DreamAvatar: Text-and-Shape Guided 3D Human Avatar Generation via Diffusion Models
(*Supplementary Material*)

Yukang Cao[1*] Yan-Pei Cao[2*] Kai Han[1†] Ying Shan[2] Kwan-Yee K. Wong[1]

[1]The University of Hong Kong [2]ARC Lab, Tencent PCG

Contents

List of Tables

List of Figures

*Equal contributions † Corresponding authors ‡ Webpage: https://yukangcao.github.io/DreamAvatar/

Table 1. **Hyper-parameters of DreamAvatarw.**

Camera setting	Camera distance range	(1.0, 1.5)
	Radius	1.0
	Elevation range	(-10, 45)
	FoV range	(40, 70)
Render setting	Resolution for 0-5k iters	(64, 64)
	Resolution for 5k-10k iters	(512, 512)
	num steps sampled per ray	512
Diffusion setting	Guidance scale	7.5
	Guidance scale Lora	1.0
	t range	(0.02, 0.98)
	Minimal step percent	0.02
	Maximal step percent for 0-5k iters	0.98
	Maximal step percent for 5k-10k iters	0.5
	$\omega(t)$	$\sqrt{\alpha_t}(1-\alpha_t)$
Training objectives	λ for vsd	1.0
	λ for Lora	1.0
	λ for sparsity	10.0
Hardware	GPU	1 $\times$ NVIDIA A40 (48GB)

A. Implementation details

Our network is built upon the Threestudio open-source 3D generative project (Threestudio [2]). To achieve this, we utilize Hash embedding, which maps the input vector $\mathbf{x} \in \mathbb{R}^3$ to a higher-frequency dimension, resulting in $\gamma(\mathbf{x}) \in \mathbb{R}^{32}$. In the Hash embedding process, we employ 16 levels and assign 16 features to each level.

For the initial 5000 epochs, we set the resolution of the rendered image to $64 \times 64 \times 3$. However, for the subsequent 5000 epochs, we increase the resolution to $512 \times 512 \times 3$. This change is feasible because the empty space has been effectively pruned after the first 5000 epochs, and rendering 512×512 images does not excessively consume VRAM. This design enhances both training efficiency and the quality of the final generated results.

The multilayer perceptron (MLP) within our NeRF model consists of three layers with dimensions [32, 64, 64, 3+1+3]. Here, the channels '3', '1', and '3' correspond to predicted normals, density values, and RGB color values, respectively. Additionally, we employ a similar MLP architecture, [32, 64, 64, 3], to address non-rigid motion.

To ensure easy reproducibility, we have included all the hyperparameters used in our experiments in Tab 1. The other hyper-parameters are set to be the default of Threestudio [2].

A.1. Background scene

Following DreamFusion [9], we also implement an environment map MLP that takes the positionally-encoded ray direction as input and similarly predicts the RGB color value $C_{bg}(\mathbf{r})$ for each rendered ray. We will then composite the previously acquired RGB color value $C(\mathbf{r})$ for each rendered ray on top of this background feature with the accumulated alpha value:

$$C'(\mathbf{r}) = C(\mathbf{r}) + (1 - \sum_i W_i) C_{bg}(\mathbf{r}), \tag{1}$$

where the weight W_i follows the definition in Eq. (2) in the main paper.

B. Video results

B.1. Video results of DreamAvatar

To better visualize the generated results, we offer an improved demonstration of our method through rotated videos in the supplementary materials. To access this demonstration, please open the file named **"index.html"** provided in the supplementary.

B.2. Video comparisons with text-to-3D baselines

Additionally, we offer visual comparisons with existing text-to-3D baselines in the form of rotated videos. These evaluations encompass diverse text prompts, and the corresponding video can be found in the supplementary materials under the file name **"comparison-baselines.mp4"**.

B.3. Video comparisons with avatar-specified baselines

We present video comparisons involving avatar-specific baselines. The rotated evaluations can be found in **"comparison-avatar.mp4"** in the supplementary materials.

B.4. Video comparisons under different poses

Considering that Latent-NeRF [8] can incorporate 3D priors with diverse poses for 3D generation, and TEXTure [10]'s ability to generate texture for 3D human shapes in different poses, we extend our analysis by providing video comparisons with these methods. These visualizations can be found in the supplementary materials under the file name **"comparison-pose.mp4"**.

