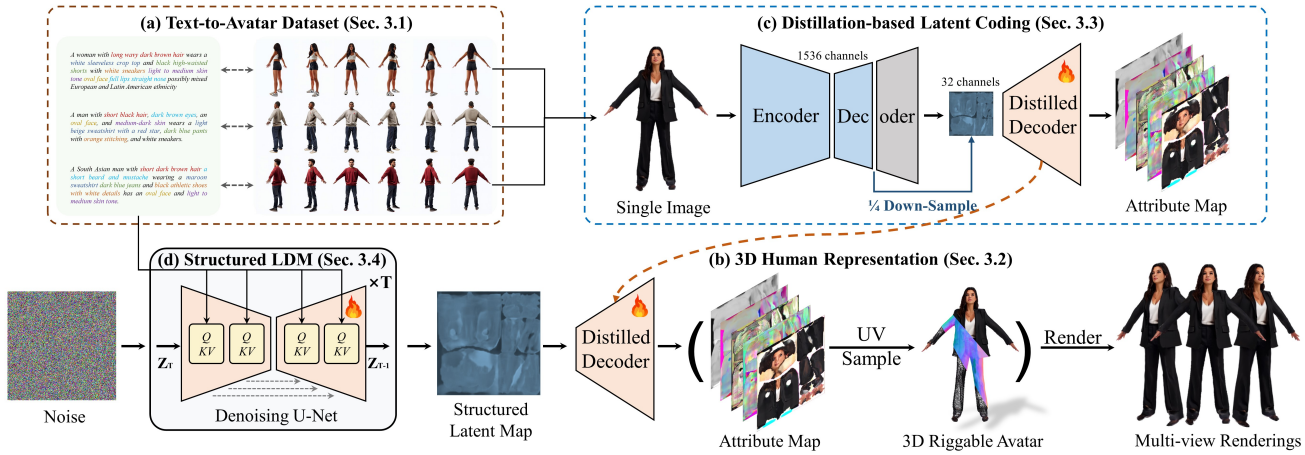

Figure 2. Overall method. (a) Given the annotated multi-view human dataset, we train a text conditioned 3D avatar generative model. (b) The model is established upon a structured 3D human representation. The model training includes two stages: (c) firstly, a decoder is required by distilling a pretrained 3D human reconstruction model; (d) secondly, a structured latent diffusion model (LDM) is trained to generate structured latent maps from noises.

where $J(\beta)$ represents the positions of the key joints, and W denotes the skinning weights. The canonical mesh $T(\beta, \theta, \psi)$ is obtained using the following formula:

$$T(\beta, \theta, \psi) = T_c + B_s(\beta) + B_e(\psi) + B_p(\theta) \quad (2)$$

where T_c is the template human mesh, and $B_s(\beta)$, $B_e(\psi)$, $B_p(\theta)$ represent shape-dependent, expression-independent, and pose-dependent deformations, respectively.

3D Gaussian Splatting. 3D Gaussian is an explicit representation for 3D scenes, composed of a set of 3D Gaussian primitives that can be real-time rendered via differentiable rendering. Each primitive consists of the following four properties: position μ, opacity α, color c, and covariance matrix Σ. In practice, the covariance matrix is typically assumed to be $\Sigma = RSS^TR^T$, where S represents the size of the Gaussian ellipsoid, and R is its rotation matrix.

By applying a view transformation, the 3D Gaussian primitives are projected onto the imaging plane, resulting in a set of 2D Gaussian ellipses. The final imaging process is as follows:

$$c(p) = \prod_{i \in N} \left(c_i \sigma_i \prod_{j=1}^{i-1} (1 - \sigma_j) \right), \quad \sigma_i = \alpha_i G(p, \mu_i, s_i, r_i), \quad (3)$$

where p is the query point position, and μ_i, s_i, r_i, c_i, and α_i represent the position, scale, rotation, color, and opacity of the i-th Gaussian, respectively. $G(p, \mu_i, s_i, r_i)$ represents the value of the i-th Gaussian at the point p.

