Supplementary Material of GAvatar: Animatable 3D Gaussian Avatars with Implicit Mesh Learning

Ye Yuan* Xueting Li* Yangyi Huang Shalini De Mello Koki Nagano Jan Kautz Umar Iqbal
NVIDIA

https://nvlabs.github.io/GAvatar

In this supplementary document, we will discuss more implementation details (Sec. A), show more baseline comparisons (Sec. B) and qualitative results (Sec. C), and list the prompts used for user study (Sec. D).

A. Implementation Details

Camera sampling. During optimization, we randomly sample camera poses to render full-body avatars from different views as well as zoom-in images of various body parts. Specifically, we randomly sample camera poses from a spherical coordinate system with radius 3.5, elevation range $[-10°, 45°]$, and y-axis field of view range $[-26°, 45°]$ for full-body renderings. To encourage detailed body parts generation, we manipulate cameras to render zoom-in images for the face, back head, arms, upper body, and lower body. During training, we evenly sample different body parts and the full body renderings.

Training. For each prompt, we optimize the avatar for 20000 iterations with the Adam optimizer. The learning rates for different learnable parameters discussed in Sec. 4.3 of the main paper are listed in Table 1 below. We train the avatar in natural pose θ_N for 3000 iterations before introducing random pose θ_A sampled from the CMU motion capture database[1] using the SMPL-X parameters from AMASS [8]. Starting from the 5000th iteration, we manipulate cameras to render zoom-in images for specific body parts (*e.g.*, face, hands, upper body, *etc.*) to facilitate learning intricate detail in these parts. The total training takes approximately 3 hours for each avatar on an NVIDIA RTX 3090Ti.

Parameter	Learning rate
Gaussian local positions $\{p_k^i\}$	0.00016
Gaussian attribute field $\mathcal{H}_\phi$	0.001
SDF $\mathcal{S}_\psi$	0.0001
opacity kernel parameters $\{\gamma, \lambda\}$	0.001
primitive motion corrective networks $\delta P_\omega, \delta R_\omega, \delta S_\omega$	0.0001
the SMPL-X shape parameters β	0.0003

Table 1. **Learning rates for different parameters.**

[1] http://mocap.cs.cmu.edu/

Network architecture. For the implicit Gaussian attribute field discussed in Sec. 4.1 in the main paper, we adopt a hash-encoded feature grid with 8 levels, where the base resolution is $16 \times 16 \times 16$. The feature grid is followed by three MLP layers that output a 55-dim vector including the scaling, rotation, and spherical harmonics features of the 3D Gaussian. For the SDF discussed in Sec. 4.2 in the main paper, we utilize a similar design as the Gaussian attribute field. Specifically, we use another hash-encoded feature grid with 16 levels and a base resolution of $16 \times 16 \times 16$. The feature grid is followed by three MLP layers that output the SDF value of the 3D Gaussian, which is then converted to its opacity value using the opacity kernel κ. During training, we initialize each primitive with 64 Gaussians lying on a $4 \times 4 \times 4$ grid within the primitive and use the densification process (see Sec. 4.3 in the main paper) to adaptively change the total Gaussian number as discussed in [5]. We also pretrain the Gaussian implicit fields to have an initial scale of 4mm in the world coordinate system.

B. Additional Baseline Comparison

We provide additional qualitative comparisons with DreamWaltz [3], AvatarCraft [4], AvatarCLIP [2], DreamGaussian [9], Fantasia3D [1], TADA [7] and DreamHuman [6] in Fig. 1, 2, 3, 4, 5, 6 and 7, respectively. We note that DreamWaltz, DreamGaussian, TADA and DreamHuman are all concurrent text-to-3D avatar works. To ensure the best performance of the baselines, we use publicly available code and default hyper-parameters for each baseline except for DreamHuman, whose code is not available yet. Thus, we compare with the avatars downloaded from the project website[2]. Overall, our method is not only more robust to various prompts, but also shows more intricate and realistic details compared to all the baseline methods.

C. Additional Qualitative Results

We showcase more characters generated by GAvatar in Fig. 8 and 9, demonstrating the robustness and generalization of the proposed method.

