ANIME-READY: CONTROLLABLE 3D ANIME CHARACTER GENERATION WITH BODY-ALIGNED COMPONENT-WISE GARMENT MODELING

Jiachen Qian[1,2], Hongye Yang[2], Youtian Lin[3], Tianhao Zhao[2], Feihu Zhang[2]*, Yao Yao[3]†, Hengshuang Zhao[1]
[1]The University of Hong Kong, [2]DreamTech, [3]Nanjing University

ABSTRACT

Automated generation of 3D anime characters has become increasingly important in digital entertainment, including animation production, virtual reality, gaming, and virtual influencers. Compared to realistic human modeling, anime-style modeling requires exaggerated proportions, stylized surface details, and artistically consistent garments, posing unique challenges for automated 3D generation. Existing anime-style approaches often suffer from low-quality meshes, blurry textures, and lack of inner skeletons, which limits their usability in animation. In this work, we present a novel framework for high-quality 3D anime character generation to overcome these limitations by combining the expressive power of the Skinned Multi-Person Linear (SMPL) model with precise garment modeling. We extend the SMPL model to Anime-SMPL to better capture the distinctive features of anime characters, which enables unified skeleton generation and blendshape-based facial expression control, rendering the generated characters animation-ready. To complement the body model, we introduce a body-aligned component-wise garment generation pipeline, which models hairstyles, upper garments, lower garments, and accessories as structured components aligned with the underlying body geometry. Furthermore, our method produces high-quality skin and facial textures, as well as detailed garment textures, enhancing the visual fidelity of the generated characters. Experimental results demonstrate that our framework significantly outperforms baseline methods in terms of mesh quality, texture clarity, and garment-body alignment, making it well-suited for a wide range of applications in anime content creation and interactive media.

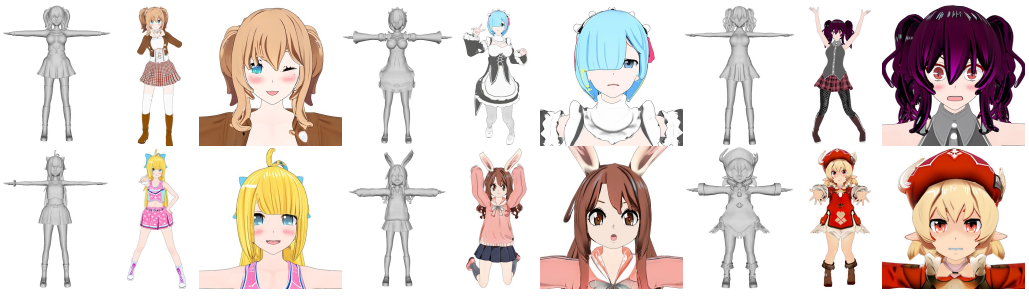

Figure 1: Our Anime-Ready generates high-quality, controllable 3D anime characters from text or a single image, with fine-grained control over actions such as finger movements and facial expressions.

*Project leader.
†Corresponding author.
This research was supported by DreamTech, and the IP belongs to DreamTech.

1 INTRODUCTION

While recent advances in 3D human modeling have led to breakthroughs in reconstructing realistic human avatars, generating high-quality, animatable 3D characters in anime style remains a highly challenging task. Anime characters often feature stylized anatomy, typically exhibit more complex garments, diverse hairstyles, and exaggerated facial features, which challenges traditional reconstruction pipelines.

Recent methods (Peng et al., 2024; He et al., 2024b; Qian et al., 2025; Dong et al., 2025) reconstruct anime characters from images using large models (Hong et al., 2023), generating plausible geometry and textures. However, these approaches often fail to capture fine details, such as hand poses and hair structure. Besides, the results of these models are typically unrigged and lack consistent topologies, which is unsuitable for parametric animation. In contrast, parametric body models such as SMPL (Loper et al., 2023) and SMPL-X (Pavlakos et al., 2019) are widely used for realistic and animatable human bodies (Cao et al., 2024; Dong et al., 2024; Hong et al., 2022; Huang et al., 2023a;b; Pan et al., 2024; Wang et al., 2024a;b; Yang et al., 2024; Zhang et al., 2022; Dong et al., 2023; Xiu et al., 2022; 2023; Huang et al., 2024a; Liao et al., 2024). These models provide well-defined skeletons and consistent mesh topologies for easy animation. However, as shown in Figure 7 , they are designed for realistic human proportions and cannot represent the exaggerated features of anime-style characters.

In this paper, we present a novel pipeline that combines the controllability of SMPL-based models with the generative power of 3D diffusion models to synthesize high-quality anime-style 3D characters that are animatable. Unlike previous methods (Peng et al., 2024; Qian et al., 2025) that directly generate the entire character, our approach decomposes the character into modular components: body, hair, upper garments, lower garments, and accessories. Each part is generated as a separate high-resolution textured mesh for the final assembly.

To make the generated 3D anime characters animatable, we introduce Anime-SMPL, a parametric model designed for anime-style characters with high-quality meshes suitable for animation. Anime-SMPL retains the core SMPL structure while incorporating modifications for exaggerated eyes and anime-style proportions. Our Anime-SMPL also provides a consistent UV layout across characters, enabling direct texture generation in UV space. Conditioned on both a character image and a textual description, we adopt a multi-view diffusion model to predict UV textures for six semantic regions: body skin, facial skin, eyebrows, eyelashes, and the left and right eyes.

To generate meshes for hair, garments, and accessories, we propose a Multi-Shape Diffusion Transformer (DiT) architecture incorporating a Mixture-of-Experts (MoE) (Jacobs et al., 1991) module. This unified model dynamically routes inputs to specialized branches according to garment types to produce distinct meshes. We further guide the generation to be consistent with the body geometry by sampling points from the Anime-SMPL surface as an explicit geometry prior. As a result, our method produces well-aligned, separate meshes for hair, garments, and accessories.

In addition, we use a diffusion model to generate images of each garment component, conditioned on a reference image of the full body. Equipped with a multi-component self-attention mechanism, this model can selectively focus on relevant regions in the reference image. The resulting component images are then processed individually by an MVAdapter (Huang et al., 2024b) to produce the final textures of the garments.

The experiments demonstrate that our method achieves state-of-the-art performance in 3D anime character generation. Our method significantly outperforms previous approaches in both geometry and texture quality, while maintaining the highest fidelity to the input images.

Our main contributions are summarized as follows.

