# Text2Relight: Creative Portrait Relighting with Text Guidance

**Junuk Cha[1,2*], Mengwei Ren[2], Krishna Kumar Singh[2], He Zhang[2], Yannick Hold-Geoffroy[2], Seunghyun Yoon[2], HyunJoon Jung[2], Jae Shin Yoon[2†], Seungryul Baek[1†]**

[1] UNIST
[2] Adobe Research

## Abstract

We present a lighting-aware image editing pipeline that, given a portrait image and a text prompt, performs single image relighting. Our model modifies the lighting and color of both the foreground and background to align with the provided text description. The unbounded nature in creativeness of a text allows us to describe the lighting of a scene with any sensory features including temperature, emotion, smell, time, and so on. However, the modeling of such mapping between the unbounded text and lighting is extremely challenging due to the lack of dataset where there exists no scalable data that provides large pairs of text and relighting, and therefore, current text-driven image editing models does not generalize to lighting-specific use cases. We overcome this problem by introducing a novel data synthesis pipeline: First, diverse and creative text prompts that describe the scenes with various lighting are automatically generated under a crafted hierarchy using a large language model (*e.g.,* ChatGPT). A text-guided image generation model creates a lighting image that best matches the text. As a condition of the lighting images, we perform image-based relighting for both foreground and background using a single portrait image or a set of OLAT (One-Light-at-A-Time) images captured from lightstage system. Particularly for the background relighting, we represent the lighting image as a set of point lights and transfer them to other background images. A generative diffusion model learns the synthesized large-scale data with auxiliary task augmentation (*e.g.,* portrait delighting and light positioning) to correlate the latent text and lighting distribution for text-guided portrait relighting. In our experiment, we demonstrate that our model outperforms existing text-guided image generation models, showing high-quality portrait relighting results with a strong generalization to unconstrained scenes.

**Project page —**
https://junukcha.github.io/project/text2relight/

## Introduction

Light, as a physical phenomenon, has been extensively studied in computer graphics to model its behavior as accurately as possible from real-world observations using a number of parameters such as color, intensity, and direction. The advent of powerful generative AI technologies such as a denoising diffusion model (Ho, Jain, and Abbeel 2020) and large language models (Radford et al. 2019) have shifted the research paradigm from how *accurately* to how *creatively* one can model such a physical phenomenon. Is it possible to model the lighting behavior as a function of our emotions? For example, how does a *joyful* lighting look like? In this paper, we break the boundary between the physical and creative space by introducing Text2Relight, a lighting-specific foundational model that can perform relighting (*i.e.,* remove the original lighting and apply a novel one) of a single portrait image driven by a text prompt as shown in Figure 1. Our main assumption is that there exists unbounded creativeness in language, and correlating this with lighting enables the mapping from a physical space to a creative one.

One simple way to achieve Text2Relight is to utilize existing foundational models for text-guided image editing such as InstructPix2pix (Brooks, Holynski, and Efros 2023). However, such a general foundational model which never learns from lighting-specific data, *e.g.,* identical scenes captured under different lighting, encodes a weak correlation of the text with lighting, producing a large distortion of the original contents as shown in Figure 2. This necessitates a new foundational model that can control the image from the lighting space that is completely decomposed from the contents space (*e.g.,* shape and intrinsic appearance).

However, developing such a lighting-specific model is challenging due to the lack of data pairs for relighting, *i.e.,* the images of identical scene and main subject captured under different lighting conditions, associated by a text description. While existing methods (Pandey et al. 2021; Saito et al. 2023) have captured the relighting data using expensive infrastructure such as lightstage system, such lab-controlled data are often not scalable (particularly for the axis of human identities) and the rendering of image relighting is often applicable to only foreground human region where the background scene is simply composed with a part of preset panorama images. Those limited imaging data, in turn, restricts the diversity of the labeled text prompts as well.

