

---

# Supplementary Material for DreamLight: Towards Harmonious and Consistent Image Relighting

---

Anonymous Author(s)

Affiliation

Address

email

## A User Study

We conduct user study involving 25 volunteers to evaluate the results of different models on 30 samples for each setting. For both image-based and text-based relighting, we show users the generated results of the corresponding methods. Users are unaware of the source model generating the images, and the image sequences in different samples are randomly shuffled. Participants are asked to rank the outputs based on their **perceptual quality**, **lighting rationality**, and **subject consistency**, respectively. In detail, perceptual quality means which looks better. Lighting rationality refers to whose foreground’s light or shadows fit the background better. Subject consistency indicates whose foreground’s appearance is best preserved. Results shown in Table 1 demonstrate the superiority of our approach. The Ranking is computed by averaging the rankings of all participants over all samples and thus a lower Ranking is better.

Table 1: User study of both image-based and text-based settings. The lower average ranking is better. HN, PP, and BN denote Harmonizer [1], PowerPaint [2], and BrushNet [3], respectively.

|                     | <i>Image-based Relighting</i> |          |        |             |             |
|---------------------|-------------------------------|----------|--------|-------------|-------------|
|                     | HN[1]                         | INR[4]   | PCT[5] | IC-Light[6] | Ours        |
| Perceptual Ranking  | 3.81                          | 3.66     | 3.97   | 2.17        | <b>1.39</b> |
| Lighting Ranking    | 3.72                          | 3.71     | 4.06   | 1.85        | <b>1.66</b> |
| Consistency Ranking | 2.87                          | 2.76     | 2.96   | 4.18        | <b>2.23</b> |
|                     | <i>Text-based Relighting</i>  |          |        |             |             |
|                     | PP-V1[2]                      | PP-V2[2] | BN[3]  | IC-Light[6] | Ours        |
| Perceptual Ranking  | 4.46                          | 2.92     | 3.22   | 2.72        | <b>1.67</b> |
| Lighting Ranking    | 3.87                          | 3.45     | 3.64   | 2.16        | <b>1.88</b> |
| Consistency Ranking | 3.73                          | 2.58     | 2.87   | 3.54        | <b>2.28</b> |

## B More Visualization Results

### B.1 Image-based Results

Figure 1 illustrates more image-based relighting results. It can be seen that our model has an excellent perception of lights of different directions and color tones. Furthermore, we perform relighting on the same foreground object by background images with lights from left to right in Figure 2. It can be observed that the foreground undergoes sequential illumination variations: initially exhibiting brighter intensity on the left side with progressive darkening towards the right, subsequently transitioning into a backlit condition, followed by a complete reversal of lighting configuration characterized by enhanced brightness on the right side and diminished intensity on the left.