- We introduce Anime-SMPL, a unified human body template designed specifically for anime-style 3D characters. This template not only enables high-quality, animatable body mesh generation, but also facilitates direct texture synthesis in UV space.
- We propose a Multi-Shape DiT architecture incorporating a MoE mechanism that leverages character geometry as guidance. This design allows a single unified model to generate distinct meshes for each garment component while ensuring mesh compatibility with the underlying body shape, significantly reducing interpenetration artifacts.

- We propose a high-resolution, component-wise texture generation pipeline that employs a diffusion model to disentangle garment components from the input image. This enables independent texture synthesis for each component while mitigating color bleeding across different regions.
- Our method surpasses previous approaches in anime-style 3D character generation, achieving higher fidelity and overall quality. Moreover, our approach makes the generation of anime-style 3D characters practical for real-world applications, enabling downstream tasks such as garment retargeting, motion control, and facial expression control.

2 Related Work

2.1 3D Object Generation

In recent years, diffusion models have made remarkable breakthroughs in image generation. However, compared to massive 2D images, the scarcity of 3D data makes it challenging to directly train highly generalizable 3D generative models. To alleviate this issue, Poole et al. (2022) introduced the Score Distillation Sampling (SDS) loss, which leverages pre-trained 2D diffusion models to guide 3D generation. While SDS-based approaches (Chen et al., 2023; Lin et al., 2023; Poole et al., 2022; Qian et al., 2023; Raj et al., 2023; Tsalicoglou et al., 2024; Wang et al., 2023b;a) have significantly improved 3D generation performance, the resulting geometry often lacks fidelity, as SDS does not inherently enforce geometric consistency. Some methods (Lu et al., 2024; Long et al., 2024; Wu et al., 2024a) attempted to deal with this problem by reconstructing 3D geometry from multi-view normal maps, which can be inferred from a single image or synthesized by learned view generation, but they may suffer from limited volumetric consistency and difficulty in handling complex topologies.

With the increasing availability of 3D data, an emerging trend in recent methods is to move beyond SDS-based supervision. LRM-based approaches(Hong et al., 2023; Wang et al., 2024c; Li et al., 2023; Xu et al., 2024a;b; Tang et al., 2024; Zhang et al., 2024a;b) predicted triplane representations from limited inputs to reconstruct 3D models. Other methods (Wu et al., 2024b; Cui et al., 2024) applied diffusion models directly on 3D data to generate triplanes for 3D synthesis. In addition to triplane representations (Ren et al., 2024; Xiang et al., 2024), alternative formats such as sparse voxel grids (Xiang et al., 2024) and VecSet-based representations (Zhang et al., 2023; 2024c; Li et al., 2024; Chen et al., 2024b; Zhao et al., 2025; Li et al., 2025) are also gaining traction in 3D generation.

2.2 Garment Generation

Sewing patterns have been widely adopted in garment generation methods (He et al., 2024a; Liu et al., 2023) due to physical realism and alignment with realistic clothing construction. These representations provide clear semantic structure and high controllability, allowing for fine-grained adjustment of garment fit, style, and layout. However, generating accurate patterns requires expert knowledge, and may struggle with stylized or unconventional garments. Besides, inferring sewing patterns from images or text remains a challenge due to structural and alignment constraints. In addition to pattern-based generation, Qiu et al. (2023) reconstructed detailed 3D dynamic clothing from monocular video, Sarafianos et al. (2024) leveraged Long et al. (2024) to generate garments. Luo et al. (2025) proposed a body-aligned generation method of wearable assets based on native 3D diffusion. However, post-processing is still required after generation, since there are still small penetrations between the garment and body.

2.3 Character Generation

Previous works have achieved significant progress in realistic 3D human body generation. Xiu et al. (2022) leveraged implicit representations from normal maps to reconstruct detailed clothed human surfaces while Xiu et al. (2023) performd explicit normal integration for high-quality geometry recovery. Huang et al. (2023a) introduced Pixel-Semantics Difference-Sampling to optimize body and clothes, and Cao et al. (2024)introduced a dual-observation-space design consisting of a canonical space and a posed space related by a learnable deformation field for jointly optimization. Dong et al. (2024) proposed a layer-wise generation strategy to get better clothing structure and higher realness. Liao et al. (2024) further leveraged the synergy of a 2D diffusion model and a parametric body model

to generate digital characters. Several prior works have also explored 3D character generation in the anime style. Peng et al. (2024) used a large reconstruction model to generate anime-style 3D characters, but lack of details in many aspects. He et al. (2024b) and Dong et al. (2025) adopted a component-wise generation strategy to improve overall quality, and Qian et al. (2025) utilized sparse voxel to increase the resolution of generated meshes, enhancing quality of detailed regions like face and hands. However, due to limitations in fine-grained details, all these methods still fall short of producing truly usable 3D anime characters.

3 METHOD

Our pipeline, illustrated in Figure 2, first generates an image of the character in a canonical pose (e.g., A-pose) from either a textual description or an input image. Based on this canonical-posed image and the Anime-SMPL template, we estimate the Anime-SMPL parameters of the target character. Analogous to SMPL, we use a joint regressor to obtain joint locations. Subsequently, we sample points on the surface of the estimated Anime-SMPL mesh and encode them into body latent tokens using a VecSet VAE (Zhao et al., 2025). These body latent tokens, together with noised latent tokens, garment label tokens, and the canonical-pose image, are fed into our MoE-structured Multi-Shape DiT network to generate 3D garment meshes. Finally, we generate texture maps for both the body and individual garments.

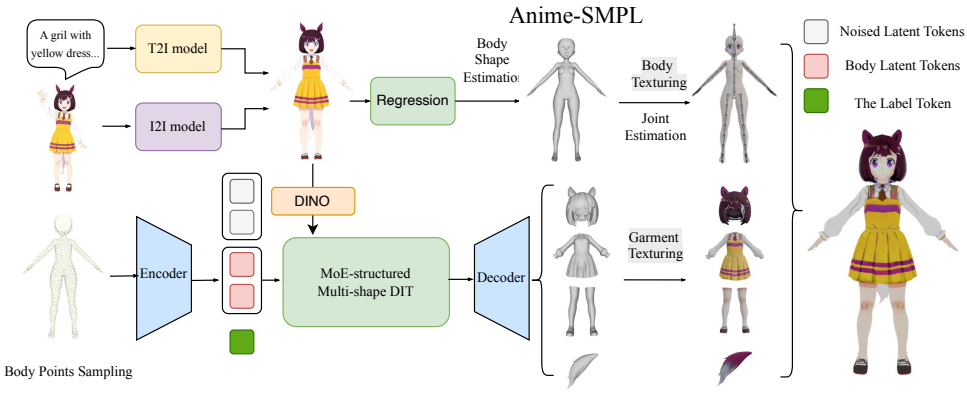

Figure 2: Pipeline for controllable 3D character generation based on Anime-SMPL and Body-Aligned Component-Wise Garment Modeling.

