AvatarVerse: High-Quality & Stable 3D Avatar Creation from Text and Pose

Huichao Zhang[1*], **Bowen Chen**[1*], **Hao Yang**[1], **Liao Qu**[1,2], **Xu Wang**[1]
Li Chen[1], **Chao Long**[1], **Feida Zhu**[1], **Daniel Du**[1], **Min Zheng**[1]

[1]ByteDance, Beijing, China.
[2]Carnegie Mellon University, PA, USA
{zhanghuichao.hc, chenbowen.cbw, wangxu.ailab, chenli.phd, longchao, zhufeida, dukang.daniel,
zhengmin.666}@bytedance.com, liaoq@andrew.cmu.edu, yanghao.alexis@foxmail.com

Abstract

Creating expressive, diverse and high-quality 3D avatars from highly customized text descriptions and pose guidance is a challenging task, due to the intricacy of modeling and texturing in 3D that ensure details and various styles (realistic, fictional, etc). We present *AvatarVerse*, a stable pipeline for generating expressive high-quality 3D avatars from nothing but text descriptions and pose guidance. In specific, we introduce a 2D diffusion model conditioned on DensePose signal to establish 3D pose control of avatars through 2D images, which enhances view consistency from partially observed scenarios. It addresses the infamous Janus Problem and significantly stablizes the generation process. Moreover, we propose a progressive high-resolution 3D synthesis strategy, which obtains substantial improvement over the quality of the created 3D avatars. To this end, the proposed Avatar-Verse pipeline achieves zero-shot 3D modeling of 3D avatars that are not only more expressive, but also in higher quality and fidelity than previous works. Rigorous qualitative evaluations and user studies showcase AvatarVerse's superiority in synthesizing high-fidelity 3D avatars, leading to a new standard in high-quality and stable 3D avatar creation. Our project page is: https://avatarverse3d.github.io/ .

1 Introduction

The creation of high-quality 3D avatars has garnered significant interest due to their widespread applications in domains such as game production, social media and communication, augmented and virtual reality (AR/VR), and human-computer interaction. Traditional manual construction of these intricate 3D models is a labor-intensive and time-consuming process, requiring thousands of hours from skilled artists possessing extensive aesthetic and 3D modeling expertise. Consequently, automating the generation of high-quality 3D avatars using only natural language descriptions holds great research prospects with the potential to save resources, which is also the goal of our work.

Recently, great efforts have been made in reconstructing high-fidelity 3D avatars from multi-view videos (Jiang et al. 2022; Li et al. 2023d; Zheng et al. 2023; Wang et al. 2023a; Isik et al. 2023) or reference images (Wang et al. 2021;

Xiu et al. 2022). These methods primarily rely on limited visual priors from videos or reference images, leading to constrained ability to generate creative avatars with complex text prompts. Diffusion models has been applied to various realms (Ho et al. 2022; Li et al. 2023c; Ruan et al. 2022; Li et al. 2023b; Nie et al. 2022) in recent years. Specifically, in 2D image generation, they (Rombach et al. 2021; Zhang and Agrawala 2023; Saharia et al. 2022) illustrate considerable creativity, primarily due to the availability of large-scale text-image pairs. Nevertheless, the limited diversity of 3D models present challenges to effectively training a 3D diffusion model. Several studies (Poole et al. 2022; Cao et al. 2023; Huang et al. 2023; Kolotouros et al. 2023) have investigated the use of pre-trained text-image diffusion models to optimize Neural Radiance Fields (NeRF) (Mildenhall et al. 2020) for generating high-fidelity 3D models. Yet, stable creation of high-quality 3D avatars exhibiting various poses, appearances, and shapes remains a difficult task. Employing common score distillation sampling (SDS) (Poole et al. 2022) to guide NeRF optimization without additional control tends to bring in the Janus (multi-face) problem. The avatars produced by current approaches tend to exhibit noticeable blurriness and coarseness, leading to the absence of high-resolution local texture details, accessories, and other relevant features.

To cope with these weaknesses, we propose *AvatarVerse*, a novel framework designed for generating high-quality and stable 3D avatars from textual descriptions and pose guidances. We first train a new ControlNet with human DensePose condition (Güler, Neverova, and Kokkinos 2018) over 800K images. SDS loss conditinal on the 2D DensePose signal is then implemented on top of the ControlNet. Through this way, we obtain precise view correspondence between different 2D views as well as between every 2D view and the 3D space. Our approach not only enables pose control of the generated avatars, but also eliminates the Janus Problem suffered by most existing methods. It thus ensures a more stable and view-consistent avatar creation process. Additionally, benefiting from the accurate and flexible supervision signals provided by DensePose, the generated avatars can be highly aligned with the joints of the SMPL model, enabling simple and effective skeletal binding and control.

While relying solely on DensePose-conditioned Control-Net may result in local artifacts and blurry details, we in-

---

Figure 1: High-quality 3D avatars generated by AvatarVerse based on a simple text description.

troduce a progressive high-resolution generation strategy to enhance the fidelity and the detail of local geometry. To alleviate the coarseness of the generated avatar, we incorporate a smoothness loss, which regularizes the synthesis procedure by encouraging a smoother gradient of the density voxel grid within our computationally efficient explicit Neural Radiance Fields (NeRF).

