Relightful Harmonization: Lighting-aware Portrait Background Replacement

Supplementary Material

7. Additional Results

We note that the all testing input portrait images shown in our paper are sampled from Unsplash or Adobe Stock.

Comparison with benchmarks To supplement Fig. 4c, we present additional visual comparison with benchmarks on the real world data in Fig. 8, 9 and 10. To supplement Fig. 4a and Fig. 4b, we show full benchmark comparison on the light stage test dataset in Fig. 11, and on the natural image test set in Fig. 12.

Ablation We present additional visual comparison among our ablation models on the natural image test and light stage test set in Fig. 14 and Fig. 15 respectively. The ablation comparison on the real test set is shown in Fig. 13 as a supplement to Fig. 6. In Fig. 16, we include additional visualization of the feature norms to illustrate the affects of the alignment module to supplement Fig. 7.

Real world testing results Fig. 17 shows reference based harmonization example as in Fig. 5e. Fig. 18 shows harmonization results when we flip the background image as in Fig 5a. Fig. 19 shows the results under spatially and temporally changing lighting as in Fig. 5d.

8. Additional Implementation Details

8.1. Network Architecture

Lighting-conditioned diffusion is built on the InstructPix2Pix [3] backbone. The core rationale for selecting the pretrained InstructPix2Pix model [3] as the foundation, rather than the stable diffusion model, is on its capability to incorporate an additional input image channel. Therefore, at the beginning of our training, the input and output image will be identital (i.e., no editing on the input image), and it will gradually incorporate the lighting conditioning from the extra lighting representation. We use a dummy editing prompt *'portrait'* during our training and inference.

The lighting conditioning branch architecture follows ControlNet [73], where an encoder structure identical to the diffusion UNet backbone is applied and the intermediate feature maps are added to the UNet encoder at respective resolutions. The lighting representation is extracted from a 4-layer CNN. We train our model with the input resolution of 512×512 (for both input image and the background image), and the lighting representation is a tensor with shape $64 \times 64 \times 320$. We empirically found that training with a higher resolution (e.g., 768×768) led to better identity preservation, but performed worse in terms of the relighting. We speculate that this is related to the stable diffusion pretraining, which is on 512×512 resolution.

The Alignment Network is an encoder-decoder architecture built with Residual blocks. The encoder is composed of three sequential residual blocks. Each of these blocks is coupled with a subsequent downsampling layer. The decoder is symmetrical to the encoder, with three residual blocks, and each of them followed by an upsampling layer. The input and output dimensions of the alignment network are consistent, maintaining a shape of $64 \times 64 \times 320$.

Ablation Models Specifics `Model#0` is a baseline diffusion model without lighting conditioning and its implementation follows InstructPix2Pix [3] with the text prompt fixed as 'Portrait'. `Model#1` takes the background image as the conditional input, which is resized to 512×512. `Model#2` shares the same architecture as `Model#1` but replaces the conditional input to the LDR environment map. `Model#3` introduces the alignment module after the conditional branch from `Model#1`. The Unet backbone from `Model#2` is used as diagramed in Fig. 2. `Model#4` finetunes on `Model#3` with the synthetic data.

8.2. Transformer relighting model

To train a relighting baseline on our light stage dataset, we built a transformer based encoder-decoder network. The network input is a concatenated input image, foreground mask, and the parsing mask, which is divided into patches of 4×4. A hierarchical Transformer encoder is applied to obtain multi-level features at $\{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$ of the original resolution. A decoder with transpose convolution is then followed to get the final result with the same resolution as the input. The target LDR environment map is concatenated at the bottleneck latent space in a similar manner as [55].

9. Failure case and analysis

We illustrate several example failure cases in Fig.20. In our training approach, since we do not impose constraints on the subject's identity, there are instances where the model struggles to retain identity-specific details. For instance, as shown in Fig.20a, the color of the subject's clothing is inaccurately altered during the color harmonization process. Similarly, in Fig. 20b, there is a notable change in hair color (*middle*). Furthermore, in scenarios where the input skin tone is not clearly indicated (*right*), our model occasionally produces ambiguous results in skin tone modification. Additionally, our method does not incorporate intermediate steps like albedo estimation, which can be crucial in handling complex lighting conditions. As a result, in inputs with pronounced cast shadows, our model sometimes fails to eliminate these shadows effectively.

