# DiffPortrait360: Consistent Portrait Diffusion for 360 View Synthesis

## Supplementary Material

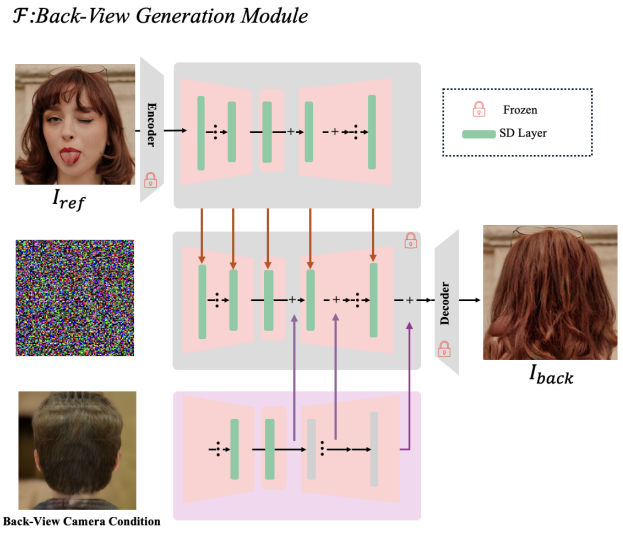


Figure 1. Illustration of our back-view generation module $\mathcal{F}$. Given an arbitrary back-view camera condition and a reference image, we follow the methods of the reference module and ControlNet to decode a specific camera view. During inference, we set the back-view camera generation condition to 180 degrees to maximize the capture of appearance information from the back view.

We provide additional implementation details in Section A, highlight more comparisons and results in Section B, and provide additional limitations in Section C.

# A. Implementation Details

## A.1. Back-view Generation Module

We illustrate the framework of our back-view generation module $\mathcal{F}$ in Fig. 1 along with its training data examples in Fig. 2, which is used to enhance the generative capabilities of $\mathcal{F}$. Specifically, we employ Stable Diffusion SD1.5 [13] as the backbone. A reference network [3] is utilized to inject information from the front face, and camera control is managed by ControlNet [15]. In practice, we set a fixed camera view of the back head. In particular, we uniformly generate a fixed view that captures the maximum back appearance information, allowing the entire back-view generation module to focus on back-view appearance generation.

For the stylized augmentation dataset used to train the back-view generation module $\mathcal{F}$, we generate 2,000 subjects that are various stylized front portraits using [5] and further produce ground truth back-view by [14]. All data were processed with a cropped resolution of $512 \times 512$. Some of the lower-quality data were filtered out using [4].

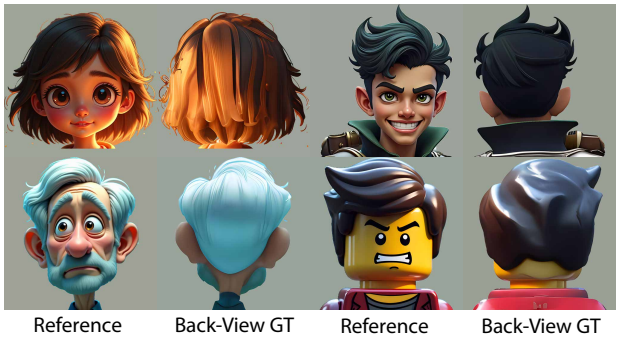


Figure 2. Examples in our stylized data augmentation which includes extensive generation of stylized back appearances. Compared to training back-view generation module $\mathcal{F}$ solely on real or synthetic networks, such augmentation helps our back-view generation module achieve greater generalizability.

## A.2. Training and Evaluations

Our model keeps the weights of the original Stable Diffusion model frozen during training. The training was conducted in stages, where we sequentially integrate and train modules $\mathcal{R}$, $\mathcal{C}$, and $\mathcal{V}$ mentioned in main paper Figure 2. All training were conducted on 6 NVIDIA RTX A6000 ADA GPUs with a learning rate of $10^{-5}$; we performed 60,000 iterations with the enhanced dataset for dual appearance and an additional 60,000 iterations for the camera control stages and the sequential training stage each, using only the PanoHead data due to unachievable camera intrinsics of the back view and a limited number of sparse camera views. For test inference, we collect 200 challenged portraits from Midjourney and Pexels [10, 12], containing a wide variation in appearance, expression, camera perspective, and style. For comparison in RenderMe360[11], we use another unseen 500 multi-view image pairs with different expressions.

