SynthLight: Portrait Relighting with Diffusion Model by Learning to Re-render Synthetic Faces

Sumit Chaturvedi[1] Mengwei Ren[2] Yannick Hold-Geoffroy[2] Jingyuan Liu[2]
Julie Dorsey[1] Zhixin Shu[2]

[1]Yale University [2]Adobe

Supplementary Material

A. Additional Results

We present additional results on input portraits from various stock websites such as Adobe Stock [1], Unsplash [3] and Pexels [2] as well as from our internal light stage captures.

In-the-wild Test Portraits We demonstrate portrait relighting in the presence of strong sunlight to produce effects such as strong cast shadow from facial features, rim-effects in hair and specular highlights in Fig. 1. In Fig. 2, we demonstrate applying a studio environment map on in-the-wild test portraits to accentuate prominent features such as facial contours and expressions in the portraits. In Fig. 3, we showcase that SynthLight generalises to several challenging cases such as a 2D cartoon, a boy with face paint and a full body portrait, beyond the diversity present in the synthetic training data.

Comparison with Baselines We evaluate SynthLight against several baseline methods on in-the-wild portraits. As shown in Fig. 4, SynthLight achieves lighting effects, such as the rim-light effect in hair and subsurface scattering in the ears. Additionally, Fig. 5 illustrates specular highlights on darker skin tones.

Ablations Fig. 9 showcases additional examples from our ablation study, illustrating the contribution of each component to the final qualitative results. The *Base* model struggles with identity preservation and fails to capture key details present in the input portrait. Adding either *Base + Multi-Task* or *Base + Inference Adaptation* improves details but remains insufficient for reproducing complex accessories, materials, and textures. For example, in Fig. 9, the cigarette in the input portrait (top) and the specularity of the choker necklace or the accurate dress color (bottom) are not faithfully replicated. In contrast, our method successfully addresses these challenges, achieving superior results.

We train an additional model, *Ours + Light Stage*, where light stage-rendered data is combined with the synthetic dataset for relighting. The light stage data is same as in Relightful Harmonization [10], and consists of roughly 6000 light stage captures, rendered under 100 environment maps. Fig. 10 illustrates overexposure issues. SwitchLight [6], trained on light stage data has overexposure artifacts, i.e., unnatural yellowish skin tones. *Ours + Light Stage* reduces this issue due to the inclusion of physically-based rendered synthetic data, though some overexposure persists. In contrast, our method trained on physically-based rendered synthetic data avoids this problem, producing natural and balanced skin tones.

Comparison with Background-Conditioned Models In Fig. 11, we compare SynthLight, trained on our synthetic physically-based rendered data using environment maps with comprehensive 360° lighting information, to a background-conditioned variant of SynthLight, and IC-Light [14]. SynthLight excels at capturing nuanced lighting effects, such as cast shadows from self-occlusion, due to its precise environmental lighting inputs, whereas, the background-conditioned model generates inaccurate lighting. Although generating strong cast shadows caused by self-occlusion is challenging for background-conditioned relighting methods such as IC-Light, our background-conditioned model is able to do so by leveraging our synthetic dataset.

B. Dataset

Synthetic Dataset In Fig. 14 we show more examples from our synthetic dataset of subjects rendered under different environment maps. Each group of 4 visualizes a subject rendered under 4 lighting conditions.

LAION Data Filtration We filter a subset of LAION [12] by first running a face detector. Since this results in a large number of false positives, we additionally curate a set of query phrases whose matching images we seek to avoid. We filter the set of images further by evaluating the CLIP [9] score of each image against the query words and retaining only those images whose CLIP score is below a threshold. Emperically, we set this threshold to 0.15.