C. Further Analysis

C.1. Effects of non-rigid motion

The deformation field that we have implemented serves a crucial role in constructing and optimizing our dual-observation spaces. It comprises two distinct components: articulate deformation and non-rigid motion. In order to showcase the effectiveness of our design, we disable the non-rigid motion and present the comparisons in Fig. 1. The results indicate that non-rigid motion can be instrumental in improving the robustness of pose control.

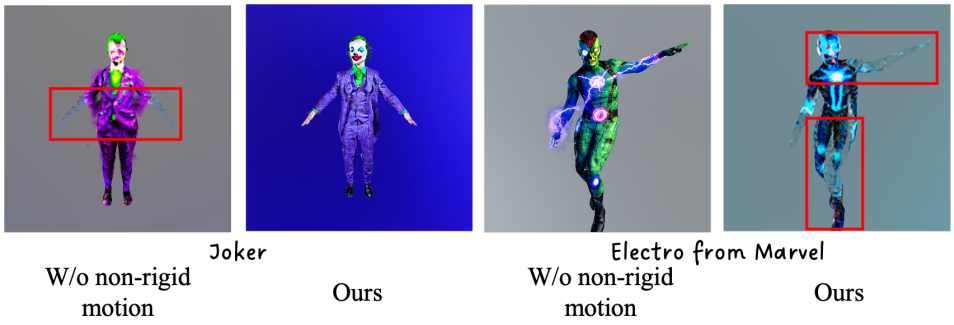

Figure 1. **Effects of non-rigid motion.**

D. Additional qualitative comparisons

D.1. Additional qualitative comparisons with text-to-3D baselines

In addition to the comparisons in the main paper, we further compare our method with four text-to-3D baseline methods, i.e., DreamFusion [1, 6, 8, 9]. The results of this comparison, presented in Fig. 2-Fig. 4, indicate that our method consistently achieves topologically and structurally correct geometry and texture compared to baseline methods, and outperforms the avatar-specified generative methods with much better and higher-resolution texture and geometry.