Structured Gaussians for 3D Human. We represent the 3D human body using a structured Gaussian attribute map, where each Gaussian's attributes are stored in a UV space aligned with the SMPL-X mesh. Initially, the position of each 3D Gaussian $\hat{\mu}_k$ is set to the densified SMPL-X mesh vertices. The scale $\hat{s}_k$ is defined by the relative distance to neighboring Gaussians, and the rotation $\hat{r}_k$ is aligned with the local tangent frame of the 3D surface. A neural network then predicts offset values $\{\delta_{\mu_k}, \delta_{r_k}, \delta_{s_k}\}$ for position, rotation, and scale, as well as the color c_k and opacity α_k of each Gaussian. The final attributes of the Gaussians are computed as:

$$\mu_k = \hat{\mu}_k + \delta\mu_k \quad (4)$$

$$r_k = \hat{r}_k \cdot \delta_{r_k} \quad (5)$$

$$s_k = \hat{s}_k \cdot \delta_{s_k} \quad (6)$$

The complete set of attributes, including μ_k, r_k, s_k, c_k, and α_k, are stored in a multi-channel attribute map within the UV space of the SMPL-X mesh. As shown in Fig. 2 (b), this attribute map is transformed into 3D Gaussian Human through UV sampling and then rendered into images. In our approach, the attribute map serves as the output of the generative network, enabling flexible editing and direct animation of 3D avatars.

3.3. Distillation-based Latent Coding

Our text-to-avatar generative model is based on the Latent Diffusion Model (LDM) framework [60]. LDM generates images by reversing a noising process iteratively in a distilled latent space, reducing computational complexity while preserving semantic information. Recent works [33, 42, 75, 76] have validated its effectiveness on 3D generative tasks. In this paper, we pioneer the integration of UV-structured priors into text-to-avatar generation, thereby establishing the first LDM for text-to-3D avatar generation.

The typical training process of a text-conditioned LDM comprises two stages. A Variational Autoencoder (VAE) is trained in the first stage to establish a latent space. In the second stage, a text-conditioned diffusion model is trained to generate latent maps within this space, which are subsequently decoded by the trained decoder to produce the final results. Through experimental observations, we found that directly training a VAE for complex 3D human models is prone to instability and demands substantial computational resources. Therefore, we propose a distillation-based decoding method that constructs the latent space based on a pre-trained, large-scale reconstruction model. This approach is not only robust but also requires significantly less computational resources.

Concretely, we leverage IDOL [97], a large reconstruction model with an encoder-decoder architecture, to encode the latent space. IDOL directly reconstructs a 3D human model from a single input image and naturally constructs a generalizable and uniform feature space that maps the input image to the 3D human representation. Its model consists of three main components: 1) The first part is a pre-trained high-resolution human foundation model, primarily responsible for capturing human poses and fine-grained appearance details from high-resolution human images. The output feature has a spatial resolution of 64×64 with 1536 channels. This feature space is aligned with the human image space and represents shallow features of the network, which are insufficient to capture 3D human structural information. 2) The second part is a UV-align transform, which aligns the previously obtained human image features into the UV feature space. The output consists of 9216 tokens, with a dimension of 1536. While this feature space contains rich 3D human structural and appearance information, its high dimensionality makes fitting its distribution using generative models challenging. 3) The third part is the UV Decoder, which converts the UV tokens from the previous part into a UV feature of size 1536×1536 with 32 channels. This UV feature is further decoded into 3D human Gaussians and rendered as an RGB image. Upon observation, this feature space exhibits good structural properties and relatively shallow feature representation close to final output, making it easier for the LDM to learn.

However, this UV feature cannot be directly utilized to train LDM due to its high resolution. Therefore, we propose a distillation phase to construct a more compact representation from the original UV feature space. Specifically, the uv feature maps from IDOL are down-sampled to a resolution of 256×256. A streamlined convolutional distillation network, consisting of upsampling and convolution operations, is then trained to restore these features to a 1024×1024 resolution. Subsequently, two separate convolutional networks decode the geometry-related and color-related attributes of the 3D Gaussians into UV maps. Finally, the Gaussian attributes are obtained through UV sampling. The detailed network architecture is provided *in the supplementary material*.

