CharacterGen: Efficient 3D Character Generation from Single Images with Multi-View Pose Calibration

ANONYMOUS AUTHOR(S)
SUBMISSION ID: 916

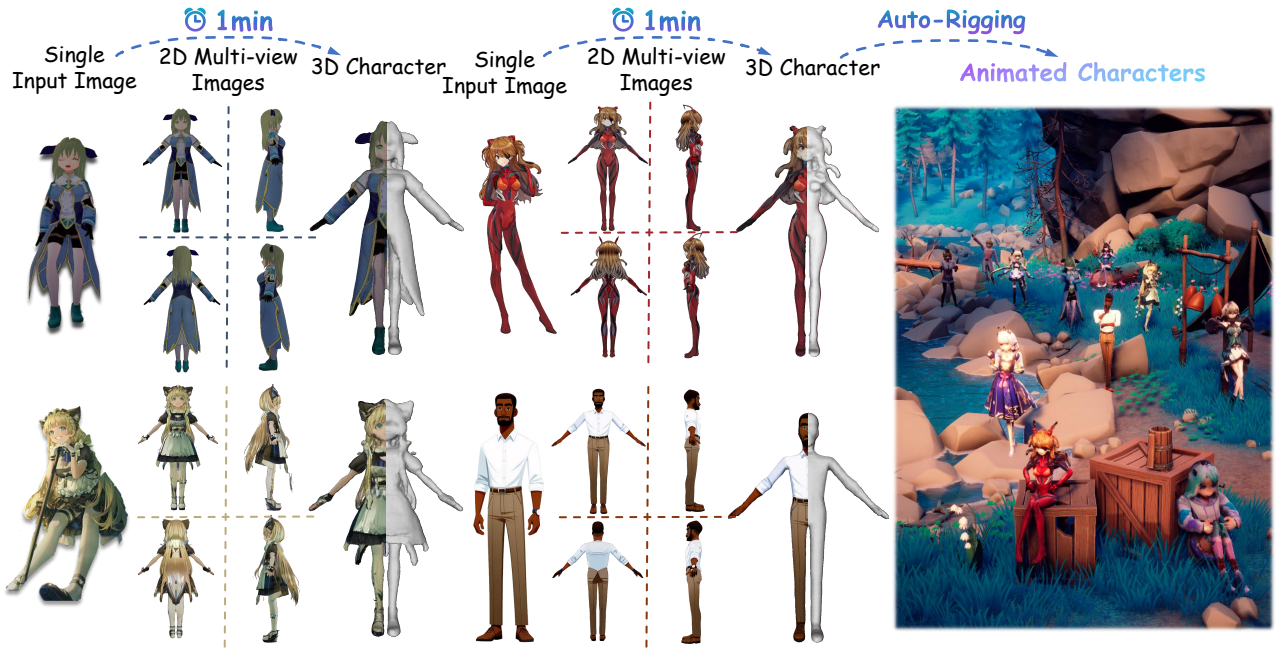

Fig. 1. In this paper we present CharacterGen, an efficient 3D character generation framework. CharacterGen takes a single input image and generates 3D pose-unified character meshes with high-quality and consistent appearance, which can be directly utilized in downstream rigging and animation workflows.

In the field of digital content creation, generating high-quality 3D characters from single images is challenging, especially given the complexities of various body poses and the issues of self-occlusion and pose ambiguity. In this paper, we present CharacterGen, a framework developed to efficiently generate 3D characters. CharacterGen introduces a streamlined generation pipeline along with an image-conditioned multi-view diffusion model. This model effectively calibrates input poses to a canonical form while retaining key attributes of the input image, thereby addressing the challenges posed by diverse poses. A transformer-based, generalizable sparse-view reconstruction model is the other core component of our approach, facilitating the creation of detailed 3D models from multi-view images. We also adopt a texture-back-projection strategy to produce high-quality texture map. Additionally, we have curated a dataset of anime characters, rendered in multiple poses and views, to train and evaluate our model. Our approach has been thoroughly evaluated through quantitative and qualitative experiments, showing its proficiency in generating 3D characters with high-quality shapes and textures, ready for downstream applications such as rigging and animation.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Shape modeling**; **Image manipulation**.

Additional Key Words and Phrases: Image-Driven Generation, 3D Avatar Generation, Avatar Pose Calibration

ACM Reference Format:
Anonymous Author(s). 2024. CharacterGen: Efficient 3D Character Generation from Single Images with Multi-View Pose Calibration. *ACM Trans. Graph.* 1, 1 (January 2024), 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM 0730-0301/2024/1-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The digital industry's rapid evolution has made the creation of high-quality 3D content a pivotal aspect across various domains, including film, video gaming, online streaming, and virtual reality (VR). Although manually modeled 3D content can attain exceptional quality, the significant time and labor investment required presents a substantial bottleneck. Addressing this, there has been a notable influx of exciting research [Qian et al. 2023; Wang and Shi 2023] focused on generating 3D models from single images. This approach

substantially lowers the barrier to entry for novice users, democratizing access to 3D content creation and potentially revolutionizing the field.

3D character models often feature complex articulations, leading to frequent self-occlusions in 2D images that pose significant challenges in reconstruction, generation, and animation. Moreover, these characters may assume a range of body poses, including some that are rare and challenging to accurately interpret, leading to a diverse yet imbalanced data domain. These complexities hinder the effective generation, rigging, and animating of such models. As a result, general 3D generation techniques [Chen et al. 2023a; Poole et al. 2023; Wang et al. 2023b] and single-view 3D reconstruction methods [Qian et al. 2023; Wang and Shi 2023] often fall short in delivering optimal outcomes. Prior research [Cao et al. 2023; Huang et al. 2023a; Kolotouros et al. 2023] has explored the use of parametric models of 3D human bodies [Alldieck et al. 2021; Loper et al. 2015] as 3D priors. However, these methods are predominantly tailored to realistic human proportions and relatively tight clothing, limiting their applicability. This constraint is especially noticeable in the context of stylized characters, known for their exaggerated body proportions and complex clothing designs, which challenge the adaptability and effectiveness of these approaches.