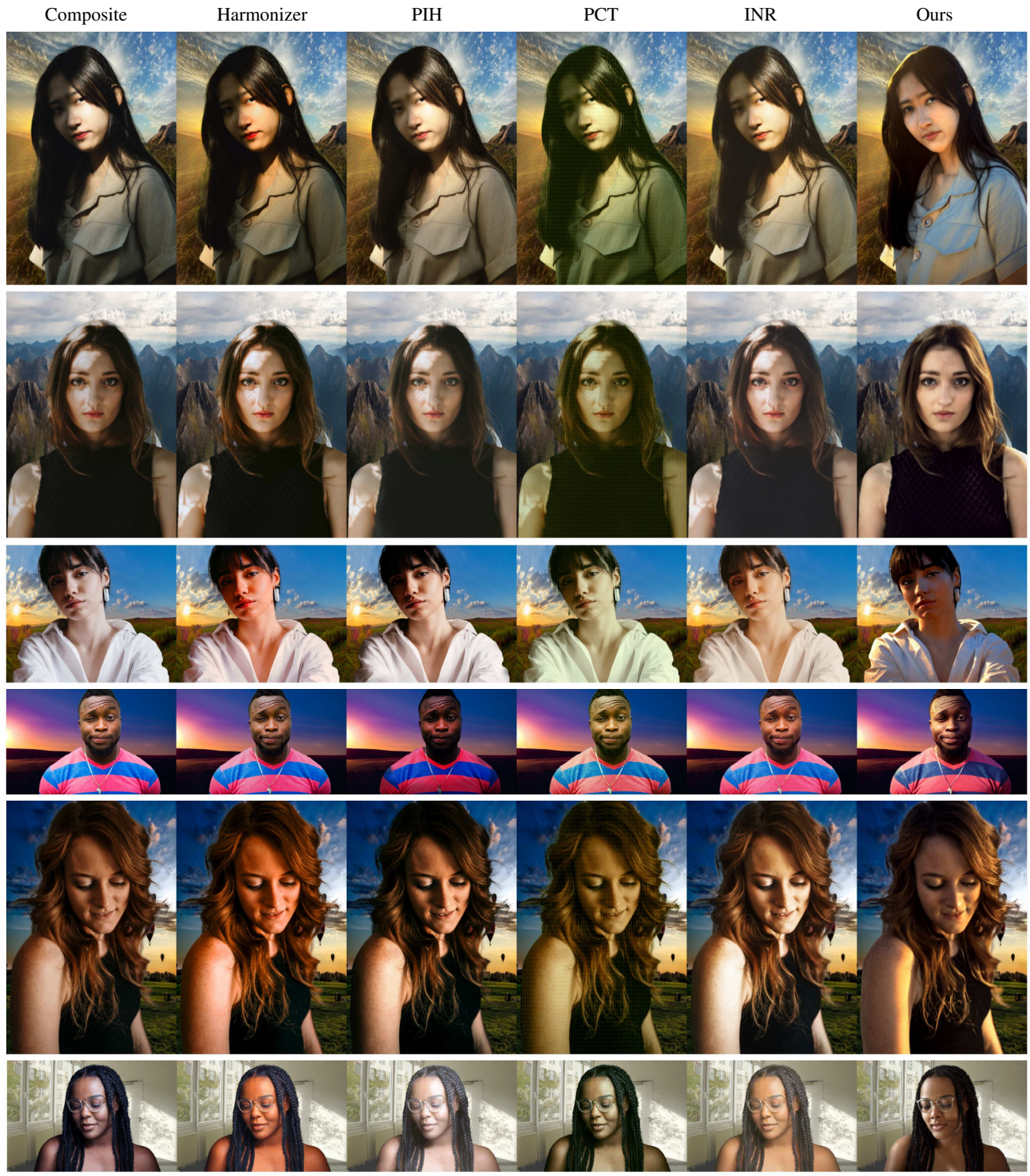


Figure 8. Example comparison results on the real world test set to supplement Fig. 4c in the main paper.

Figure 9. Example comparison results on the real world test set to supplement Fig. 4c in the main paper.

Figure 10. Example comparison results on the real world test set to supplement Fig. 4c in the main paper.