## A.3. Dataset Details

Our dataset has 800 unique subjects: 150 from RenderMe360 and 600 from PanoHead/SphereHead. We use 150 from RenderMe360, excluding those with complex head accessories to avoid unwanted artifacts. RenderMe360 is not used for sequential training due to sparse camera views. For the back-head generator, we use 1,800 subjects: 150 from RenderMe360 (real-world), 650 from PanoHead/SphereHead (synthetic), and 1,000 from Unique3D (stylized).


Figure 3. Ablation study on camera conditioning.

# B. More Results

## B.1. More Ablation Study

As shown in Fig. 3, our experiments show image-based view control is more accurate than naive camera pose representation via text embedding. The effectiveness of alternative latent pose representations remains unclear, and explicit pose estimation requires extensive high-quality data, which we lack. However, we compare both methods using the POSE metric in Tab. 1 of DiffPortrait3D [14]. (POSE↓ Our Image Cond.: **0.0018**, Cam Param.: 0.0081).

## B.2. More Qualitative Comparisons

We conduct more qualitative comparisons with PanoHead [1], SphereHead [8], Unique3D [14], Rome [7] and Itseez3D[1] on stylized portraits in Fig. 6, where our method shows superior generalization capability on diverse styles. We also show additional qualitative comparisons with PanoHead [1], SphereHead [8], Zero-1-to-3 [9], Unique3D [14] and DiffPortrait3D [6] on RenderMe360[11] with paired ground truth results in Fig. 7.

## B.3. Comparing DiffPortrait3D on More Views

Please find more comparisons with DiffPortrait3D [6] on more views in Fig. 4.

## B.4. More Results

Finally, more visual results of our method are introduced in Fig. 8, Fig. 9 and Fig. 10.

# C. Limitation and Future Work

Our experiments prove unlike the temporal fusion layer of DiffPortrait3D trained on random views, our approach using continuous improves local consistency which is crucial for 3D tasks. To this end, we also show that our emphasis on achieving view consistency is crucial for constructing a NeRF representation, enabling real-time free-viewpoint rendering from any camera position. While our method outperforms all state-of-the-art techniques, it still exhibits small inconsistencies for certain portraits. This is due to the inherent 2D nature of stable diffusion generation, which remains frozen during training to maintain stability. In the future, we plan to either incorporate geometric priors explicitly into the diffusion-based view synthesis process or extend our framework by directly encoding multi-view images into visual patches similar to, e.g., SORA[2] and explore the use of differentiable rendering techniques and efficient radiance field representations, such as 3D Gaussian Splats. Additionally, our method currently struggles with certain types of headgear, such as various hats and unseen hairstyles shown in Fig. 11, due to a biased distribution in our training data. We believe that additional data collection is likely to improve the performance. Finally, as our generated heads are currently static, future directions include making them animatable and relightable, and the cropping size of the head area should be expanded to accommodate scenarios involving longer hair.

[1] https://itseez3d.com/

# References

[1] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y. Ogras, and Linjie Luo. Panohead: Geometry-aware 3d full-head synthesis in 360deg. In *CVPR*, pages 20950–20959, 2023. 2

[2] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. *N/A*, 2024. 2

[3] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. *arXiv preprint arXiv:2304.08465*, 2023. 1

[4] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5154–5163, 2020. 1

[5] Flux AI. Flux ai: Design for hardware engineers. https://flux-ai.io/?gad_source=1&gclid=CjwKCAiA3Na5BhAZEiwAzrfagAN-MUBK_2j_msa4PEJA7TqbF8_CkYQEFis4H8biY4osqRvQ1EeQRBoCJlsQAvD_BwE. Accessed: 2024-11-15. 1

[6] Yuming Gu, Hongyi Xu, You Xie, Guoxian Song, Yichun Shi, Di Chang, Jing Yang, and Linjie Luo. Diffportrait3d: Controllable diffusion for zero-shot portrait view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10456–10465, 2024. 2, 3

[7] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. Realistic one-shot mesh-based head avatars. In *European Conference on Computer Vision*, pages 345–362. Springer, 2022. 2

[8] Heyuan Li, Ce Chen, Tianhao Shi, Yuda Qiu, Sizhe An, Guanying Chen, and Xiaoguang Han. Spherehead: Stable 3d full-head synthesis with spherical tri-plane representation, 2024. 2


Figure 4. More comparisons with DiffPortrait3D [6] on more views.