In this paper, we introduce CharacterGen, a new approach for pose-unified 3D character generation from single images. Our method stands out significantly from previous ones by allowing any input body pose in the input image and outputting a clean 3D character model. The foundational principle of CharacterGen hinges on simultaneously unifying body poses and producing consistent multi-view images during the generation process. This is achieved by transforming each pose into a canonical "A-pose", a stance widely utilized in 3D character modeling, while concurrently ensuring image consistency across multiple views. This dual approach effectively addresses the challenges of self-occlusion and ambiguous human poses, significantly streamlining subsequent reconstruction, rigging, and animation stages.

Our 3D character generation approach is structured into two tightly interconnected stages: initially, lifting a single image to multiple viewpoints while simultaneously calibrating the input pose to a canonical one; following this, we proceed to reconstruct 3D characters under this canonical pose. This method is supported by *two key insights*: *firstly*, it incorporates established principles and successful techniques from recent advancements in controllable image generation [Ye et al. 2023; Zhang and Agrawala 2023]; *secondly*, it aims to alleviate the challenges associated with sparse-view reconstruction for 3D characters. By focusing on the canonical pose, where the geometric and textural structures are more clearly defined and self-occlusion is minimized, our approach simplifies the task of reconstructing both geometry and texture from limited views. The first stage involves a diffusion-based, image-conditioned multi-view generation model [Liu et al. 2023a; Shi et al. 2023], adept at capturing and translating both the global and local features character features from the input image to the canonical pose, which further facilitates the generation of consistent canonical pose images across multiple views. Following this, in the second stage, we employ a transformer-based, generalizable sparse-view reconstruction model [Hong et al. 2023b]. This model is key to generating a coarse textured 3D character model from the images produced in the first stage. We further refine the model's texture resolution through projective texture mapping and differentiable rendering, achieving a detailed final model. Importantly, generating characters under a canonical pose also significantly benefits downstream applications, such as rigging and animation. Our whole generation process can be finished in 1 mintue.

To train our pipeline, we compiled a multi-pose and multi-view character dataset, focusing on anime characters due to their widespread availability online, notably on platforms such as VRoid Hub [1]. We amassed a collection of 13,746 characters and rendered these from various viewpoints across multiple body poses. This extensive collection has been organized into a dataset that we refer to as Anime3D.

In summary, our paper presents the following key contributions:

- We propose an image-conditioned diffusion model that effectively generates multi-view consistent images of characters in a controlled canonical pose from varying input poses. This approach addresses challenges such as self-occlusion and pose ambiguity.
- We introduce a streamlined pipeline combining our diffusion model for multi-view image generation and a transformer-based reconstruction model. This pipeline efficiently transforms single-view inputs into detailed 3D character models.
- A curated dataset of 13,746 anime characters, rendered in multiple poses and views, providing a diverse training and evaluation resource for our model and future research in 3D character generation.

2 RELATED WORKS

In this section, we mainly discuss related works on diffusion-based 3D object and avatar generation domains. Besides these work, CharacterGen also relies on transformer-based reconstruction models [Hong et al. 2023b; Li et al. 2023] for efficient 3D character generation.

2.1 Diffusion-Based 3D Object Generation

Diffusion methods have shown a strong ability to guide 3D object generation tasks in the past year. The pioneering work DreamFusion [Poole et al. 2023] and SJC [Wang et al. 2023a] introduces score distillation sampling (SDS) to provide gradient guidance from pre-trained 2D diffusion models to 3D NeRF training. Magic3D [Lin et al. 2023a] and Fantasia3D [Chen et al. 2023a] propose to utilizes implicit tetrahedral field to support rendering with high resolution in the refinement stage. ProlificDreamer [Wang et al. 2023b] proposes VSD to distill gradient scores from a LoRA network to better learn the distribution of 3D objects. Zero-123 [Liu et al. 2023b] presents a novel diffusion model to generate multi-view images that are conform to the input image with given camera poses. Magic123 [Qian et al. 2023] combines both SDS and Zero-123 guidance in generating 3D objects with image prompts and adopts reconstruction loss to enhance front-view texture quality. MVDreamer [Shi et al. 2023] and ImageDream [Wang and Shi 2023] utilize multi-view diffusion models to provide highly consistent guidance in 3D object generation process. SyncDreamer [Liu et al. 2023a] utilizes 3D-aware

[1]

attention modules to achieve synchronized multi-view image generation. Some other works [Erkoç et al. 2023; Hui et al. 2022; Jun and Nichol 2023; Müller et al. 2023; Nichol et al. 2022; Zeng et al. 2022; Zhang et al. 2023b] employ 3D data to train diffusion models for direct 3D object generation, which can achieve fast generation speed but is struggled with generation quality.

2.2 3D Avatar Generation

With strong human-body priors like SMPL [Loper et al. 2015] and SMPL-X [Pavlakos et al. 2019], it is possible to generate high quality human avatars based on the general 3D generation methods. EVA3D [Hong et al. 2023a] utilizes a GAN backbone along with a pose-guided sampling method to generate high quality 3D human avatars. AvatarCLIP [Hong et al. 2022] first solves text-to-human generation by utilizing pre-trained CLIP model to provide guidance in training the geometry and color networks of implicit NeuS field. Dreamavatar [Cao et al. 2023] and AvatarCraft [Jiang et al. 2023] utilize SMPL to initialize the implicit human geometry in the diffusion-guided generation process. DreamHuman [Kolotouros et al. 2023] adopts ImGHum [Alldieck et al. 2021] as body priors and proposes focus rendering mechanism to better reconstruct the detail geometry of generated avatars. DreamWaltz [Huang et al. 2023a] first generates static A-pose avatar directly and then utilizes ControlNet [Zhang and Agrawala 2023] to provide pose guidance to finetuning the animatable representation. AvatarVerse [Zhang et al. 2023a] and AvatarStudio [Zhang et al. 2023c] all utilize a DensePose-guided ControlNet in the generation process to circumvent Janus problem and to support part geometry optimization. TeCH [Huang et al. 2023b] support image-prompt powered avatar generation by training an additional DreamBooth model [Ruiz et al. 2023] on the input image as the guidance model of SDS. TADA [Liao et al. 2023] directly distill scores from diffusion models to optimize the normals and displacement of SMPL body mesh.