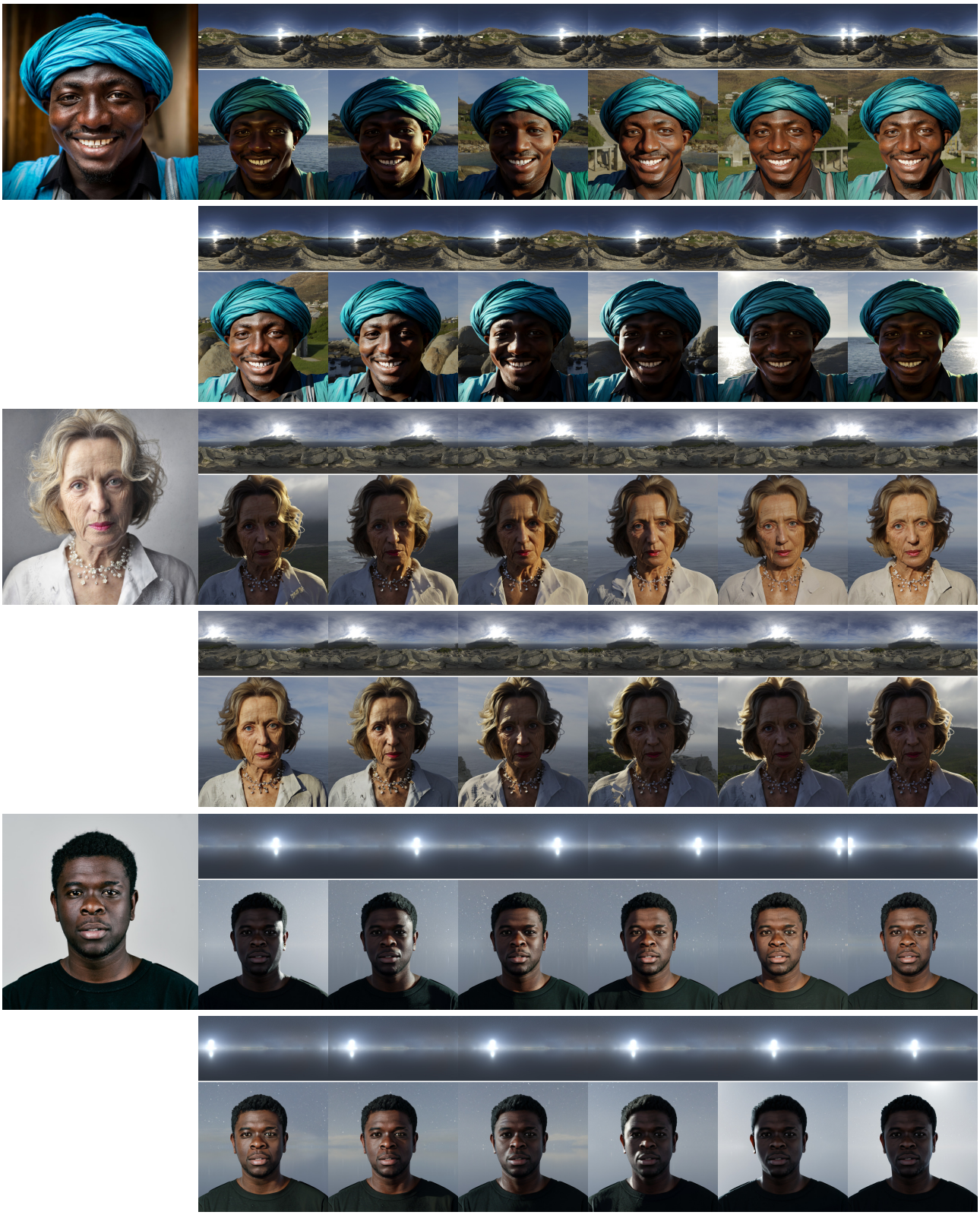

Figure 1. In order to demonstrate portrait lighting effects in the presence of strong sunlight such as strong cast shadows by facial features, rim-effects in hair and specular highlights, we show in-the-wild portraits relit using outdoor environment maps.