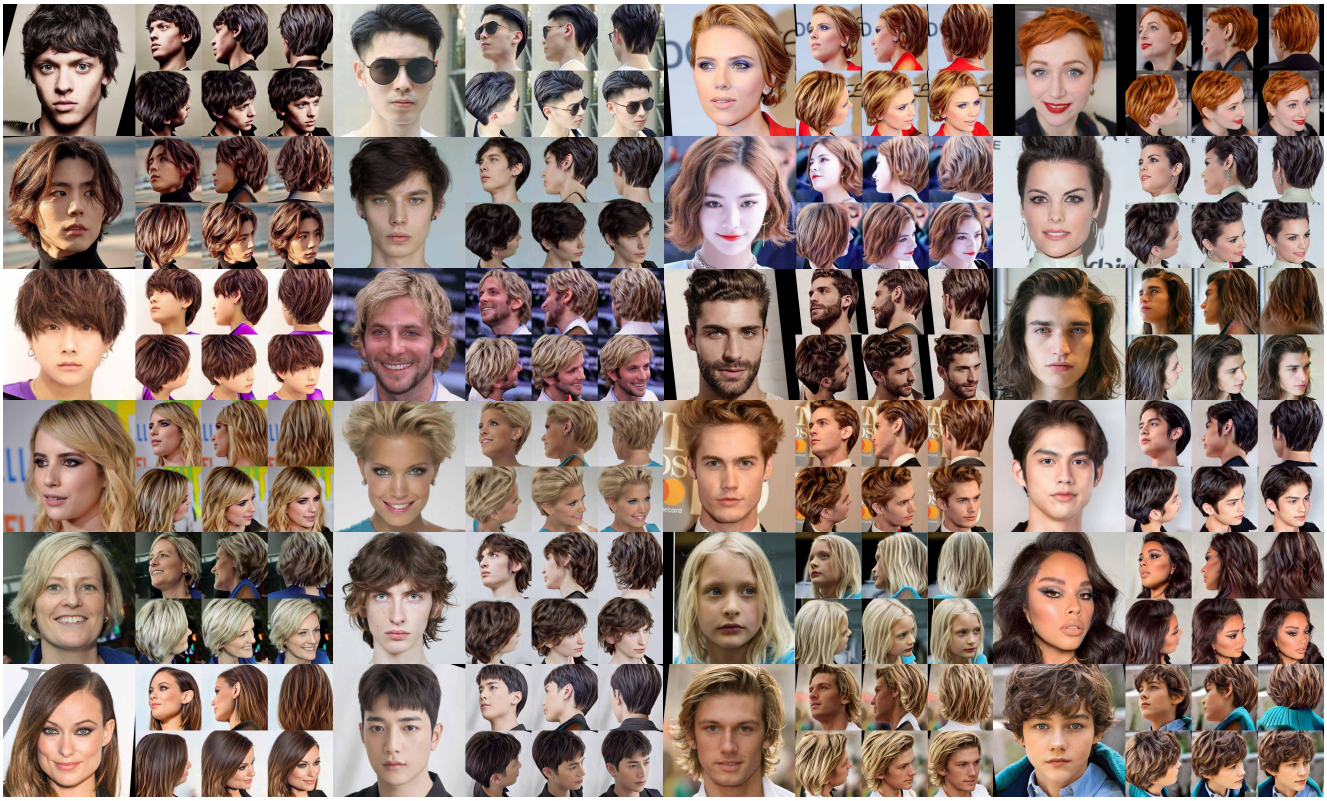

Figure 5. More real-world results.

