DiffPortrait3D: Controllable Diffusion for Zero-Shot Portrait View Synthesis

Supplementary Material

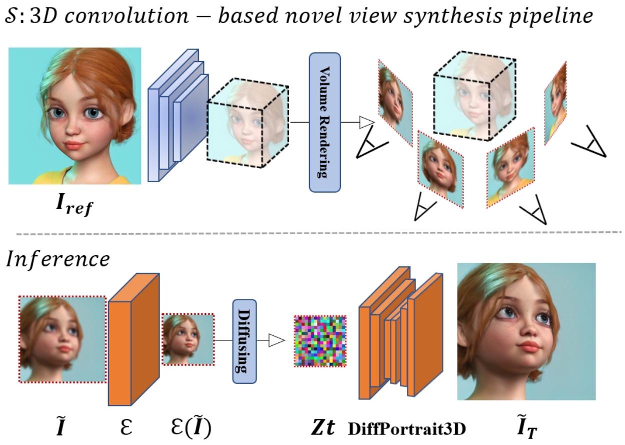

Figure 1. shows how our 3D convolution-based novel view synthesis pipeline $\mathcal{S}$ works. In practice, a 3D-convolution-based network first maps the reference image into a 3D feature volume. Then, given a conditioned camera view, we follow the volume rendering to integrate the 3D features into a 2D feature map which is further decoded to the final RGB image $\tilde{I}$ with a 2D convolution network. During the inference phase, we enhance 3D awareness by commencing with noise generated via a 1000-step forward diffusion process applied to $\mathcal{E}(\tilde{I})$, which serves as the initial noise for our DiffPortrait3D pipeline.

In this supplementary paper, we provide additional implementation details in Section A, showcase more visual results and numerical comparisons in Section B, and discuss limitations & ethics consideration in Section C and Section D .

A. Implementation Detail

A.1. 3D-Aware Noise

In Figure 1 and Figure 2, we illustrate the framework of generating our "3D-aware" noise. Specifically, we build a 3D convolution-based novel view synthesis pipeline (denoted as $\mathcal{S}$), trained as a multi-view image reconstruction task. Similar to [10] and [9], we first employ a 3D appearance feature extraction network to map the reference image to a 3D appearance feature volume. To synthesize an image at a novel view, we follow the volume rendering as in NeRF [7] to integrate the 3D features into a 2D feature map which is further decoded to the final RGB image $\tilde{I}$ with a deep 2D convolutional network. The network modules are trained with image reconstruction losses against ground-truth multi-view images, including pixel-aligned L_2 and VGG perceptual losses [4].

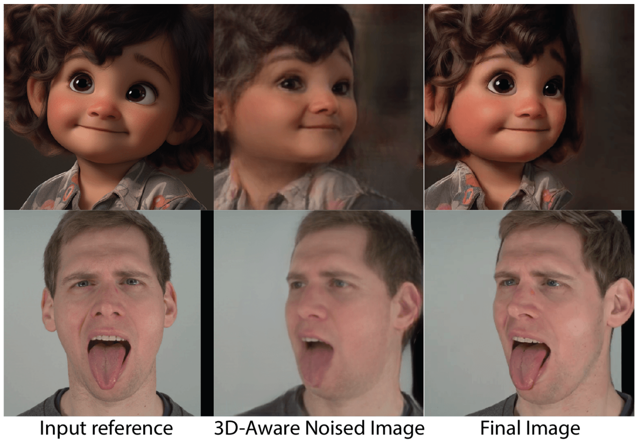

Figure 2. Our 3D-Aware Noise effectively helps strengthen the novel view synthesis result.

During inference, given a reference image, we first employ our trained 3D novel view synthesis network $\mathcal{S}$ to generate a proxy rendering $\tilde{I}$ at the target view. While being blurry, $\tilde{I}$ contains rich 3D structural semantics and acts as a good guidance to the diffusion process in our DiffPortrait3D. We incorporate this 3D awareness by generating the starting noise using the forward noising process of 1000 steps applied to the latent map of $\tilde{I}$, i.e., $\mathcal{E}(\tilde{I})$. Better reconstruction and consistency are observed with our proposed 3D-aware noise, as evidenced in Table 1 numerically and in Figure 5 of the main paper visually.