The overall contributions are as follows:

- We present *AvatarVerse*, a method that can automatically create a high-quality 3D avatar according to nothing but a text description and a reference human pose.

- We present the DensePose-Conditioned Score Distillation Sampling Loss, an approach that facilitates pose-aware 3D avatar synthesis and effectively mitigates the Janus problem, thereby enhancing system stability.

- We bolster the quality of the produced 3D avatars via a progressive high-resolution generation strategy. This method, through a meticulous coarse-to-fine refining process, synthesizes 3D avatars with superior detail, encompassing elements like hands, accessories, and beyond.

- AvatarVerse delivers exceptional performance, excelling in both quality and stability. Rigorous qualitative eval-

uations, complemented by comprehensive user studies, underscore AvatarVerse's supremacy in crafting high-fidelity 3D avatars, thereby setting a new benchmark in stable, zero-shot 3D avatar creation of the highest quality.

2 Related Work

2.1 Text-Guided 3D Content Generation

The success in text-guided 2D image generation has paved the way for the development of text-guided 3D content generation methods. DreamFields (Jain et al. 2021), and CLIP-Mesh (Khalid et al. 2022) utilize the CLIP model (Radford et al. 2021) to optimize underlying 3D representations such as meshes and NeRF. DreamFusion (Poole et al. 2022) first proposes score distillation sampling (SDS) loss to get supervision from a pre-trained diffusion model during the 3D generation. Latent-NeRF (Metzer et al. 2022) improves upon DreamFusion by optimizing a NeRF that operates the diffusion process in a latent space. ProlificDreamer (Wang et al. 2023b) proposes variational score distillation and produces high-resolution and high-fidelity results. Despite their promising performance in 3D general content generation, these methods often produce suboptimal results when gener-

ating avatars, exhibiting issues like low quality, Janus (multi-face) problem, and incorrect body parts. In contrast, our AvatarVerse enables an accurate and high-quality generation of 3D avatars from text prompts.

2.2 Text-Guided 3D Avatar Generation

Avatar-CLIP (Hong et al. 2022) first initializes 3D human geometry with a shape VAE network and utilizes CLIP (Radford et al. 2021) to facilitate geometry sculpting and texture generation. DreamAvatar (Cao et al. 2023) and AvatarCraft (Jiang et al. 2023) employ the SMPL model as a shape prior and utilize pretrained text-to-image diffusion models to generate 3D avatars. DreamFace (Zhang et al. 2023) introduces a coarse-to-fine scheme to create personalized 3D facial structures. HeadSculpt (Han et al. 2023) generates 3D head avatars by leveraging landmark-based control and a learned textual embedding representing the back view appearance of heads. Concurrent with our work, DreamWaltz (Huang et al. 2023) presents 3D-consistent occlusion-aware score distillation sampling, which incorporates 3D-aware skeleton conditioning for view-aligned supervision. Constrained by the original training data, the skeleton-conditioned diffusion model may still exhibit view inconsistencies such as failing to generate the backside of desired avatars or struggling to generate specific body parts when provided with partial skeleton information. Furthermore, the sparse nature of the skeleton makes it challenging for the model to determine avatar contours and edges, leading to low-quality results. On the contrary, our proposed DensePose-conditioned ControlNet ensures high-quality, view-consistent image generation of various viewpoints and body parts, including full body, legs, head, and more, guaranteeing superior avatar quality.

2.3 High-Quality 3D Avatar Generation

Recently, there has been a growing focus on achieving high-quality or high-fidelity 3D generation and reconstruction. Some methods attempt to generate high-fidelity 3D human avatars from multi-view RGB videos (Jiang et al. 2022; Li et al. 2023d; Zheng et al. 2023; Wang et al. 2023a; Isik et al. 2023). There has also been work (Lin et al. 2022) explored a coarse-to-fine methodology, specifically by optimizing a high-resolution latent diffusion model to refine a textured 3D mesh model. In parallel to our work, DreamHuman (Kolotouros et al. 2023) zooms in and renders a 64 × 64 image for 6 important body regions during optimization. However, limited by the computation needs of Mip-NeRF-360, it can only produce low-resolution avatars without high-resolution details. Also, DreamHuman use SMPL shape for direct geometric supervision, which tends to provide skin-tight avatars. Our method, on the other hand, is more controllable and flexible, allowing for the creation of a wider range of accessories, clothing, and other features. Our AvatarVerse introduces a progressive high-resolution generation strategy. This involves gradually decreasing the camera's radius and focusing on distinct body parts, which facilitates the creation of a diverse range of accessories, clothing, and other elements. Our use of progressive grid also ensures a fine-grained generation.

3 Methodology

In this section, we present AvatarVerse, a fully automatic pipeline that can make a realistic 3D avatar from nothing but a text description and a body pose. After introducing some preliminaries, we first explain the DensePose-conditioned SDS loss, which facilitates pose-aware 3D avatar synthesis and effectively mitigates the Janus problem. We then introduce novel strategies that enhance the synthesis quality: the progressive high-resolution generation strategy and the avatar surface smoothing strategy.