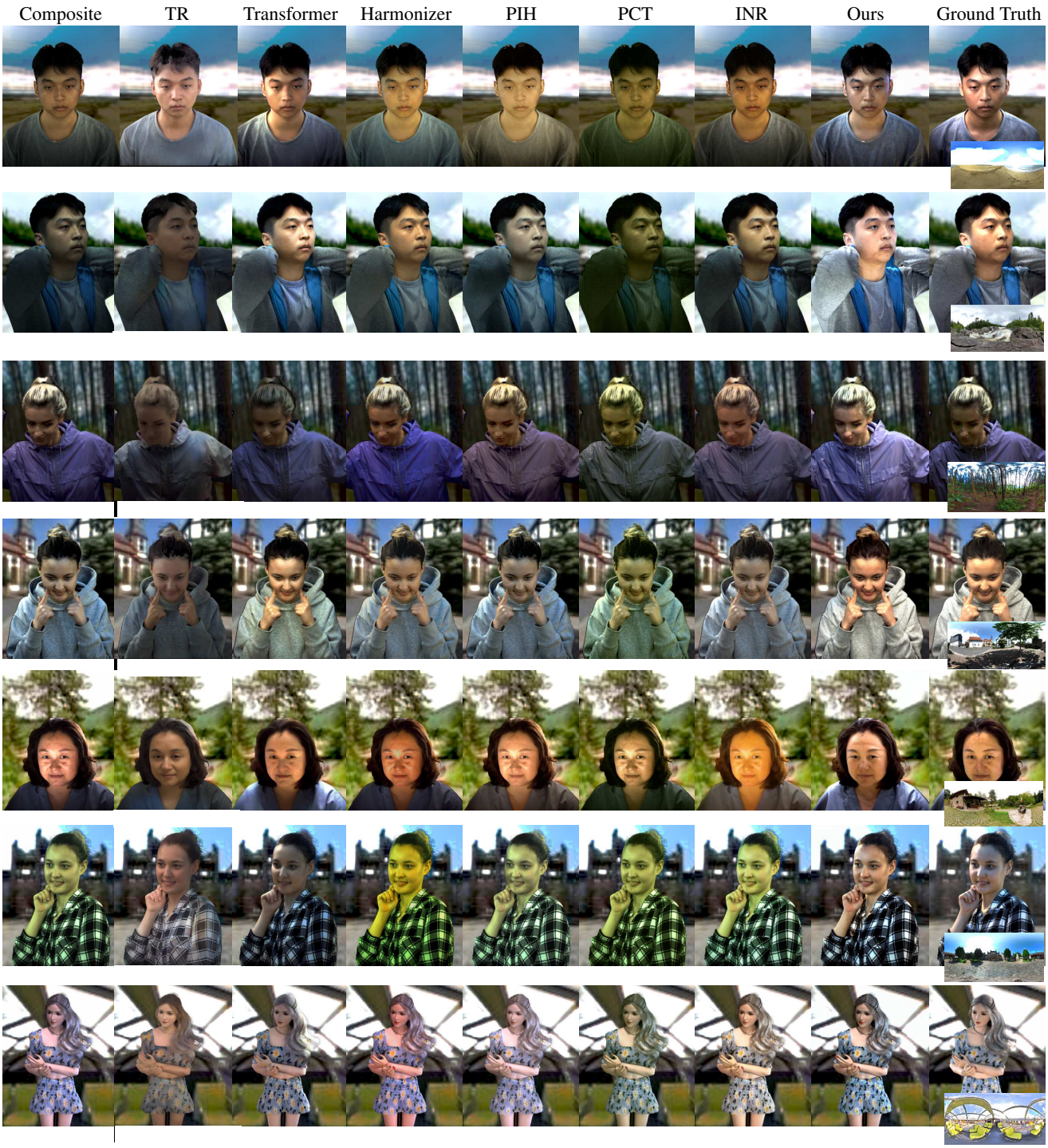


Figure 11. Example comparison results on the light stage test set to supplement Fig. 4a in the main paper. The environment map is shown at the bottom of the ground truth image.