Most of the aforementioned methods mainly focus on text-to-3D character generation and lacks the ability to utilize image prompt, which is necessary for stylized character generation. The DreamBooth-based methods will face severe Janus problem due to the strong front-view priors caused by overfitting the diffusion model on the single input image.

3 METHOD

In this section, we will demonstrate the whole framework of our CharacterGen. The main target of CharacterGen is to efficiently generate A-pose 3D characters from ambiguous posed images to aid downstream applications such as rigging and animation. In Sec. 3.1, We introduce our Anime3D dataset to show how we organize our data to extend diffusion's ability in 3D spatial understanding and character pose calibration. Then we illustrate how CharacterGen generates high-consistent multi-view pose-unified character images in Sec. 3.2. Finally we show our efficient 3D reconstruction pipeline in Sec. 3.3.

3.1 Anime3D Dataset

To further improve diffusion models' ability to understand 3D characters and to alleviate Janus problem, we present the Anime3D dataset with 13,746 subjects. This dataset has been carefully curated through manual selection to exclude anomalous models.

Fig. 2. We sample a character from our Anime3D dataset with four different camera poses to demonstrate how we organize the image pairs in the training process to extend UNet's ability in canonical pose calibration.

3.1.1 Data Acquisition. Existing large 3D object datasets like Objaverse [Deitke et al. 2023] or OmniObject3D [Wu et al. 2023] do not contain enough 3D stylized character objects for our training purposes. Inspired by PAniC3D [Chen et al. 2023b], we first collect a large dataset of nearly 14500 anime characters from the VRoidHub [VRoid 2022] and then filter non-human-like data. After filtering, it remains 13,746 character models.

3.1.2 Data Processing. We need to render all the objects into image format to finetune 2D diffusion models. Then, we utilize three-vrm [Pixiv 2019] framework to render the posed characters. We obtain A-pose characters and posed characters to generate canonical pose and random pose image pairs. The details of pose settings can be found in our Appendix. We normalize the coordinates of the character model to $[-0.5, 0.5]^3$ and render the images with ambient light and directional light.

In the training process, we propose to use four A-pose images and a single posed image as a pair because four images of orthographic views already contains sufficient appearance information of a 3D character. Therefore it is natural to render all objects with azimuth angles of $\{0°, 90°, 180°, 270°\}$ and an elevation angle of $0°$. To enhance the model's understanding of spatial body layout, we also render three additional groups with random initial azimuth, as depicted in Fig. 2. We also render 4 additional views with totally random azimuth and elevation to finetune the generalizable reconstruction model (see Sec. 3.3).

3.2 Multi-view Image Generation and Pose Calibration

In this stage, we will demonstrate how to generate high-consistent multi-view images with the given character image. The whole framework is shown in Fig. 3. We design IDUNet to transfer patch-level appearance features from the input image to Multi-view denoising UNet. We also introduce a pose embedding network to provide