3.1 DensePose SDS Loss

Prior research (Poole et al. 2022; Lin et al. 2022) predominantly employs supplementary text prompts, such as "front view" or "overhead view", to enhance view consistency. However, reliance solely on text prompts proves inadequate for accurately conditioning a 2D diffusion model on arbitrary views. This inadequacy engenders instability in 3D model synthesis, giving rise to issues like the Janus problem. As a solution, we propose the utilization of DensePose (Güler, Neverova, and Kokkinos 2018) as a more robust control signal, as depicted in Figure 2.

We choose DensePose as the condition because it delivers precise localization of 3D body parts in 2D images, affording intricate details and boundary conditions that may be overlooked by skeletal or other types of conditions. Notably, it exhibits resilience in challenging scenarios, facilitating accurate control even when body parts are partially concealed.

We first train a ControlNet (Zhang and Agrawala 2023) conditioned by DensePose part-labeled annotations using the DeepFashion (Liu et al. 2016) dataset. Figure 3 illustrates the capabilities of our ControlNet in generating high-quality view-consistent images, including various viewpoints and body parts such as full body, legs, head, and more. Given a specific camera viewpoint and pose P, we generate the DensePose condition image c by rendering the part-labeled SMPL model with the corresponding pose P. The conditioned SDS loss is shown in the following equation:

$$\nabla_\theta \mathcal{L}_{P-SDS}(\phi, x = g(\theta, P)) = E_{t,\epsilon}\left[w(t)(\hat{\epsilon} - \epsilon)\frac{\partial x}{\partial \theta}\right] \quad (1)$$

$$\hat{\epsilon} = \epsilon_\phi(z_t; y, t, c = h(SMPL, P)) \quad (2)$$

Here, g and h represent the NeRF render function and SMPL render function, respectively. The NeRF model and the SMPL pose model share identical camera viewpoints. This alignment of viewpoints enables coherent and consistent representations between the scene captured by NeRF and the corresponding human pose modeled by SMPL, allowing for better avatar generation. Our DensePose-conditioned ControlNet can generate various non-skin-tight realistic and fictional avatars as shown in Figure 3 (c).

3.2 Progressive High-Resolution Generation

Previous studies commonly apply SDS loss over the entire body, such global guidance often fails to produce high-quality details, especially for areas like hands, face, etc. These approaches lack effective guidance mechanisms to ensure the generation of high-quality, detailed geometry and

(a) Avatar Generation

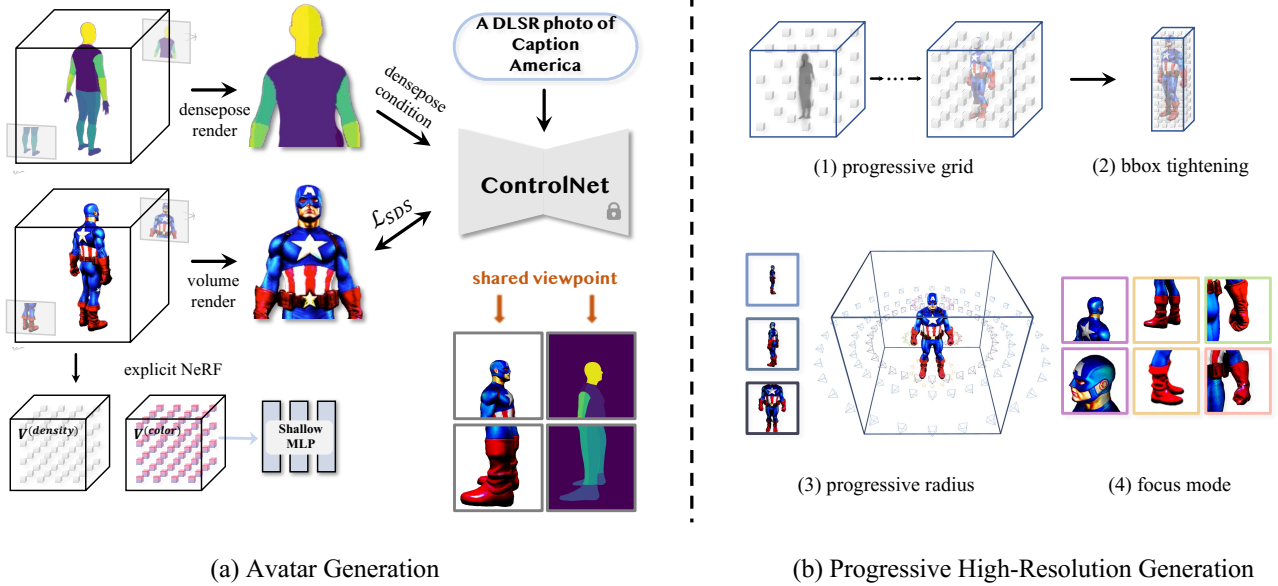(b) Progressive High-Resolution Generation

Figure 2: The overview of AvatarVerse. Our network takes a text prompt and DensePose signal as input to optimize an explicit NeRF via a DensePose-COCO pre-trained ControlNet. We use strategies including progressive grid, progressive radius, and focus mode to generate high-resolution and high-quality 3D avatars.

textures. To address this limitation, we propose a variety of guidance strategies aimed at promoting the generation of accurate and detailed representations, including progressive grid, focus mode, and progressive radius.