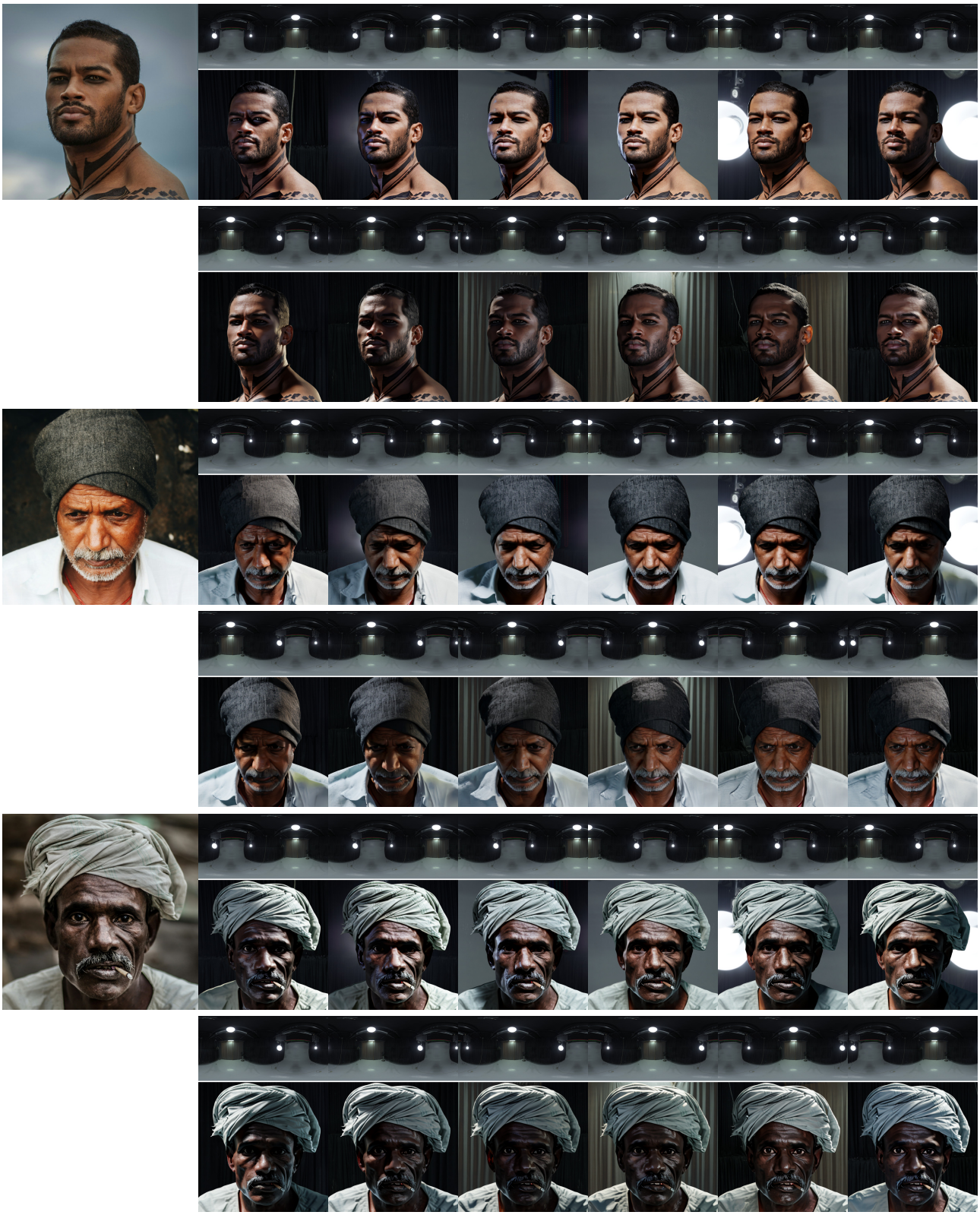

Figure 2. To demonstrate SynthLight's ability to enhance portraits with studio-style lighting, we present in-the-wild portraits relit using a studio environment map, where the studio lights accentuate prominent features such as facial contours and expressions.

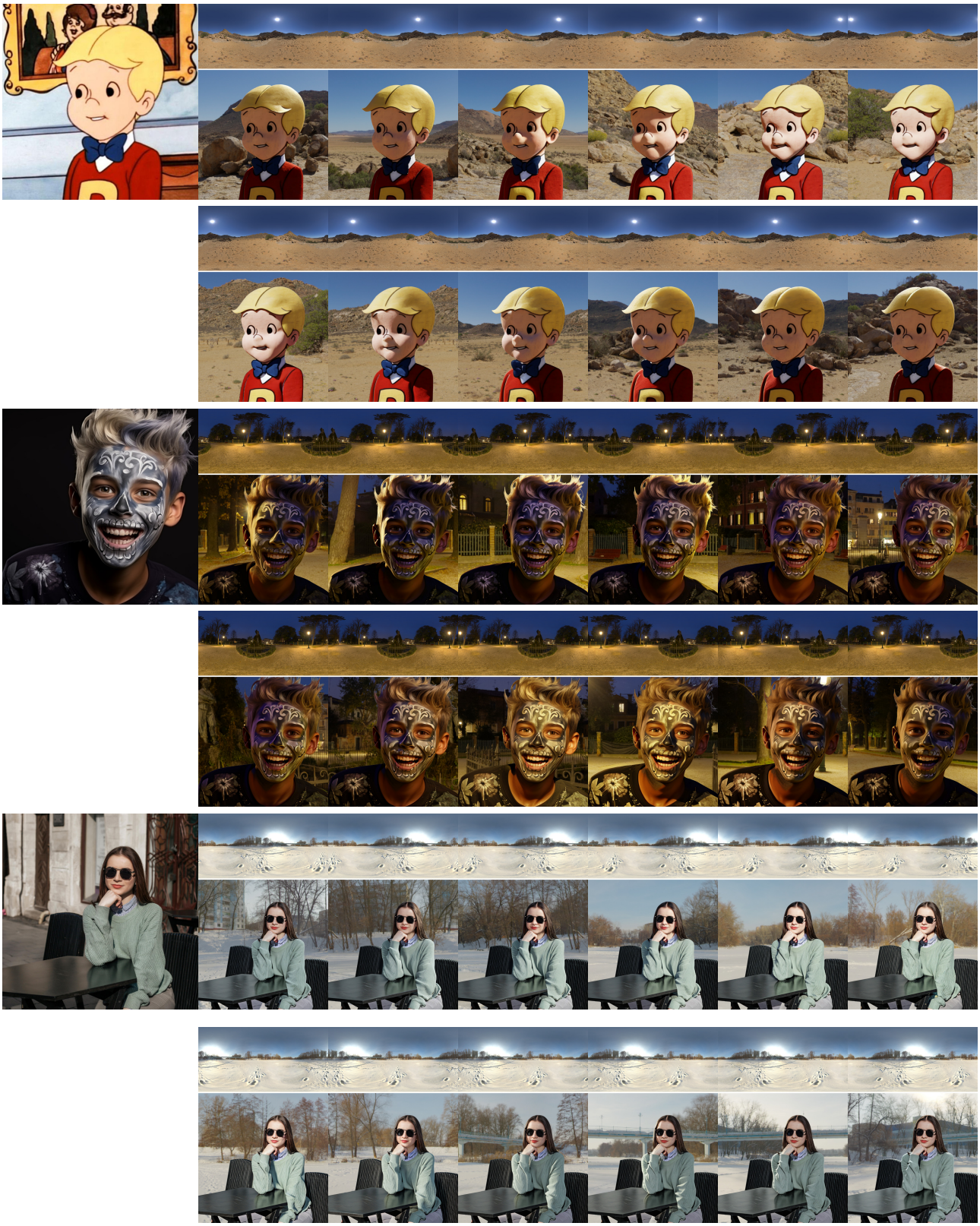

Figure 3. We show challenging in-the-wild portraits featuring 2D cartoon characters, child wearing face paint and a full body portrait, demonstrating that our method can generalize beyond the synthetic dataset seen during training.