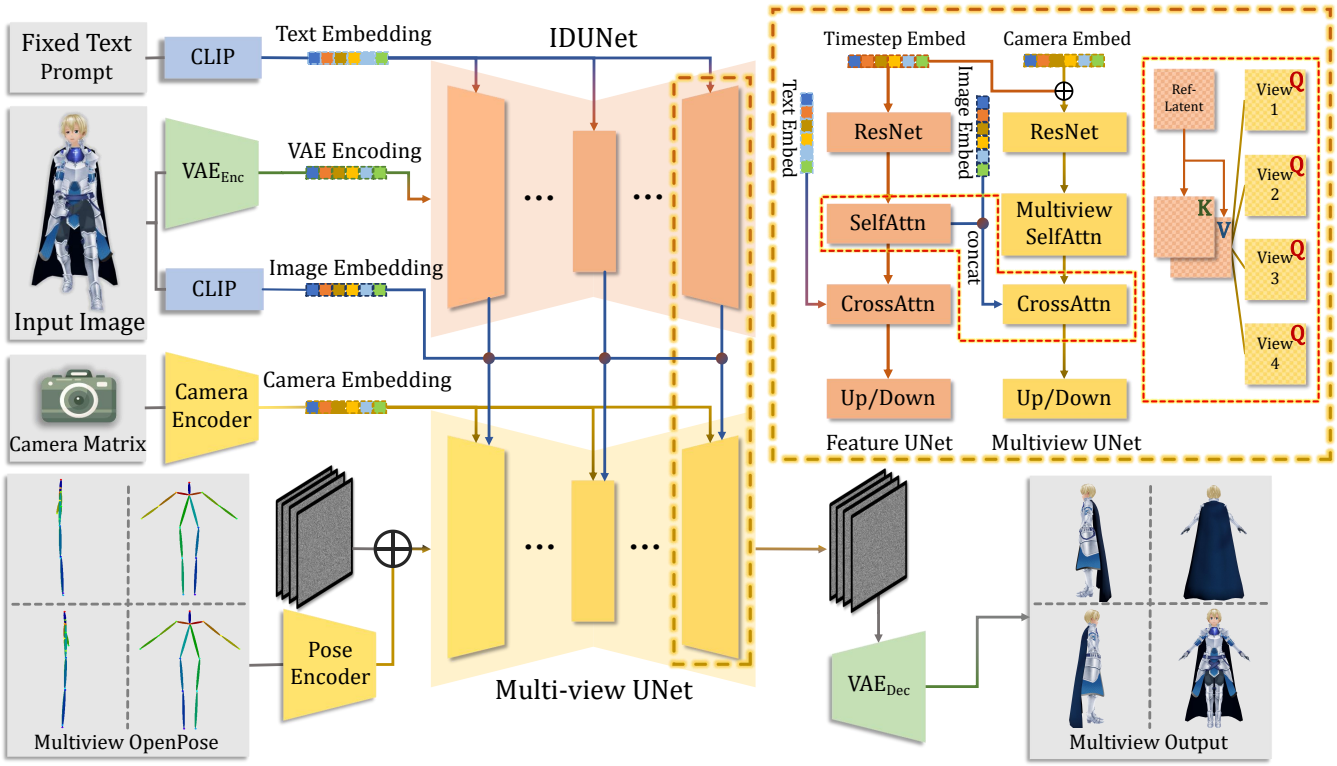

Fig. 3. We show the whole pipeline of how we generate four views of consistent images. We also demonstrate the details of how IDUNet extract local pixel-level features to strengthen the multi-view UNet.

more character layout information to further aid the canonical pose calibration task.

3.2.1 IDUNet. The design of IDUNet aims to retain as many features as possible from the original posed image and ensure high consistency across generated four views. Previous work IP-Adapter [Ye et al. 2023] adds adapter modules in diffusion UNet strcuture. The appearance information of input images can be transferred to generated images via the cross-attention mechanism between input condition image features and latent features. However, in practice, we observe IP-Adapter cannot fully capture the detailed texture of input images. Because IP-Adapter-like methods only utilize the global CLIP embeddings of the condition image, which loses pixel-level details during the image encoding process and lead to dissimilar results.

To better incorporate features of the condition image, we propose IDUNet to introduce pixel-level guidance in the generation process. Inspired by ControlNet [Zhang and Agrawala 2023], the structure of IDUNet is identical to the base diffusion model (Multiview UNet). Different from ControlNet, to ensure local patch-level interaction between all patches in both the denoising image and condition image, we leverage the cross-attention between the latent tokens and condition image tokens rather than merely adding them together.

It is worth noting that the IDUNet is used to provide pixel-level features into the denoising UNet, and adding noise to the input condition image will severely diminish the detail texture of 3D characters. In contrast, traditional denoising UNet is applied on a noisy image to predict noise according to the timesteps. Thus, we directly adopt VAE to encode the non-noise input image.

3.2.2 Multi-view UNet. The target of multi-view UNet is to generate high appearance consistency multi-view A-pose images with single posed input image. Within the Multi-view UNet, we simultaneously apply the denoising process on the four-view latents $x_{4v} \in \mathrm{R}^{B\times 4\times N\times D}$. The transformer block in our Multi-view UNet consists of a spatial self-attention module and a cross-attention module. Like MVDream [Shi et al. 2023], the spatial self-attention module directly takes tokens of all four image latents x_{4v} and corresponding camera view embeddings. In the spatial self-attention layer, x_{4v} is reshaped into $(B, 4N, D)$ for patch-level cross-view interaction. This design allows the denoising UNet to capture the global relationships across different views, ensuring image generation with high consistency.

In each following cross-attention module, the condition features come from IDUNet f_{ID} will be concatenated with CLIP-encoded image features f_{CLIP}. Then, cross-attention layer will be applied to introduce patch-level interactions into x_{4v}. The whole process is shown in Eq. 1 and is visualized in the right above part of Fig. 3.

$$f_{\mathrm{Cond}} = \mathrm{concat}(f_{\mathrm{ID}}, f_{\mathrm{CLIP}}) \tag{1}$$

$$x_{4v} = \mathrm{Cross\_Attn}(x_{4v}, f_{\mathrm{Cond}}) \tag{2}$$

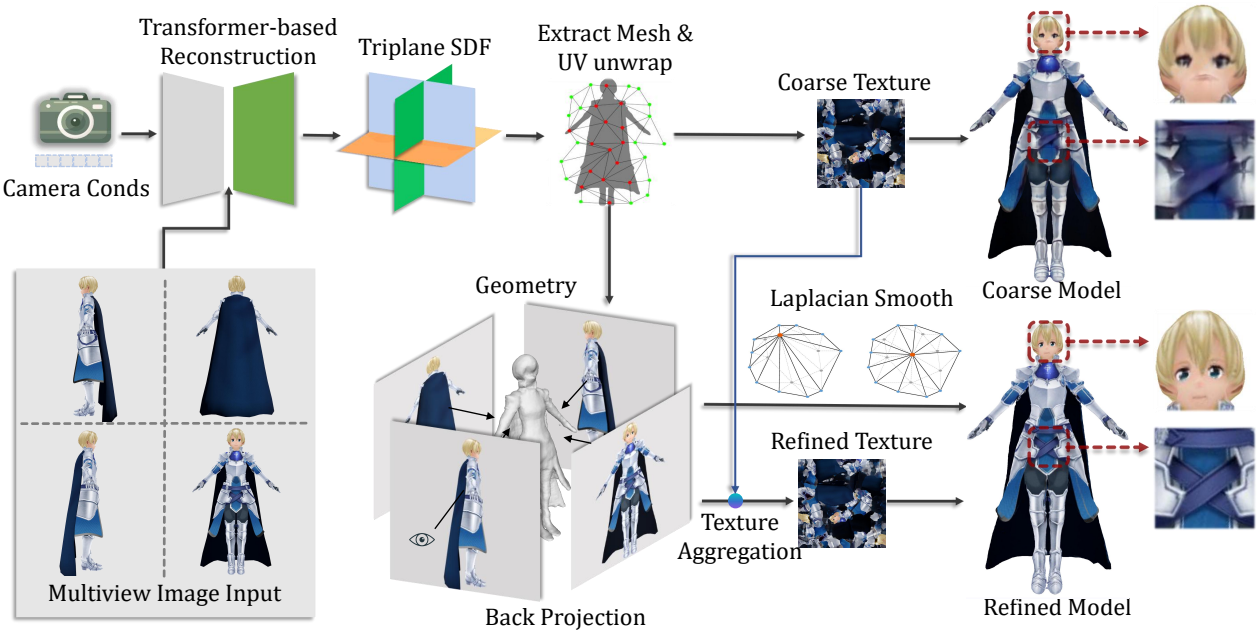

Fig. 4. We show the pipeline on how we generate final refined character meshes from generated multi-view images. In the first stage, we utilize a deep transformer-based network to generate an character with a coarse texture and then adopt our texture back-projection strategy to enhance the appearance of the generated mesh.