[2] https://dream-human.github.io/

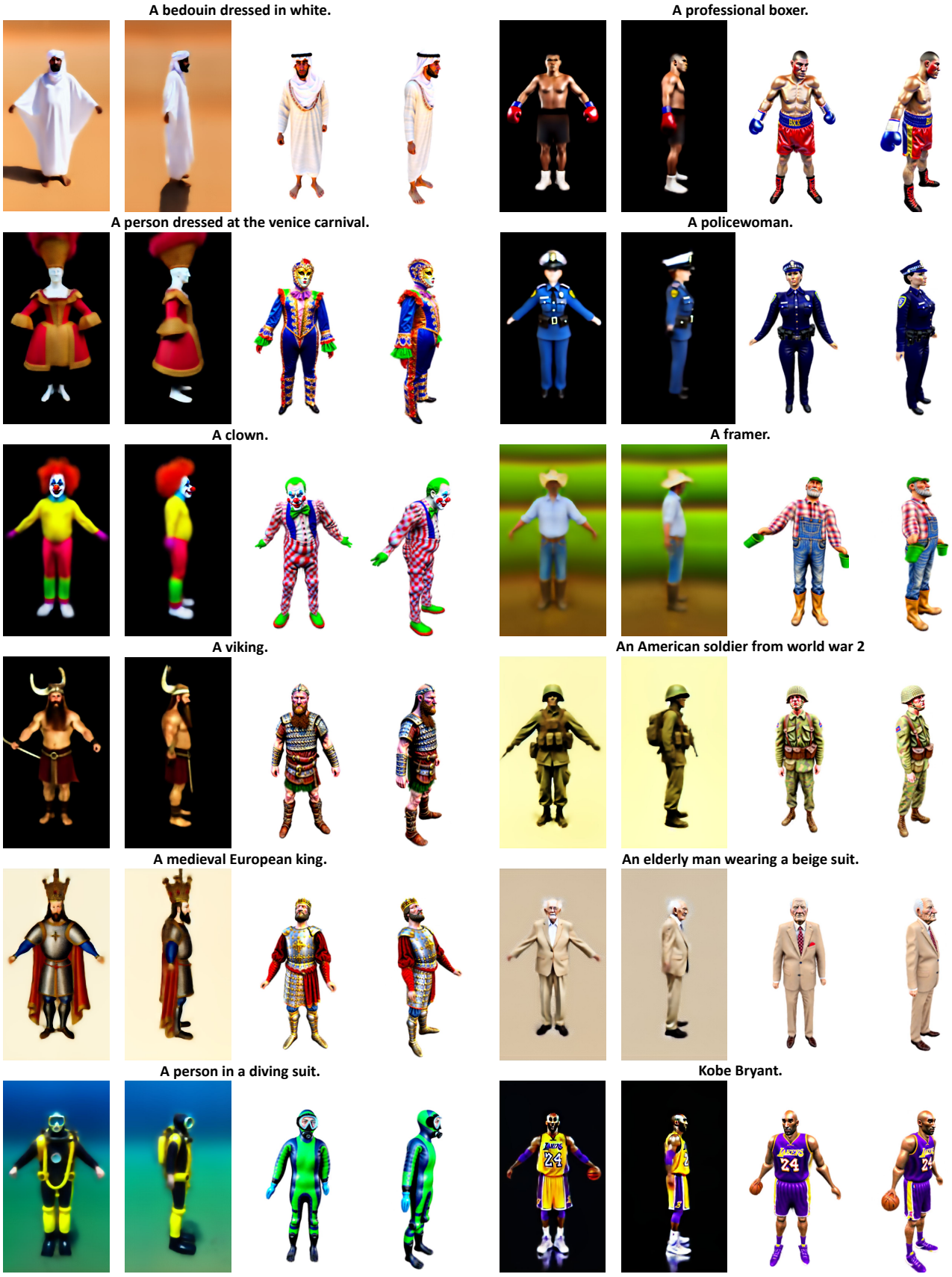

Figure 1. More comparisons with DreamWaltz [3].

Figure 2. More comparisons with AvatarCraft [4].

Figure 3. More comparisons with AvatarCLIP [2].

Figure 4. More comparisons with DreamGaussian [9].