3.1 2D CANONICAL POSE CHARACTER GENERATION

As shown in Figure 2, our framework generates a front-view image of a character in a canonical pose from either a textual description or an image of the character in an arbitrary pose.

Text-to-Image Synthesis. We fine-tune PixArt-Σ (Chen et al., 2024a) using paired data consisting of textual descriptions and corresponding front-view images of characters in a canonical pose.

Image-to-Image Synthesis. Similar to CharacterGen (Peng et al., 2024) and StdGen (He et al., 2024b), we employ a ReferenceUNet and a CLIP encoder (Radford et al., 2021) as image feature extractors, and incorporate pose information using a general a-pose skeleton image as an additional condition. To train our image-to-image model, we render diverse anime characters in full-body views under various camera viewpoints, poses, and facial expressions. To further improve the generalization performance of the model, we apply data augmentation strategies, including varying illumination conditions, modifying contour line styles, and upper-body-focused augmentations commonly used in pose estimation tasks (Xiao et al., 2018; Sun et al., 2019).

3.2 ANIME-SMPL

The SMPL model (Loper et al., 2015), has been instrumental in human shape estimation and 3D reconstruction, due to its ability to capture various body shapes and poses. However, the application of SMPL to anime characters is limited, owing to substantial geometric and stylistic differences, as illustrated in Figure 7. To overcome this limitation, we introduce Anime-SMPL, a parameterized body model tailored for anime-style characters, designed to better capture their unique proportions and stylistic characteristics.

Anime-SMPL Parameterization. In contrast to SMPL (Loper et al., 2015), which parameterizes both pose and shape, our framework focuses exclusively on shape modeling, as our aim is to reconstruct 3D anime characters from input images depicting a fixed canonical pose. As in SMPL, we apply Principal Component Analysis (PCA) to model shape variations in anime characters. We perform PCA on a dataset of 20,000 characters, each represented by 12,489 vertices, and retain the top 98 principal components to capture the dominant modes of shape variation. Similarly to SMPL, we estimate the joint regressor matrix $\mathbf{J}$ using non-negative least squares (NNLS), by solving the following optimization problem:

$$\underset{\mathbf{J}\geq 0}{\text{minimize}} \ \|\mathbf{JV} - \mathbf{B_V}\|_F^2 \quad \text{s.t.} \quad \mathbf{J}\mathbf{1}_N = \mathbf{1}_K$$

where $\mathbf{V} \in \mathbb{R}^{N\times 3}$ is the vertex matrix, $\mathbf{B_V} \in \mathbb{R}^{K\times 3}$ is the target joint location matrix, $\mathbf{1}_N$ is an all-ones vector of length N, and $\mathbf{1}_K$ is an all-ones vector of length K. The problem is solved using the Splitting Conic Solver (SCS). The linear blend skinning (LBS) weights are directly adopted from pre-defined values in our dataset.

Shape Parameters Estimation. To estimate shape parameters for arbitrary anime characters, we train a shape prediction network on front-view images of characters exhibiting diverse clothing styles and body shapes. The network is optimized using the mean squared error (MSE) between the predicted parameters $\hat{\boldsymbol{\beta}}$ and the ground truth $\boldsymbol{\beta}$. In practice, we employ a ResNet-based architecture (He et al., 2016) as our shape prediction network.

3.3 MOE-STRUCTURED MULTI-SHAPE DIT FOR GARMENT GENERATION

Despite recent progress in 3D generation algorithms, synthesizing high-quality anime characters remains challenging. The difficulty arises primarily from the complex geometric structures inherent to anime characters, such as diverse hairstyles and intricate costumes. Previous approaches typically generate the entire character, making it difficult to precisely control fine-grained details in each part. To mitigate these challenges, we propose a component-wise generation strategy. Specifically, for each character, we decompose the outfit, excluding the body, into four components: *hairstyles*, *upper garments*, *lower garments*, and *accessories*. We employ the VecSet Diffusion Model (Zhang et al., 2023; 2024c; Li et al., 2024; Chen et al., 2024b; Zhao et al., 2025; Li et al., 2025) as the core generative model to synthesize these components.

MoE-structured DiT Block. The original VecSet Diffusion Model cannot directly generate garment components from a single image. To address this limitation, we design a MoE-structured Multi-Shape DiT architecture that enables independent generation of each garment component. Figure 3 illustrates the architecture of our MoE-structured Multi-Shape DiT.

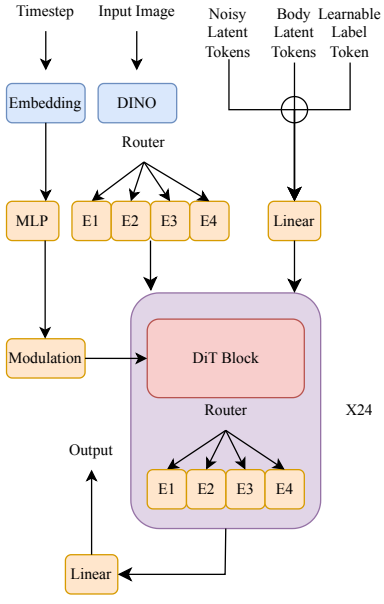

Figure 3: Overview of our MoE-structured Multi-Shape DiT.

In the VecSet Diffusion Model, the input image is first encoded by DINOv2 (Oquab et al., 2023) to produce conditioning tokens for the DiT backbone network. We adopt a Mixture-of-Experts (MoE)

design comprising four MLP expert branches, each dedicated to a specific garment component. In our MoE-structured Multi-Shape DiT, all parameters are shared except those of the four MLP experts, enabling precise, component-aware generation with minimal parameter overhead. We use a learnable label token to guide the network in generating the corresponding component.

Body-Aligned Garment Generation. Relying solely on image-based conditioning is insufficient for generating body-aligned garment components with the VecSet Diffusion Model, as it often leads to shape misalignments and noticeable interpenetration artifacts. In this section, we introduce an additional geometric condition to guide the shape and spatial alignment of the garment components.