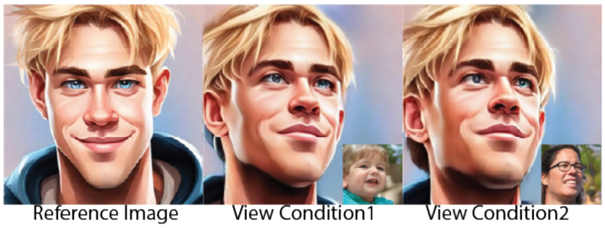

Figure 3. Ablation on view conditional images.

A.2. Metrics

A.2.1 Identity Similarity

Our identity similarity score (ID) is calculated based on the cosine similarity of the face embeddings with a pre-trained face recognition module[1] as ,

$$ID = f_{g'} \cdot f_g \tag{1}$$

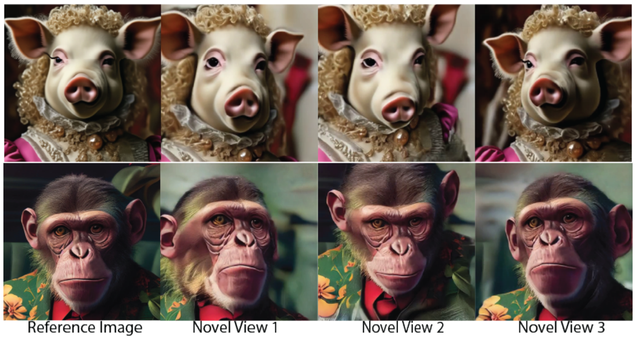

Figure 4. Novel view synthesis of anthropomorphic animals.

where $f_{g'}$, and f_g are the feature embedding of the generated image and ground-truth image respectively.

A.2.2 Pose Accuracy

We evaluate the pose accuracy (POSE) with the assistance of an off-the-shelf face reconstruction model [2]. We detect pitch, yaw, and roll from the generated novel view images, then compute the L_2 loss against the camera poses estimated from ground truth images.

A.3. Baselines

A.4. EG3D-Pivot Tuning Inversion

For our baseline EG3D-PTI, we follow the standard procedure as described in [8], where for each reference image, we first optimize the latent noise for 500 iterations and further fine-tune the generator weights for another additional 250 iterations. Once completed, we used the optimized latent noise finetuned 3D-aware generator to synthesize the image at novel views.

A.5. Zero-1-to-3

Zero-1-to-3 [6] is one of the state-of-the-art novel diffusion-based view synthesis works designed for general 3D objects. Nevertheless, we compare to it for a thorough evaluation of existing works on novel portrait synthesis. In Table 2, to maximize its performance on our task, our numerical results are all based on portraits with removed backgrounds. Additionally, due to the differences in 3D camera coordinates, we only report the FID and identity similarity (ID) of the novel view synthesis results for a fair comparison.

A.6. PanoHead-Pivot Tuning Inversion

PanoHead extends the EG3D framework by enabling novel view synthesis in 360°. However, owing to the inherent limitations of GAN-based architecture, PanoHead, like EG3D, necessitates time-consuming instance-specific optimization (pivot-tuning) while still suffers from limited perceptual quality and identity loss, especially for portraits with out-of-domain styles or extreme expressions (as shown in Figure 6 compared to our results in Figure 4 of the main paper).

	3D-Aware noise	Random noise
LPIPS ↓	**0.27**	0.32
SSIM ↑	**0.68**	0.65
DIST ↓	**0.18**	0.21
ID ↑	**0.70**	**0.70**
FID ↓	**25.37**	26.81

Table 1. Quantitative ablation of 3D-Aware Noise and Random Noise results of novel view synthesis of NeRSemble [5] at the resolution of 512x512

B. More Experiment results

B.1. Alignments

Our model was trained with EG3D-aligned reference and target images. However, our method does not restrict the reference images to be cropped and aligned, nor with the camera condition images. In Figure 5, we showcase that with differently aligned reference images, our method synthesizes close novel view results.