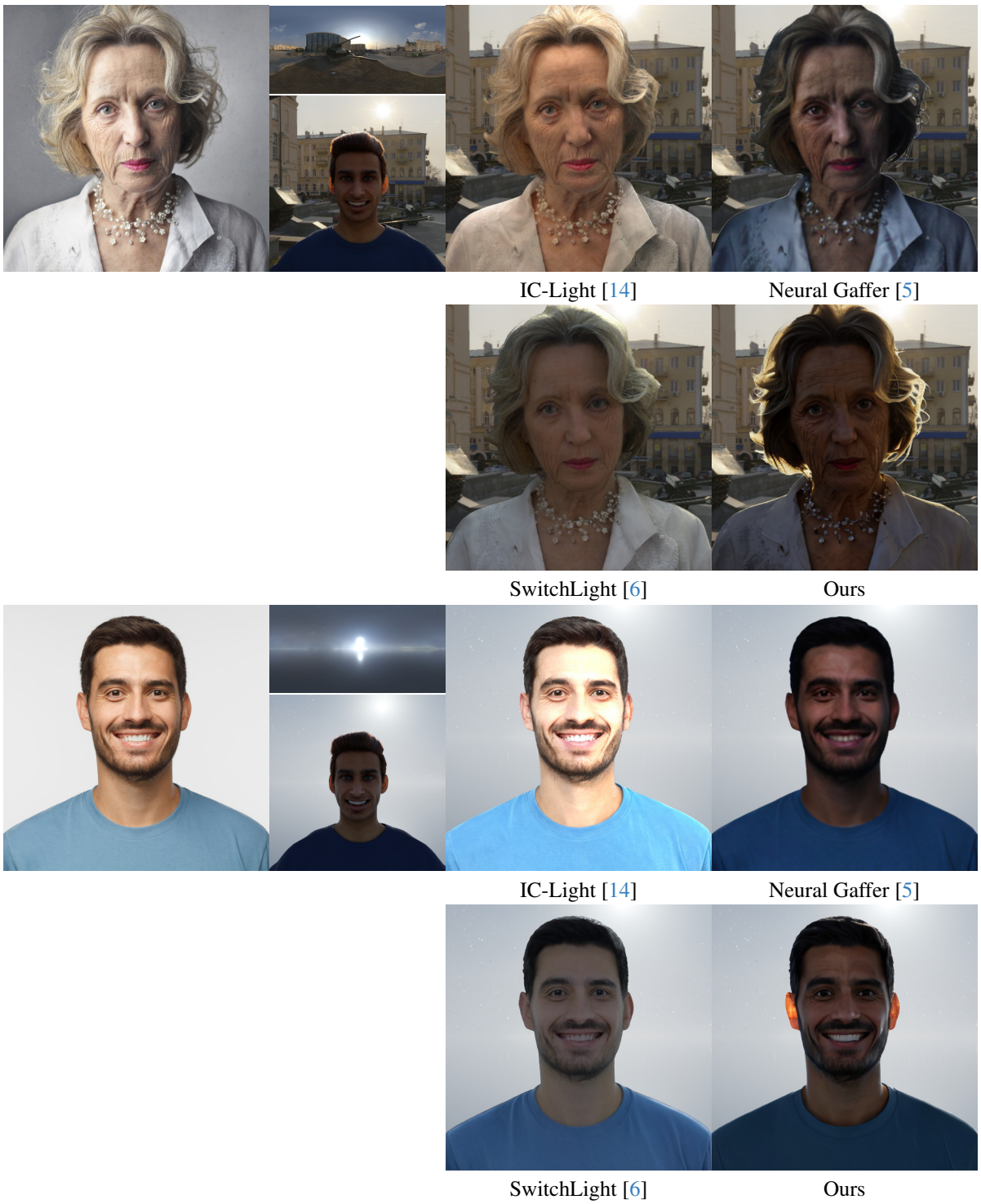

Figure 4. We show the input portrait, the environment map used to relight and a reference synthetic data rendering from Blender (left) and results from our method and baselines (right). SynthLight achieves lighting effects such as rim-light on hair (top) and subsurface scattering in ears (bottom).