Specifically, we perform point cloud sampling on the 3D body surface and encode it using the VecSet VAE encoder to obtain body latent tokens. The body latent tokens are then concatenated with the noised garment component tokens as input to the VecSet Diffusion Model. To reduce computational overhead, we encode the body latent tokens at lower resolution: the garment component tokens have a length of 3072, while the body latent tokens are set to 512. With this conditioning, which directly encodes spatial information of the body surface, the model is able to generate garment components accurately aligned with the underlying character geometry, which greatly reduces penetration issues and improves overall fidelity. The garment components are represented as Signed Distance Fields (SDFs) as output and are finally converted to 3D meshes via the marching cubes algorithm.

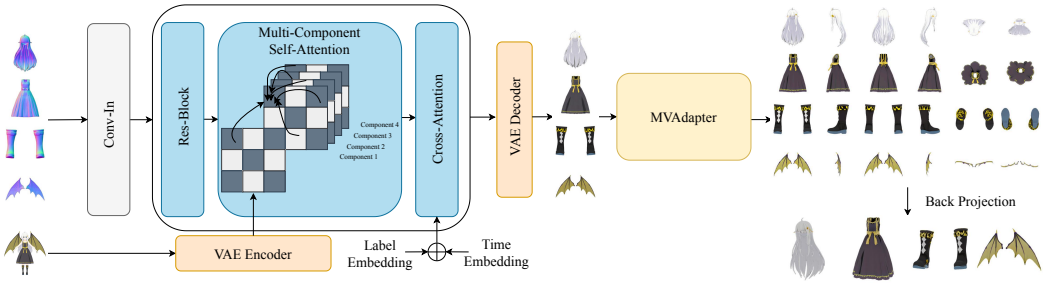

Figure 4: Pipeline of our Component-Wise High-Resolution Texture Generation.

3.4 Texture Generation

Previous works, such as MVAdapter (Huang et al., 2024b), have demonstrated strong performance in multiview generation. Motivated by these advances, we adopt MVAdapter as a core component in our texture generation pipeline. Specifically, our method divides texture generation into two stages: one for the body and the other for the individual garment components.

Body Texture Generation. Leveraging the unified UV layout of Anime-SMPL across all characters, our method facilitates texture generation in UV space. We employ a multiview diffusion model (Huang et al., 2024b) as the texture synthesis network, which semantically decomposes the body texture into six regions: body skin, facial skin, left eye, right eye, eyebrows and eyelashes. The diffusion model takes region-specific text prompts as input and is conditioned on the character image to generate the UV texture for each region. A representative result is shown in Figure 5.

Input Body Skin Face Skin Left Eye Right Eye Eyebrow Eyelash

Figure 5: An example of the body texture we generated.

Component-wise Garments Texture Generation. Given the meshes of all garment components, we initially attempt to use normal maps as input, and employ MVAdapter (Huang et al., 2024b),

conditioned on the canonical-pose character image, to synthesize multi-view images for each garment component. However, this approach results in color bleeding, where the appearance of each component is adversely affected by neighboring regions.

To address this issue, we design a component-wise high-resolution texture generation pipeline, as illustrated in Figure 4. Following Dong et al. (2025), we first train a diffusion model to decompose the full-body image into enlarged and independent views of individual garment components. Specifically, we use the normal maps of all garment components as input and condition the model on the canonical-pose character image. A multi-component self-attention mechanism is employed to facilitate information exchange across components, and the resulting features are fused with label and timestep embeddings via cross-attention. This design enables the generation of high-quality images for these garment components. These segmented images are subsequently fed into the MVAdapter to synthesize multiview renderings, which are finally projected onto the 3D surface to obtain the complete texture map for each component.

Thanks to our component-wise garment generation strategy, self-occlusion is significantly alleviated when projecting textures from six canonical views (front, back, left, right, top, and bottom). Moreover, generating textures for each component allows for higher-resolution allocation, allowing for the generation of high-quality texture maps.

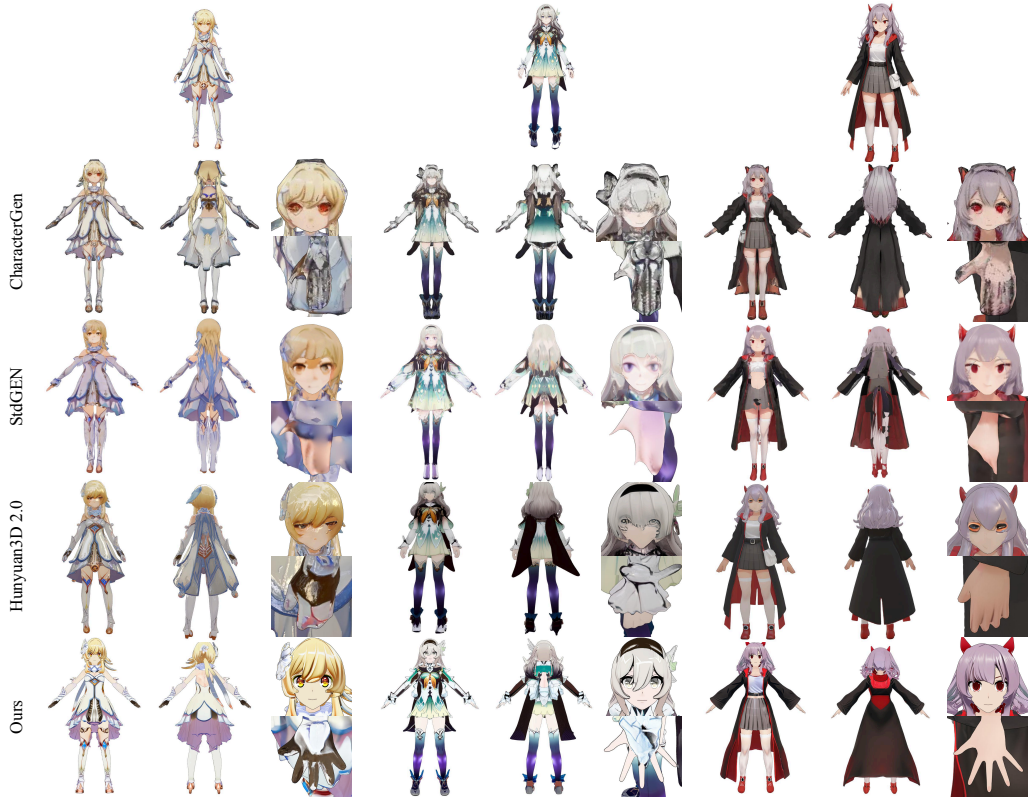

Figure 6: Qualitative comparisons between our method and previous state-of-the-art methods

4 Experiments

4.1 Implement Details

Training. For the Anime SMPL shape prediction network, training is conducted on a single NVIDIA L20 GPU and takes approximately 4 hours. For the MoE-structured Multi-Shape DiT, we use 16 NVIDIA A100 GPUs with AdamW optimization and a learning rate of 1×10^{-4}, which requires about 10 days of training. For the remaining modules, 2D Canonical Pose Character Generation, Body

Table 1: User study results. The score ranges from 1 to 5, where 5 is the best score.