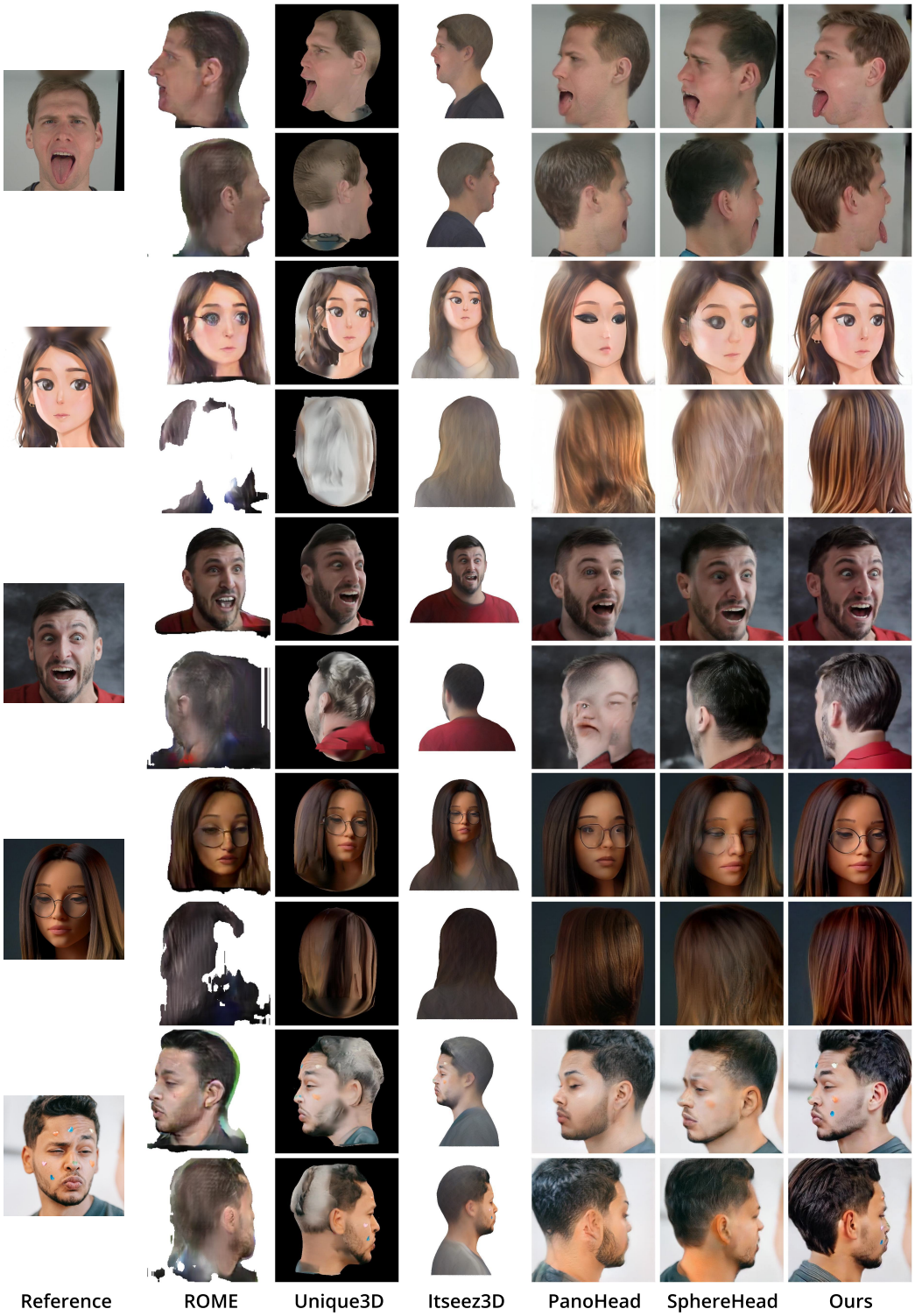


Figure 6. More qualitative comparisons with existing methods on in the stylized portraits. Our method shows superior generalization capability to novel view synthesis of wild portraits with unseen appearances, expressions, and styles, even without any reliance on fine-tuning.