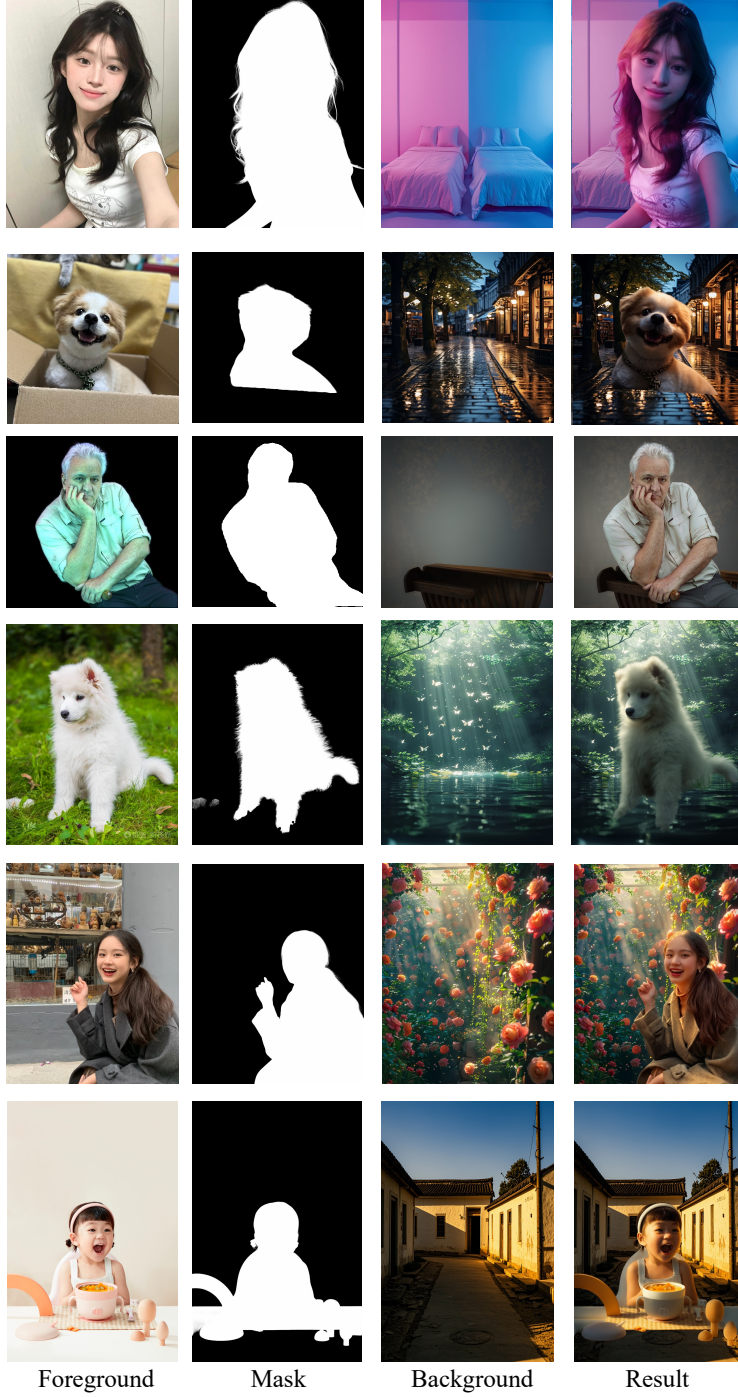


Figure 1: Visualization of the image-based relighting results.

## 21 B.2 Text-based Results

22 Figure 3 shows the relighting results with the guidance of text prompt. The text prompt can specify  
 23 either the light or the scene and our model is capable of generating the harmonious results. Besides,  
 24 we also test the adaption ability of our model to different seeds and the results are shown in Figure 4.  
 25 The prompt is “beach and palm trees in Hawaii”. The figures are not refined by the fixer for better  
 26 illustration. It can be seen that our model is robust to different seeds.

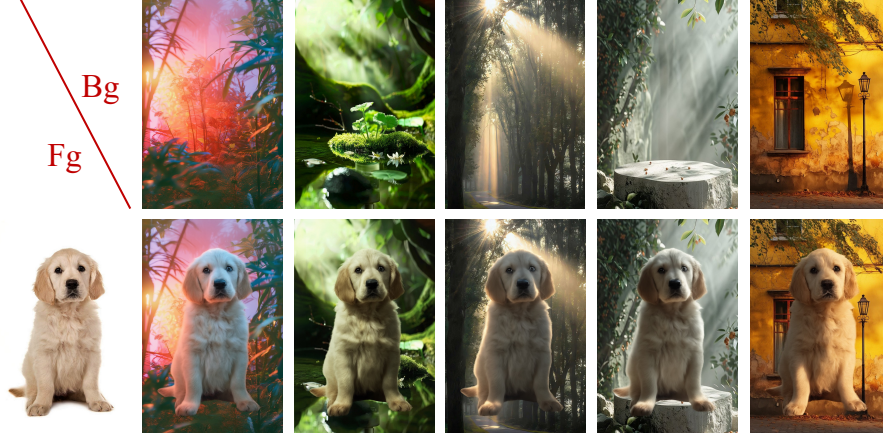


Figure 2: Image-based relighting results with light from different directions.