Methods	Mesh Quality↑	Texture Quality↑	Fidelity↑
CharacterGen (Peng et al., 2024)	2.58	2.14	2.51
StdGEN (He et al., 2024b)	2.69	2.23	2.52
Huanyuan3D 2.0 (Zhao et al., 2025)	3.14	3.49	3.42
Ours	**3.83**	**3.75**	**3.74**

Texture Generation, and Component-wise Garment Texture Generation, each is trained independently on 8 NVIDIA A100 GPUs for approximately 2 days.

Ineference. The image generation stage takes 5 seconds, ANIME-SMPL parameter prediction takes 2 seconds, the MoE-Structured Multi-Shape DiT requires a total of 40 seconds, body texture generation takes 10 seconds, and garment texture generation takes 360 seconds.

Dataset. All models are trained on our private dataset, which consists of 20k anime characters aligned to a unified body template. Specifically, each character shares the same body mesh topology, including identical vertex count, vertex ordering, and face connectivity. For every character, garments are consistently divided into four components: hairstyle, upper garment, lower garment, and accessories. For characters missing any of these components, the corresponding part is left unassigned.

4.2 Comparisons with State-of-The-Art Methods

To ensure a fair comparison with state-of-the-art anime-style character generation methods, we evaluate our model using both front-view images of characters in arbitrary poses collected from the Internet, as well as additional synthesized samples with diverse poses. The qualitative results are presented in Figures 6 and 12. These results demonstrate that our method produces highly detailed and realistic facial and hand regions of the characters.

Since our method and the baselines are trained on different datasets, we note that CharacterGen and StdGen are trained on Anime3D (Peng et al., 2024), whereas Hunyuan3D 2.0 is trained on large-scale datasets including ObjaverseXL (Deitke et al., 2023). In contrast, our method is trained on a proprietary dataset. Consequently, reconstruction-based metrics such as PSNR, SSIM, and LPIPS are not directly comparable. Instead, we conduct a user study that better reflects the perceptual quality of the generated results rather than simply computing losses against the ground truth. Specifically, we randomly sample 16 anime-style characters with diverse poses from the Internet as well as from synthesized images, and generate corresponding 3D models. Thirty participants are then asked to assess the visual quality of the generated textures and meshes, as well as their fidelity to the input images. To mitigate potential bias, participants are instructed to focus solely on the perceptual quality of the generated avatars, regardless of pose variations. As summarized in Table 1, the results show that our method consistently outperforms all competing approaches.

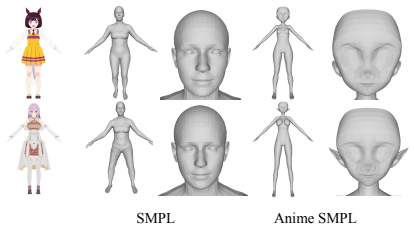

Figure 7: Comparison between SMPL and Anime-SMPL on Anime Character Mesh Reconstruction

4.3 Pose Canonicalization Comparison

The results of the pose canonicalization experiment are presented in Figure 13 (Appendix). Our method demonstrates superior generalization compared to prior approaches (Peng et al., 2024; He et al., 2024b; Zhao et al., 2025). We attribute this to several data augmentation strategies employed during training, including random cropping of the lower body, variations in lighting conditions, and adjustments to character contour line thickness.

4.4 Ablation Study

Anime-SMPL vs. SMPL. We compare the body mesh predicted by the SMPL model with that generated by our Anime-SMPL model. As shown in Figure 7, the mesh produced by our method aligns more closely with the input image in terms of ear shape, facial contour, and relative thigh and calf thickness. In contrast, the mesh predicted by the original SMPL model exhibits noticeable discrepancies in these areas. These results demonstrate that Anime-SMPL more effectively captures the distinctive geometric characteristics of anime-style characters.

Body-Aligned Garment Generation.

To verify the effectiveness of the proposed Body-Aligned Garment Generation strategy, we compare the garment-to-body alignment of models with and without body latent tokens, as illustrated in Figure 8. The results demonstrate that although our MoE-structured Multi-Shape DiT can roughly infer the spatial layout of each garment without explicit body geometry, the resulting garments often exhibit suboptimal alignment and increased susceptibility to interpenetration artifacts. In contrast, the incorporation of body latent tokens provides explicit geometric guidance, allowing the model to generate garments that better conform to the body surface. This effect is especially pronounced for tight-fitting garments, such as swimsuits, where the generated clothing aligns precisely with minimal interpenetration.

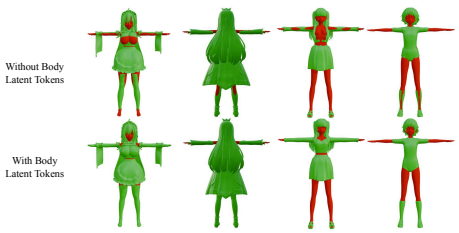

Figure 8: Comparison of Garment Fitting with and without Body Latent Tokens.

MoE-structured Multi-Shape DiT. To demonstrate the effectiveness of the MoE design in our MoE-structured Multi-Shape DiT, we conducted a comparative experiment evaluating the generation performance of our MULTI-SHAPE DiT with and without MoE layers. As illustrated in Figure 9, the left shows the input image, the middle displays the upper garments generation results from DiT without MoE layers, and the right presents the up garments generation results from DiT with MoE layers. Our MoE-structured Multi-Shape DiT significantly outperforms the variant without MoE layers in terms of both generation quality and image-geometry alignment, thereby validating the effectiveness of our proposed MoE layer design.

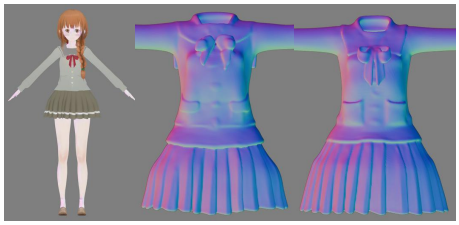

Figure 9: Comparison of the generation performance of our Multi-Shape DiT with and without MoE layers.

5 Application

Leveraging our Anime-SMPL model and the Body-Aligned Component-Wise Garment Modeling framework, the generated 3D anime characters can be directly applied to various downstream tasks, including garment retargeting, motion control, and facial expression control.