To address this data challenge, we propose a scalable data simulation pipeline that can synthesize the relighting data of a portrait scene for both foreground and background, and

*This research was conducted when Junuk Cha was an intern at Adobe Research.
†These authors are co-last authors.


Figure 1: **Text2Relight** generates the image of a relighted portrait (right) as a condition of a text prompt while keeping original contents in an input image (left).


Figure 2: Results from existing text-guided image editing models (left: (Brooks, Holynski, and Efros 2023), right: (Fu et al. 2023)) which largely distorts the input images by generating new contents.

associated text prompts. Our synthesis pipeline is designed with bottom-up fashion: text generation, text-aware lighting image generation, and image-based relighting. A large language model automatically generates diverse and creative text prompts based on our crafted language hierarchy to describe lighting environment of a scene. Text-guided image generation models generate an RGB image or an HDR panorama map as a condition of a text prompt, which are used as a lighting image. Finally, the lighting distribution of the generated lighting image is transferred to a portrait image using various image-based relighting methods; where a number of factors including data availability, algorithm maturity, and scene complexity lead us to approach the image-based relighting differently for background and foreground.

For foreground relighting, we develop an end-to-end portrait relighting model that can control the lighting of an input image as a function of a background image; but when OLAT (One-Light-at-A-Time) images are available from lightstage, we apply HDR rendering techniques using the generated panorama map. Since our method handles the case with only a single image, it can be applicable to any in-the-wild portrait scene.

For background relighting, we represent the scene lighting as a set of point lights (Kocsis, Sitzmann, and Nießner 2023) and introduce a robust and efficient lighting optimization method. The positions of the point lights are initialized with a distance-based localization algorithm; and they are jointly optimized with other learnable variables (*e.g.,* intensity and diffusion parameters) by minimizing the photometric difference from the lighting image. We relight a background image by transporting the optimized light sources using an inverse rendering techniques.

Using our large simulation data, we develop a lighting-specific foundational model by repurposing of an existing text-guided image editing model (Brooks, Holynski, and Efros 2023). In training time, the model jointly learns with an auxiliary task such as portrait shadow removal and text-guided light positioning to improve the geometric awareness and better intrinsic appearance modeling.

In our experiments, we show the analysis that our hierarchical text generation is indeed useful to push the distribution of the text diversity. Our method outperforms existing text-guided foundational models in terms of pixel-wise perceptual distance, AI-based semantic score, and user preference score. We also provide in-depth ablation studies on the crafted language hierarchy, various data sources, and conditional variables to validate the effectiveness of our data simulation and model training pipeline. As applications, we demonstrate portrait shadow removal, light positioning, and background harmonization.

In summary, our main contributions include:

- Text2Relight, a new formulation of a lighting-specific foundational model for text-guided portrait relighting.
- A novel and scalable data simulation pipeline to synthesize relighted images and associated text prompts.
- A method to generate largely distributed light-aware text prompts with crafted hierarchy.
- Performance enhancement by joint training with auxiliary tasks (shadow removal and light positioning).

## Related Work

**Portrait Relighting** refers to the process of adjusting the lighting conditions in a portrait image including direction, intensity and color, so as to simulate different lighting scenarios, or creatively manipulate the lighting to achieve desired aesthetic effects.

Image-based human relighting has been extensively studied since it requires a minimum requirement (*e.g.,* only a single image) for practical real-world application. While many

