# DiffusionRig: Learning Personalized Priors for Facial Appearance Editing
Supplementary Material

## A. More Results on Personalized Editing

We provide more results on using physical buffers to rig/drive facial appearance generation in Figure 1.

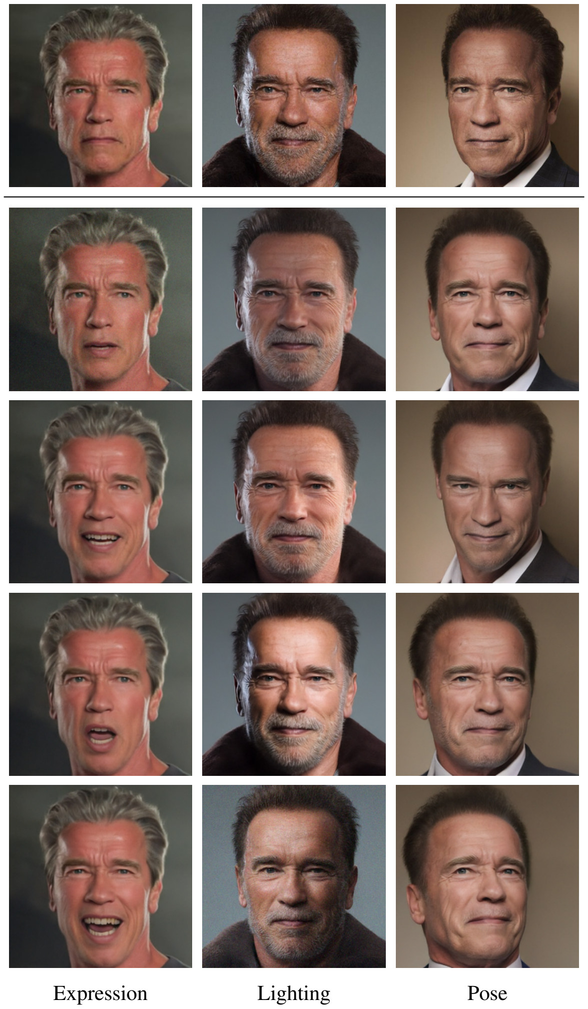

Figure 1. **More results on using physical buffers to rig the facial appearance.** The physical buffers (not shown) are used to edit the input images (top row) in terms of facial expression (left column), lighting (middle column), and head pose (right column).

## B. Extreme Lighting Editing

During the training of DiffusionRig, we rely on the SH-based lighting model from DECA, which is limited in modeling high-frequency lighting. At inference time, we can use a different lighting representation that can model directional lighting with cast shadow (through ray casting). We show one such extreme lighting example and another RGB lighting example in Figure 2, for which our model regresses slightly towards less extreme lighting but still produces reasonable results.

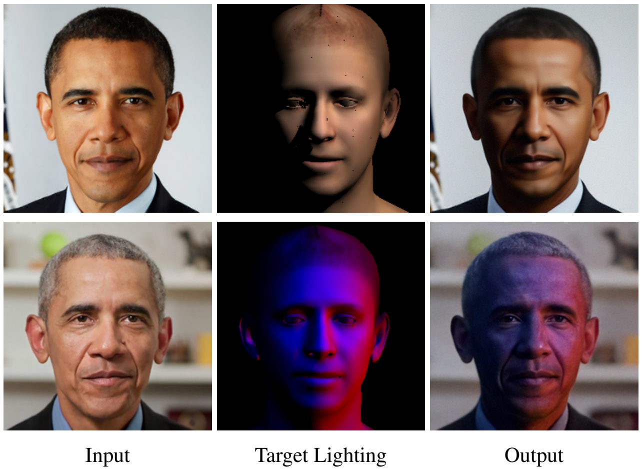

Figure 2. Stress test with difficult directional and RGB lighting.

## C. Personal Photo Collections

In Figure 3, we show two sets of images we used to train Stage 2. For celebrities, we crawl the photos from the internet; for non-celebrities, we use everyday photos. In comparison, MyStyle requires 92–279 images for finetuning. When using only 20 images as we do, MyStyle cannot learn personalized priors well as shown in the main paper.

## D. Neural Network Architecture Details

For our global encoder, we modify the ResNet-18 model by replacing the final classification layer with a feature extraction layer. More specifically, we change the last layer into a linear layer that outputs our latent code.