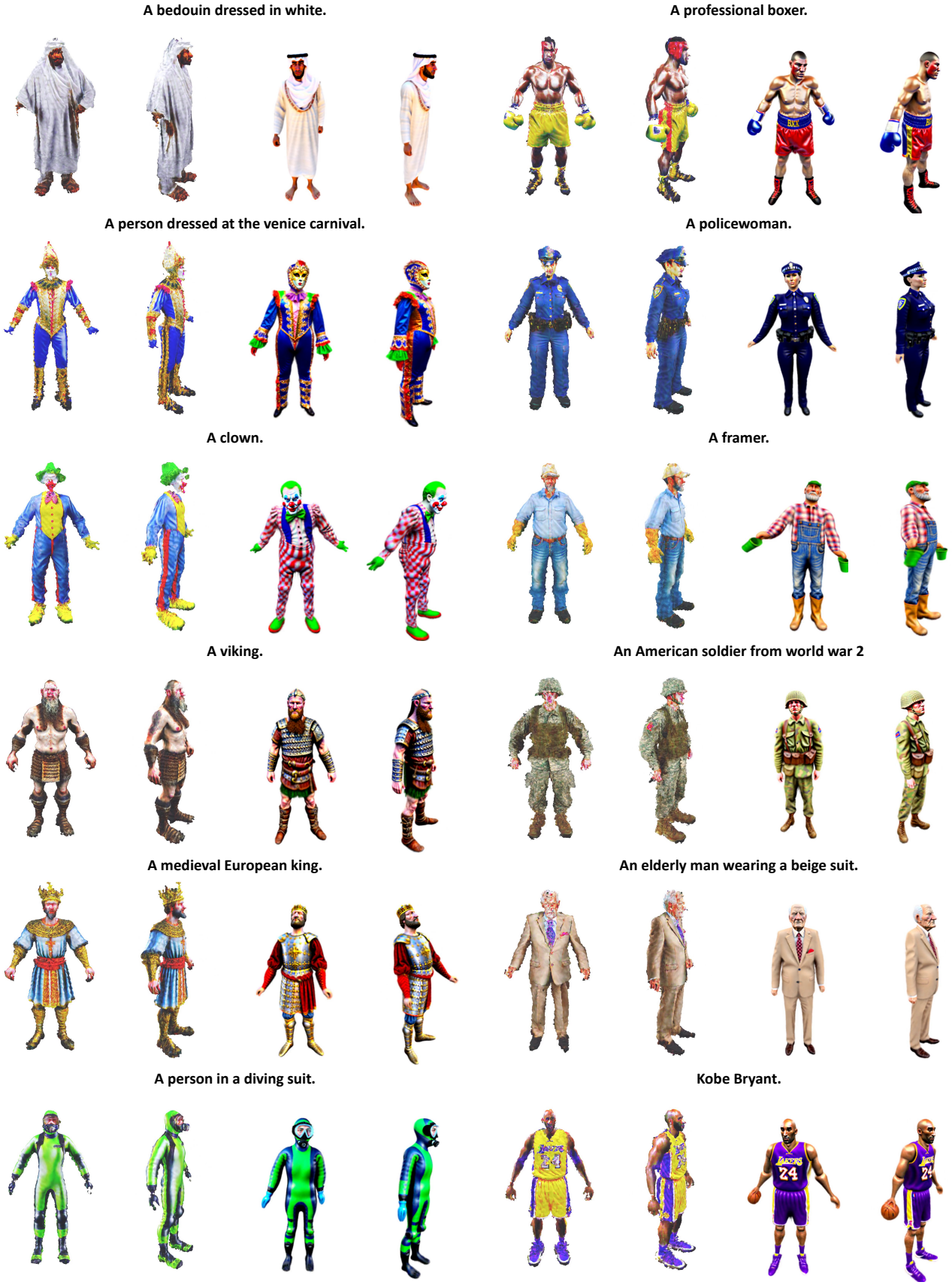

Figure 5. More comparisons with Fantastia3D [1].