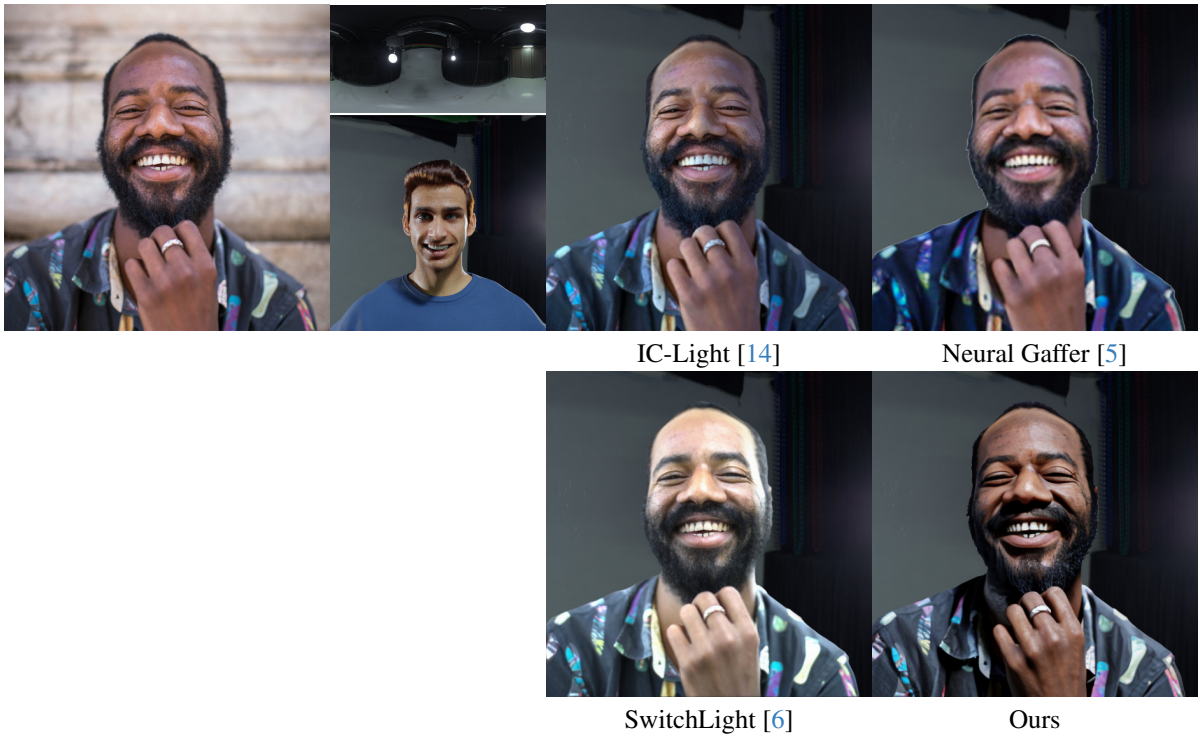

Figure 5. We demonstrate lighting effects that our method achieves such as specular highlights.

	Test Synthetic				Test Light Stage			
Method	LPIPS↓	SSIM↑	PSNR↑	FN↓	LPIPS↓	SSIM↑	PSNR↑	FN↓
Ours (init SD 1.5)	0.069	0.937	28.299	0.195	0.177	0.808	19.317	0.188
Ours (init IC-Light)	**0.063**	**0.945**	**29.572**	**0.165**	**0.165**	**0.813**	**19.698**	**0.173**

Table 1. We evaluate our method (initialized with IC-Light) [14] with a variant initialized with SD 1.5 [11]. All tables in both main paper and the supplementary, including non-inference specific ablations, are generated with classifier-free guidance parameters, $\lambda_T = 2$, $\lambda_I = 3$. See main paper for detailed descriptions of them.

C. Additional Implementation Details

Network Architecture The inputs to SynthLight are a portrait image and an environment map, both with a resolution of 512×512. The environment map is transformed from high-dynamic range to low-dynamic range through the following sequence of operations: clipping to range $[0, 65536]$, normalization to range $[0, 1]$, and exponentiation by $\frac{1}{2.2}$. These inputs are encoded into latents of shape $64 \times 64 \times 4$ using the VAE from Stable Diffusion.

SynthLight extends Stable Diffusion 1.5 by adding 8 additional channels to the first convolutional layer of the Unet, yielding a total of 12 channels (4 each for the denoising latent, input portrait, and environment map). The weights for these extra channels are initialized to 0.

Training and Inference We evaluate the performance of training with SD 1.5 initialization compared to IC-Light initialization (see Tab. 1 and Fig. 7). While IC-Light initialization yields slightly better test set performance—prompting us to report it as our primary method—in Fig. 7, even without IC-Light, our method generates advanced lighting effects, such as strong cast shadows and subsurface scattering in the ear. Conversely, without our training and inference procedures, IC-Light alone cannot produce the nuanced lighting effects (e.g. rim-effects, subsurface scattering and specular highlights) as illustrated in Fig. 4 and Fig. 5.

During training, a foreground mask is applied to the input portrait. Each condition—input portrait, environment map, and text prompt—is randomly dropped with a probability of 0.1. For inference, classifier-free guidance is applied with $\lambda_I = 3$, $\lambda_T = 2$, and the prompt "A nice person."

Figure 6. We show the input portrait, the environment map used to relight and a reference synthetic data rendering from Blender (left) and results from our method and ablations (right). We demonstrate the impact of fine-tuning with our synthetic dataset. The base model, IC-Light [14], without this fine-tuning, is unable to relight images using an environment map.

Ablation Details *Base* serves as the baseline model, trained solely on the synthetic dataset. During inference, it omits *inference adaptation*, meaning no classifier-free guidance is applied to the input portrait. *Base + Multi-Task* incorporates additional training with LAION data using a text-to-portrait task, where the input portrait and environment maps are dropped. The relighting and text-to-portrait tasks are mixed in a 7:3 ratio. *Base + Inference Adaptation* applies classifier-free guidance on input portrait, while keeping the same training configuration as *Base*. Finally, *Ours* combines both strategies. We train an additional model where light stage-rendered data complement the synthetic dataset for relighting – *Ours + Light Stage*.