B.2. View-consistent novel view synthesis

We show more challenging results in Figure 8 , 9 and 10. Our model is able to generalize well to arbitrary face portraits with unposed camera views, extreme facial expressions, and diverse artistic depictions. Please also refer to our supplementary video for more high-resolution results.

B.3. Ablation on view condition images

Our method effectively disentangles the control of camera views from appearance. As shown in Figure 3, visual differences are hardly noticeable between the synthesized novel views when using two view conditions generated at the same camera pose but with distinct appearance seeds. For quantitative assessment, we perform novel view synthesis across all our test images using two sets of view conditional images generated under the same camera pose but featuring different appearances. We calculate the differences in image pixels (LPIPS ↓ 0.09) and camera poses (POSE↓ 0.0041). Note that the LPIPS difference is partially attributed to slight structural shifting.

B.4. Anthropomorphic animals

While being trained sorely on real human images, our method is empowered with strong domain generalization

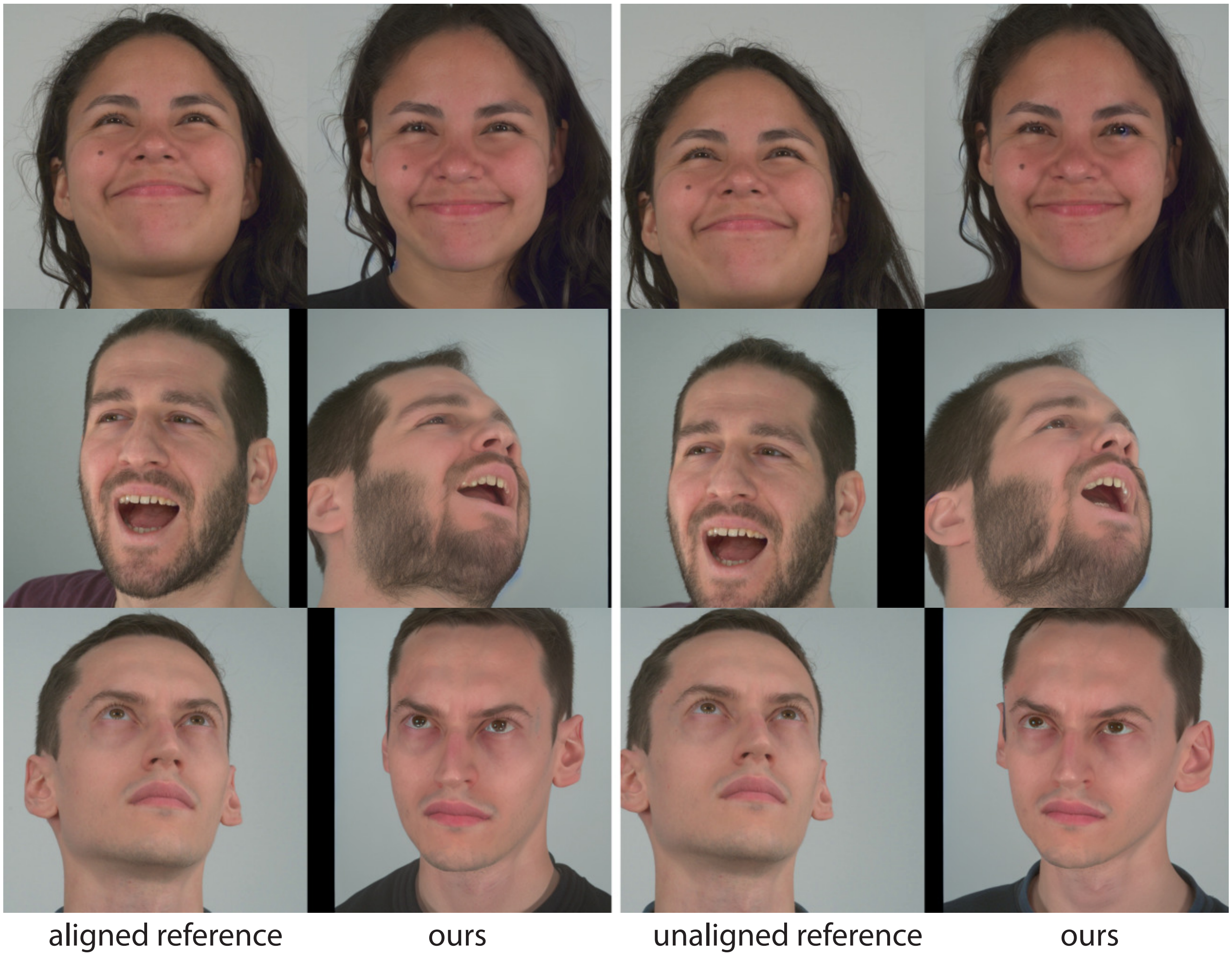

Figure 5. DiffPortrait3D effectively derives appearance features from the reference image, without strict restriction to its image alignment. Similar novel view synthesis results are achieved using EG3D-aligned and non-aligned reference images.

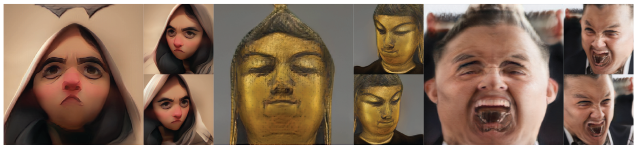

Figure 6. Novel view synthesis with PanoHead-PTI.

capability (e.g., Figure 4) by leveraging the generative prior of a pre-trained stable diffusion model. However, we acknowledge that visual artifacts are possible due to the appearance bias originating from the training data distribution.

C. Limitation and Future Work

While the image coherence is largely strengthened with our cross-view module and 3D-aware generation, we still observe occasionally flickering artifacts in unobserved regions. We leave the exploration of longer-range consistent view manipulation as future work. In this work, the appearance is formulated to be sourced from the reference