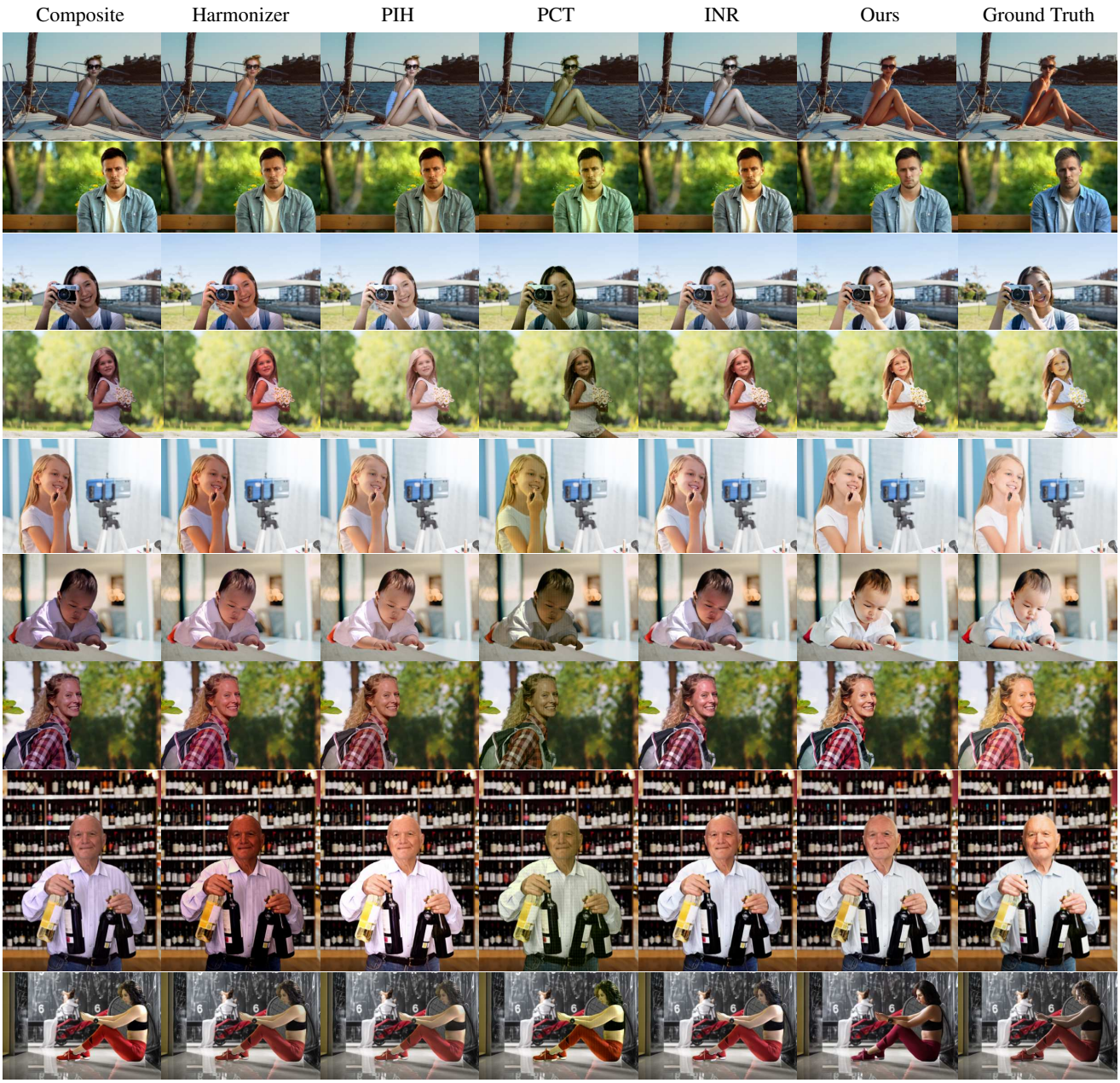


Figure 12. Example comparison results on the natural image test set to supplement Fig. 4b in the main paper.

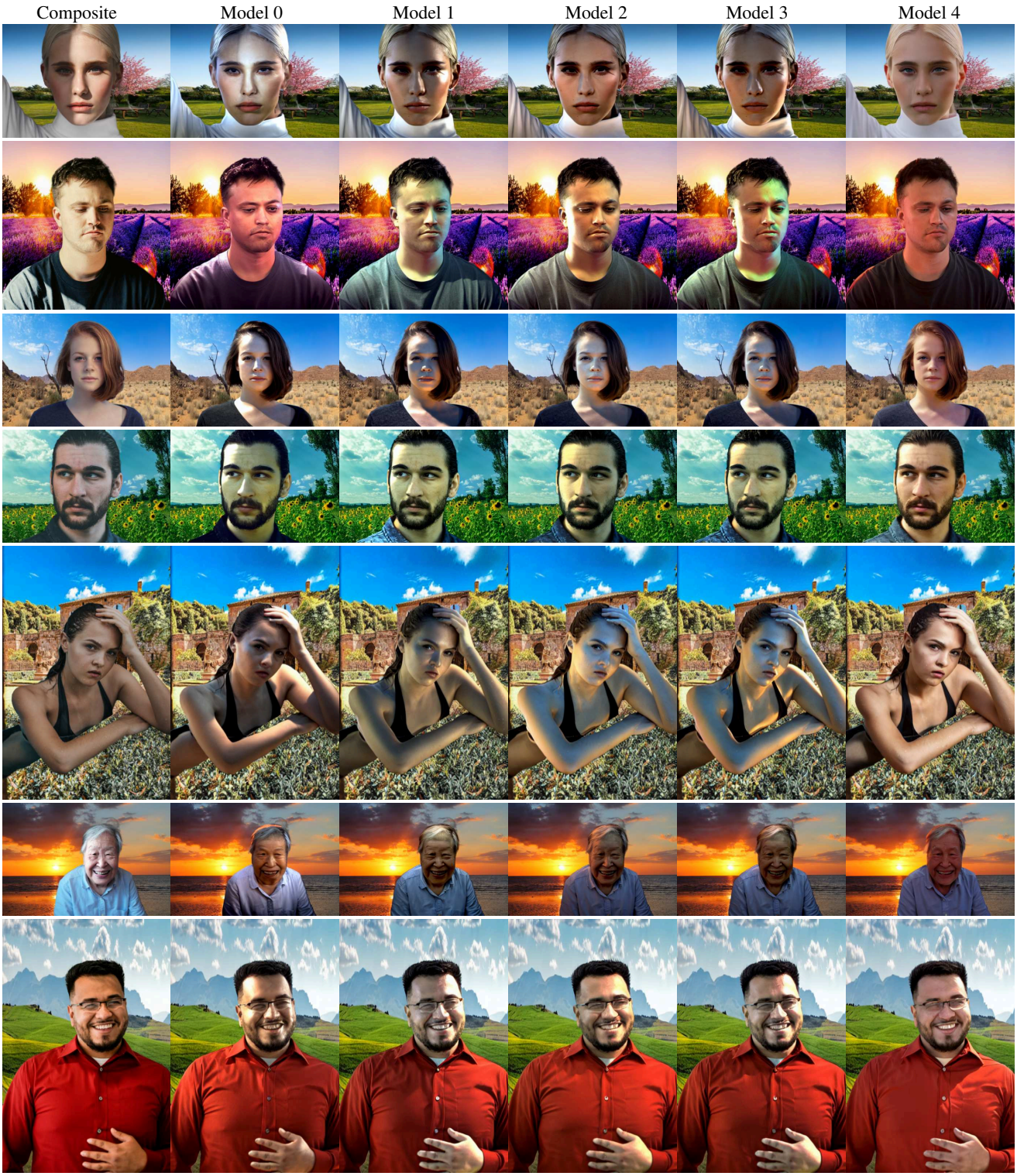


Figure 13. Example testing results from our ablation on the real image test set. Model 0 to Model 4 correspond to the configurations in Table 3. Our final model (Model 4) presents the best visual quality while maintaining plausible lighting effects.