D. User Study

We provide additional details about our user study. Screenshots illustrating the setup can be found in Fig. 12 and Fig. 13. The user study is conducted in three phases, with each phase focusing on a specific aspect of evaluation:

Phase 1: Visual Quality In the first phase, participants are asked to specify their preference between our method and the baseline in terms of *visual quality*. Each comparison is presented as a 2 alternative forced choice.

Phase 2: Lighting In the second phase, participants evaluate the *lighting* of the renderings. To aid their judgment, we provide a synthetic reference rendered in Blender under

Figure 7. We show the input portrait, the environment map used for relighting, and a 3D model rendered in Blender for reference (left). On the right, we present results with IC-Light and SD 1.5 initialization for finetuning on our synthetic dataset. We note that while IC-Light initialization yields slightly better performance on our light stage test set, both are comparable in terms of visual quality and achieve realistic lighting effects such as shadows and subsurface scattering.

the same environment map. This phase also uses a 2 alternative forced choice format.

Phase 3: Identity In the final phase, participants assess the *identity* of the renderings. A reference input portrait is provided, and users judge which option better preserves the subject's identity. As with the previous phases, this is conducted as a 2 alternative forced choice task.

General Instructions Participants are instructed to choose at random if making a selection is too difficult. At the beginning of each phase, a tutorial question is presented, where the answer is obvious. For example, in these cases:

- One example has severe degradation in visual quality.
- The lighting in one example is clearly incorrect.
- One rendering fails to match the reference identity.

The correct answer and the reasoning are explained to participants to familiarize them with the task.

Study Statistics The study consists of 30 questions in total, including three tutorial questions (one per phase). Participants can opt to exit the study at any time. In total, we collected 482 responses from 20 participants over a one-week period.

E. Limitations

Fig. 15 highlights some limitations observed with our method. We notice minor loss of detail, particularly in small or intricate facial features. This can be attributed to limited camera pose diversity in our synthetic dataset, i.e. headshot-only renderings, and the reliance on Stable Diffusion 1.5, which causes our method to inherit image reconstruction artifacts from Stable Diffusion's VAE. These issues can be mitigated by leveraging larger models with with better VAEs, such as those in Flux or Stable Diffusion 3, and incorporating greater camera pose variation in our synthetic dataset.

Fig. 15 illustrate another failure mode where our method struggles with accurately capturing cloth textures. While this limitation is rare, it arises from the restricted range of materials and textures used for clothing in the synthetic dataset. Expanding the diversity and quality of the dataset's cloth-related materials could effectively address this issue and is left for future work.

Figure 8. We show additional comparisons against baselines, illustrating that our method produces accurate lighting, that qualitatively matches the rendering of a 3D model in Blender, while preserving identity and maintaining high visual quality.

Figure 9. We show the input portrait, the environment map used to relight and a reference synthetic data rendering from Blender (left) and results from our method and ablations (right). Examples show the contributions of each component in our proposed method. The *Base* model, without multi-task training or inference adaptation, struggles with identity preservation and detail reproduction. *Base + Multitask* and *Base + Inference Adaptation* improve details but fail to replicate complex features like accessories and textures. Our method successfully preserves identity and reproduces intricate details, such as the cigarette (top) and specularity of the necklace (bottom).

Figure 10. Overexposure issues. SwitchLight, trained on light stage data, has overexposure artifacts and produces unnatural skin tones. *Ours + Light Stage* reduces this issue but retains some artifacts while *Ours*, trained on synthetic data alone, doesn't exhibit these artifacts.