These two networks form the distilled decoder, which reconstructs the 3D Gaussian human representation from the low-resolution latent feature. We randomly select four orthogonal views per avatar during training from the dataset as ground truth. Specifically, we input the front-view image into the IDOL encoder to obtain the corresponding UV features, which are then decoded by the distilled decoder into a 3D human. This 3D representation is rendered from the selected views, and the rendered results I_{pred} are supervised by the corresponding ground truth images I_{gt}. The total training loss L_{dist} consists of two components: the image loss (L2 loss and a VGG loss) and the L2 regularization term for the Gaussian offsets:

$$L_{\text{dist}} = \sum_{i=1}^{N} \left(\|I_{\text{pred}} - I_{\text{gt}}\|^2 + \lambda_{\text{vgg}} L_{\text{vgg}}(I_{\text{pred}} - I_{\text{gt}}) \right) + \lambda_{\text{offset}} \|G_{\text{offset}}\|^2 \quad (7)$$

where $\lambda_{\text{L2}} = 20$, $\lambda_{\text{vgg}} = 20$ and $\lambda_{\text{offset}} = 1$

3.4. Structured Latent Diffusion Model

We use Latent Diffusion to fit the structured UV latent distribution. To introduce text control, we employ CLIP as the text encoder and train the diffusion model using Classifier-Free Guidance. This enables the model to generate text-aligned structured UV features from noise.

Text Conditioning. For the text annotations in the dataset, we follow Stable Diffusion and use CLIP[58] as the text encoder. After encoding the text with CLIP, we obtain 77 tokens of 768 channels, which are then mapped through a small linear layer and injected into the cross attention block of the diffusion model as the text condition.

Classifier-free Guidance. We employ classifier-free guidance for text control. Expressly, during the training of the Diffusion model, we randomly set the text annotations to null for 20% of the data, enabling joint training of both conditioned and unconditioned generation. During inference, the network outputs for the conditioned and unconditioned cases are linearly combined with a weight w to produce the final output [28].

Diffusion Model. Diffusion is a probabilistic model that fits the dataset distribution by progressively denoising Gaussian noise. The denoising process is an inverse discrete-time Markov chain of length t. In the forward noising process, Gaussian noise $e \sim \mathcal{N}(0, I)$ is gradually added to the samples from the dataset at each time step t. At the t-th step, the noisy sample x_t is given by

$$x_t := \alpha(t)x + \sigma(t)e \quad (8)$$

where both α and σ are part of the noise scheduling. After T steps of noising, the sample is fully transformed into Gaussian noise. In the reverse denoising process, the Diffusion model starts with Gaussian noise at step T and progressively denoises until it recovers the clean sample z_0 at step 0.

We adopt an x_0-prediction approach when training our Structured Latent Diffusion model. First, a feature f_0 is extracted from the structured latent feature as a sample. A time step t is randomly chosen from the range 1 to T, and the corresponding α_t and σ_t are generated using a noise scheduler. The feature f_t is then obtained using the equation for x_t. The network is then tasked with predicting the clean feature $\hat{f}_0$ corresponding to f_t, and is supervised using the mean squared error (MSE) loss:

$$L_{\text{diff}} = \|\hat{f}_0 - f_0\|_2^2 \tag{9}$$

3.5. Structure-Aware Editing

Recently, SMPL-X-aligned approaches (e.g., IDOL [97]) have facilitated texture editing of avatars by modifying the SMPL-X UV texture maps and controlling their shape via SMPL-X coefficients. However, due to the increased complexity of clothing geometry and texture, these methods generally struggle with tasks such as clothing replacement.

Benefiting from our effective distillation of IDOL, the latent space we obtain is exceptionally well-structured. Consequently, editing the generated 3D avatars by manipulating the structured latent representation is straightforward. Nevertheless, because an avatar's clothing is often strongly correlated with its identity, directly swapping the corresponding regions of the structured latent between two avatars can result in severe artifacts, such as unnatural edge transitions. Therefore, we choose to leverage the powerful inpainting capability of diffusion models to complete the regions of the structured latent corresponding to the clothing to be replaced, thereby producing a natural and plausible clothing swap effect.

Specifically, since we use latent diffusion to generate 3D digital humans and our latent space is well-structured, it is natural that we can perform virtual try-on by editing the structured latent through diffusion inpainting. Following Avrahami *et al.*'s work [7], we denote the latent corresponding to the 3D digital human as the background part L_{bg}, which needs to be preserved, and execute a specific denoising process to generate the modified foreground L_{fg} as follows. First, we randomly sample the Gaussian noise to obtain the noise L_{fg}^T. At each denoising step t, we predict the noisy latent L_{fg}^{t-1} for the previous step under the control of the target text, then add noise to the clean background latent L_{bg} via the noise scheduler to get L_{bg}^{t-1} for step $t-1$. By using a preprocessed foreground mask $mask_{fg}$, we combine L_{bg}^{t-1} and L_{fg}^{t-1} to obtain L^{t-1}. L^{t-1} is then used as input for the network in the next denoising step to predict L_{fg}^{t-2}. This process is repeated until $t = 0$, at which point the latent after the clothing change is obtained. The resulting latent is then passed into the decoder to generate the 3D human with the new clothing.