Figure 7. More qualitative comparisons of novel view synthesis on RenderMe360 [11]. Our method achieves effective appearance control for novel synthesis under substantial change of camera view for synthesis.


[9] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. 2

[10] Midjourney. midjourney. *https://www.midjourney.com*, 2024. 1

[11] Dongwei Pan, Long Zhuo, Jingtan Piao, Huiwen Luo, Wei Cheng, Yuxin Wang, Siming Fan, Shengqi Liu, Lei Yang, Bo Dai, et al. Renderme-360: a large digital asset library and benchmarks towards high-fidelity head avatars. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 5

[12] Pexels. pexels. *https://www.pexels.com/*, 2024. 1

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz,

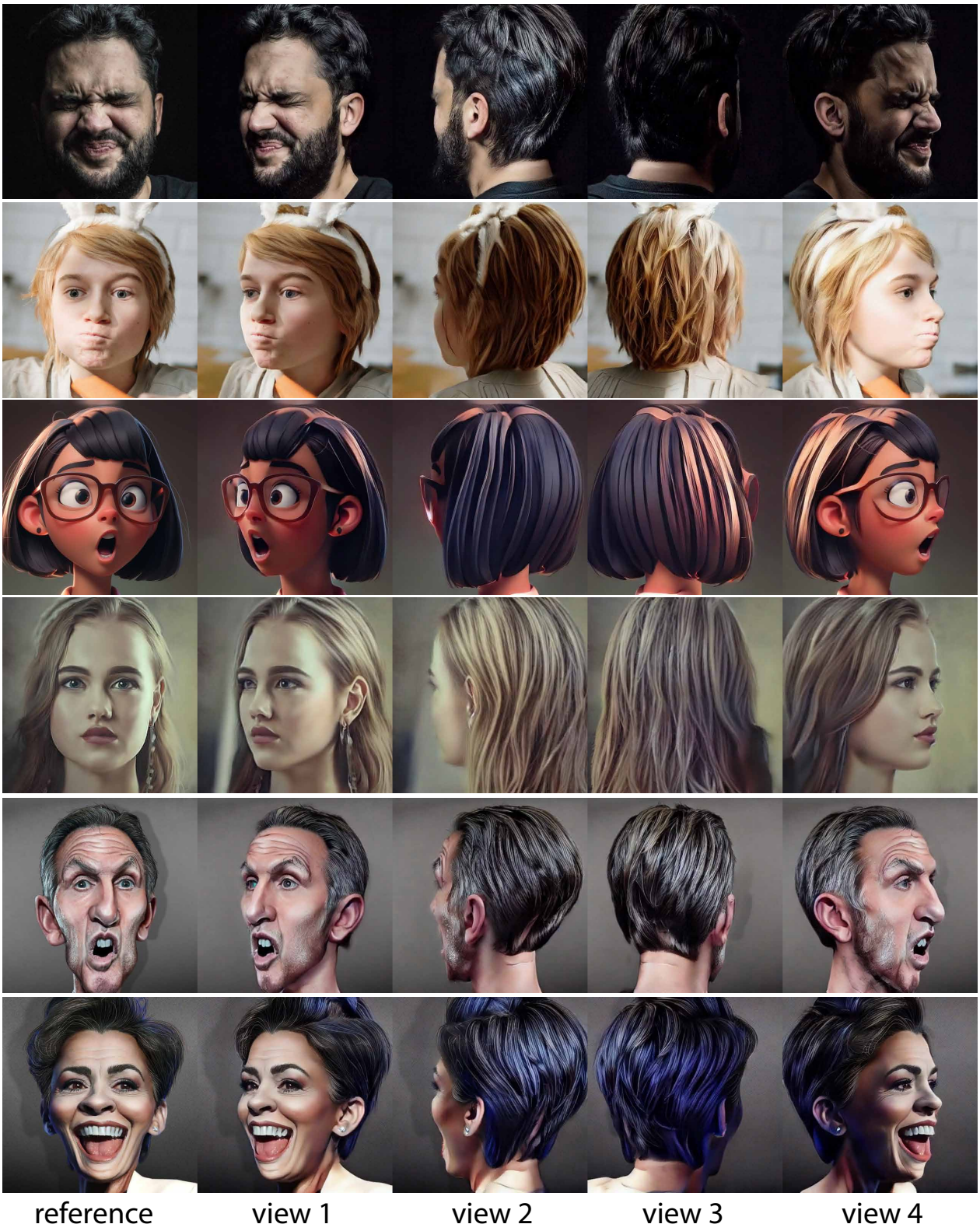


Figure 8. More qualitative results of our method.