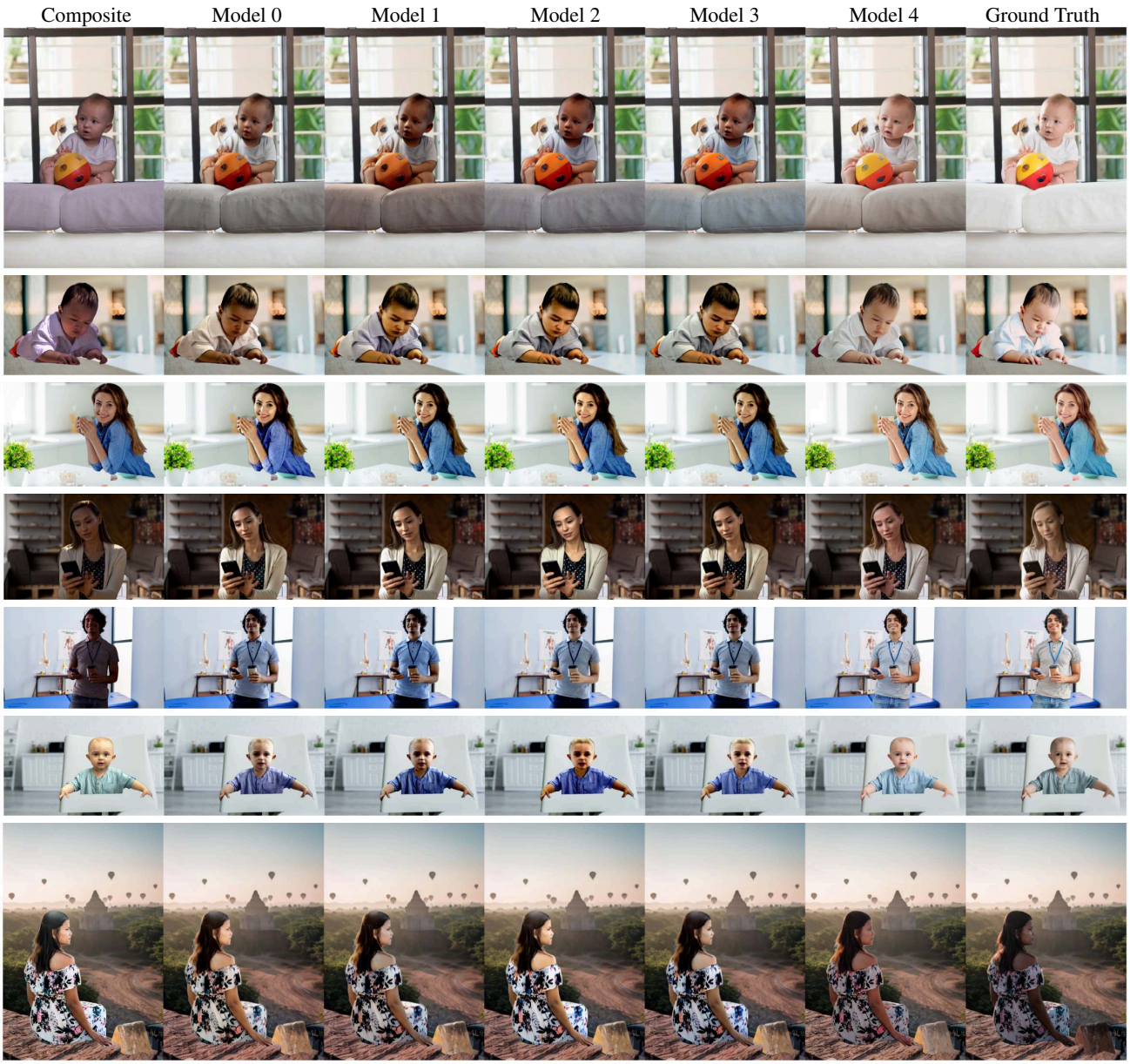


Figure 14. Example testing results from our ablation on the natural image test set. Model 0 to Model 4 correspond to the configurations in Table 3. Our final model (Model 4) presents the best visual quality while maintaining plausible lighting effects.