images only. This could result in some loss of identity given the limited appearance context. In the future, we would like to extend our framework such that the identity can be multi-sourced from e.g., text and personalized Loras [3]. As discussed above, we also include visualizations of failure cases in Figure 7. Rows (a) through (f) display artifacts in areas not observed, accompanied by changes in identity, particularly noticeable in (b), (c), and (d). In cases (a), (e), and (f), it is evident that the model struggles to accurately replicate secondary elements from the reference image, such as sunglasses in (f), hands and flowers in (e), and leaves in (a), leading to inconsistent outcomes. One potential issue we've identified stems from the use of a 2D diffusion backbone that integrates 3D-Aware information. This approach, while innovative, may lead to minor inconsistencies and deviations, especially in challenging depictions. Addressing

	Ours	Zero1to3	PanoHead-PTI
LPIPS ↓	**0.04/0.19/0.04**	0.28/-/0.34	0.22/0.28/0.11
SSIM ↑	**0.90/0.74/0.88**	0.70/-/0.52	0.60/0.53/0.76
DIST ↓	**0.05/0.15/0.04**	0.16/-/0.19	0.18/0.26/0.12
ID ↑	**0.94/0.70/0.92**	0.82/0.09/0.70	0.47/0.38/0.12
FID ↓	**6.5/32.6/11.6**	61.4/113.8/57.5	56.53/60.4/90.47

Table 2. Quantitative comparison of our method, PanoHead-PTI and Zero-1-to-3, showing numerical results of reconstruction/novel view synthesis of NeRSemble [5], and reconstruction of in-the-wild test images(from left to right). For a fair comparison to Zero-1-to-3, the evaluation is performed with the removed backgrounds at the resolution of 512×512. PanoHead-PTI remains the same setting as EG3D-PTI in the main paper.

these limitations is an important area that should be addressed in future work.

D. Ethic Consideration

We acknowledge the profound capabilities of the diffusion model as a powerful generative model. The framework proposed in our paper could, theoretically, be utilized to compromise multi-perspective facial recognition systems. We assert that the model and the accompanying research code are intended exclusively for advancing scientific research and must not be used for illicit purposes.