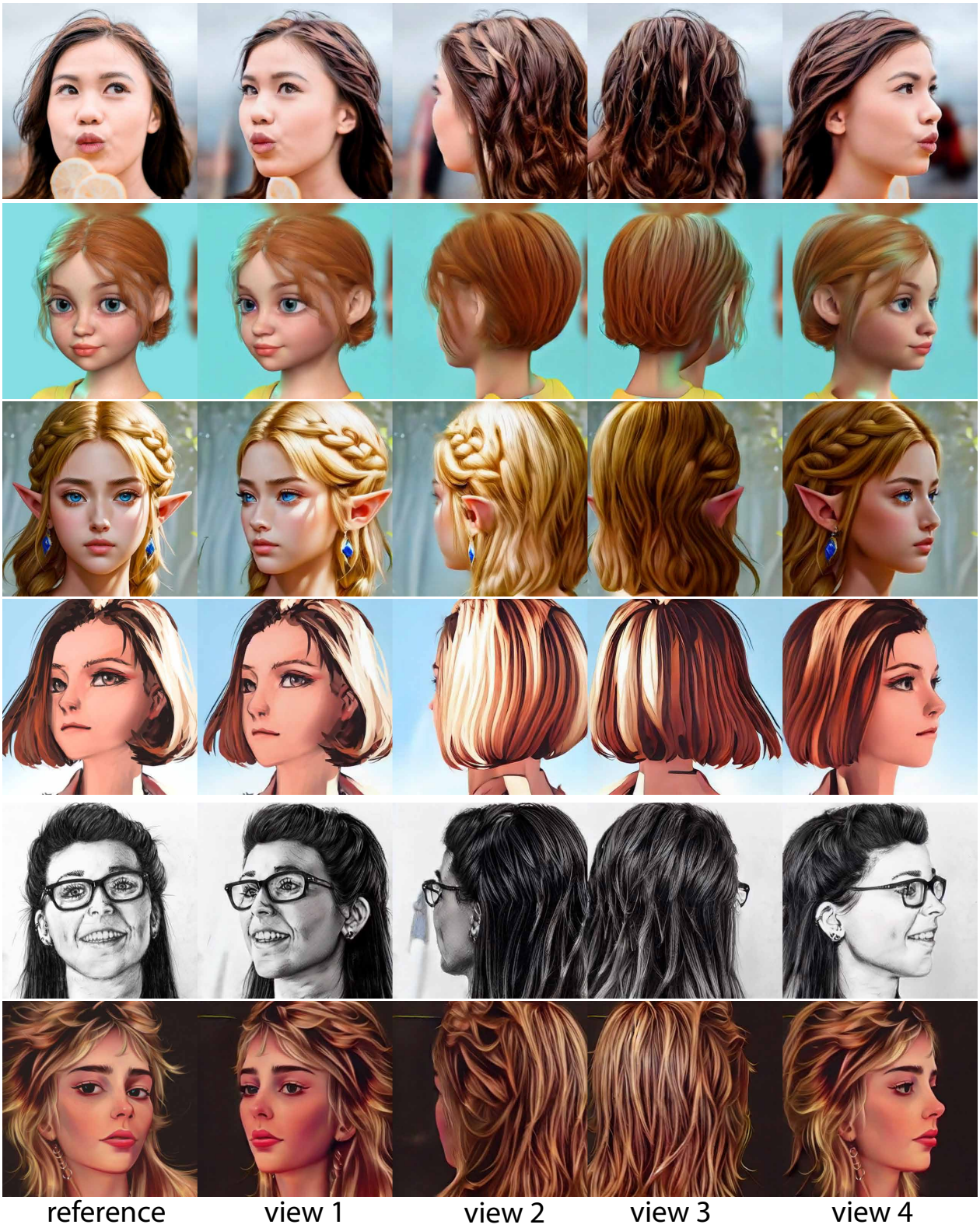


Figure 9. More qualitative results of our method.