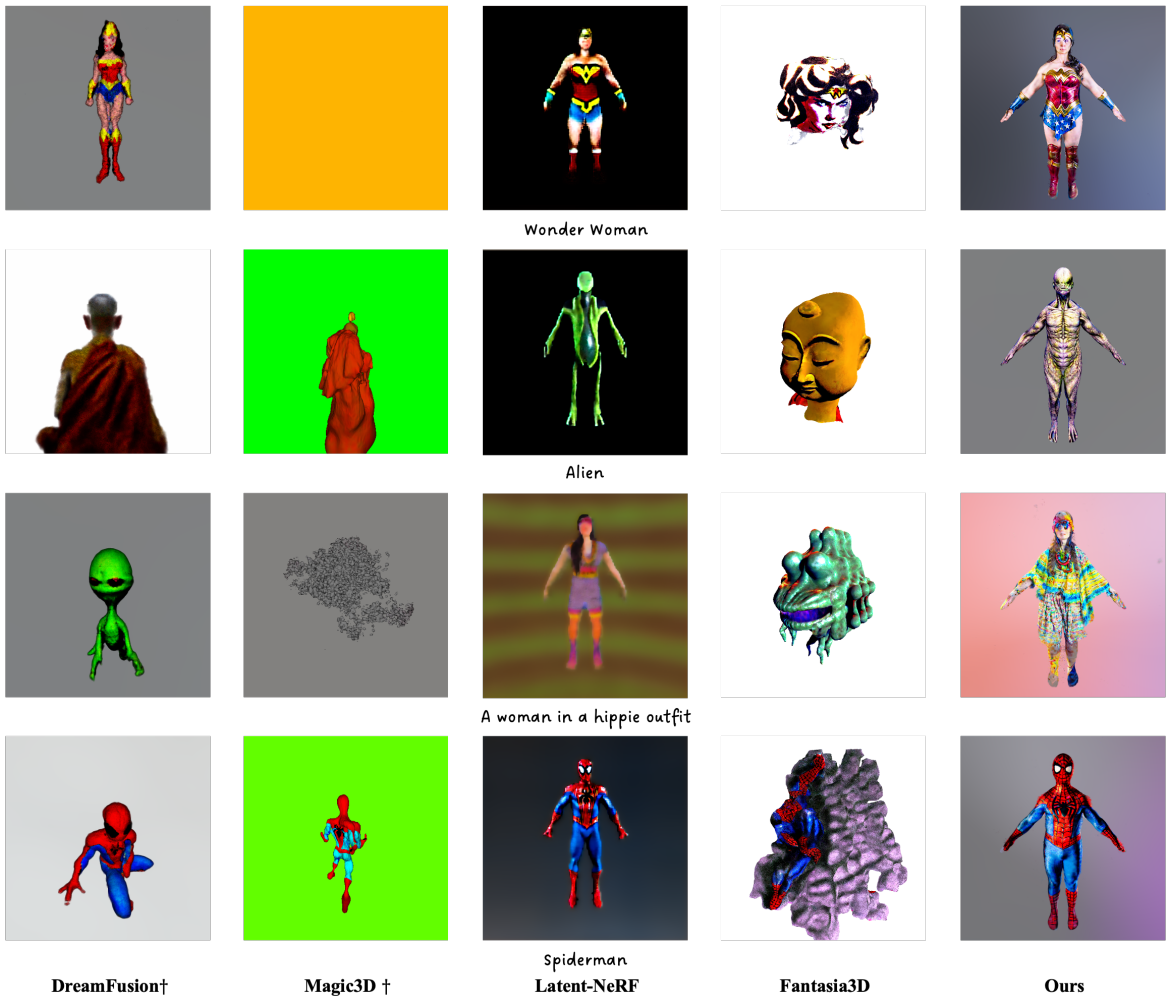

Figure 2. **Additional qualitative comparisons with text-to-3D baselines (Part I)**