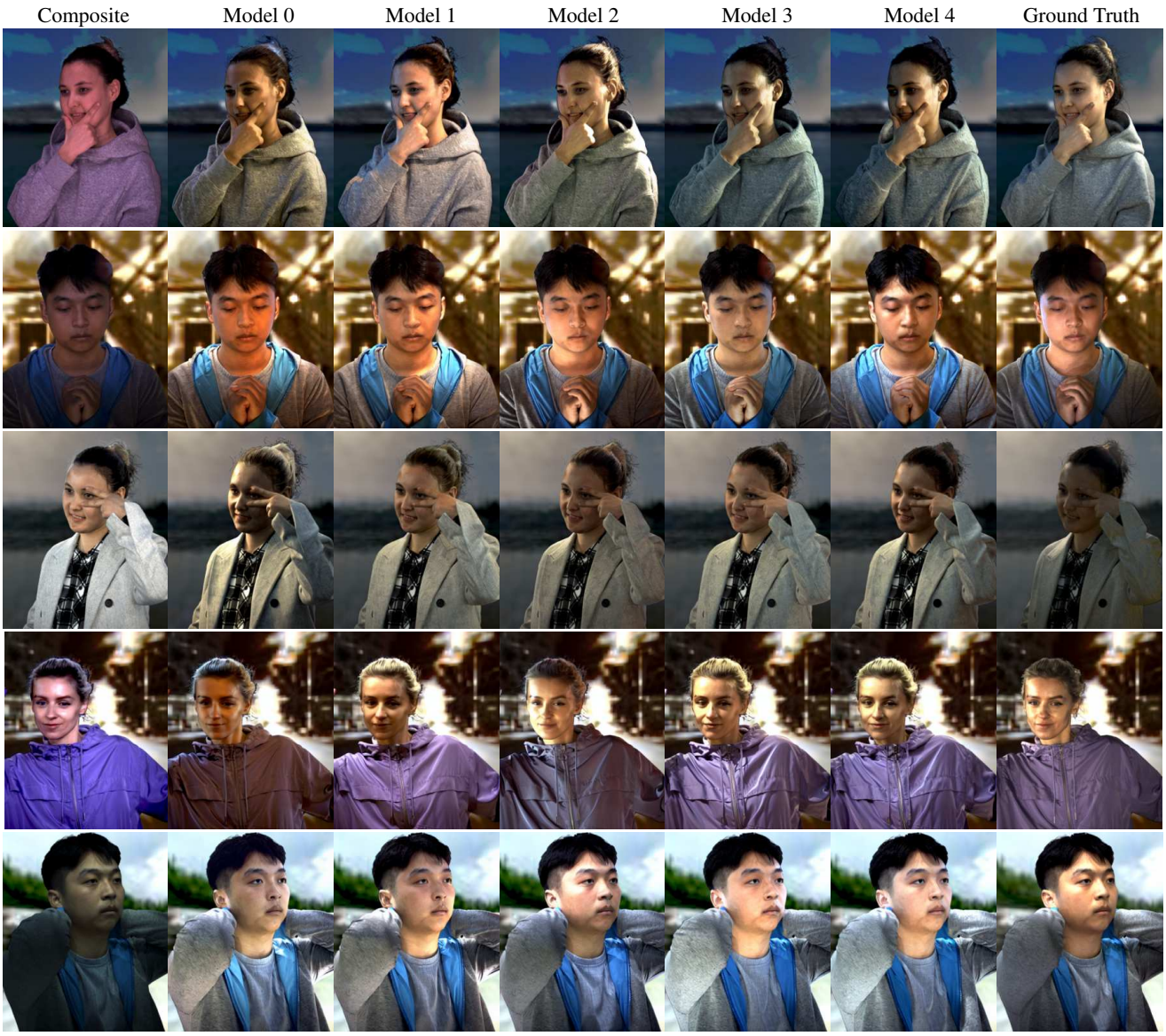


Figure 15. Example testing results from our ablation on the light stage test set. Model 0 to Model 4 correspond to the configurations in Table 3. Our final model (Model 4) presents the best visual quality while maintaining plausible lighting effects.