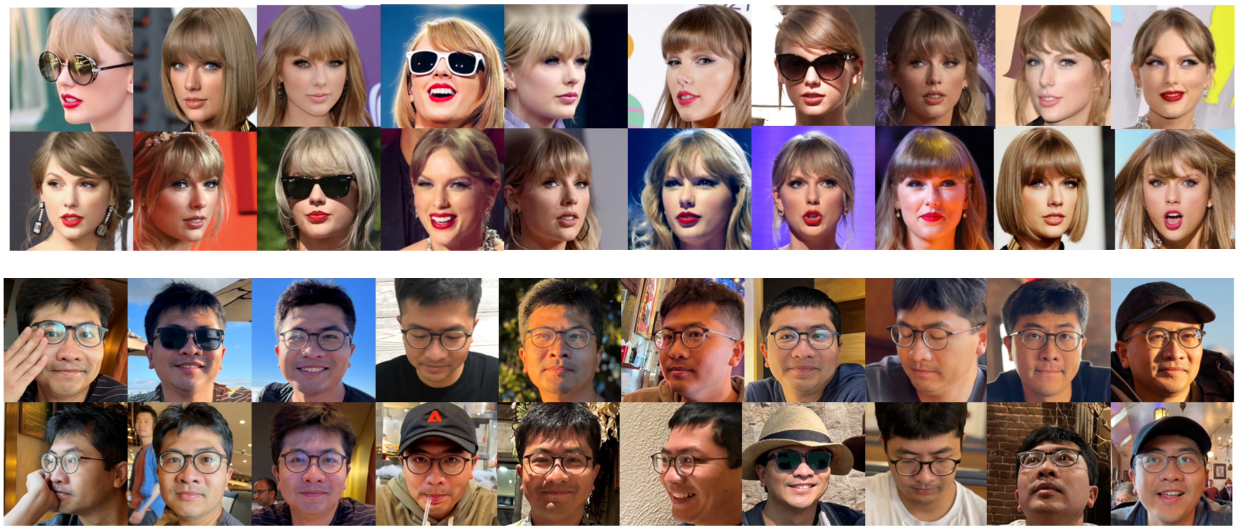

Figure 3. Personal photo collections used for training Stage 2: Taylor Swift and a non-celebrity.

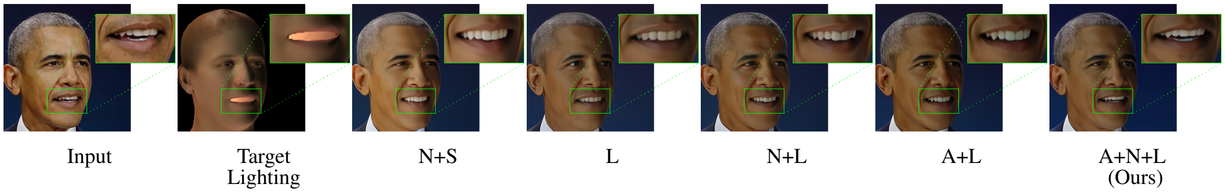

Figure 4. **Ablation on different input conditions.** N: Normals, S: SH layers, L: Lambertian rendering, A: Albedo.

Our diffusion model is based on the architecture presented in Guided-Diffusion [2]. We modify the architecture so that the model can take the global latent code as another condition (in addition to the concatenation of physical buffers and the noise image). This global latent code is used for scaling and shifting the features. Our model architecture details can be found in Table 1, where we provide hyperparameters for both our $256 \times 256$ and $512 \times 512$ models.

## E. Different Types of Pixel-Aligned Buffers

We ablate different pixel-aligned buffers in Figure 4. In our method, we use three kinds of physical buffers from DECA which are Normals (N), Albedo (A) and Lambertian rendering (L). With Lambertian rendering being the only physical buffer that contains lighting information, we include it in all our ablation studies except for the "N+S" where we use Normals and Spherical Harmonics with SH rendered on all-white albedo (i.e., shading), so it doesn't contain albedo information. We can see with Normals, Albedo, and Lambertian rendering, the results preserve details (e.g., mouth) better, while N+S cannot render accurate lighting due to the missing albedo.

| | $256 \times 256$ | $512 \times 512$ |
|---|---|---|
| Diffusion Steps | 1000 | 1000 |
| Channels | 128 | 128 |
| Channels Multiple | 1, 1, 2, 2, 4, 4 | 0.5, 1, 1, 2, 2, 4, 4 |
| Heads Channels | 128 | 128 |
| Attention Resolution | 16 | 16 |
| Dropout | 0.1 | 0.1 |
| P2_gamma† | 1.0 | 1.0 |
| P2_k† | 1.0 | 1.0 |
| Optimizer | Adam | Adam |
| Weight Decay | 0.0 | 0.0 |
| Batch Size (S1) | 256 | 64 |
| Batch Size (S2) | 4 | 2 |
| Iterations (S1) | 50k | 200k |
| Iterations (S1) | 5k | 20k |
| Learning Rate (S1) | $10^{-4}$ | $10^{-4}$ |
| Learning Rate (S2) | $10^{-5}$ | $10^{-5}$ |

Table 1. **DiffusionRig architecture details**. S1 and S2 denote Stage 1 and Stage 2, respectively. Refer to Guided-Diffusion [2] for more details. † are two hyperparamters defined in prior work [1].