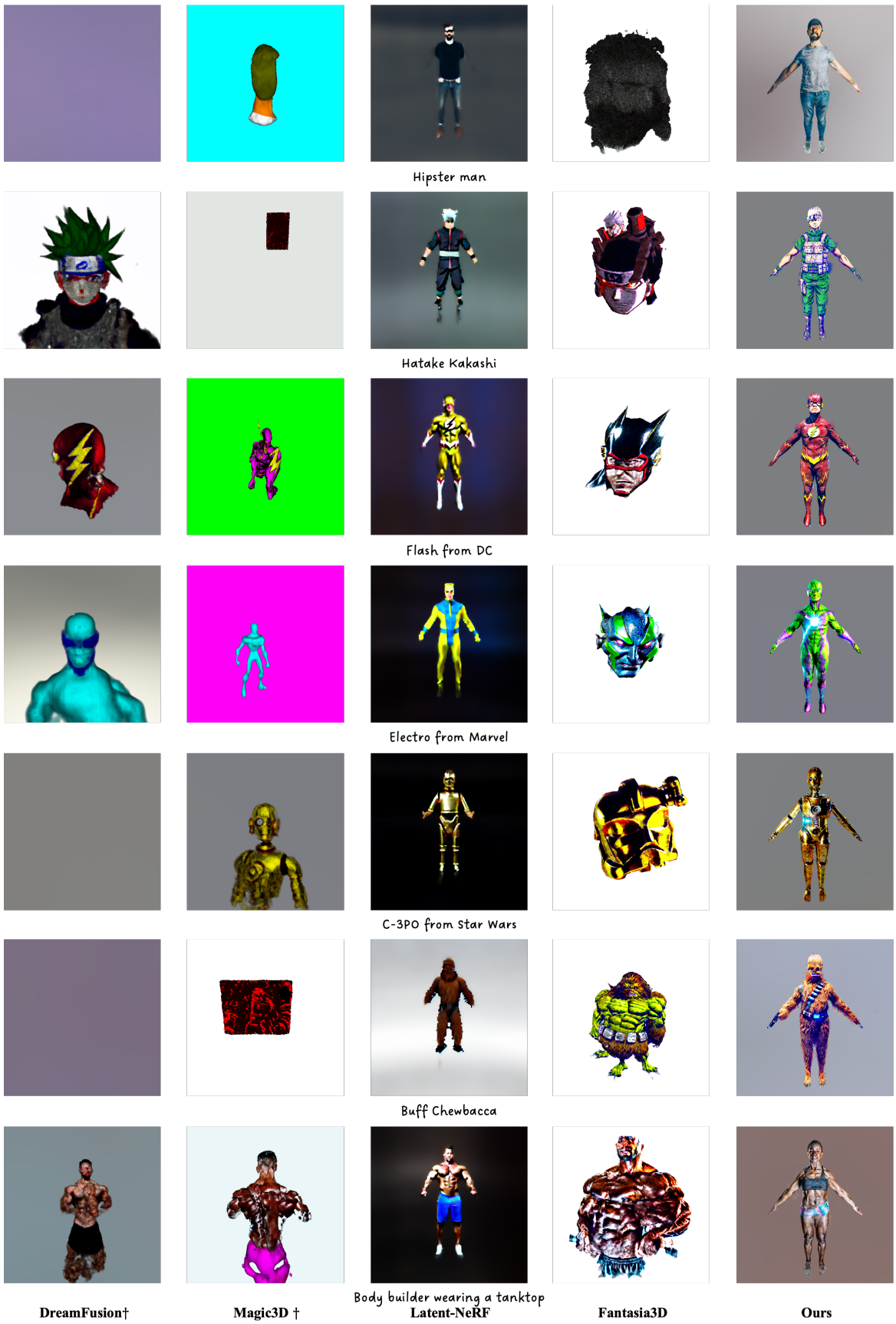

Figure 3. **Additional qualitative comparisons with text-to-3D baselines (Part II)**

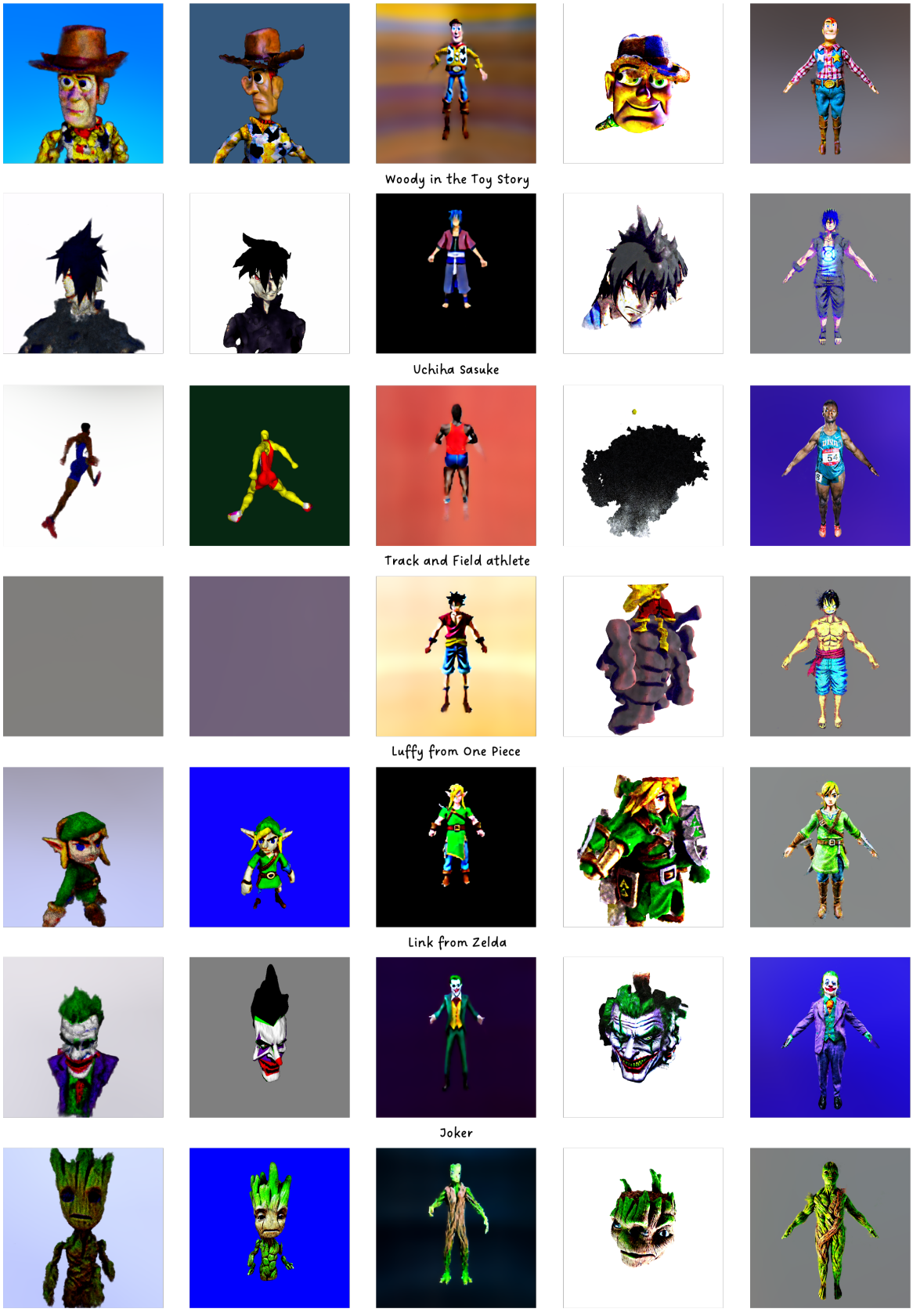

Figure 4. **Additional qualitative comparisons with text-to-3D baselines (Part III)**