Progressive Grid Progressive training strategy is commonly used in 2d generation and 3d reconstruction method (Karras et al. 2019; Liu et al. 2020; Sun, Sun, and Chen 2021), while we find it critical in our method for neat and efficient 3d avatar generation. We set a predetermined number of voxels N_v as the final model resolution and double the voxel number after certain steps of optimization. The voxel size s_v is updated accordingly. During the early stage of training, we only need to generate a rough avatar shape. By allocating fewer grids, we can reduce the learning space and minimize floating artifacts. This strategy enables a gradual refinement of the avatar throughout the optimization process, allowing the model to adaptively allocate computational resources.

Also, the early stage of NeRF optimization is dominated by free space (i.e., space with low density). Motivated by this fact, we aim to find the areas of coarse avatar and allocate computational and memory resources to these important regions. To delineate the targeted area, we employ a density threshold to filter the scene and use a bounding box (bbox) to tightly enclose this area.

Let d_x, d_y, d_z represent the lengths of the tightened bbox, he voxel size can be computed as $s_v = \sqrt[3]{\frac{d_x \times d_y \times d_z}{N_v}}$. By shrinking the lengths of the bbox, the voxel size decreases, enabling high-resolution and more voxel around the avatar. This would enhance the model's ability to capture and model

intricate details, such as fine-grained body contours, facial features, and clothing folds.

Progressive Radius Let `pg_ckpt` be the set of checkpoint steps. When reaching the training step in `pg_ckpt`, we decrease the radius of the camera by 20%. This allows for gradual rendering of finer details stage by stage. By applying the conditioned SDS loss to smaller regions of the avatar, the model can capture and emphasize intricate features, ultimately producing more realistic and visually appealing outputs.

Focus Mode Similarly, to generate better intricacy in specific body parts, we introduce a focus mode (as illustrated in Fig. 2 (b)) during both the coarse/fine stage. Thanks to the SMPL prior, we can easily compute the raw body parts positions for any given pose. By placing the camera close to important body parts, loss calculation can be performed in a very small avatar region with 512×512 resolution. Owing to the stable performance of our DensePose ControlNet, as shown in Fig. 2, partial body can be generated without additional computational resources. Focus mode thus facilitates the creation of high-quality avatar details.

Mesh Refinement To render fine-grained high-resolution avatars within reasonable memory constraints and computation budgets, we further incorporate deformable tetrahedral grids (Lin et al. 2022; Shen et al. 2021) to learn textured 3D meshes of the generated avatars. Similar to (Lin et al. 2022), we use the trained explicit NeRF as the initialization for the mesh geometry, and optimize the mesh via backpropagation using the DensePose conditioned SDS gradient (Eq. 1).
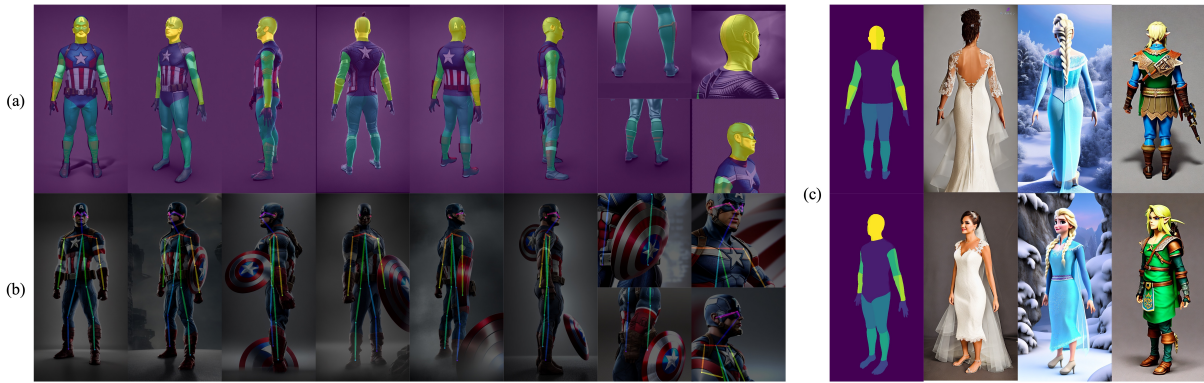
Figure 3: Qualitative results of our DensePose-conditioned ControlNet. (a) 10 generated images controlled by DensePose with varying viewpoints and body parts. (b) 10 corresponding images with the same viewpoints controlled by human pose (Openpose) signals. It often fails to generate the backside of the avatar (4-th (b)) and struggles with part generation (the last two columns). (c) non-skin-tight generation results in both realistic and fictional avatars.
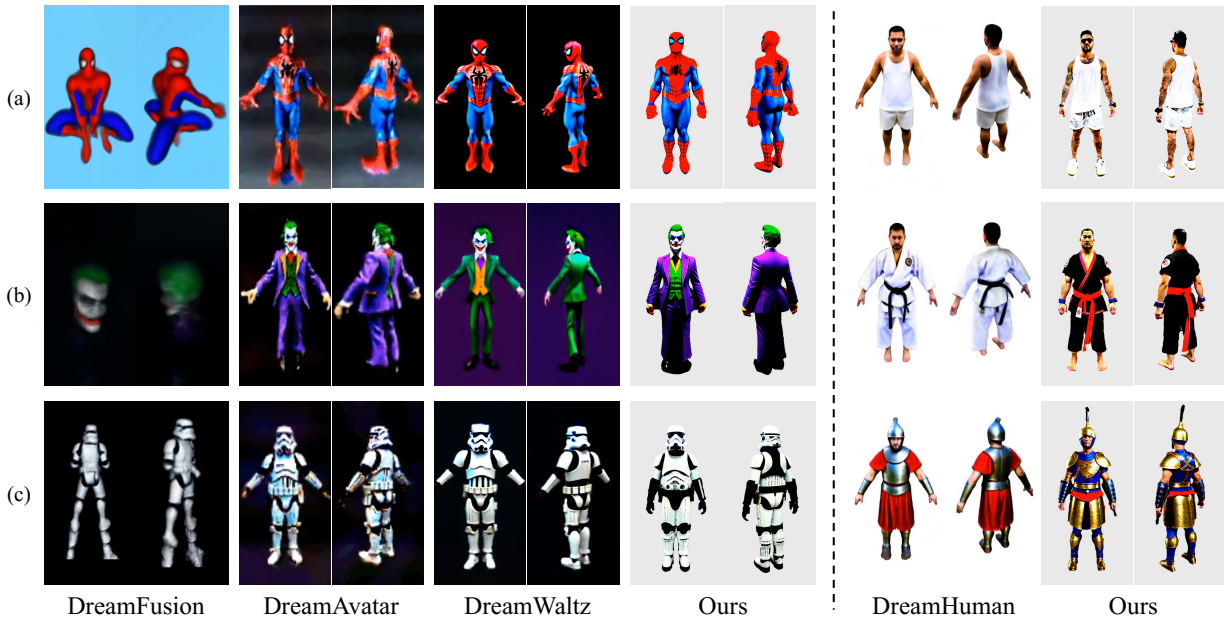