4. Experiments

4.1. Implementation Details

Our training set comprises 70,000 pairs of multi-view images, each annotated with SMPL-X parameters and accompanying text descriptions, as detailed in Sec. 3.1. For each individual, we utilize 24 views at a resolution of 896×640 for training, with accompanying text descriptions ranging from 12 to 40 words.

The distillation-based latent decoder is trained on 4 NVIDIA RTX A6000 GPU with a batch size of 2. For each person, four orthogonal views are randomly selected for both rendering and supervisory signals. The structured latent diffusion model is trained on 4 NVIDIA RTX 3090 GPU with a batch size of 8, utilizing the DDPM noise scheduler with 1000 steps. The entire training process takes approximately 90 hours to complete. In the inference phase, we apply the DDPM sampling method with 100 denoising steps to refine latent noise into a 3D representation. The output is then decoded by the Auto Decoder, completing the process in 12 seconds on an NVIDIA RTX 3090 GPU.

4.2. Text-Guided 3D Human Generation

Baselines. We compare our proposed method with existing state-of-the-art text-to-3D human generation methods, including TADA [44], X-oscar [52], HumanGaussian [47], and HumanNorm [34]. All these baseline methods are SDS-based 3D avatar generative models.

Qualitative comparison. The rendered results of all methods are shown in Fig. 3. Under prompts with everyday clothing, SDS-based methods generally fail to generate realistic avatars. Specifically, HumanGaussian, TADA, and X-OSCAR, which directly distill Stable Diffusion using SDS Loss, exhibit overly saturated and unrealistic colors. Furthermore, due to the lack of real human geometric supervision, TADA and X-OSCAR produce avatars with disproportionately small heads, thin arms, and overly long legs. HumanGaussian generates flat and disproportionate human figures. HumanNorm introduces Normal Diffusion and Multi-step SDS Loss, partially mitigating body proportions and color oversaturation issues. However, due to the inherent bias of SDS Loss, discrepancies remain between the distilled knowledge and realistic human distributions, leading to artifacts in the face, forearms, hands, and feet.

In contrast, our method directly learns the distribution of real human bodies using diffusion, avoiding the issues of color oversaturation and unrealistic geometry in other

Figure 3. Qualitative comparison of 3D avatars generated from five text prompts using our method and four baseline methods. The baselines often exhibit over-saturated colors, artifacts, and unrealistic body proportions. HumanNorm shows improved texture realism but struggles with accurate human proportions. Our method generates photorealistic avatars with natural textures and proper geometry.

methods. Additionally, our single-pass generation process leverages the diffusion denoising process without iterative optimization, achieving significantly higher efficiency with an inference time of 12 seconds on an NVIDIA RTX 3090 GPU, compared to several hours required by the baseline methods.

Quantitative comparison. We adopt the CLIP Score [27] and user study to evaluate the five methods objectively and subjectively. The test prompts for evaluation are generated by ChatGPT with random appearance, and the results are reported in Tab. 1

For objective comparison, CLIP score is leveraged to assess the consistency between the input text description and the output renderings. Our method achieves the second-highest CLIP Score among all approaches. Although X-OSCAR attains a marginally higher CLIP Score, we attribute this to its direct incorporation of CLIP loss during training. In contrast, our method generates avatars of notably superior visual quality. This observation is further corroborated by user study results, where our method received the highest preference score, showcasing exceptional realism and alignment with textual descriptions.

For subjective comparison, we invited 28 participants to evaluate different methods based on three criteria: text consistency (Tex.), visual quality (Vis.), and realism (Real.), using questionnaire rating from 0 to 5. The results reveal that the TeRA model excels by achieving the highest score across all three questions, markedly surpassing the runner-up, thereby demonstrating our model's superior text-appearance consistency, enhanced realism, and improved rendering quality.