### 27 B.3 Training data illustration

28 Figure 5 shows the high-quality examples of the generated training data. The first two lines are  
 29 generated through our relighting ominicontrol [7] lora. The third row is result generated by 3D  
 30 rendering. The last row is the image generated by IC-Light [6]. It can be seen that the generated  
 31 data is rich in light and color variations. However, automatic generation processes often result in the  
 32 presence of low-quality data. For example, the data created using relight lora will have problems  
 33 such as unreasonable lighting and inconsistent foreground. Therefore we need to filter low quality  
 34 data using aesthetic models as well as MLLMs. This is one of the reasons why we need our model  
 35 and do not directly utilize existing data generation pipeline for relighting.

### 36 B.4 Comparison with methods that require environment map

37 Figure 6 illustrates the comparison between our DreamLight and Graffer [8] that requires environment  
 38 map as input. Results in the red boxes are generated by our DreamLight. Our method is performed  
 39 on environment map by cropping the background image from it. It can be seen that since our method  
 40 is not trained on environment maps, it performs inferior in the consistency of light changes than  
 41 methods that require environment maps. However, Graffer is prone to unnatural lighting effects and  
 42 is greatly influenced by whether there is a strong light source. As a contrast, our method achieves  
 43 more natural relighting results.

## 44 C More Details

### 45 C.1 Training Process of the Fixer

46 : In this part we elaborate on the training process of Spectral Foreground Fixer (SFF). As described  
 47 in the main paper, we utilize the MSE loss and perceptual loss to provide supervision for SFF. This  
 48 process can be formulated as:

$$\begin{aligned}\mathcal{L}_{recon}(x, y) &= ||x - y||_2 + \lambda * \mathcal{L}_{percep}(x, y), \\ \mathcal{L}_{percep}(x, y) &= ||F_x - F_y||_2,\end{aligned}\tag{1}$$

49 where  $F$  denotes the encoding network such as VGG.  $x$  and  $y$  indicate the prediction and correspond-  
 50 ing ground truth. The supervision is applied on both the predicted high-frequency part  $HQ'_{in}$  and  
 51 entire output image  $I'_{out}$ . Thus, the final loss for SFF is defined as:

$$\mathcal{L}_{SFF} = \mathcal{L}_{recon}(HQ'_{in}, HQ^{gt}) + \mathcal{L}_{recon}(I'_{out}, I^{gt}).\tag{2}$$

### 52 C.2 Evaluation data generation process

53 The images are rendered via Arnold Renderer. First, we generate a rectangle mesh with the same  
 54 aspect ratio as the input photo. Then we use the paid Switchlight API to split the photo into albedo,



Figure 3: Visualization of the text-based relighting results.

normal, roughness and specular maps. We construct the Arnold standard surface material and apply it to the mesh by using the above maps. To get better shading effect, we use the height maps generated by DeepBump as the displacement maps, which produces a smoother transition at edges and leads to better render results. Finally, we render the relighting photo by using an Arnold orthographic camera to ensure the position of each pixel is the same as the source image. Figure 7 illustrates some generated evaluation cases. It can be seen that the evaluation set contains samples with diverse lighting, *e.g.*, strong neon effects as well as natural sunlight.

## D Limitations

Although the proposed method achieves harmonious relighting and guarantees the consistency of the subject, it has limitations in some specific scenarios. Specifically, since PGLA is focusing attention to image’s left, right, top, and bottom sides, it may struggle to handle lighting conditions where the source is outside the camera’s field of view. The quality of the corresponding results are relied on the





Figure 4: Visualization of the text-based relighting results of different seeds.



Figure 5: Examples of the generated training data.

67 training data and learning prior. In addition, since there exists no training data for gradual changes  
 68 in light, the model is slightly less consistent when dealing with this situations such as providing  
 69 environment map.

## 70 E Broader Impacts

71 The proposed model is designed for achieving harmonious and consistent relighting. It is capable of  
 72 empowering non-professionals (*e.g.*, content creators, small businesses) to achieve high-quality visual  
 73 consistency, democratizing advanced photo/video editing tools. The potential negative impact is the  
 74 risks misuse for creating hyper-realistic manipulated media (*e.g.*, falsifying scenes, impersonations),  
 75 exacerbating misinformation and trust crises. Therefore, users should follow ethical and legal usage  
 76 norms.