| DreamFusion | DreamAvatar | DreamWaltz | Ours | DreamHuman | Ours |

Figure 4: Qualitative comparisons with four SOTA methods. We show several non-cherry-picked results generated by Avatar-Verse. Our method generates higher-resolution details and maintains a fine-grained geometry compared with other methods. (a): "Spiderman"; " a man wearing a white tanktop and shorts", (b): "Joker"; "a karate master wearing a Black belt", (c): "Stormtrooper"; "a Roman soldier wearing his armor".

3.3 Avatar Surface Smoothing

Maintaining a globally coherent avatar shape for explicit grids is challenging due to high degree of freedom and lack of spatial coherence. Individual optimization of each voxel limits information sharing across grids, resulting in a less smooth surface for the generated avatar and some local minima. To address this problem, we follow the definition of the Gaussian convolution $\mathcal{G}$ in (Wu et al. 2022) and include a modified smoothness regularization formulated as:

$$\mathcal{L}_s(V) = \|\mathcal{G}(V, k_g, \sigma_g) - V\|_2^2 \qquad (3)$$

$$\mathcal{L} = \mathcal{L}_{P-SDS} + \lambda * \mathcal{L}_s(V) \qquad (4)$$

Here, k_g and σ_g represent the kernel size and the standard deviation. We apply this smoothness term to the gradient of the density voxel grid V, yielding a gradient smoothness loss $\mathcal{L}_s(\nabla V^{(density)})$. This encourages a smoother surface and mitigates the presence of noisy points. The overall loss is defined in (4), where λ denotes the smooth coefficient.

4 Experiments

In this section, we illustrate the effectiveness of our proposed method. We demonstrate the efficacy of each proposed strategy and provide a detailed comparison against recent state-of-the-art methods.

4.1 Implementation Details

We follow (Sun, Sun, and Chen 2021) to implement the explicit NeRF in our method. For each text prompt, we train AvatarVerse for 5000 and 4000 iterations in the coarse stage and mesh refinement stage, respectively. The whole generation process takes around 2 hours on one single NVIDIA A100 GPU. We include initialization, densepose training and progressive high-resolution generation details in this section. For more comprehensive experiment details, we refer the reader to our Supplementary Material.

Initialization To aid in the early stages of optimization, we adopt a technique inspired by (Poole et al. 2022) and introduce an ellipsoidal density "blob" around the origin. The dimensions of the "blob" in the XYZ axes are determined based on the range of coordinates in the SMPL pose model. Furthermore, we incorporate additional SMPL-derived density bias (Cao et al. 2023) to facilitate avatar generation.

DensePose Training We annotate the DeepFashion dataset (Liu et al. 2016) using a pretrained DensePose (Güler, Neverova, and Kokkinos 2018) model, resulting in over 800K image pairs. The ControlNet is trained using these image pairs with BLIP2-generated text prompt (Li et al. 2023a). The diffusion model employed is SD1.5.

Progressive High-Resolution Generation For progressive grid, we double the number of voxels at 500, 1500, and 2000 iterations at the coarse stage. After 3000 steps, we shrink the bounding box to the region where the density exceeds 0.1. Our progressive radius consists of three stages, where the camera radius ranges from 1.4 to 2.1, 1 to 1.5, and 0.8 to 1.2 respectively. We reduce the radius at 1000 and 2000 iterations across both stages. Our focus mode starts from the 1000-th step in the coarse stage and is consistently employed throughout the mesh refinement phase.