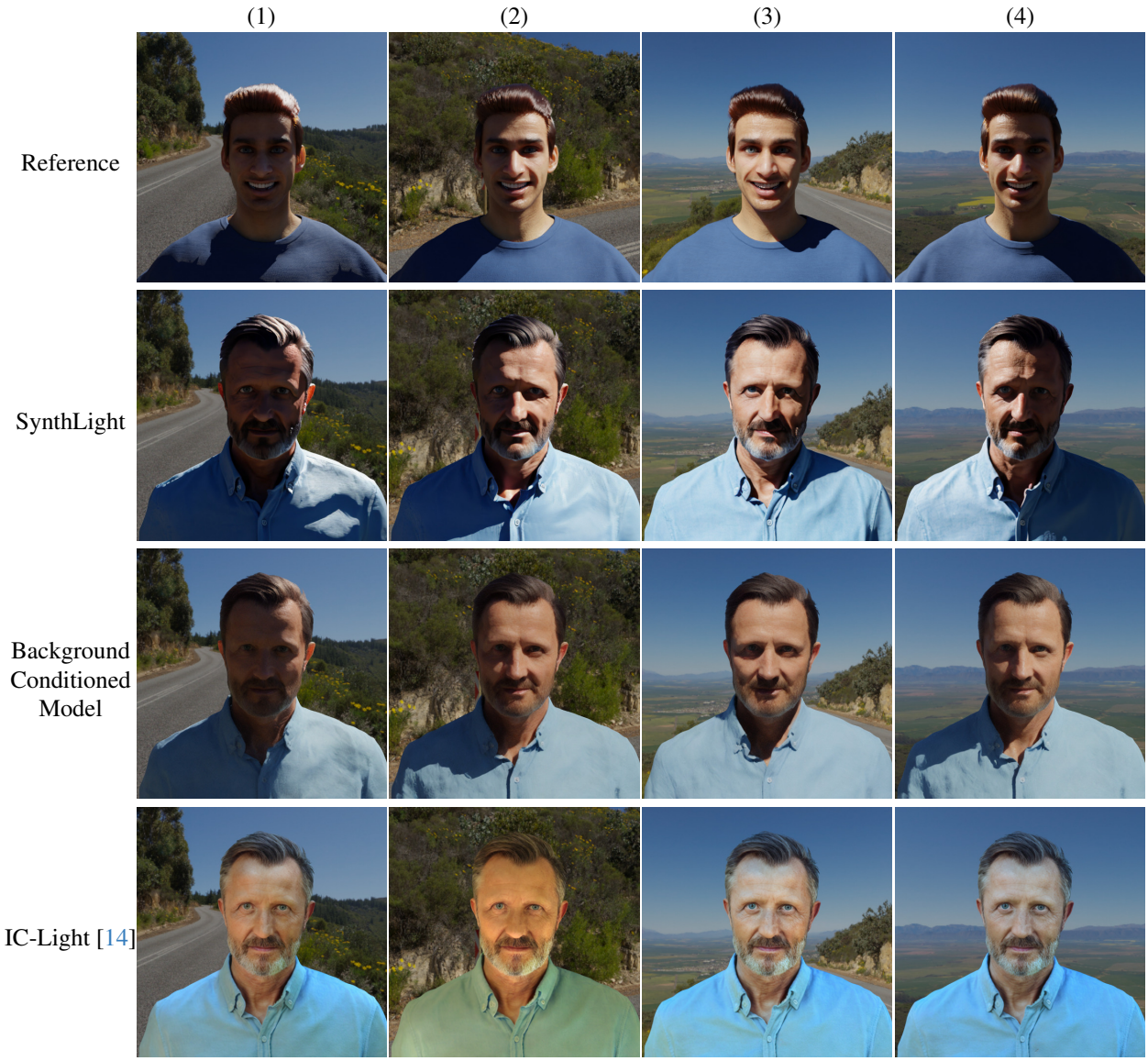

Figure 11. *Background vs Environment Map as Lighting Condition*: We compare SynthLight with a background conditioned variant of SynthLight and IC-Light and show a reference model rendered in Blender (top row). Background contains insufficient lighting cues, causing a background conditioned model to generate inaccurate lighting (columns 3-4). While challenging for prior background-conditioned relighting methods like IC-Light, by leveraging our synthetic dataset, the background conditioned model can still generate lighting effects like strong cast shadows.

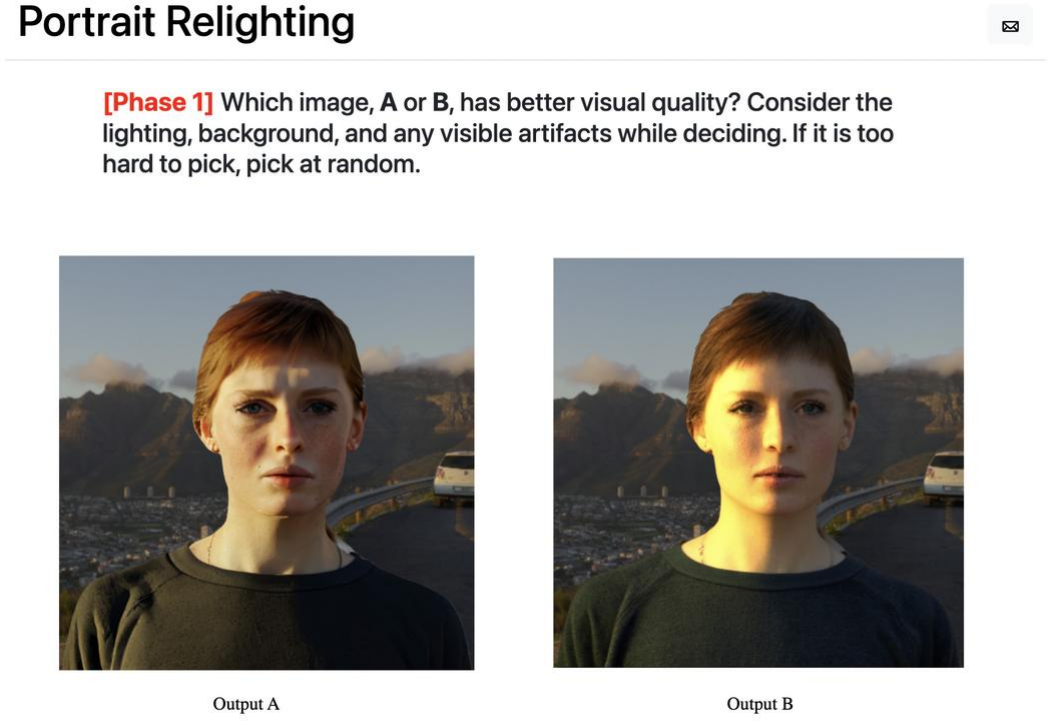

Figure 12. *User Study*: We ask users to pick between our method and baseline on visual quality of image (top) and lighting, with a given reference (bottom).