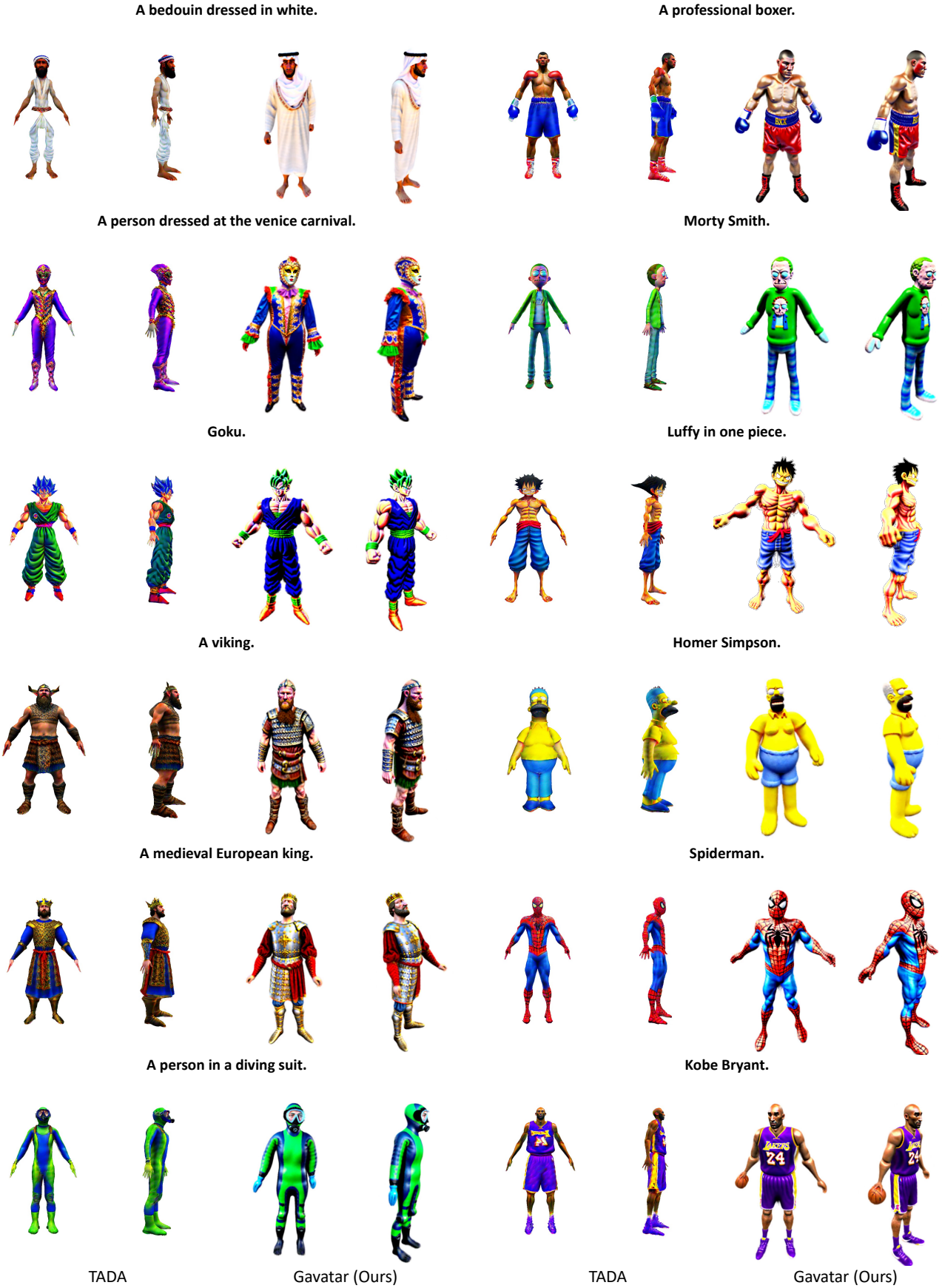

Figure 6. More comparisons with TADA [7].

Figure 7. More comparisons with DreamHuman [6].

Figure 8. More results by GAvatar.

Figure 9. More results by GAvatar.

D. User Study Prompts

For fair comparisons, we use the following 24 prompts commonly used by various baselines in the user study.

```
A professional boxer.
Morty Smith.
A person in a diving suit.
An American soldier from World War 2.
Goku.
Rick Sanchez.
A person dressed at the Venice carnival.
A medieval European king.
An elderly man wearing a beige suit.
Kobe Bryant.
A man wearing a white tank top and shorts.
A policewoman.
A black female surgeon.
A viking.
Oprah Winfrey.
A bedouin dressed in white.
A framer.
A clown.
Jane Goodall.
Homer Simpson.
Kristoff in Frozen.
Luffy in one piece.
Spiderman.
Jeff Bezos.
```

References

[1] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation. In *International Conference on Computer Vision (ICCV)*, 2023. 1, 6

[2] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. AvatarCLIP: Zero-Shot Text-Driven Generation and Animation of 3D Avatars. *Transactions on Graphics (TOG)*, 41(4):1–19, 2022. 1, 4

[3] 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. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 1, 2

[4] Ruixiang Jiang, Can Wang, Jingbo Zhang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Avatarcraft: Transforming text into neural human avatars with parameterized shape and pose control. In *International Conference on Computer Vision (ICCV)*, 2023. 1, 3

[5] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 2023. 1

[6] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. *arXiv preprint:2306.09329*, 2023. 1, 8

[7] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J Black. Tada! text to animatable digital avatars. In *International Conference on 3D Vision (3DV)*, 2024. 1, 7

[8] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, pages 5442–5451, 2019. 1

[9] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 1, 5