4.2 Qualitative Results

Comparison with SOTA Methods We present qualitative comparisons with DreamFusion (Poole et al. 2022), DreamAvatar (Cao et al. 2023), DreamWaltz (Huang et al. 2023), and DreamHuman (Kolotouros et al. 2023) in Fig. 4. Our method consistently outperforms them in terms of both geometry and texture quality. The surface of the avatars generated by our method is exceptionally clear, owing to our progressive high-resolution generation strategy. In comparison to DreamHuman, the avatars produced by our method exhibit a richer array of details across all cases, encompassing skin, facial features, clothing, and more.

Flexible Avatar Generation In Fig. 5, we demonstrate the capability of our method in generating 3D partial avatars, which is not achievable by other existing methods due to the absence of the DensePose control. Our method enables

Figure 5: Flexible Avatar Generation. (a) Partial Generation. Results are generated with the same prompt "Stormtrooper" and "Batman". (b) Arbitrary Pose Generation.

the partial generation by directly modifying the input DensePose signal, eliminating the need for extra descriptive information such as "The head of..." or "The upper body of...". This allows us to generate partial avatars of various types thanks to the attached semantics, including full-body, half-body, head-only, hand-only, etc. Additionally, our AvatarVerse is capable of generating avatars in various poses, showcasing our stable control over view consistency.
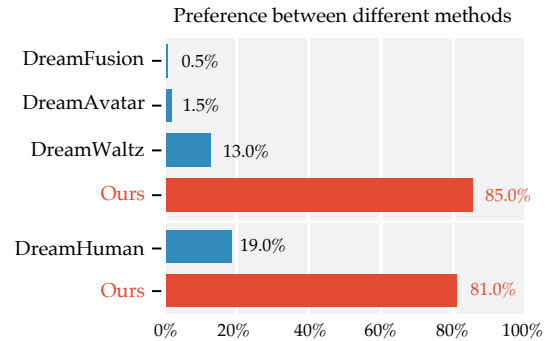

Figure 6: Quantitative results of user study.

4.3 User Study

To further assess the quality of our generated 3D avatars, we conduct user studies comparing the performance of our results with four SOTA methods under the same text prompts. We randomly select 30 generated outcomes (presented as rendered rotating videos) and ask 16 volunteers to vote for their favorite results based on geometry and texture quality. In Fig. 6, we compare AvatarVerse with DreamFusion (Poole et al. 2022), DreamAvatar (Cao et al. 2023), and DreamWaltz (Huang et al. 2023), demonstrating a significant preference for our method over other approaches.

We also compare our method with DreamHuman (Kolotouros et al. 2023) in terms of realistic human. A remarkable 81% of volunteers voted in favor of our AvatarVerse.

4.4 Ablation Study

Effectiveness of Progressive Strategies To evaluate the design choices of AvatarVerse, we conduct an ablation study on the effectiveness of b) the progressive grid, c) the progressive radius, d) the focus mode, and e) the mesh refinement. We sequentially add these components and report the results in Fig. 7. The initial result lacks detail (e.g., no sword in the back, no armguards) and exhibits numerous floating artifacts. The overall quality is blurry and unclear. Upon incorporating the progressive grid, more voxels are gathered around the avatar region, this introduces more details into the avatar. By progressively narrowing the camera distance, the model can leverage the detail inherent in the latent diffusion, thereby eliminating a large number of floating artifacts and enhancing local details, such as the sword in the back. The focus mode further zooms in and utilizes a resolution of 512×512 to target and optimize certain body parts, generating high-definition and intricate local details. The mesh refinement further optimize 3D mesh of the coarse avatar, resulting in finer avatar texture.
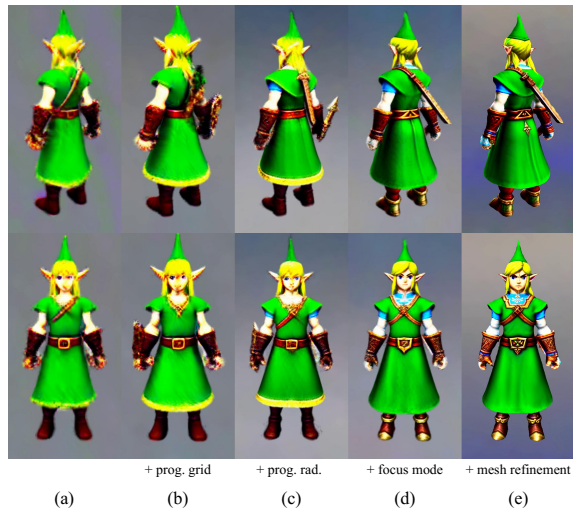

Figure 7: Impact of progressive strategies. (a) none progressive strategy; (b) add progressive grid; (c) add progressive radius upon (b); (d) add focus mode upon (c); (e) add mesh refinement, our full method.

Effectiveness of DensePose Control Figure 8 illustrates the influence of various control signals. When conditioned by the skeleton, the model can generate avatars that more closely resemble human figures. However, the avatar's edges appear blurry and still face severe Janus problem. By incorporating DensePose control into our framework, we achieve more precise avatar boundaries, intricate details, and stable avatar control, resulting in a substantial improvement in the overall quality and appearance of the generated avatars.