References

[1] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 1

[2] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE Computer Vision and Pattern Recognition Workshops*, 2019. 2

[3] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[4] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 1

[5] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics*, 2023. 2, 4

[6] 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

[7] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1

[8] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics*, 42(1):1–13, 2022. 2

[9] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019. 1

[10] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1

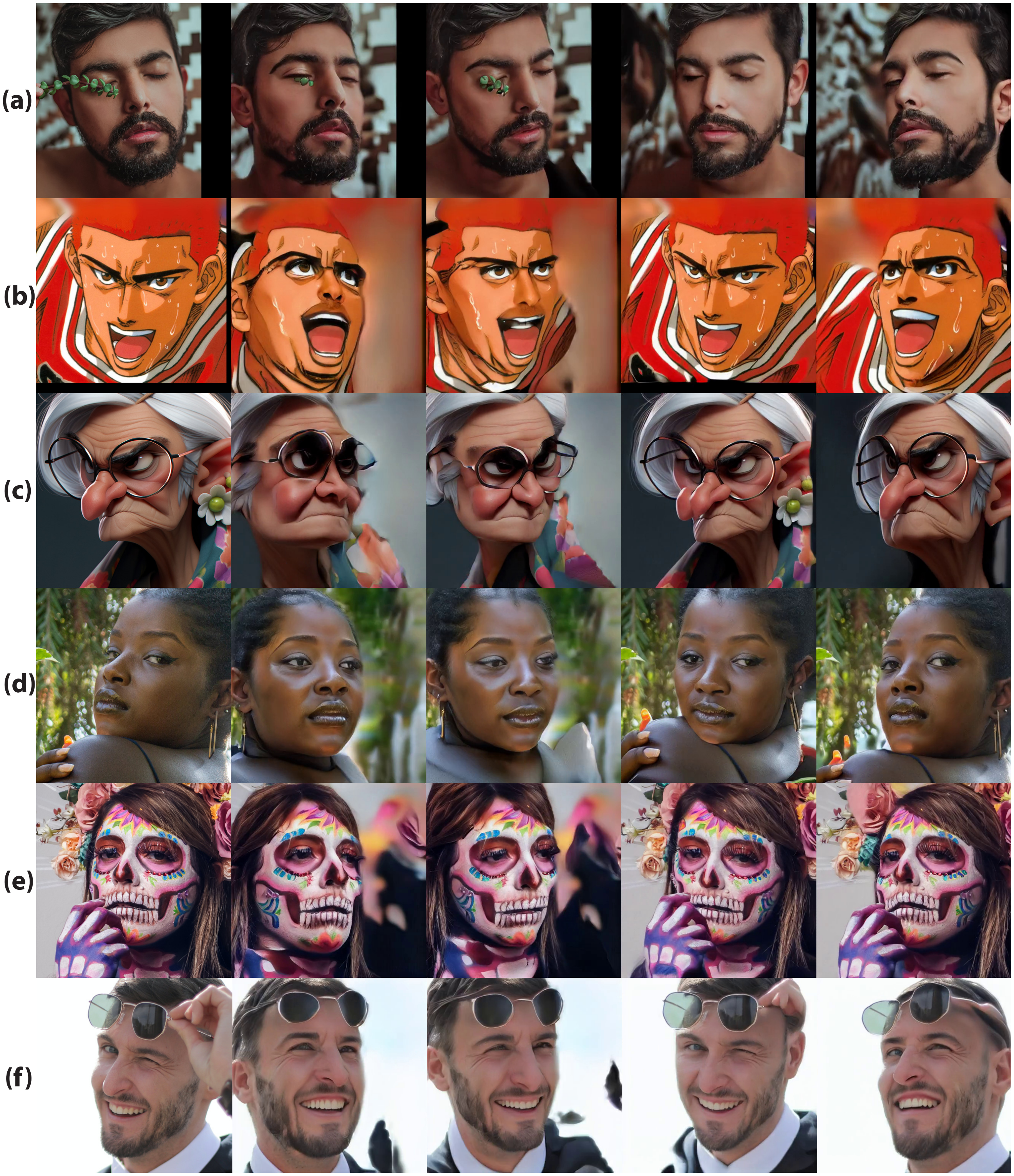

Figure 7. Failure Case