Following the previous study [Lin et al. 2023b] on the diffusion process, applying zero sample-noise-ratio (SNR) at the last timestamp T in the training steps will improve final generation quality because the input noise in the inference stage is pure Gaussian noise. To achieve zero-SNR in the training stage, we manually set SNR_T to zero and linearly scale other Guassian distribution parmaeters β_s. We set the UNet to directly output velocity $v_{pred} \in \mathrm{R}^{(B,4,N,D)}$ of the multi-view latents and convert it to the noise ϵ_{pred}. The final optimization target is shown in Eq. 3.

$$L_{4v} = ||\epsilon_{4v} - \epsilon_{pred}||_2^2. \tag{3}$$

Here ϵ_{4v} is the noise added to the multi-view A-pose images in the diffusion forward pass. We employ Stable Diffusion 2.1 as the base model to trade off generation quality and memory overheads.

3.2.3 Pose Calibration. Combined with IDUNet, our Multi-view diffusion model can successfully generate high-consistent orthographic view images while maintaining sufficient features of the prompt image. To enable the diffusion model to achieve character pose calibration in the generation process, we jointly train the two UNets with the image pairs from our Anime3D dataset. However, simply training the diffusion network without extra pose constraints will lead to character layout misplacement and emergence of unrelated body parts. To tackle these problems, we introduce character layouts to the diffusion models for generating A-pose character images via adopting OpenPose [Cao et al. 2017] to predict pose embedding as an additional condition. The generated embedding are directly added to the latent noise to aid diffusion learn the relationships between character joints and generated character layout. In the inference stage, we leverage three sets of OpenPose images from our Anime3D dataset and select the one with the highest CLIP score as input pose condition.

3.3 3D Character Generation

In this section, we will introduce how to efficiently generate 3D characters with four-view images generated from our multi-view pose calibration diffusion model. As shown in Fig. 4, we adopt a coarse-to-fine process for the 3D character generation task. We first utilize a two-stage transformer-based network to reconstruct geometry and coarse appearance of the character following the design of LRM [Hong et al. 2023b]. Subsequently, we employ a texture back-projection strategy to quickly improve the texture quality with generated high-resolution four-view images. Finally we utilize differentiable rendering techniques [Laine et al. 2020] to further refine seams on the texture map.

3.3.1 Character Reconstruction with Coarse Texture. Inspired by LRM [Hong et al. 2023b], We utilize a deep transformer network to efficiently reconstruct characters from the four-view images generated by the multi-view diffusion model. While LRM is trained using the Objaverse dataset [Deitke et al. 2023] to achieve versatile 3D object generation, it does not sufficiently capture the intricacies of human character layouts. To retain reconstruction network's ability in processing both general 3D objects and stylized characters, we initially pre-train our transformer network on the Objaverse dataset. Then we finetune the model with our Anime3D dataset to introduce more priors of human body structure.

Original LRM proposed to mainly train with NeRF [Mildenhall et al. 2022] reprentation. However, directly extracting geometry from NeRF often yields noisy surface geometry, which can be problematic for the subsequent use of character meshes in downstream graphics pipelines. Alternatively, we utilize a two-stage finetuning strategy for our reconstruction network. The first stage involves using a triplane NeRF representation similar to LRM, to establish the character's coarse geometry and appearance. In the second stage, we modify the decoder module of our reconstruction network to predict SDF rather then density fields, which allows CharacterGen to achieve smoother and more precise surface geometry.

In addition to MSE loss, we also incorporate mask loss and LPIPS loss [Zhang et al. 2018] to supervise reconstruction appearance. Mask loss is adopted on the rendering opacity to accelerate the convergence speed and LPIPS loss is used to extract perceptual information of input images. The final training target can be organized in Eq. 4.

$$L_{recon} = \lambda_1 L_{mse} + \lambda_2 L_{mask} + \lambda_3 L_{LPIPS} \tag{4}$$

Here, λ_1, λ_2 and λ_3 are hyperparameters, which are set as 1, 0.1 and 0.5 by default.

3.3.2 3D Character Refinement. Our reconstruction network can rapidly reconstruct the 3D implicit representation of an character and we can extract the final mesh along with a coarse UV map from the reconstructed tri-plane with the DMTet [Shen et al. 2021]. However, the generated mesh still lacks texture details because the DMTet-based extraction will lose appearance information during the UV unwrapping process. To tackle this problem, we propose to further utilize the generated four-view images to improve the quality of the generated texture maps with differentiable rendering techniques. For efficient rasterization in this step, we employ NvDiffRast [Laine et al. 2020] as the renderer. We find that directly optimizing all pixels of the generated texture map like NvDiffRec [Munkberg et al. 2022] with generated sparse input views will cause a texture downsampling problem, resulting in severe degradation of the output refined texture map. To circumvent this problem, we project the four-view images into texels in the texture space and employ a depth test to filter out occluded texels. To mitigate noise along the character's silhouette, we implement a normal-based filtering approach. Then we directly optimize the texture without projected texels with MSE loss to alleviate seams in final textures.

4 EXPERIMENTS

4.1 Data Preparation

For the training stage, we first divide our Anime3D dataset into a train set and a test set following a 0.98/0.02 split ratio. In the inference stage, we not only use our test set but also incorporate images from the Internet. Please refer to the supplementary materials for more implementation details of data processing and model training.