Effectiveness of Surface Smoothing Avatar surface smoothing plays a critical role in the AvatarVerse framework, as it guarantees the generated avatars exhibit compact geometry and smooth surfaces. As shown in Figure 9, by
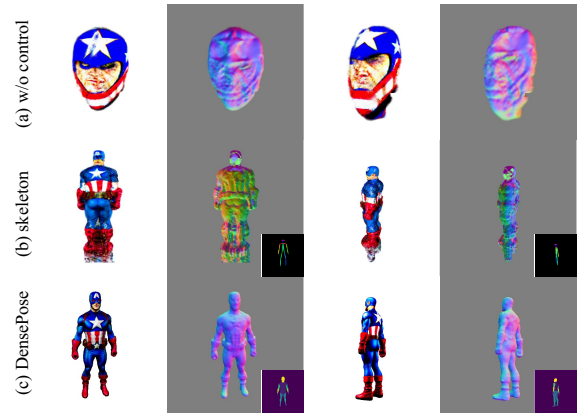

Figure 8: Impact of control signal. (a) without additional control; (b) with skeleton control; (c) with our DensePose control. For each type, we show the RGB, normal, depth, and the corresponding control signal.

(a) w/o surface smoothing
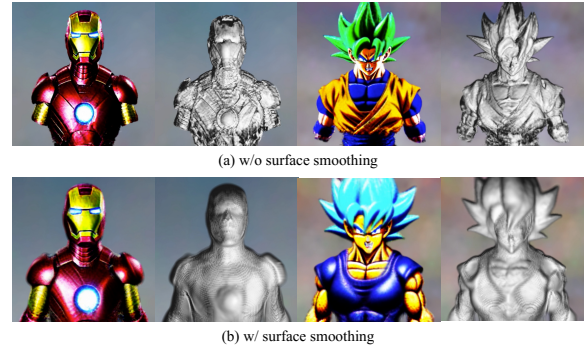
(b) w/ surface smoothing

Figure 9: Impact of surface smoothing strategy. (a) without surface smoothing; (b) with surface smoothing. Results are generated with the same text prompt.

finding a balance between the smooth loss and the conditioned SDS loss, the visual quality and realism of the avatars are greatly improved.

5 Conclusion

In this paper, we introduce AvatarVerse, a novel framework designed to generate high-quality and stable 3D avatars from textual prompts and poses. By employing our trained DensePose-conditioned ControlNet, we facilitate stable partial or full-body control during explicit NeRF optimization. Our 3D avatar outcomes exhibit superior texture and geometry quality, thanks to our progressive high-resolution generation strategy. Furthermore, the generated avatars are easily animatable through skeletal binding, as they exhibit high alignment with the joints of the SMPL model. Through comprehensive experiments and user studies, we demonstrate that our AvatarVerse significantly outperforms previous and contemporary approaches. We believe that our approach renews the generation of high-quality 3D avatars in the neural and prompt-interaction era.

References

Cao, Y.; Cao, Y.-P.; Han, K.; Shan, Y.; and Wong, K.-Y. K. 2023. DreamAvatar: Text-and-Shape Guided 3D Human Avatar Generation via Diffusion Models. *ArXiv*, abs/2304.00916.

Güler, R. A.; Neverova, N.; and Kokkinos, I. 2018. Dense-Pose: Dense Human Pose Estimation in the Wild. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7297–7306.

Han, X.; Cao, Y.; Han, K.; Zhu, X.; Deng, J.; Song, Y.-Z.; Xiang, T.; and Wong, K.-Y. K. 2023. HeadSculpt: Crafting 3D Head Avatars with Text. *ArXiv*, abs/2306.03038.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022. Video Diffusion Models. *ArXiv*, abs/2204.03458.

Hong, F.; Zhang, M.; Pan, L.; Cai, Z.; Yang, L.; and Liu, Z. 2022. AvatarCLIP: Zero-Shot Text-Driven Generation and Animation of 3D Avatars. *ACM Trans. Graph.*, 41: 161:1–161:19.

Huang, Y.; Wang, J.; Zeng, A.; Cao, H.; Qi, X.; Shi, Y.; Zha, Z.; and Zhang, L. 2023. DreamWaltz: Make a Scene with Complex 3D Animatable Avatars. *ArXiv*, abs/2305.12529.

Isik, M.; Rünz, M.; Georgopoulos, M.; Khakhulin, T.; Starck, J.; de Agapito, L.; and Nießner, M. 2023. HumanRF: High-Fidelity Neural Radiance Fields for Humans in Motion. *ArXiv*, abs/2305.06356.

Jain, A.; Mildenhall, B.; Barron, J. T.; Abbeel, P.; and Poole, B. 2021. Zero-Shot Text-Guided Object Generation with Dream Fields. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 857–866.

Jiang, R.; Wang, C.; Zhang, J.; Chai, M.; He, M.; Chen, D.; and Liao, J. 2023. AvatarCraft: Transforming Text into Neural Human Avatars with Parameterized Shape and Pose Control. *ArXiv*, abs/2303.17606.