## F. Higher-Resolution Results ($512 \times 512$)

DiffusionRig can be trained at $512 \times 512$ resolution. We show these higher-resolution results in Figures 5 and 6 on two new celebrities.

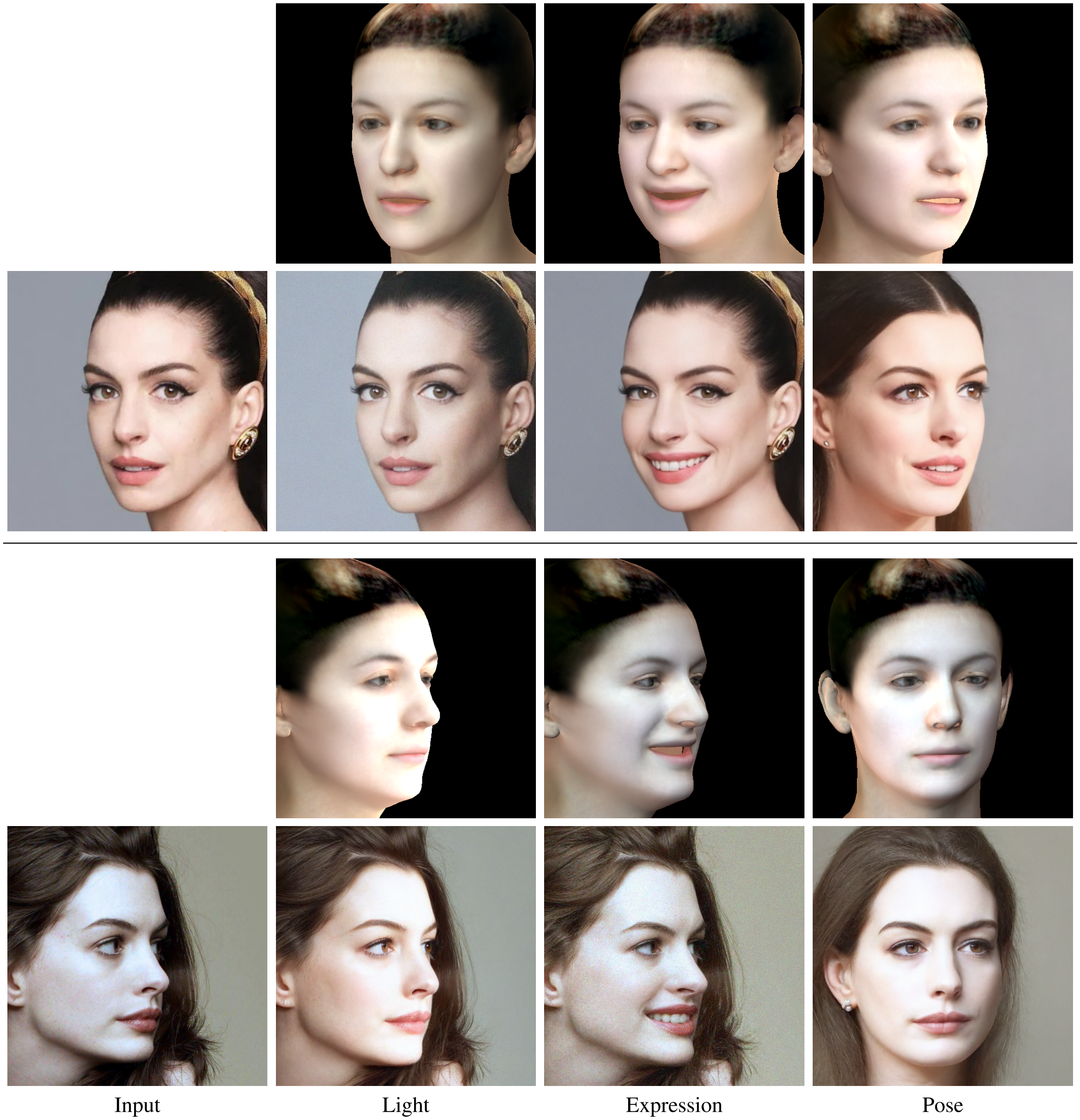

Figure 5. **512×512 Facial Appearance Editing Results.** Two groups of results are presented here with the first row of each group being the physical buffers that drive the editing.