The runtimes of different methods are also reported in Table 1. As all four other methods are SDS-based methods requiring iterative optimization for each generation, their runtimes are typically more than 1 hour on a single Nvidia RTX 3090 GPU. In contrast, our method boasts a single generation time of just 12 seconds, representing a significant improvement of two orders of magnitude in speed compared to other methods.

4.3. Ablation Study

As illustrated in Fig. 4, we perform an ablation study focusing on two key modules: the resolution of the distilled structured latent space and our novel inpainting strategy designed for virtual try-on.

Latent Space Resolution. In Fig. 4-(a), we compare the performance of structured latent space with resolutions of 128×128 and 256×256. The results demonstrate that the

Method	CLIP Score↑	User Study Tex.↑	User Study Vis.↑	User Study Real.↑	Time↓
TADA [44]	29.86	3.27	2.25	2.11	2.3h
X-Oscar [52]	**32.46**	3.56	2.54	2.26	2.0h
HumanGaussiann [47]	29.31	3.74	2.49	2.28	1.0h
HumanNorm [34]	29.94	3.79	3.01	3.04	4.0h
TeRA (Ours)	30.17	**4.54**	**4.33**	**4.35**	**12s**

Table 1. Quantitative comparison of CLIP Score, User Study results and time cost for text-to-3D human generation methods. The **best** and second-best scores are marked.

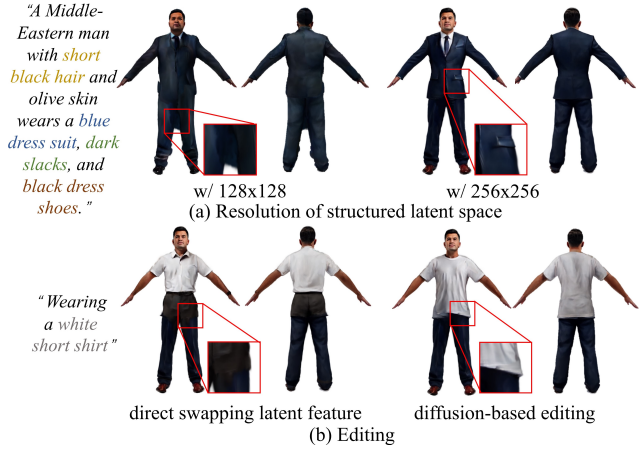

Figure 4. Ablation study on key components of TeRA . (a) Comparison of structured latent space resolutions (128×128 vs. 256×256) showing that higher resolution provides richer details and fewer artifacts. (b) Evaluation of the proposed inpainting method for virtual try-on, demonstrating that inpainting on the structured latent space yields smoother transitions and fewer artifacts compared to direct feature swapping.

higher resolution of 256×256 provides richer details and significantly reduces artifacts in the generated 3D avatars. The results validate our choice of a 256×256 resolution as an optimal balance, providing high-quality outputs without excessively increasing the training cost of the diffusion model.

Inpainting Strategy. We evaluate the effectiveness of our proposed inpainting method for virtual try-on, as shown in Fig. 4-(b). We compare our method to a baseline approach that directly swaps the latent features in the specified region with newly generated features guided by the new prompt. As evidenced in the figure, our proposed inpainting technique applied to the structured latent space results in smoother transitions and significantly reduces artifacts compared to the direct feature swapping method.

4.4. Application

Our proposed model generates 3D avatars by learning structured latent representations through diffusion, enabling versatile downstream applications such as *editing* and *animation*, as illustrated in Fig. 5. Since our method directly generates the structured latent representation using diffusion, it supports inpainting operations on the generated latent space, allowing seamless 3D avatar editing such as virtual try-on. Additionally, representing the 3D human using a combination of SMPL-X and Gaussian Attribute Maps enables flexible texture editing through color map modifications and shape editing by altering SMPL-X parameters. Furthermore, this design facilitates straightforward animation by directly driving the generated avatars using SMPL-X pose sequences, eliminating the need for post-processing and ensuring efficient and realistic motion control.