5.1 3D Character Garment Retargeting

Leveraging the unified body template of our Anime-SMPL model, we facilitate straightforward and effective garment retargeting across characters. For each vertex on the source garment, we find its five nearest anchor points on the canonical body, compute a weighted relative displacement and apply it to the corresponding anchors on the target character to accurately transfer garment geometry. Sample retargeting results are presented in Figure 17.

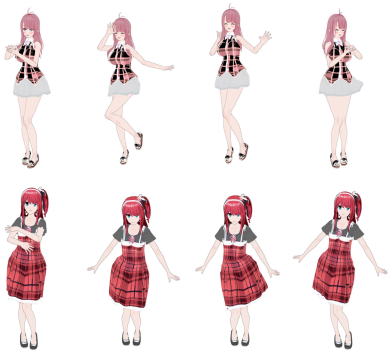

Figure 10: Animation results.

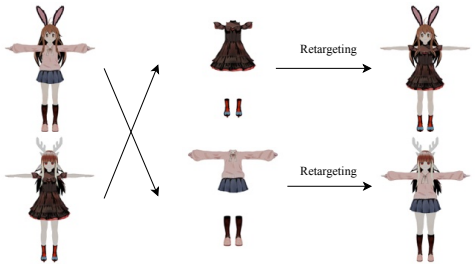

Figure 11: Garment Retargetting results.

5.2 Motion Control

Leveraging the precise skeletal structure and skinning weights of our Anime-SMPL model, high-quality motion control can be readily achieved. For garments without predefined skinning weights, we use a hybrid approach: body-hugging clothes inherit weights via nearest-neighbor sampling from the body, while garments with greater separation, such as skirts, are animated using physics-based simulation. Animation results are presented in Figure 1 and Figure 10. As illustrated in these figures, our generated 3D anime characters capture both natural full-body articulation and fine-grained movements, such as delicate finger gestures, thereby underscoring the precision and expressiveness enabled by our pipeline.

5.3 Facial Expression Control

Leveraging the unified body template provided by our Anime-SMPL model, facial expressions can be controlled by manipulating facial vertices using blend shapes. Representative facial expression control results are shown in Figure 1. Furthermore, real-time facial expression control can be achieved via a face tracker, as demonstrated in Figure 18.

6 Limitations and Discussion

Although our method enables the generation of 3D anime characters from either text prompts or a single image, several limitations remain. First, the pose canonicalization step in the 2D Canonical-Pose Character Generation stage struggles with characters exhibiting complex poses or multiple accessories, often resulting in distorted outputs. Second, garment meshes extracted from SDFs via the marching cubes algorithm are inherently double-sided, which can impair physical simulation. Additionally, our texture generation using a multi-view diffusion model suffers from misalignment between projected images and the underlying geometry, as well as cross-view inconsistencies, which negatively affect texture quality.

For future work, mesh generation approaches (Chen et al., 2024c; Hao et al., 2024) that operate directly on vertices and faces may help mitigate the double-surface artifact. Moreover, generating textures directly in 3D space could alleviate cross-view inconsistency and enhance overall coherence.

7 Conclusion

In this work, we present a novel pipeline for generating high-quality, animatable anime-style 3D characters. We decompose each character into a set of componentsincluding body, hair, upper garments, lower garments, and accessorieswhich are generated independently. By integrating Anime-SMPL, a MoE-structured Multi-Shape DiT, and a high-resolution texture generation framework, our pipeline enables fine-grained modeling of both geometry and appearance. This design allows the production of high-quality, fully textured, animation-ready anime-style 3D characters.

ETHICS STATEMENT

This work complies with the ICLR Code of Ethics. Our research focuses on 3D anime character generation from text and images. The study does not involve human subjects, biometric data, or any sensitive personal information. We acknowledge the potential risks of misuse of generative technologies, such as producing inappropriate or misleading content. To mitigate these concerns, our research is restricted to academic purposes, and we encourage responsible use of our models aligned with ethical guidelines in the community. No conflicts of interest, sponsorship biases, or legal compliance issues are associated with this work.

ACKNOWLEDGEMENTS

This work is supported by the Hong Kong Research Grant Council General Research Fund (No. 17213925), National Natural Science Foundation of China (No. 62422606) and Gusu Innovation & Entrepreneurship Leading Talents Program (ZXL2024361).

REFERENCES

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 958–968, 2024.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024a.

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22246–22256, 2023.

Rui Chen, Jianfeng Zhang, Yixun Liang, Guan Luo, Weiyu Li, Jiarui Liu, Xiu Li, Xiaoxiao Long, Jiashi Feng, and Ping Tan. Dora: Sampling and benchmarking for 3d shape variational auto-encoders. *arXiv preprint arXiv:2412.17808*, 2024b.

Yiwen Chen, Tong He, Di Huang, Weicai Ye, Sijin Chen, Jiaxiang Tang, Xin Chen, Zhongang Cai, Lei Yang, Gang Yu, et al. Meshanything: Artist-created mesh generation with autoregressive transformers. *arXiv preprint arXiv:2406.10163*, 2024c.

Ruikai Cui, Xibin Song, Weixuan Sun, Senbo Wang, Weizhe Liu, Shenzhou Chen, Taizhang Shang, Yang Li, Nick Barnes, Hongdong Li, et al. Lam3d: Large image-point clouds alignment model for 3d reconstruction from single image. *Advances in Neural Information Processing Systems*, 37: 4454–4480, 2024.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36:35799–35813, 2023.

Junting Dong, Qi Fang, Zehuan Huang, Xudong Xu, Jingbo Wang, Sida Peng, and Bo Dai. Tela: Text to layer-wise 3d clothed human generation. In *European Conference on Computer Vision*, pp. 19–36. Springer, 2024.

Junting Dong, Mingze Sun, Qi Fang, Yan-Pei Cao, Jingbo Wang, and Bo Dai. Clothing-disentangled 3d character generation from a single image. *OpenReview Preprint*, 2025.

Zijian Dong, Xu Chen, Jinlong Yang, Michael J Black, Otmar Hilliges, and Andreas Geiger. Ag3d: Learning to generate 3d avatars from 2d image collections. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 14916–14927, 2023.

Zekun Hao, David W Romero, Tsung-Yi Lin, and Ming-Yu Liu. Meshtron: High-fidelity, artist-like 3d mesh generation at scale. *arXiv preprint arXiv:2412.09548*, 2024.