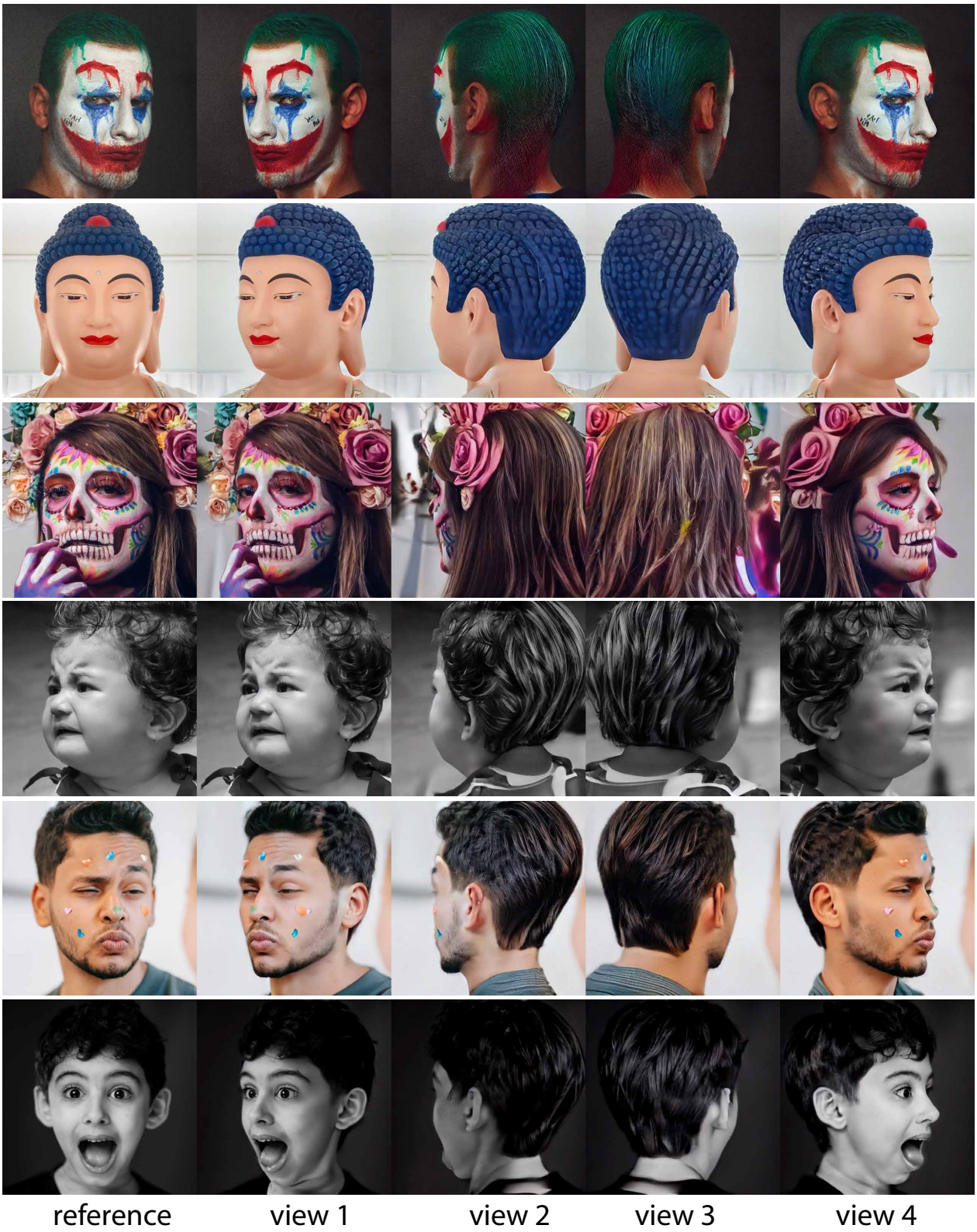


Figure 10. More qualitative results of our method.