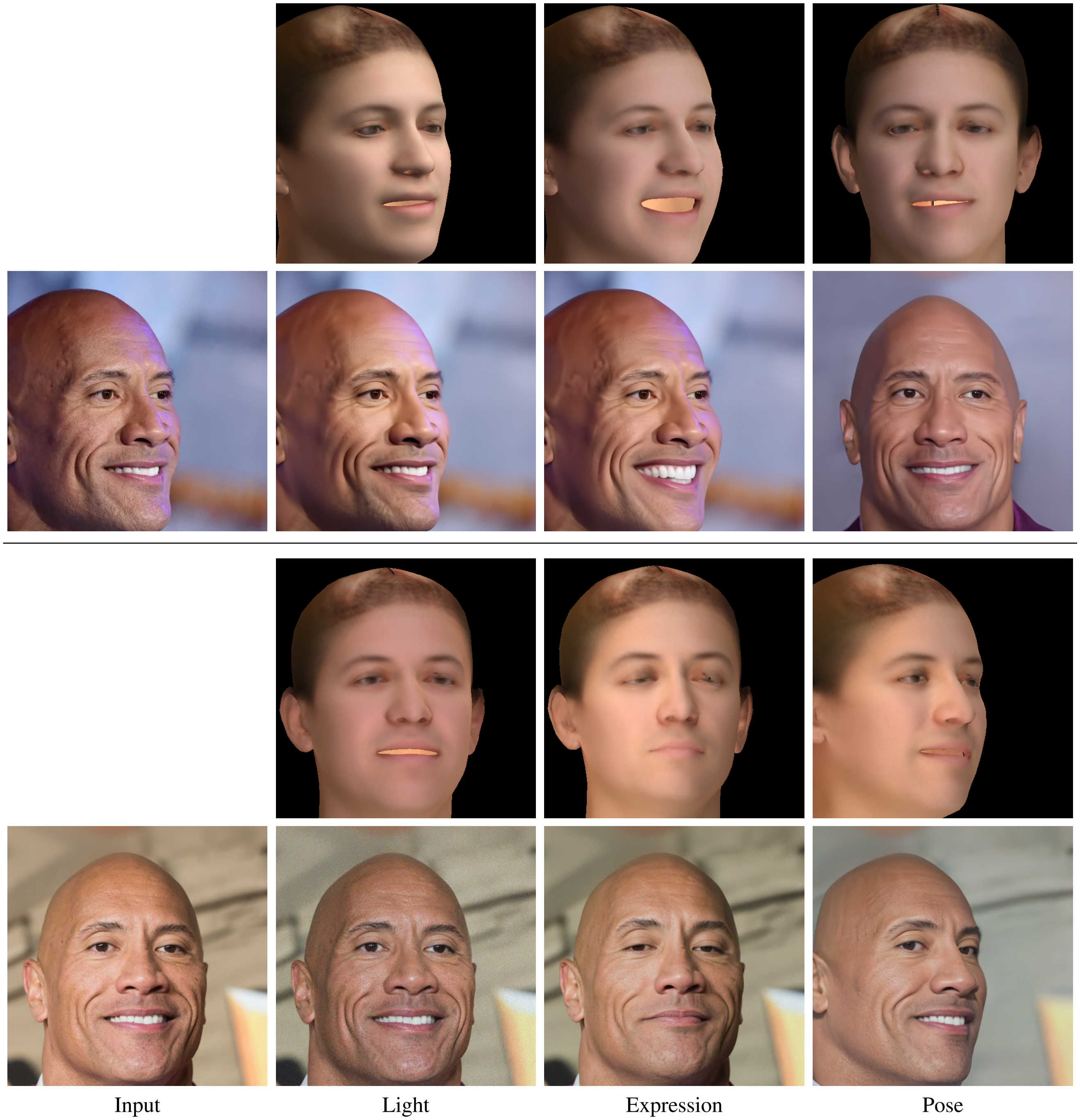

Figure 6. **512×512 Facial Appearance Editing Results.** Two groups of results are presented here with the first row of each group being the physical buffers that drive the editing.

## References

[1] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11472–11481, 2022. 2

[2] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2