Kai He, Kaixin Yao, Qixuan Zhang, Jingyi Yu, Lingjie Liu, and Lan Xu. Dresscode: Autoregressively sewing and generating garments from text guidance. *ACM Transactions on Graphics (TOG)*, 43 (4):1–13, 2024a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Yuze He, Yanning Zhou, Wang Zhao, Zhongkai Wu, Kaiwen Xiao, Wei Yang, Yong-Jin Liu, and Xiao Han. Stdgen: Semantic-decomposed 3d character generation from single images. *arXiv preprint arXiv:2411.05738*, 2024b.

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.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023.

Shuo Huang, Zongxin Yang, Liangting Li, Yi Yang, and Jia Jia. Avatarfusion: Zero-shot generation of clothing-decoupled 3d avatars using 2d diffusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 5734–5745, 2023a.

Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. Tech: Text-guided reconstruction of lifelike clothed humans. In *2024 International Conference on 3D Vision (3DV)*, pp. 1531–1542. IEEE, 2024a.

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. *Advances in Neural Information Processing Systems*, 36:4566–4584, 2023b.

Zehuan Huang, Yuan-Chen Guo, Haoran Wang, Ran Yi, Lizhuang Ma, Yan-Pei Cao, and Lu Sheng. Mv-adapter: Multi-view consistent image generation made easy. *arXiv preprint arXiv:2412.03632*, 2024b.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023.

Weiyu Li, Jiarui Liu, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. Craftsman: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. *arXiv preprint arXiv:2405.14979*, 2024.

Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, et al. Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *arXiv preprint arXiv:2502.06608*, 2025.

Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J Black. Tada! text to animatable digital avatars. In *2024 International Conference on 3D Vision (3DV)*, pp. 1508–1519. IEEE, 2024.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 300–309, 2023.

Lijuan Liu, Xiangyu Xu, Zhijie Lin, Jiabin Liang, and Shuicheng Yan. Towards garment sewing pattern reconstruction from a single image. *ACM Transactions on Graphics (TOG)*, 42(6):1–15, 2023.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9970–9980, 2024.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 2015.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 851–866. 2023.

Yuanxun Lu, Jingyang Zhang, Shiwei Li, Tian Fang, David McKinnon, Yanghai Tsin, Long Quan, Xun Cao, and Yao Yao. Direct2. 5: Diverse text-to-3d generation via multi-view 2.5 d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8744–8753, 2024.

Zhongjin Luo, Yang Li, Mingrui Zhang, Senbo Wang, Han Yan, Xibin Song, Taizhang Shang, Wei Mao, Hongdong Li, Xiaoguang Han, et al. Bag: Body-aligned 3d wearable asset generation. *arXiv preprint arXiv:2501.16177*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Panwang Pan, Zhuo Su, Chenguo Lin, Zhen Fan, Yongjie Zhang, Zeming Li, Tingting Shen, Yadong Mu, and Yebin Liu. Humansplat: Generalizable single-image human gaussian splatting with structure priors. *Advances in Neural Information Processing Systems*, 37:74383–74410, 2024.

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*, pp. 10975–10985, 2019.

Hao-Yang Peng, Jia-Peng Zhang, Meng-Hao Guo, Yan-Pei Cao, and Shi-Min Hu. Charactergen: Efficient 3d character generation from single images with multi-view pose canonicalization. *ACM Transactions on Graphics (TOG)*, 43(4):1–13, 2024.

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

Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023.

Jiachen Qian, Hongye Yang, Shuang Wu, Jingxi Xu, and Feihu Zhang. High-quality text-to-3d character generation with sparsecubes and sparse transformers. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

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

Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. Dreambooth3d: Subject-driven text-to-3d generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2349–2359, 2023.

Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4209–4219, 2024.

Nikolaos Sarafianos, Tuur Stuyck, Xiaoyu Xiang, Yilei Li, Jovan Popovic, and Rakesh Ranjan. Garment3dgen: 3d garment stylization and texture generation. *arXiv preprint arXiv:2403.18816*, 2024.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5693–5703, 2019.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2024.

Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. Textmesh: Generation of realistic 3d meshes from text prompts. In *2024 International Conference on 3D Vision (3DV)*, pp. 1554–1563. IEEE, 2024.

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*, pp. 12619–12629, 2023a.

Jionghao Wang, Yuan Liu, Zhiyang Dou, Zhengming Yu, Yongqing Liang, Cheng Lin, Rong Xie, Li Song, Xin Li, and Wenping Wang. Disentangled clothed avatar generation from text descriptions. In *European Conference on Computer Vision*, pp. 381–401. Springer, 2024a.

Yi Wang, Jian Ma, Ruizhi Shao, Qiao Feng, Yu-Kun Lai, and Kun Li. Humancoser: Layered 3d human generation via semantic-aware diffusion model. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 436–445. IEEE, 2024b.

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. *Advances in Neural Information Processing Systems*, 36:8406–8441, 2023b.

Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. In *European Conference on Computer Vision*, pp. 57–74. Springer, 2024c.

Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a.

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. *Advances in Neural Information Processing Systems*, 37:121859–121881, 2024b.

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.

Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 466–481, 2018.

Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. in 2022 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13286–13296, 2022.

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*, pp. 512–523, 2023.

Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024a.

Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. In *European Conference on Computer Vision*, pp. 1–20. Springer, 2024b.

Yifan Yang, Dong Liu, Shuhai Zhang, Zeshuai Deng, Zixiong Huang, and Mingkui Tan. Hilo: Detailed and robust 3d clothed human reconstruction with high-and low-frequency information of parametric models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10671–10681, 2024.

Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions On Graphics (TOG)*, 42(4):1–16, 2023.

Chubin Zhang, Hongliang Song, Yi Wei, Chen Yu, Jiwen Lu, and Yansong Tang. Geolrm: Geometry-aware large reconstruction model for high-quality 3d gaussian generation. *Advances in Neural Information Processing Systems*, 37:55761–55784, 2024a.

Jianfeng Zhang, Zihang Jiang, Dingdong Yang, Hongyi Xu, Yichun Shi, Guoxian Song, Zhongcong Xu, Xinchao Wang, and Jiashi Feng. Avatargen: a 3d generative model for animatable human avatars. In *European Conference on Computer Vision*, pp. 668–685. Springer, 2022.

Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision*, pp. 1–19. Springer, 2024b.

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 (TOG)*, 43(4):1–20, 2024c.

Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025.

A Appendix

Acknowledgements

We sincerely thank the engineering team at DreamTech, especially Jingxi Xu, for their valuable assistance and support throughout this project.