D.2. Additional qualitative comparisons with avatar-specified baselines

In Fig. 5-Fig. 7, we provide more qualitative comparisons with avatar-specified baselines. We have not included TADA! [5] due to the unavailability of their publicly released code (we obtained the results in the main paper directly from the authors). We choose to compare ProlificDreamer [11] in this set. ProlificDreamer shares the Variational Distillation Sampling (VSD) optimization strategy, which allows us to effectively demonstrate the performance improvement achieved through our design.

Through the visualizations, we can observe that: **(1)** Our method outperforms DreamWaltz [4] in generating 3D human avatars with superior resolution in both texture and geometry; **(2)** AvatarCLIP [3] is highly constrained to the SMPL [7] prior, limiting its capability to generate true clothing topology; **(3)** ProlificDreamer lacks the proper constraints and hence always generate 3D human avatars with topologically and structurally incorrect geometry.

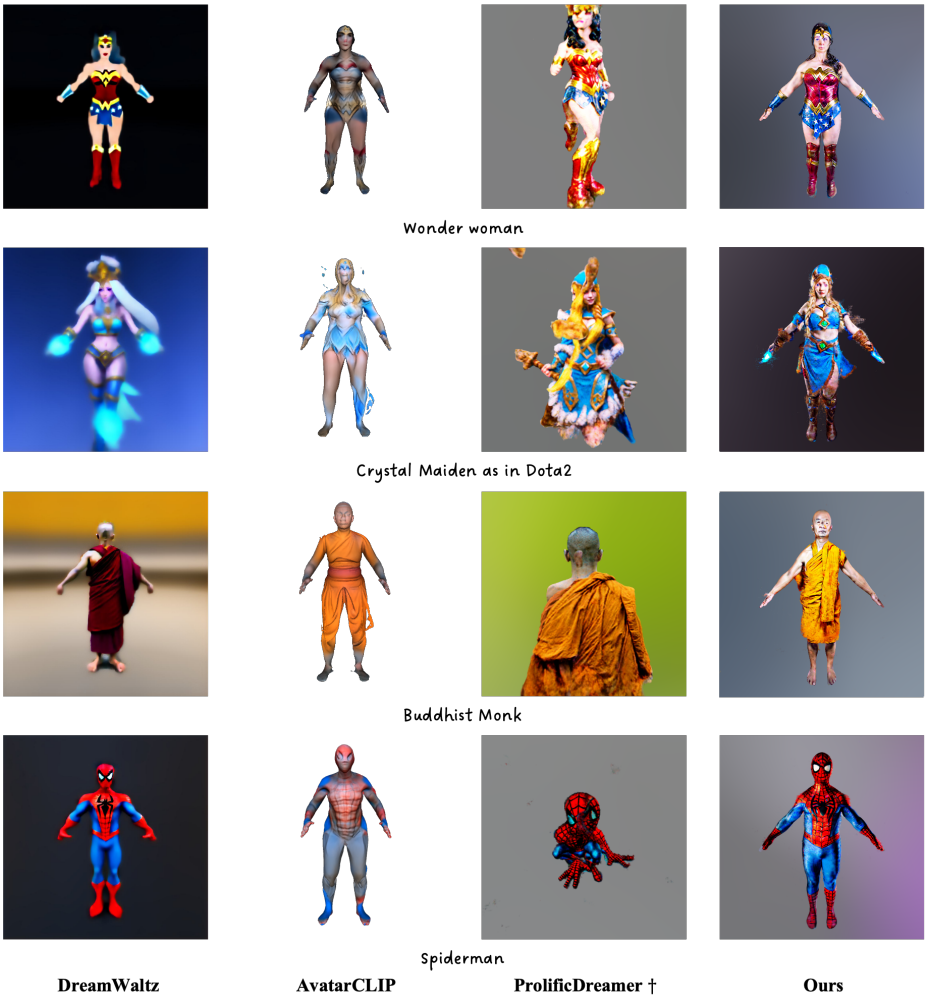

Figure 5. **Additional qualitative comparisons with avatar-specified baselines (Part I)**