Jiang, T.; Chen, X.; Song, J.; and Hilliges, O. 2022. InstantAvatar: Learning Avatars from Monocular Video in 60 Seconds. *ArXiv*, abs/2212.10550.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2019. Analyzing and Improving the Image Quality of StyleGAN. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8107–8116.

Khalid, N. M.; Xie, T.; Belilovsky, E.; and Popa, T. 2022. CLIP-Mesh: Generating textured meshes from text using pretrained image-text models. *SIGGRAPH Asia 2022 Conference Papers*.

Kolotouros, N.; Alldieck, T.; Zanfir, A.; Bazavan, E. G.; Fieraru, M.; and Sminchisescu, C. 2023. DreamHuman: Animatable 3D Avatars from Text. *ArXiv*, abs/2306.09329.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *ICML*.

Li, X.; Chen, Y.; Lin, C.-C.; Singh, R.; Raj, B.; and Liu, Z. 2023b. Completing Visual Objects via Bridging Generation and Segmentation. *ArXiv*, abs/2310.00808.

Li, X.; Wen, Y.; Yang, M.; Wang, J.; Singh, R.; and Raj, B. 2023c. Rethinking Voice-Face Correlation: A Geometry View. *Proceedings of the 31st ACM International Conference on Multimedia*.

Li, Z.; Zheng, Z.; Liu, Y.; Zhou, B.; and Liu, Y. 2023d. Pose-Vocab: Learning Joint-structured Pose Embeddings for Human Avatar Modeling. *ArXiv*, abs/2304.13006.

Lin, C.-H.; Gao, J.; Tang, L.; Takikawa, T.; Zeng, X.; Huang, X.; Kreis, K.; Fidler, S.; Liu, M.-Y.; and Lin, T.-Y. 2022. Magic3D: High-Resolution Text-to-3D Content Creation. *ArXiv*, abs/2211.10440.

Liu, L.; Gu, J.; Lin, K. Z.; Chua, T.-S.; and Theobalt, C. 2020. Neural Sparse Voxel Fields. *ArXiv*, abs/2007.11571.

Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; and Tang, X. 2016. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1096–1104.

Metzer, G.; Richardson, E.; Patashnik, O.; Giryes, R.; and Cohen-Or, D. 2022. Latent-NeRF for Shape-Guided Generation of 3D Shapes and Textures. *arXiv preprint arXiv:2211.07600*.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *ArXiv*, abs/2003.08934.

Nie, W.; Guo, B.; Huang, Y.; Xiao, C.; Vahdat, A.; and Anandkumar, A. 2022. Diffusion Models for Adversarial Purification. In *International Conference on Machine Learning*.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. DreamFusion: Text-to-3D using 2D Diffusion. *ArXiv*, abs/2209.14988.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674–10685.

Ruan, L.; Ma, Y.; Yang, H.; He, H.; Liu, B.; Fu, J.; Yuan, N. J.; Jin, Q.; and Guo, B. 2022. MM-Diffusion: Learning Multi-Modal Diffusion Models for Joint Audio and Video Generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10219–10228.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *ArXiv*, abs/2205.11487.

Shen, T.; Gao, J.; Yin, K.; Liu, M.-Y.; and Fidler, S. 2021. Deep Marching Tetrahedra: a Hybrid Representation for High-Resolution 3D Shape Synthesis. *ArXiv*, abs/2111.04276.

Sun, C.; Sun, M.; and Chen, H.-T. 2021. Direct Voxel Grid Optimization: Super-fast Convergence for Radiance Fields Reconstruction. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5449–5459.

Wang, C.; Chai, M.; He, M.; Chen, D.; and Liao, J. 2021. Cross-Domain and Disentangled Face Manipulation With 3D Guidance. *IEEE Transactions on Visualization and Computer Graphics*, 29: 2053–2066.

Wang, L.; Zhao, X.; Sun, J.; Zhang, Y.; Zhang, H.; Yu, T.; and Liu, Y. 2023a. StyleAvatar: Real-time Photo-realistic Portrait Avatar from a Single Video. *ArXiv*, abs/2305.00942.

Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2023b. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. *ArXiv*, abs/2305.16213.

Wu, T.; Wang, J.; Pan, X.; Xu, X.; Theobalt, C.; Liu, Z.; and Lin, D. 2022. Voxurf: Voxel-based Efficient and Accurate Neural Surface Reconstruction. *ArXiv*, abs/2208.12697.

Xiu, Y.; Yang, J.; Cao, X.; Tzionas, D.; and Black, M. J. 2022. ECON: Explicit Clothed humans Obtained from Normals. *ArXiv*, abs/2212.07422.

Zhang, L.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. *ArXiv*, abs/2302.05543.

Zhang, L.; Qiu, Q.; Lin, H.; Zhang, Q.; Shi, C.; Yang, W.; Shi, Y.; Yang, S.; Xu, L.; and Yu, J. 2023. DreamFace: Progressive Generation of Animatable 3D Faces under Text Guidance. *ArXiv*, abs/2304.03117.

Zheng, Z.; Zhao, X.; Zhang, H.; Liu, B.; and Liu, Y. 2023. AvatarReX: Real-time Expressive Full-body Avatars. *ArXiv*, abs/2305.04789.