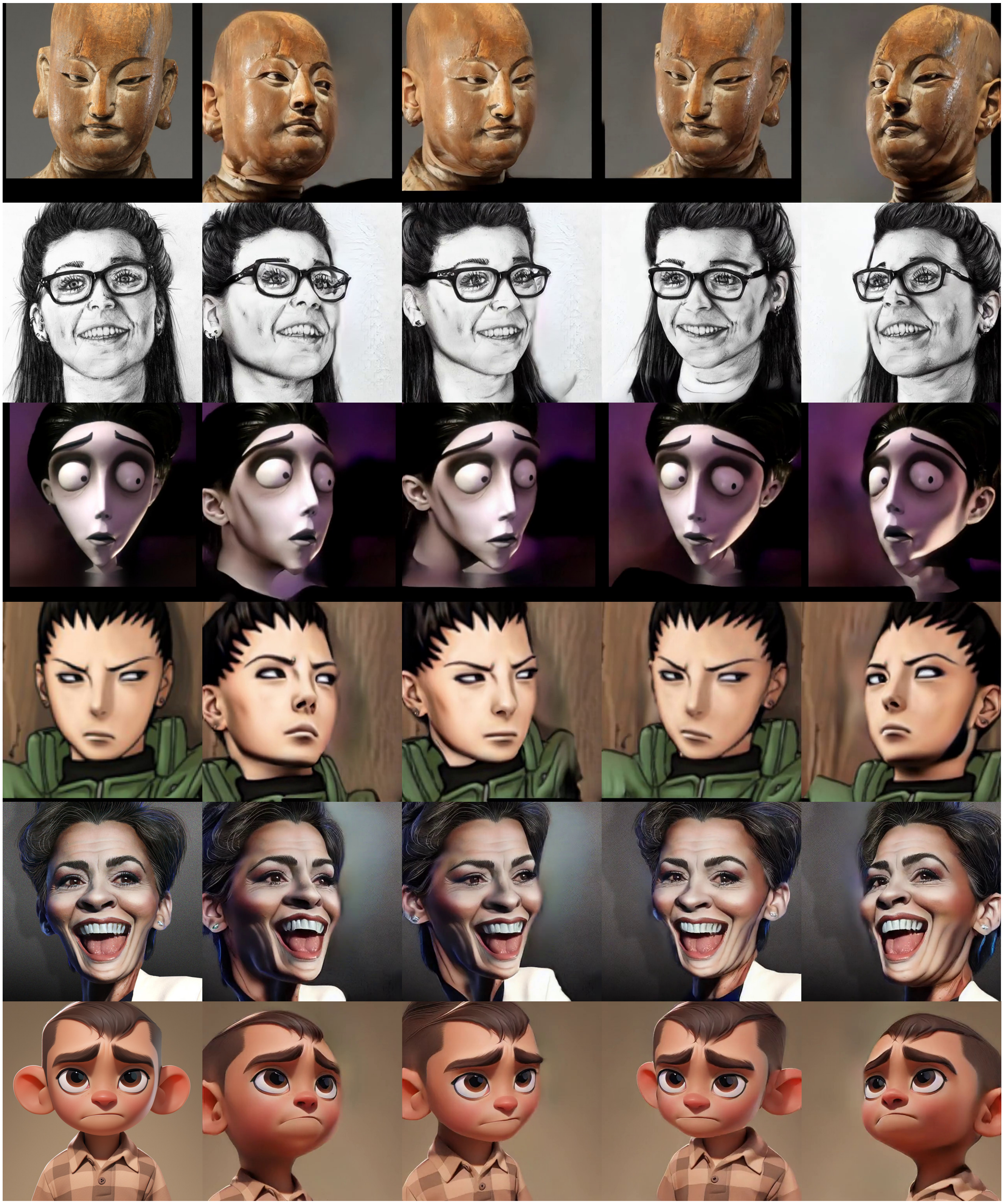

Figure 8. More novel-view consistent results

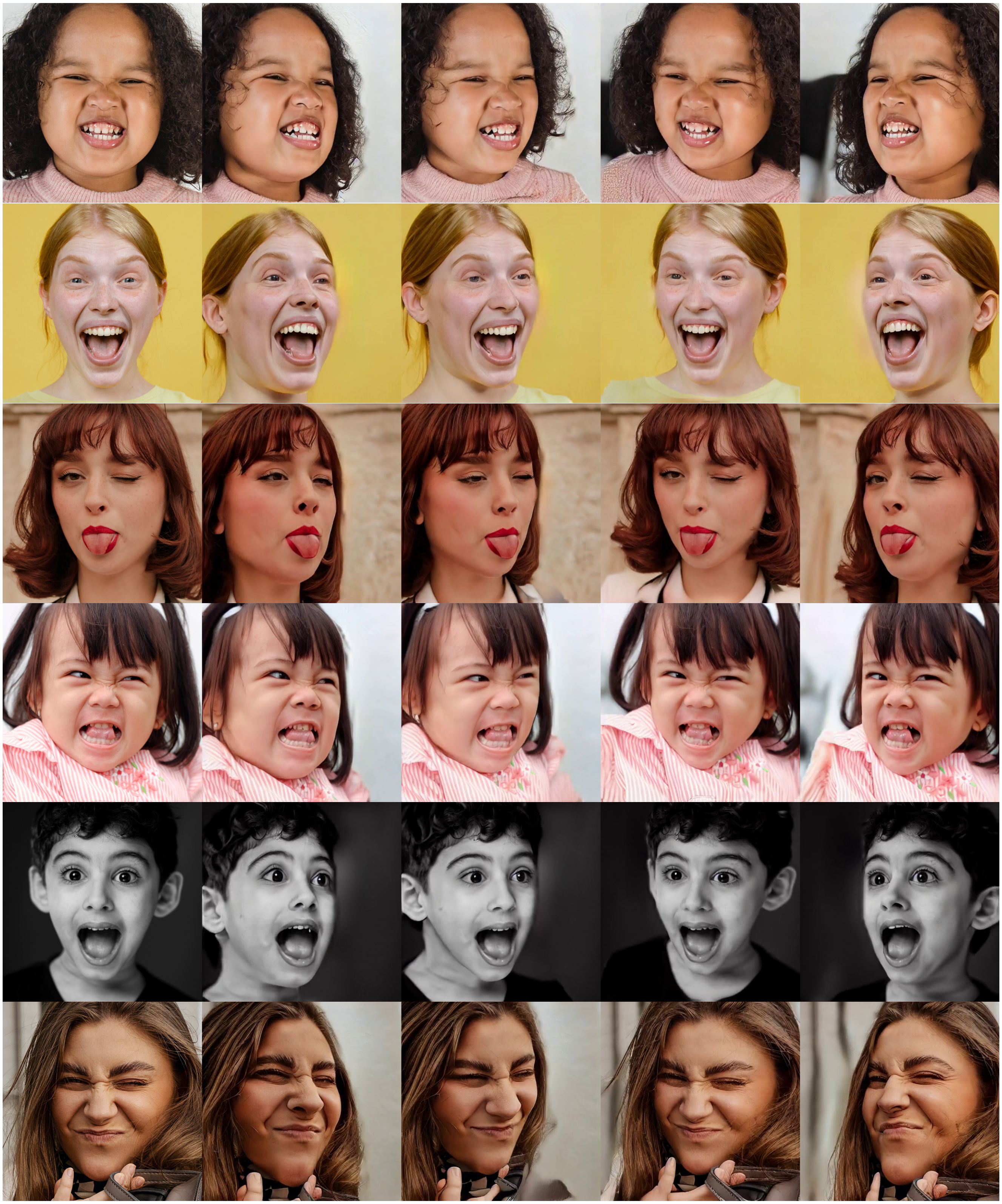

Figure 9. More novel-view consistent results

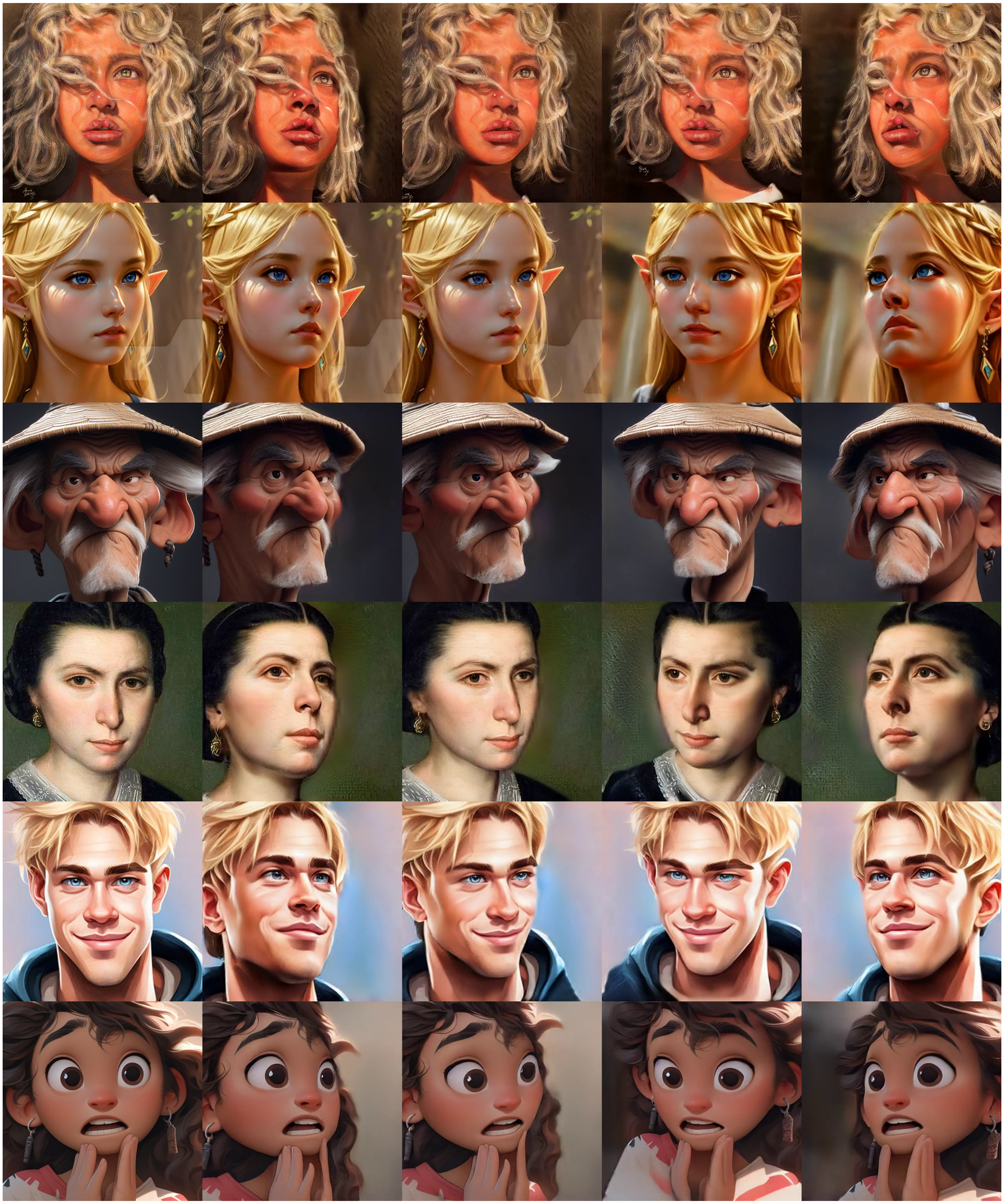

Figure 10. More novel-view consistent results