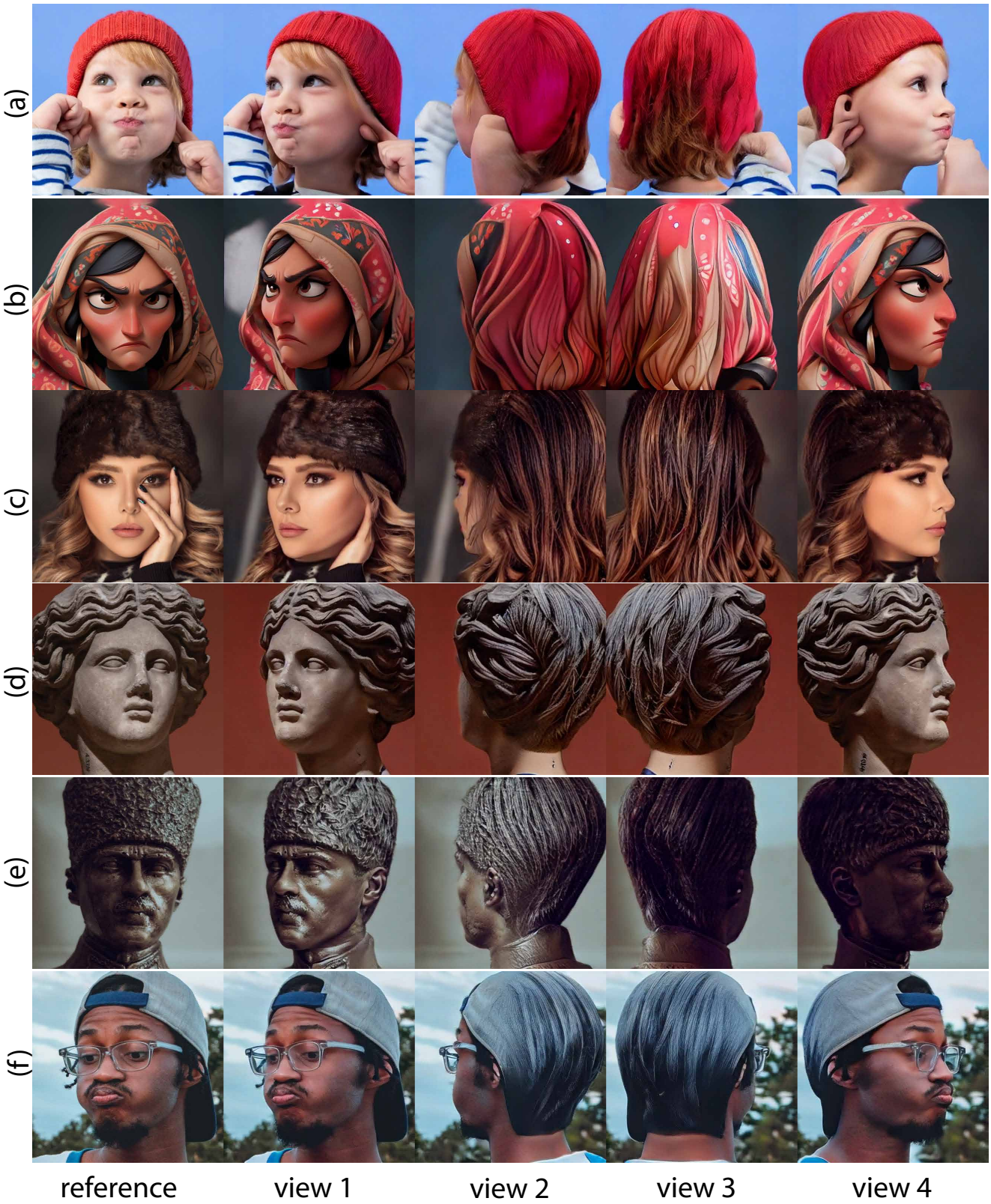


Figure 11. Limitations of our method.

Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1

[14] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *arXiv preprint arXiv:2405.20343*, 2024. 1, 2

[15] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 1