4.2 Results and Comparison

We conduct experiments on both 2D multi-view character image generation and 3D character mesh generation to evaluate the efficiency of our CharacterGen. Additionally, we conduct user study on our generated characters and list the results in the appendix.

Table 1. We show the quantitative metrics of all methods on the test split to evaluate the efficiency of our CharacterGen .

Methods	SSIM↑	LPIPS↓	FID↓
CharacterGen	0.833	0.087	0.017
Zero123 [Liu et al. 2023b]	0.768	0.224	0.142
SyncDreamer [Liu et al. 2023a]	0.807	0.194	0.396

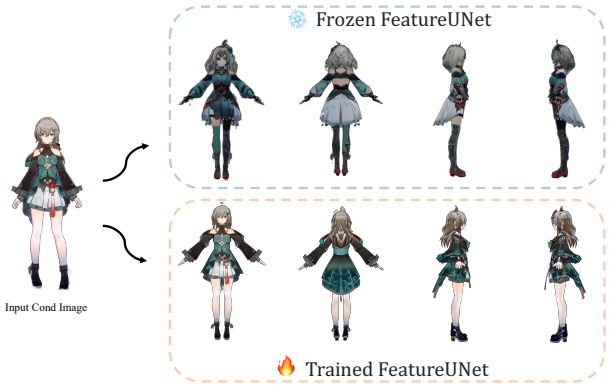

Fig. 5. Frozen IDUNet cannot extract enough appearance information from the prompt image and generates dissimilar images.

4.2.1 2D Multi-view Generation. We test our models on images from the test split of Anime3D as well as online sources and compare our results with Zero-123 [Liu et al. 2023b] and SyncDreamer [Liu et al. 2023a]. The comparison results are shown in Fig. 8. It can be observed that given some difficult body poses, Zero-123 and SyncDreamer struggle to preserve enough geometry and appearance information of generated images. Our CharacterGen adeptly performs canonical pose calibration and generates consistent character images across four views, which significantly enhances the subsequent character mesh reconstruction process.

We also conduct experiments on all character images from the test split of Anime3D and show the quality metrics in Tab. 1. Our calibrated A-pose images are benchmarked against ground-truth A-pose images, whereas images generated by other methods are compared with corresponding posed images. The results evaluate CharacterGen's superior generation quality and its consistency with the multi-view diffusion model.

4.2.2 3D Character Generation. In this section, we compare our generated 3D characters with image-prompt 3D character generation methods. ImageDream [Wang and Shi 2023] and Magic123 [Qian et al. 2023] all utilize SDS-based optimization framework. TeCH [Huang et al. 2023b] extends ECON [Xiu et al. 2023] and utilizes DreamBooth [Ruiz et al. 2023] to achieve image-prompt generation. We visualize the results in Fig. 9.

It can be observed that our CharacterGen effectively circumvents Janus problem thanks to our robust four-view reconstruction mechanism. Our generated 3D character meshes also exhibit satisfactory appearance on unseen body parts, with resourceful back-view and

Table 2. Time cost of generating a single 3D character. Models loading time is excluded for all methods.

Methods	Time
CharacterGen	1min
Magic123 [Qian et al. 2023]	70min
ImageDream [Wang and Shi 2023]	45min
TeCH [Huang et al. 2023b]	270min

side-view priors from our Anime3D. Most of 3D characters generated by other methods suffer from several mesh face cohesion problems, which makes it extremely hard to rig and animate these characters. CharacterGen can successfully generate canonical pose meshes for characters with ambiguous poses, which facilitates downstream graphic applications. We also evaluate other methods using A-pose character images calibrated by AnimateAnyone [Hu et al. 2023]. Please refer to the appendix for more details.

4.2.3 Generation Speed. We benchmark the time cost of generating a single 3D character against other image-prompted 3D generation methods, with the comparative results detailed in Tab. 2. Our method demonstrates a significant speed improvement, achieving 3D character generation up to 10 times faster than other alternatives.

4.3 Ablation Study

4.3.1 IDUNet. To demonstrate the importance of jointly training the IDUNet, we train CharacterGen network and freeze IDUNet with pre-trained diffusion models in the training phase. The generated four-view images are shown in Fig. 5. The results reveal that the generated images cannot preserve enough features of the input images, resulting in reduced similarity. This shows the necessity of jointly finetuning the IDUNet with non-noisy posed images to enhance its ability to extract detailed clothes and facial appearance.

4.3.2 Pose embedding network. Pose embedding network plays a crucial role to keep character layouts in the generated four-view images. We generate additional sets of images without pose embedding network and display the results in Fig. 6. It can be observed that the generated character images may not be located in the middle of the image in the absence of the pose embedding network. Furthermore, the lack of layout guidance can lead to the generation of extraneous clothing parts, which could hazard 3D reconstruction in the subsequent step.

4.4 Applications

CharacterGen can generate A-pose 3D characters with a refined texture maps, thereby simplifying the subsequent rigging process. We employ AccuRig [actorcore 2023] to automatically rig our generated character meshes. The rigged 3D characters can be readily utilized as 3D assets across various domains. We render the animated rigged models in Warudo [HakuyaLabs 2023] and present some results in Fig. 7.

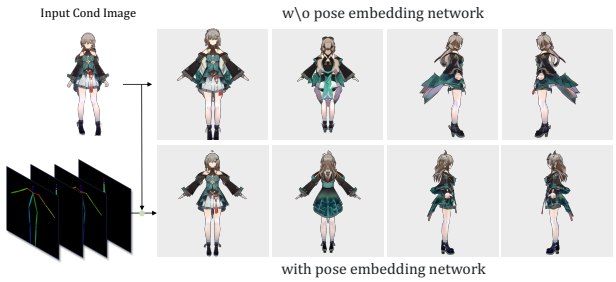

Fig. 6. We show that without the pose embedding network, the generated characters may be misplaced.

Fig. 7. We rig generated characters and utilize them as 3D assets in downstream applications.