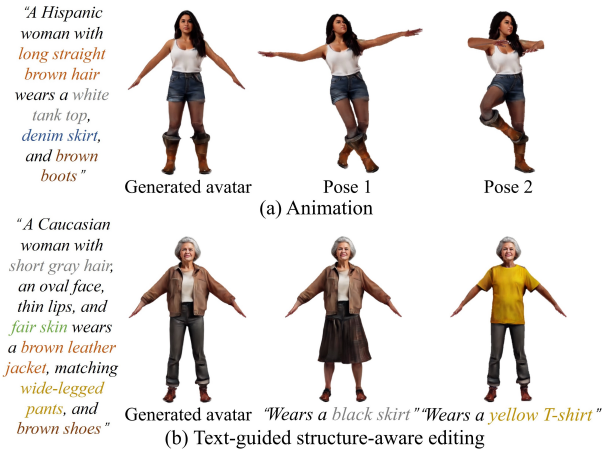

Figure 5. Illustration of TeRA 's downstream applications. The upper row shows direct avatar animation using SMPL-X poses without post-processing, while the lower row presents natural virtual try-on results via diffusion-based latent space editing.

5. Conclusion

We introduce TeRA , a text-to-3D avatar generation model that achieves fast and high-quality 3D human reconstruction. By leveraging a structured latent space through distilling a pre-trained large reconstruction model, TeRA produces photorealistic avatars with strong text-model alignment, outperforming state-of-the-art SDS-based models in both speed and visual quality. Our method enables practical applications such as text-based editing and animation, demonstrating its potential in virtual try-on, gaming, and AR/VR scenarios.

TeRA still faces certain limitations. As the training data consists of static models, it cannot model dynamic details like clothing wrinkles resulting from human movement. Furthermore, due to TeRA's reliance on the SMPL-X model for human body representation, its modeling quality is limited for loose garments such as dresses.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[2] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018. 2

[3] Oleg Alexander, Mike Rogers, William Lambeth, Jen-Yuan Chiang, Wan-Chun Ma, Chuan-Chang Wang, and Paul Debevec. The digital emily project: Achieving a photorealistic digital actor. *IEEE Computer Graphics and Applications*, 30 (4):20–31, 2010. 1

[4] Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imghum: Implicit generative models of 3d human shape and articulated pose. In *ICCV*, 2021. 3

[5] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y Ogras, and Linjie Luo. Panohead: Geometry-aware 3d full-head synthesis in 360deg. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20950–20959, 2023. 2

[6] Søren Asmussen and Michael Taksar. Controlled diffusion models for optimal dividend pay-out. *Insurance: Mathematics and Economics*, 20(1):1–15, 1997. 2

[7] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM transactions on graphics (TOG)*, 42 (4):1–11, 2023. 6

[8] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3

[9] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2

[10] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852, 2023. 2

[11] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2

[12] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. *arXiv preprint arXiv:2304.00916*, 2023. 3

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

[14] Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. Text2shape: Generating shapes from natural language by learning joint embeddings. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pages 100–116. Springer, 2019. 2

[15] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023. 2

[16] Zilong Chen, Feng Wang, and Huaping Liu. Text-to-3d using gaussian splatting. *arXiv preprint arXiv:2309.16585*, 2023. 2

[17] Shiyang Cheng, Michael Bronstein, Yuxiang Zhou, Irene Kotsia, Maja Pantic, and Stefanos Zafeiriou. Meshgan: Non-linear 3d morphable models of faces. *arXiv preprint arXiv:1903.10384*, 2019. 2

[18] Soon Yau Cheong, Duygu Ceylan, Armin Mustafa, Andrew Gilbert, and Chun-Hao Paul Huang. Boosting camera motion control for video diffusion transformers. *arXiv preprint arXiv:2410.10802*, 2024. 2

[19] Paul Debevec. The light stages and their applications to photoreal digital actors. *SIGGRAPH Asia*, 2(4):1–6, 2012. 1

[20] William Gao, Noam Aigerman, Thibault Groueix, Vova Kim, and Rana Hanocka. Textdeformer: Geometry manipulation using text guidance. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 3

[21] Yunhao Ge, Xiaohui Zeng, Jacob Samuel Huffman, Tsung-Yi Lin, Ming-Yu Liu, and Yin Cui. Visual fact checker: Enabling high-fidelity detailed caption generation. 2024. 3

[22] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 3

[23] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (ToG)*, 38(6):1–19, 2019. 1

[24] Xun Guo, Mingwu Zheng, Liang Hou, Yuan Gao, Yufan Deng, Chongyang Ma, Weiming Hu, Zhengjun Zha, Haibin Huang, Pengfei Wan, et al. I2v-adapter: A general image-to-video adapter for video diffusion models. *arXiv preprint arXiv:2312.16693*, 2023. 2