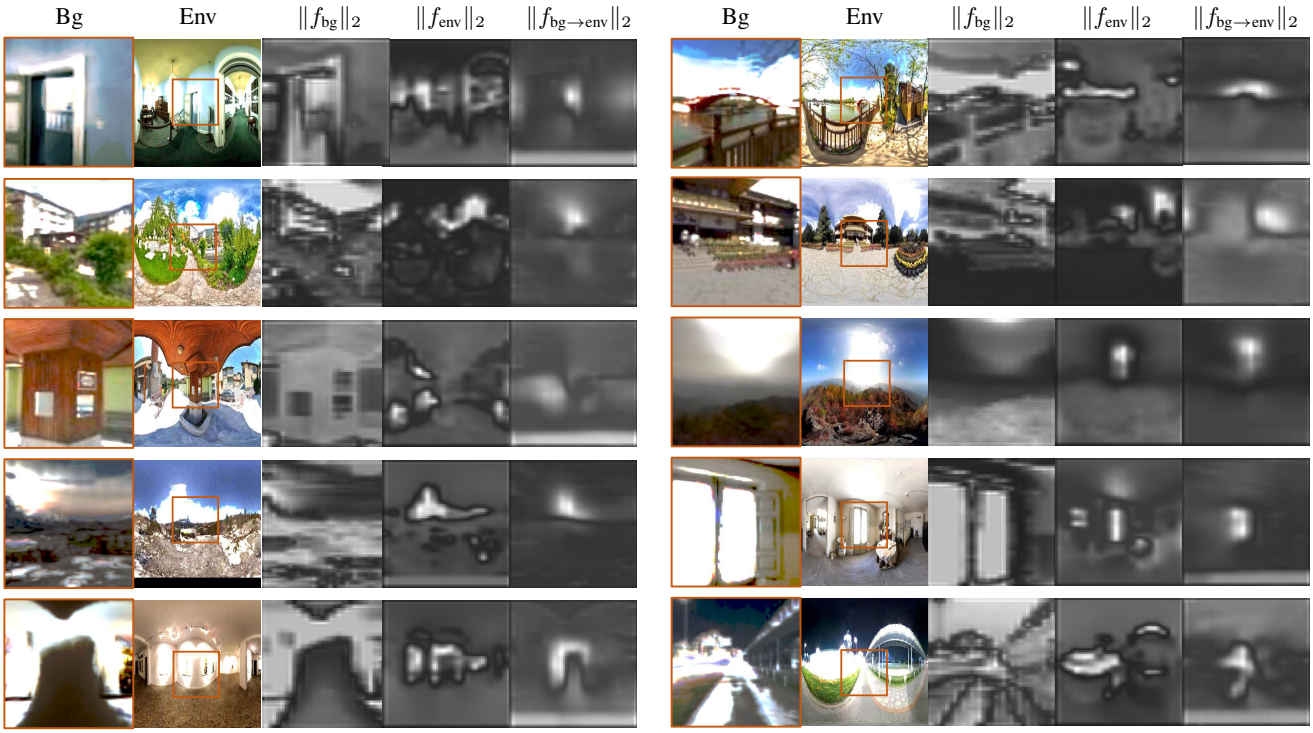


Figure 16. The L_2 norm of learned lighting representations to supplement Fig. 7. The aligned background-derived feature on the right matches the panorama much closer, indicating a better lighting representation.