Figure 6. **Additional qualitative comparisons with avatar-specified baselines (Part II)**

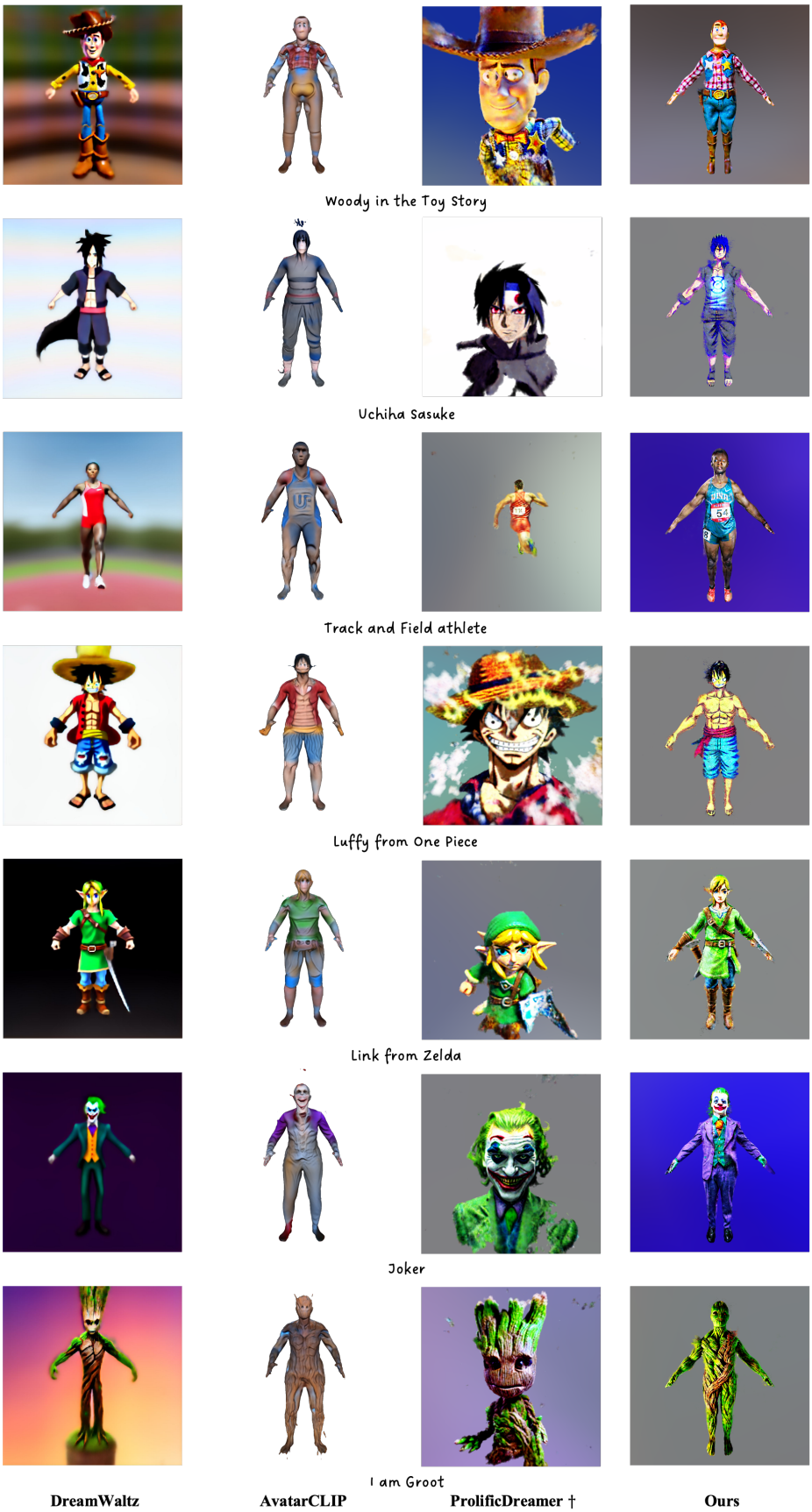

Figure 7. **Additional qualitative comparisons with avatar-specified baselines (Part III)**

D.3. Additional qualitative comparisons under different poses

Considering that (1) Latent-NeRF [8] can incorporate 3D priors with diverse poses for 3D generation, and (2) TEXTure [10] is designed to generate textures for 3D human shapes in different poses, we further provide qualitative comparisons with them under different poses. The visualizations provided in Fig. 8-Fig. 9 demonstrate that our method can robustly generate 3D human avatars under various poses, and largely outperforms existing state-of-the-arts.