5 DISCUSSIONS AND LIMITATIONS

While our method can generate 3D characters from a single input image with arbitrary poses, there still exist certain limitations. For the four-view A-pose image generation step, our method may not retain enough information when the character is in an extreme pose or is rendered from a rear view.

As for future works, the integration of additional non-photorealistic rendering (NPR) techniques into the texture refinement stage may further enhance the texture quality of the generated characters. Moreover, leveraging our trained Multi-view UNet structure, it is possible to combine SDS optimization method to achieve 3D character generation with superior geometry quality.

6 CONCLUSION

In this paper, we propose CharacterGen, a novel and efficient image-prompt character generation framework. We compile a new multi-pose, stylized character dataset Anime3D to train our pipeline. Our designs include IDUNet, which extracts patch-level features from the input condition image to guide multi-view A-pose character image generation. Subsequently, we utilize a transformer-based network to reconstruct 3D character meshes and propose to utilize texture back-projection refinement strategy to further improve the appearance of reconstructed character meshes. Experiments demonstrate that CharacterGen can generate high-quality 3D characters suitable for multiple downstream applications.

REFERENCES

actorcore. 2023. *accurig, a software for automatic character rigging.* https://actorcore.reallusion.com/auto-rig/accurig

Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. 2021. imGHUM: Implicit Generative Models of 3D Human Shape and Articulated Pose. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 5441–5450. https://doi.org/10.1109/ICCV48922.2021.00541

Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K. Wong. 2023. DreamAvatar: Text-and-Shape Guided 3D Human Avatar Generation via Diffusion Models. *CoRR* abs/2304.00916 (2023). https://doi.org/10.48550/ARXIV.2304.00916 arXiv:2304.00916

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 1302–1310. https://doi.org/10.1109/CVPR.2017.143

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. 2023a. Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation. *CoRR* abs/2303.13873 (2023). https://doi.org/10.48550/ARXIV.2303.13873 arXiv:2303.13873

Shuhong Chen, Kevin Zhang, Yichun Shi, Heng Wang, Yiheng Zhu, Guoxian Song, Sizhe An, Janus Kristjansson, Xiao Yang, and Matthias Zwicker. 2023b. PAniC-3D: Stylized Single-view 3D Reconstruction from Portraits of Anime Characters. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 21068–21077. https://doi.org/10.1109/CVPR52729.2023.02018

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023. Objaverse: A Universe of Annotated 3D Objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 13142–13153. https://doi.org/10.1109/CVPR52729.2023.01263

Ziya Erkoç, Fangchang Ma, Qi Shan, Matthias Nießner, and Angela Dai. 2023. HyperDiffusion: Generating Implicit Neural Fields with Weight-Space Diffusion. *CoRR* abs/2303.17015 (2023). https://doi.org/10.48550/ARXIV.2303.17015 arXiv:2303.17015

HakuyaLabs. 2023. *warudo, a 3D virtual image live broadcast software.* https://warudo.app/

Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. 2023a. EVA3D: Compositional 3D Human Generation from 2D Image Collections. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/pdf?id=g7U9jD_2CUr

Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. 2022. AvatarCLIP: zero-shot text-driven generation and animation of 3D avatars. *ACM Trans. Graph.* 41, 4 (2022), 161:1–161:19. https://doi.org/10.1145/3528223.3530094

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. 2023b. LRM: Large Reconstruction Model for Single Image to 3D. *CoRR* abs/2311.04400 (2023). https://doi.org/10.48550/ARXIV.2311.04400 arXiv:2311.04400

Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. 2023. Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation. *CoRR* abs/2311.17117 (2023). https://doi.org/10.48550/ARXIV.2311.17117 arXiv:2311.17117

Yukun Huang, Jianan Wang, Ailing Zeng, He Cao, Xianbiao Qi, Yukai Shi, Zheng-Jun Zha, and Lei Zhang. 2023a. DreamWaltz: Make a Scene with Complex 3D Animatable Avatars. *CoRR* abs/2305.12529 (2023). https://doi.org/10.48550/ARXIV.2305.12529 arXiv:2305.12529

Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. 2023b. TeCH: Text-guided Reconstruction of Lifelike Clothed Humans. *CoRR* abs/2308.08545 (2023). https://doi.org/10.48550/ARXIV.2308.08545 arXiv:2308.08545

Ka-Hei Hui, Ruihui Li, Jingyu Hu, and Chi-Wing Fu. 2022. Neural Wavelet-domain Diffusion for 3D Shape Generation. In *SIGGRAPH Asia 2022 Conference Papers, SA 2022, Daegu, Republic of Korea, December 6-9, 2022*, Soon Ki Jung, Jehee Lee, and Adam W. Bargteil (Eds.). ACM, 24:1–24:9. https://doi.org/10.1145/3550469.3555394

Ruixiang Jiang, Can Wang, Jingbo Zhang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. 2023. AvatarCraft: Transforming Text into Neural Human Avatars with Parameterized Shape and Pose Control. *CoRR* abs/2303.17606 (2023). https://doi.org/10.48550/ARXIV.2303.17606 arXiv:2303.17606

Heewoo Jun and Alex Nichol. 2023. Shap-E: Generating Conditional 3D Implicit Functions. *CoRR* abs/2305.02463 (2023). https://doi.org/10.48550/ARXIV.2305.02463 arXiv:2305.02463

Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. 2023. DreamHuman: Animatable 3D Avatars from Text. *CoRR* abs/2306.09329 (2023). https://doi.org/10.48550/ARXIV.2306.09329 arXiv:2306.09329

Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. 2020. Modular primitives for high-performance differentiable rendering. *ACM Trans. Graph.* 39, 6 (2020), 194:1–194:14. https://doi.org/10.1145/3414685.3417861

Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. 2023. Instant3D: Fast Text-to-3D with Sparse-View Generation and Large Reconstruction Model. *CoRR* abs/2311.06214 (2023). https://doi.org/10.48550/ARXIV.2311.06214 arXiv:2311.06214

Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J. Black. 2023. TADA! Text to Animatable Digital Avatars. *CoRR* abs/2308.10899 (2023). https://doi.org/10.48550/ARXIV.2308.10899 arXiv:2308.10899

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. 2023a. Magic3D: High-Resolution Text-to-3D Content Creation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 300–309. https://doi.org/10.1109/CVPR52729.2023.00037

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. 2023b. Common Diffusion Noise Schedules and Sample Steps are Flawed. *CoRR* abs/2305.08891 (2023). https://doi.org/10.48550/ARXIV.2305.08891 arXiv:2305.08891

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023b. Zero-1-to-3: Zero-shot One Image to 3D Object. *CoRR* abs/2303.11328 (2023). https://doi.org/10.48550/ARXIV.2303.11328 arXiv:2303.11328

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. 2023a. SyncDreamer: Generating Multiview-consistent Images from a Single-view Image. *CoRR* abs/2309.03453 (2023). https://doi.org/10.48550/ARXIV.2309.03453 arXiv:2309.03453

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.* 34, 6 (2015), 248:1–248:16. https://doi.org/10.1145/2816795.2818013

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2022. NeRF: representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2022), 99–106. https://doi.org/10.1145/3503250

Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Peter Kontschieder, and Matthias Nießner. 2023. DiffRF: Rendering-Guided 3D Radiance Field Diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 4328–4338. https://doi.org/10.1109/CVPR52729.2023.00421

Jacob Munkberg, Wenzheng Chen, Jon Hasselgren, Alex Evans, Tianchang Shen, Thomas Müller, Jun Gao, and Sanja Fidler. 2022. Extracting Triangular 3D Models, Materials, and Lighting From Images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 8270–8280. https://doi.org/10.1109/CVPR52688.2022.00810

Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. 2022. Point-E: A System for Generating 3D Point Clouds from Complex Prompts. *CoRR* abs/2212.08751 (2022). https://doi.org/10.48550/ARXIV.2212.08751 arXiv:2212.08751

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body From a Single Image. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 10975–10985. https://doi.org/10.1109/CVPR.2019.01123

Pixiv. 2019. *VRM tools of three.js.* https://github.com/pixiv/three-vrm

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2023. DreamFusion: Text-to-3D using 2D Diffusion. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/pdf?id=FjNys5c7VyY

Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. 2023. Magic123: One Image to High-Quality 3D Object Generation Using Both 2D and 3D Diffusion Priors. *CoRR* abs/2306.17843 (2023). https://doi.org/10.48550/ARXIV.2306.17843 arXiv:2306.17843

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 22500–22510. https://doi.org/10.1109/CVPR52729.2023.02155

Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. 2021. Deep Marching Tetrahedra: a Hybrid Representation for High-Resolution 3D Shape Synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 6087–6101. https://proceedings.neurips.cc/paper/2021/hash/30a237d18c50f563cba4531f1db44acf-Abstract.html

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. 2023. MVDream: Multi-view Diffusion for 3D Generation. *CoRR* abs/2308.16512 (2023). https://doi.org/10.48550/ARXIV.2308.16512 arXiv:2308.16512

VRoid. 2022. VRoid Hub. https://vroid.com/

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. 2023a. Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 12619–12629. https:

//doi.org/10.1109/CVPR52729.2023.01214

Peng Wang and Yichun Shi. 2023. ImageDream: Image-Prompt Multi-view Diffusion for 3D Generation. *CoRR* abs/2312.02201 (2023). https://doi.org/10.48550/ARXIV.2312.02201 arXiv:2312.02201

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. 2023b. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. *CoRR* abs/2305.16213 (2023). https://doi.org/10.48550/ARXIV.2305.16213 arXiv:2305.16213

Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. 2023. OmniObject3D: Large-Vocabulary 3D Object Dataset for Realistic Perception, Reconstruction and Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 803–814. https://doi.org/10.1109/CVPR52729.2023.00084

Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. 2023. ECON: Explicit Clothed humans Optimized via Normal integration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 512–523. https://doi.org/10.1109/CVPR52729.2023.00057

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. *CoRR* abs/2308.06721 (2023). https://doi.org/10.48550/ARXIV.2308.06721 arXiv:2308.06721

Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. 2022. LION: Latent Point Diffusion Models for 3D Shape Generation. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/40e56dabe12095a5fc44a6e4c3835948-Abstract-Conference.html

Biao Zhang, Jiapeng Tang, Matthias Nießner, and Peter Wonka. 2023b. 3DShape2VecSet: A 3D Shape Representation for Neural Fields and Generative Diffusion Models. *ACM Trans. Graph.* 42, 4 (2023), 92:1–92:16. https://doi.org/10.1145/3592442

Huichao Zhang, Bowen Chen, Hao Yang, Liao Qu, Xu Wang, Li Chen, Chao Long, Feida Zhu, Kang Du, and Min Zheng. 2023a. AvatarVerse: High-quality & Stable 3D Avatar Creation from Text and Pose. *CoRR* abs/2308.03610 (2023). https://doi.org/10.48550/ARXIV.2308.03610 arXiv:2308.03610

Jianfeng Zhang, Xuanmeng Zhang, Huichao Zhang, Jun Hao Liew, Chenxu Zhang, Yi Yang, and Jiashi Feng. 2023c. AvatarStudio: High-fidelity and Animatable 3D Avatar Creation from Text. *CoRR* abs/2311.17917 (2023). https://doi.org/10.48550/ARXIV.2311.17917 arXiv:2311.17917

Lvmin Zhang and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. *CoRR* abs/2302.05543 (2023). https://doi.org/10.48550/ARXIV.2302.05543 arXiv:2302.05543

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 586–595. https://doi.org/10.1109/CVPR.2018.00068

Fig. 8. We compare our generated four A-pose character images with other methods. The azimuths for all examples are set as $\{0°, 90°, 180°, 270°\}$.

Fig. 9. We compare the appearance and geometry of our generated 3D characters with other methods.