## 77 References

- 78 [1] Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and Rynson WH Lau. Harmonizer: Learning to  
 79 perform white-box image and video harmonization. In *ECCV*, pages 690–706, 2022.
- 80 [2] Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one  
 81 word: Learning with task prompts for high-quality versatile image inpainting. *arXiv preprint*  
 82 *arXiv:2312.03594*, 2023.

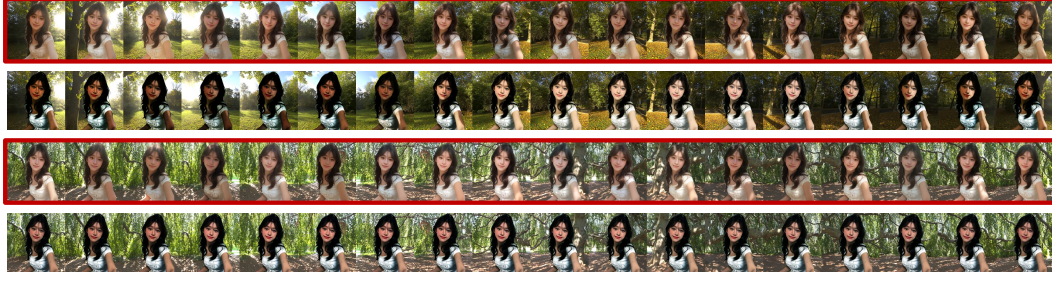


Figure 6: Relighting comparison with method that require environment map. Best viewed zoom-in.

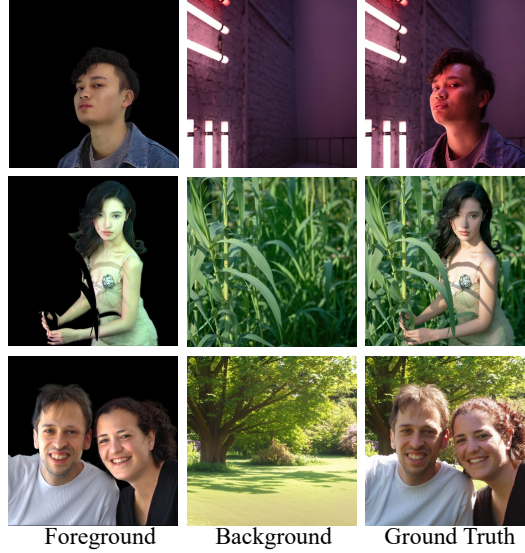


Figure 7: Examples of the generated evaluation data.

- 83 [3] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A  
 84 plug-and-play image inpainting model with decomposed dual-branch diffusion. *arXiv preprint*  
 85 *arXiv:2403.06976*, 2024.
- 86 [4] Jianqi Chen, Yilan Zhang, Zhengxia Zou, Keyan Chen, and Zhenwei Shi. Dense pixel-to-pixel  
 87 harmonization via continuous image representation. *IEEE Transactions on Circuits and Systems*  
 88 *for Video Technology*, 2023.
- 89 [5] Julian Jorge Andrade Guerreiro, Mitsuru Nakazawa, and Björn Stenger. Pct-net: Full resolution  
 90 image harmonization using pixel-wise color transformations. In *Proceedings of the IEEE/CVF*  
 91 *Conference on Computer Vision and Pattern Recognition*, pages 5917–5926, 2023.
- 92 [6] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based  
 93 illumination harmonization and editing by imposing consistent light transport. In *ICLR*, 2025.
- 94 [7] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol:  
 95 Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 2024.
- 96 [8] Haian Jin, Yuan Li, Fujun Luan, Yuanbo Xiangli, Sai Bi, Kai Zhang, Zexiang Xu, Jin Sun,  
 97 and Noah Snavely. Neural gaffer: Relighting any object via diffusion. *arXiv preprint*  
 98 *arXiv:2406.07520*, 2024.