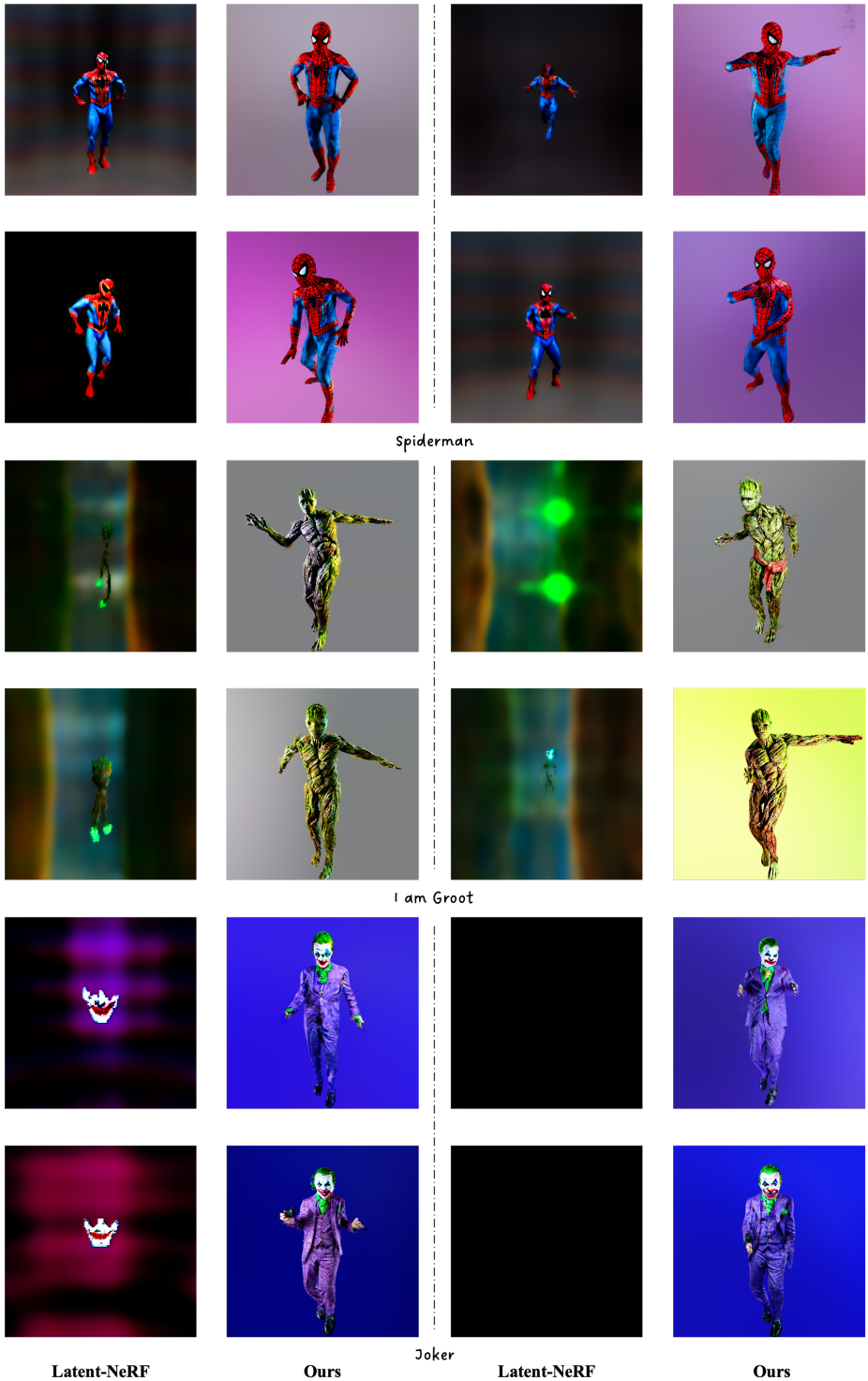

Figure 8. **Additional qualitative comparisons under different poses (Part I - *v.s.* Latent-NeRF).**