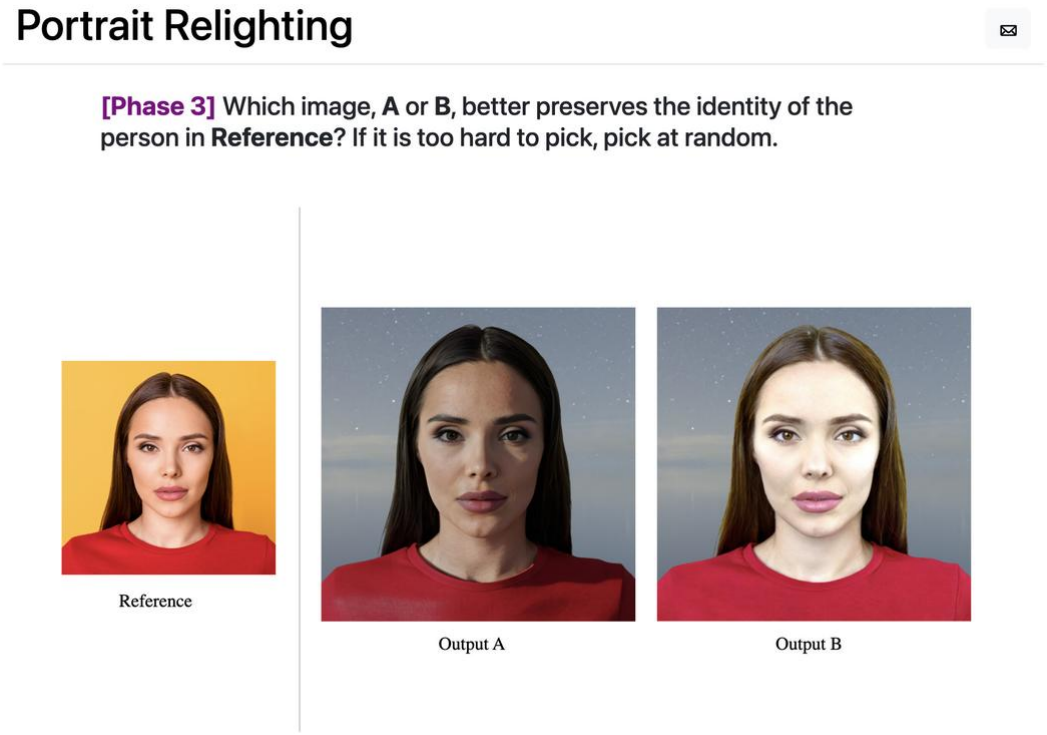

Figure 13. We ask users to judge identity preservation by providing a reference identity and asking them to select between our method and baseline.

Figure 14. More examples from synthetic dataset. Each group of four represents a subject rendered under four different lighting conditions.

(a) We observe *minor detail loss* in facial features, such as the eyes, arising from limited camera pose diversity and Stable Diffusion 1.5's VAE [11] artifacts. Mitigations include using improved VAEs (e.g., Flux [7], Stable Diffusion 3 [4]) and enhancing pose variation in the dataset.

Figure 15. Limitations of our method include minor detail loss in full-body portraits and inaccuracies in cloth texture.

References

[1] Adobe Stock. https://stock.adobe.com/. Accessed: 2025-03-22. 1

[2] Pexels. https://www.pexels.com/. Accessed: 2025-03-22. 1

[3] Unsplash. https://unsplash.com/. Accessed: 2025-03-22. 1

[4] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*. JMLR.org, 2024. 15

[5] Haian Jin, Yuan Li, Fujun Luan, Yuanbo Xiangli, Sai Bi, Kai Zhang, Zexiang Xu, Jin Sun, and Noah Snavely. Neural Gaffer: Relighting Any Object via Diffusion. In *Advances in Neural Information Processing Systems*, 2024. 5, 6, 9

[6] Hoon Kim, Minje Jang, Wonjun Yoon, Jisoo Lee, Donghyun Na, and Sanghyun Woo. SwitchLight: Co-design of Physics-driven Architecture and Pre-training Framework for Human Portrait Relighting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25096–25106, 2024. 1, 5, 6, 9, 11

[7] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 15

[8] Rohit Pandey, Sergio Orts Escolano, Chloe Legendre, Christian Häne, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Trans. Graph.*, 40(4), 2021. 9

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 1

[10] Mengwei Ren, Wei Xiong, Jae Shin Yoon, Zhixin Shu, Jianming Zhang, HyunJoon Jung, Guido Gerig, and He Zhang. Relightful Harmonization: Lighting-Aware Portrait Background Replacement. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6452–6462, 2024. 1

[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2022. 6, 8, 15

[12] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 1

[13] Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. DiLightNet: Fine-grained Lighting Control for Diffusion-based Image Generation. In *ACM SIGGRAPH 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery. 9

[14] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling In-the-Wild Training for Diffusion-based Illumination Harmonization and Editing by Imposing Consistent Light Transport. In *International Conference on Learning Representations*, 2025. 1, 5, 6, 7, 8, 9, 12