[25] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 2

[26] Yuxiao He, Yiyu Zhuang, Yanwen Wang, Yao Yao, Siyu Zhu, Xiaoyu Li, Qi Zhang, Xun Cao, and Hao Zhu. Head360: Learning a parametric 3d full-head for free-view synthesis in 360. In *European Conference on Computer Vision*, pages 254–272. Springer, 2024. 2

[27] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 7

[28] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*. 5

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020. 2

[30] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. Eva3d: Compositional 3d human generation from 2d image collections. *arXiv preprint arXiv:2210.04888*, 2022. 3

[31] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *ACM Transactions on Graphics*, 2022. 3

[32] Chen Hou, Guoqiang Wei, Yan Zeng, and Zhibo Chen. Training-free camera control for video generation. *arXiv preprint arXiv:2406.10126*, 2024. 2

[33] Tao Hu, Fangzhou Hong, and Ziwei Liu. Structldm: Structured latent diffusion for 3d human generation. In *European Conference on Computer Vision*, pages 363–381. Springer, 2024. 4

[34] Xin Huang, Ruizhi Shao, Qi Zhang, Hongwen Zhang, Ying Feng, Yebin Liu, and Qing Wang. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. In *CVPR*, 2024. 3, 6, 8

[35] Yukun Huang, Jianan Wang, Ailing Zeng, He Cao, Xianbiao Qi, Yukai Shi, Zheng-Jun Zha, and Lei Zhang. Dreamwaltz: Make a scene with complex 3d animatable avatars. *arXiv preprint arXiv:2305.12529*, 2023. 3

[36] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 867–876, 2022. 2

[37] Ruixiang Jiang, Can Wang, Jingbo Zhang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Avatarcraft: Transforming text into neural human avatars with parameterized shape and pose control. *arXiv preprint arXiv:2303.17606*, 2023. 3

[38] Wonjoon Jin, Taesung Kim, and In So Lee. Flovd: Optical flow meets video diffusion model for enhanced camera control. *arXiv preprint arXiv:2502.08244*, 2025. 2

[39] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 2, 3

[40] Tobias Kirschstein, Simon Giebenhain, Jiapeng Tang, Markos Georgopoulos, and Matthias Nießner. Gghead: Fast and generalizable 3d gaussian heads. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 2

[41] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. 2023. 3

[42] Yushi Lan, Fangzhou Hong, Shuai Yang, Shangchen Zhou, Xuyi Meng, Bo Dai, Xingang Pan, and Chen Change Loy. Ln3diff: Scalable latent neural fields diffusion for speedy 3d generation. In *European Conference on Computer Vision*, pages 112–130. Springer, 2024. 4

[43] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. pages 12888–12900, 2022. 3

[44] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J. Black. TADA! Text to Animatable Digital Avatars. In *3DV*, 2024. 3, 6, 8

[45] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 2

[46] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10462–10472, 2022. 3

[47] Xian Liu, Xiaohang Zhan, Jiaxiang Tang, Ying Shan, Gang Zeng, Dahua Lin, Xihui Liu, and Ziwei Liu. Humangaussian: Text-driven 3d human generation with gaussian splatting. 2024. 3, 6, 8

[48] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024. 2

[49] Zhengzhe Liu, Yi Wang, Xiaojuan Qi, and Chi-Wing Fu. Towards implicit text-guided 3d shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17896–17906, 2022. 2

[50] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 2015. 3

[51] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024. 2

[52] Yiwei Ma, Zhekai Lin, Jiayi Ji, Yijun Fan, Xiaoshuai Sun, and Rongrong Ji. X-oscar: A progressive framework for high-quality text-guided 3d animatable avatar generation. *arXiv preprint arXiv:2405.00954*, 2024. 6, 8

[53] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel 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*, pages 12663–12673, 2023. 2

[54] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13492–13502, 2022. 2

[55] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Clip-mesh: Generating textured meshes

from text using pretrained image-text models. In *SIGGRAPH Asia 2022 conference papers*, pages 1–8, 2022. 2

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

[57] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 1, 2

[58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. pages 8748–8763, 2021. 2, 3, 5