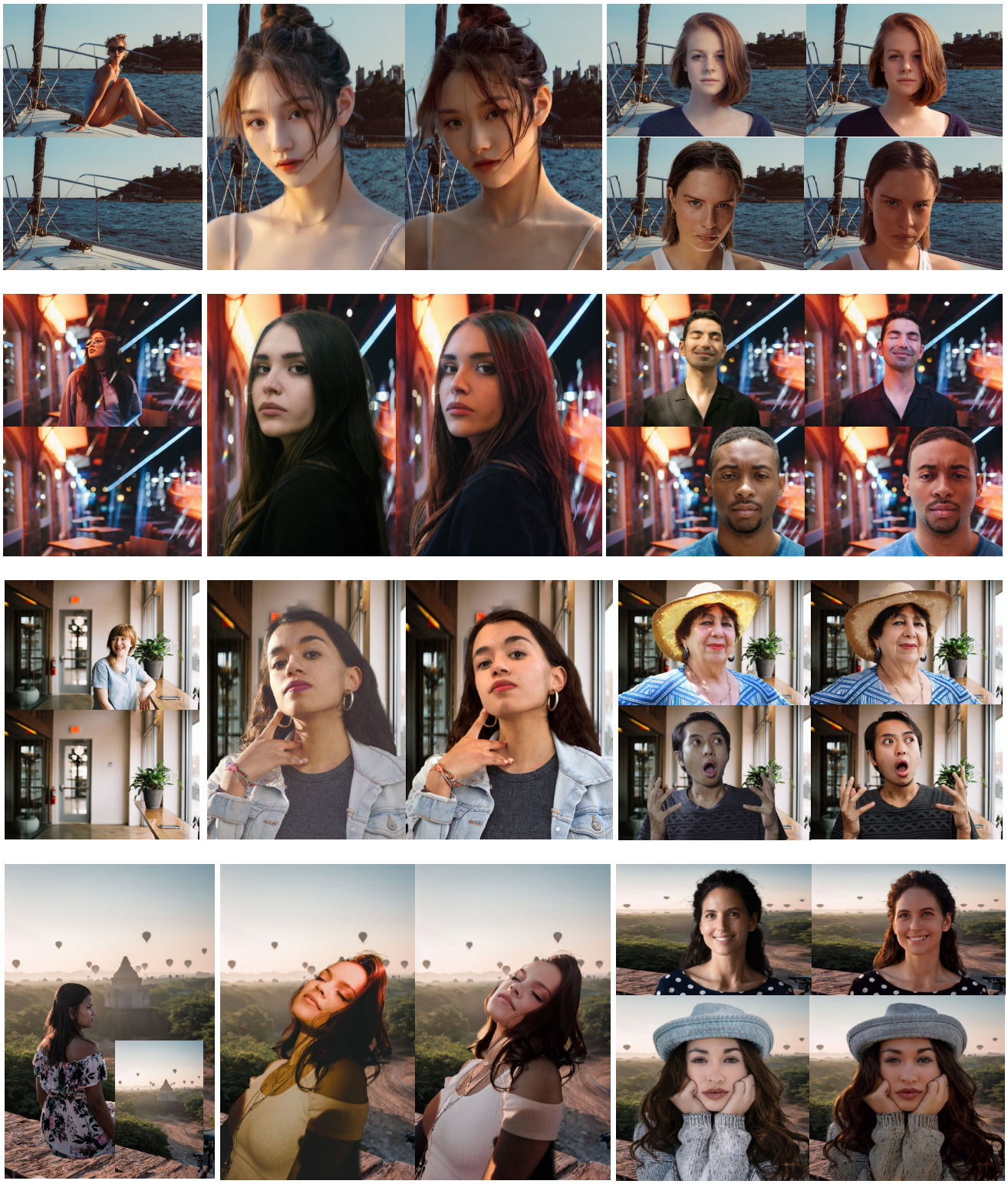


Figure 17. Visual results on the reference-based harmonization application to supplement Fig. 5e. It allows user images to be blended into scenes from real portraits. This involves removing the subject from the reference image (*upper left*) to create a background (*lower left*) for composition. The harmonized results (*right*) achieve lighting effects closely resembling those in the reference.

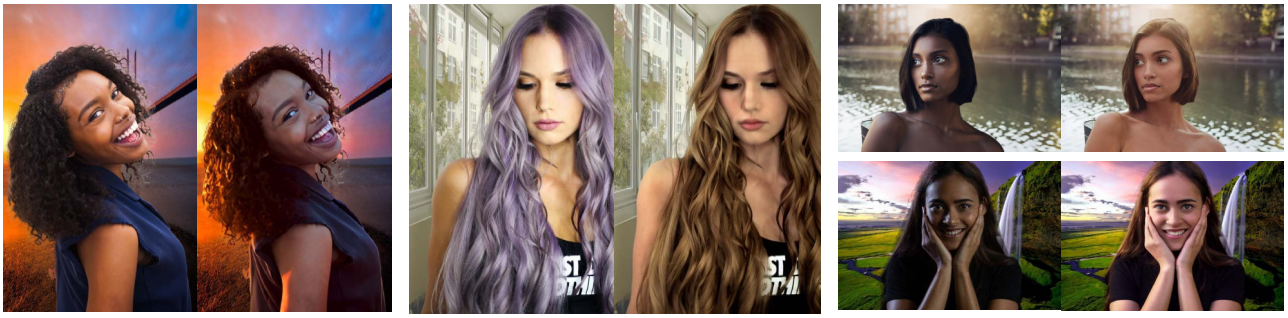
Figure 18. Harmonization results when flipping the background, to supplement Fig. 5a.

(a) We create spatially changing lighting conditions by cropping background images (*top*) from a panoramic image. Our model produces visually coherent lighting changes on different portrait images.

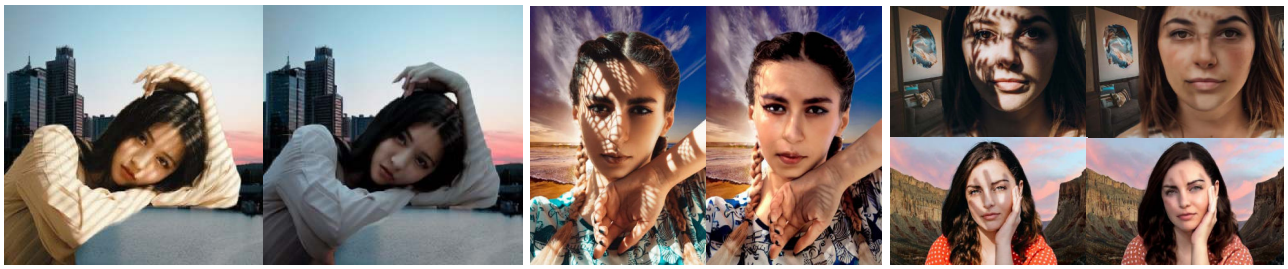
(b) We obtain temporally changing lighting conditions by taking multiple screenshots (*top*) from a timelapse video (https://www.youtube.com/watch?v=CSfri4U9w28). Our model produces visually reasonable harmonization results.

Figure 19. Harmonization results under the background images where lighting conditions are changing spatially (a) or temporally (b).

(a) In some examples, our model modified the color of the subject clothes due to its harmonization nature.



(b) In some examples, our model may not fully preserve the subject identity, such as the hair color and the skin tone, especially when the input skin tone is ambiguous (right two examples).



(c) In portraits with strong casted shadows, our model may fail to completely remove them.

Figure 20. Failure cases