Supplementary Material of AvatarCraft: Transforming Text into Neural Human Avatars with Parameterized Shape and Pose Control

1. More Implementation Details

We adopt similar augmentation strategies to those used in recent text-to-3D generative models [1, 5, 3, 2, 4]. In this section, we describe the adopted augmentations in detail.

Random Camera Extrinsic Augmentation. During the template avatar creation process, we randomize the camera extrinsic when rendering the implicit avatar $\mathcal{N}$. More specifically, we fix the camera to look at the center and sample elevation angel $\theta \sim \mathcal{U}(-\pi/6, \pi/6)$, azimuth angle $\phi \sim \mathcal{U}((-\pi/3, \pi/3) \cup (2\pi/3, 4\pi/3))$, and camera distance in $\mathcal{U}(2.0, 2.2)$.

Random Background Augmentation. We augment the rendering of implicit avatar $\mathcal{N}$ by using random background color. Three types of background are used, including 1) pure white. 2) pure black. 3) Gaussian noise $\mathcal{N}(0.5, 0.1)$.

View-dependent Prompt Augmentation. We combine the prompt augmentation technique for human [1] and for general object [5] to provide more semantically meaningful text guidance. First, we use *"the body of {}"* and *"the face of {}"* when rendering body and head bounding box of avatars, respectively. Additionally, we further augment the prompt by *"front view of {}"*, *"side view of {}"* depending on the range of azimuth angle ϕ. More specifically, we use *"Front view of {}"* when $5\pi/6 \leq \phi \leq 7\pi/6$, and use *"Back view of {}"* when $-\pi/6 \leq \phi \leq \pi/6$, otherwise we use *"Side view of {}"*. Putting it together, example prompts include "back view of the body of Captain America" and "front view of the head of Captain America". We visualize our augmentation in Fig. 1.

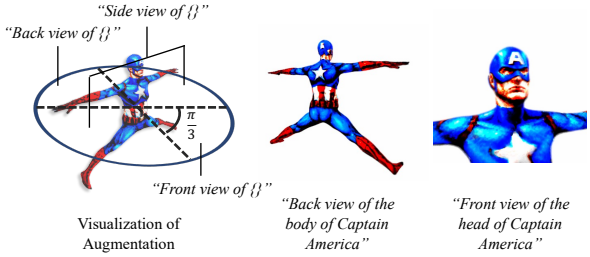

Figure 1. **Prompt Augmentation.**

2. Additional Qualitative Results

In this section, we provide more results of our proposed method, including additional generated avatars in Fig. 2 and Fig. 3, as well as pose sequences in Fig. 4.

3. Supplementary Video

We provide a supplementary video with more visual results rendered in multiple views. We highly recommend watching our supplementary video in https://avatar-craft.github.io/ to observe the user-friendliness and view consistency that our method can achieve in creating, reshaping, and animating neural human avatars.

References

[1] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*, 2022. 1

[2] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 867–876, 2022. 1

[3] Han-Hung Lee and Angel X Chang. Understanding pure clip guidance for voxel grid nerf models. *arXiv preprint arXiv:2209.15172*, 2022. 1

[4] 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. *arXiv preprint arXiv:2211.10440*, 2022. 1

[5] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1

Figure 2. **More Generated Avatars.**

Figure 3. **More Generated Avatars.**

Figure 4. **Pose Sequences.**