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

[60] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 4

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

[62] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2

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

[64] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019. 1

[65] Dong Wook Shu, Sung Woo Park, and Junseok Kwon. 3d point cloud generative adversarial network based on tree structured graph convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3859–3868, 2019. 2

[66] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In *ICLR*, 2024. 2

[67] Xi Tian, Yong-Liang Yang, and Qi Wu. Shapescaffolder: Structure-aware 3d shape generation from text. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2715–2724, 2023. 2

[68] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, , Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024. 2

[69] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3835–3844, 2022. 2

[70] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. 2

[71] Jionghao Wang, Yuan Liu, Zhiyang Dou, Zhengming Yu, Yongqing Liang, Xin Li, Wenping Wang, Rong Xie, and Li Song. Disentangled clothed avatar generation from text descriptions. *arXiv preprint arXiv:2312.05295*, 2023. 3

[72] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 2

[73] Jiacheng Wei, Hao Wang, Jiashi Feng, Guosheng Lin, and Kim-Hui Yap. Taps3d: Text-guided 3d textured shape generation from pseudo supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16805–16815, 2023. 2

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

[75] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. In *NeurIPS*, 2024. 1, 2, 4

[76] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024. 1, 2, 4

[77] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed humans optimized via normal integration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 512–523, 2023. 1

[78] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 3

[79] Taoran Yi, Jiemin Fang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussian splatting with point cloud priors. *arXiv preprint arXiv:2310.08529*, 2023. 2

[80] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *Proceedings of*

the IEEE conference on computer vision and pattern recognition, pages 7287–7296, 2018. 1

[81] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5746–5756, 2021. 1

[82] Yifei Zeng, Yuanxun Lu, Xinya Ji, Yao Yao, Hao Zhu, and Xun Cao. Avatarbooth: High-quality and customizable 3d human avatar generation. *arXiv preprint arXiv:2306.09864*, 2023. 3

[83] Huichao Zhang, Bowen Chen, Hao Yang, Liao Qu, Xu Wang, Li Chen, Chao Long, Feida Zhu, Kang Du, and Min Zheng. Avatarverse: High-quality & stable 3d avatar creation from text and pose. *arXiv preprint arXiv:2308.03610*, 2023. 3

[84] Hao Zhang, Yao Feng, Peter Kulits, Yandong Wen, Justus Thies, and Michael J Black. Text-guided generation and editing of compositional 3d avatars. *arXiv preprint arXiv:2309.07125*, 2023. 3

[85] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2, 3

[86] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics*, 2024. 1, 2

[87] Weitian Zhang, Yichao Yan, Yunhui Liu, Xingdong Sheng, and Xiaokang Yang. E 3gen: Efficient, expressive and editable avatars generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6860–6869, 2024. 3

[88] Xinchen Zhang, Ling Yang, Yaqi Cai, Zhaochen Yu, Kaini Wang, Jiake Xie, Ye Tian, Minkai Xu, Yong Tang, Yujiu Yang, and Bin Cui. Realcompo: Balancing realism and compositionality improves text-to-image diffusion models. *NeurIPS*, 2024. 2

[89] Rui Zhao, Wei Li, Zhipeng Hu, Lincheng Li, Zhengxia Zou, Zhenwei Shi, and Changjie Fan. Zero-shot text-to-parameter translation for game character auto-creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21023, 2023. 3

[90] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:11127–11150, 2023. 2

[91] Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024. 2

[92] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6): 3170–3184, 2021. 1

[93] Hao Zhu, Xinxin Zuo, Haotian Yang, Sen Wang, Xun Cao, and Ruigang Yang. Detailed avatar recovery from single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7363–7379, 2021. 1

[94] Jingyu Zhuang, Di Kang, Linchao Bao, Liang Lin, and Guanbin Li. Dagsm: Disentangled avatar generation with gs-enhanced mesh. *arXiv preprint arXiv:2411.15205*, 2024. 3

[95] Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. Mofanerf: Morphable facial neural radiance field. In *European conference on computer vision*, pages 268–285. Springer, 2022. 2

[96] Yiyu Zhuang, Yuxiao He, Jiawei Zhang, Yanwen Wang, Jiahe Zhu, Yao Yao, Siyu Zhu, Xun Cao, and Hao Zhu. Towards native generative model for 3d head avatar, 2024. 2

[97] Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. Idol: Instant photorealistic 3d human creation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2, 3, 5, 6