We acknowledge the use of Large Language Models for assistance with language editing in this manuscript. All technical content and results are the sole responsibility of the authors.

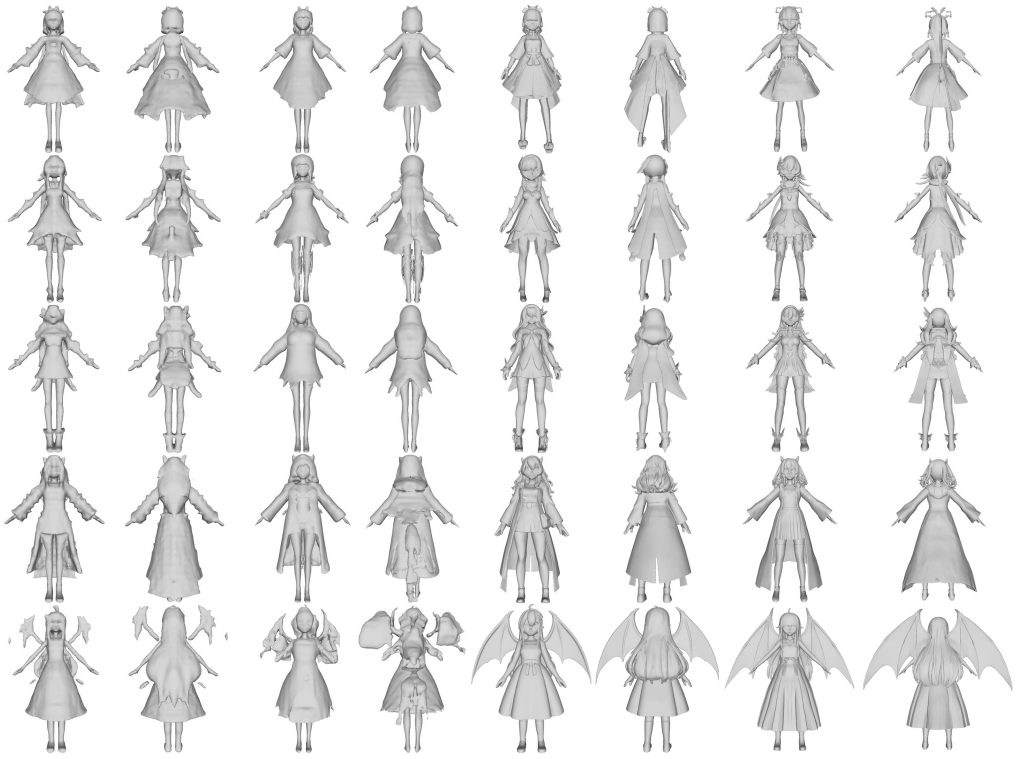

Figure 12: Mesh comparison between our method and other methods.

Figure 13: Pose Canonicalization Comparison between our method and other methods.

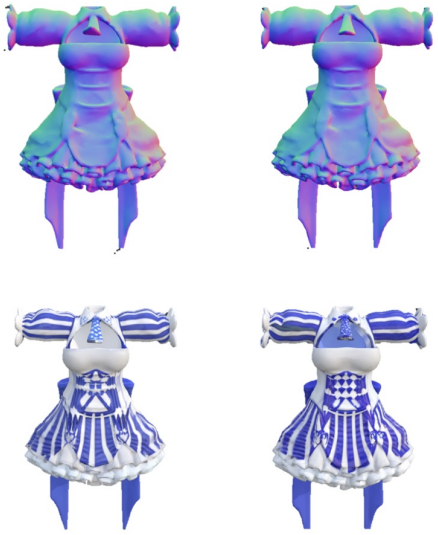

Figure 14: Since the Anime3D dataset does not have a unified body template, we can only train our component-wise garment model on it. As shown beyond, our method shows no significant difference in performance between the Anime3D dataset and our dataset.

Figure 15: Comparison with Rodin. The first row shows our results, and the second row shows the results generated by Rodin followed by auto-rigging using Mixamo. The first column presents the animation results, and the second column shows the distribution of skinning weights for the shoulder bones.

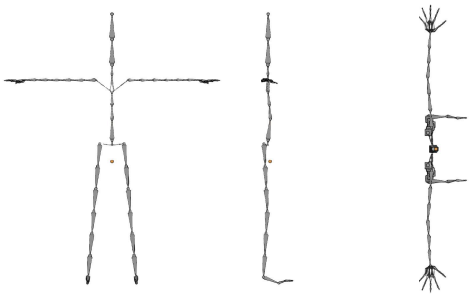

Figure 16: Visualization of our Anime-SMPL's joint structure.

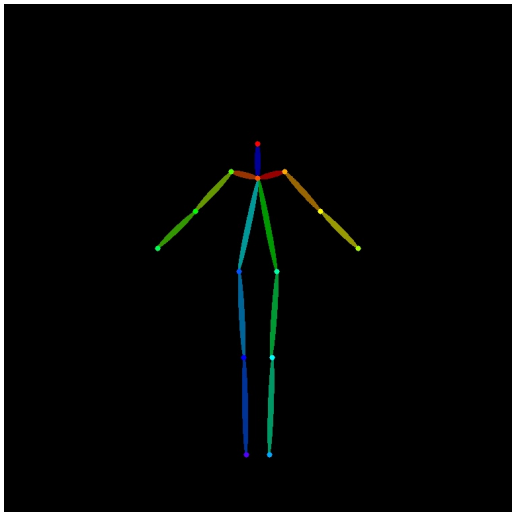

Figure 17: General a-pose skeleton image used in Image-to-Image Synthesis.

Figure 18: Real-Time facial expression control with face tracker.

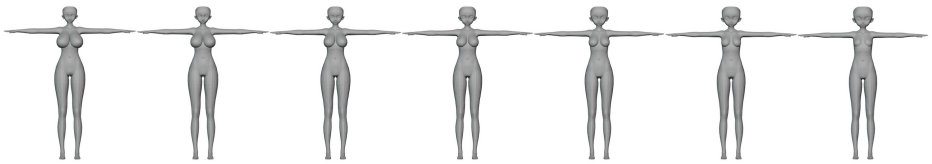

Figure 19: Visualization of the effect varying β, we vary the first parameter of β from -2 to 2 from left to right.

Figure 20: MVAdapter's Color Bleeding. The first row shows the input image on the far left and the result produced by MVAdapter on the right. In the second row, the far left displays the frontal view of the hair decomposed by our decompose model, while the right side shows the MV result obtained by using the decomposed image as input.

Figure 21: Our result on real humans.