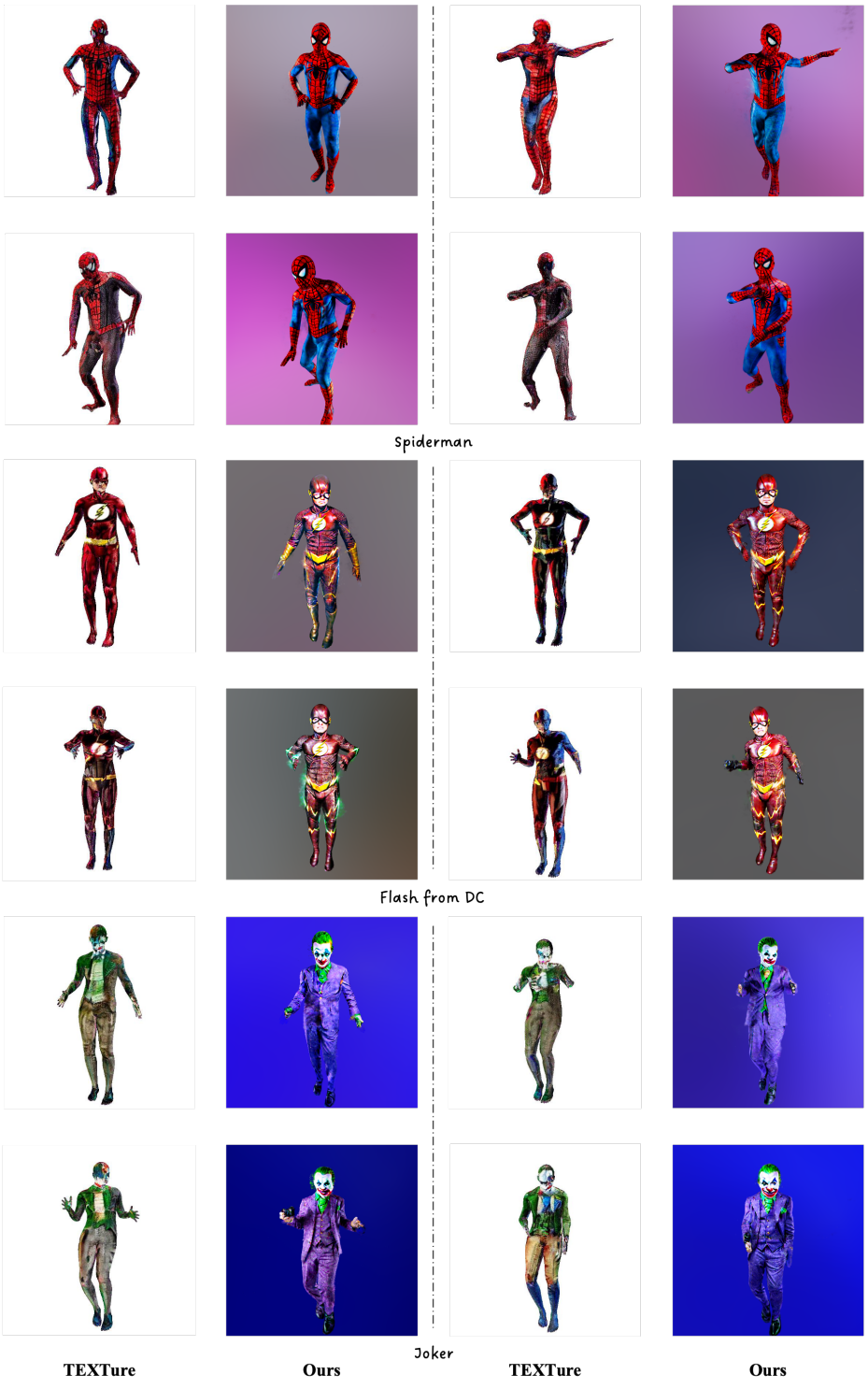

Figure 9. **Additional qualitative comparisons under different poses (Part II - *v.s.* TEXTure).**

References

[1] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. 2023. 5

[2] Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio, 2023. 2

[3] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *ACM Transactions on Graphics*, 2022. 8

[4] 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. 2023. 8

[5] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxaing Tang, Yangyi Huang, Justus Thies, and Michael J Black. Tada! text to animatable digital avatars. *arXiv preprint arXiv:2308.10899*, 2023. 8

[6] 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. 5

[7] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *TOG*, 2015. 8

[8] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. *arXiv preprint arXiv:2211.07600*, 2022. 3, 5, 11

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

[10] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. *arXiv preprint arXiv:2302.01721*, 2023. 3, 11

[11] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 8