existing works have introduced a relighting model targeting various human body parts such as face, portrait, and full body, their fundamental approaches consist of two steps: intrinsic decomposition and shading from a target lighting. Previous works first predict the intrinsic appearance of humans such as albedo, shading, and geometry from a single image using a neural network, and re-render the shading using graphics techniques such as HDR rendering from a panorama environment (Pandey et al. 2021; Wang et al. 2023; Yeh et al. 2022; Ji et al. 2022) and ray tracing (Hou et al. 2022). Some methods (Hou et al. 2021; Zhou et al. 2019) learn to generate shading as a function of a small number of spherical harmonics, and encoding the latent lighting rotation features onto the relighting model is also possible by learning from the augmented panorama environment maps with rotations (Song et al. 2021). To improve physical plausibility, some methods predict more detailed intrinsic values such as roughness and reflectivity (Kim et al. 2024) or predict more residual appearance (Nestmeyer et al. 2020; Tajima, Kanamori, and Endo 2021; Ji et al. 2022) that can not be modeled by a simple intrinsic decomposition. Conditioning explicit geometry such as 3D face model (Zhou et al. 2019; Ponglertnapakorn, Tritrong, and Suwajanakorn 2023) and 3D tri-plane (Mei et al. 2024) in the latent space further helps to improve the geometric plausibility of the lighting behavior. Data-driven lighting estimation from a user scribble (Mei et al. 2023) further enhances the controllability of the lighting. Other than two-step relighting approaches, some works introduced a method that can directly change the lighting distribution from an image to improve the relighting efficiency. A neural network jointly decodes the latent person image as well as the latent lighting parameters such as spherical harmonics (Zhou et al. 2019), latent background feature (Ren et al. 2023), or latent illumination map (Sun et al. 2019).

While promising, existing relighting methods mainly focus on the foreground portrait with the requirement of expensive conditional variables such as HDR panoarma maps. Since they do not support text-based lighting editing, it greatly limits the flexibility and creativity for adjusting the lighting in the image.

**Text-Guided Image Editing.** Relevant to our task, recent advancements in diffusion models have significantly enhanced the capacity for sophisticated image editing directed by textual descriptions. (Avrahami, Lischinski, and Fried 2022; Nichol et al. 2021) enable precise local editing within specified regions of an image using masks, such as object replacement or object style transfer. The advantage of local editing is that it allows modification of the desired regions while preserving the rest of the image. In contrast, InstructPix2Pix (Brooks, Holynski, and Efros 2023) globally edits the source image to match the text prompt. While it provides an overall natural-looking image, it sometimes modifies the context too aggressively, altering the identity of the portraits and/or the characteristics of the scenes. To address this issue, MGIE (Fu et al. 2023) utilizes a multimodal large language model to obtain more expressive instructions, which aids in precise local editing while making global adjustments. Imagic (Kawar et al. 2023) further enables complicated non-rigid editing by optimizing text embeddings aligned with the source and target images.

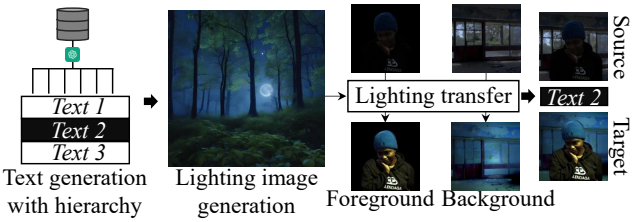


Figure 3: **Overview of the data synthesis pipeline.** We first generate a text prompt with a language hierarchy from which we generate a lighting image. Subsequently, we transfer the lighting from the lighting image to a portrait image captured from either lightstage or real world (with background inpainting). These form the *training dataset* for our Text2Relight model.

Nevertheless, current text-based editing methods are primarily designed for general use cases, lacking specific considerations in modeling the lighting distribution that may fall beyond the existing large-scale training data. Consequently, they often fail when presented with relighting-specific prompts. Additionally, to our knowledge, there are no meticulously curated prompt-image pairs tailored for fine-tuning these models for the relighting tasks.

## Method

### Problem Formulation

Given a portrait image, our goal is to control the lighting of the scene for both foreground and background driven by a text prompt, while ensuring that the original content and identity are preserved: $\tilde{I} = f_\theta(I, M, T)$, where $\theta$ denotes the learnable parameters, and $f$ is the text-guided relighting function which takes as input source image $I$, foreground mask $M$, and text prompt $T$, and generates the relighted image $\tilde{I}$.

To learn this mapping function, $f$ needs to be trained with the ground truth $\tilde{I}_{gt}$. However, no dataset provides pairs of the corresponding texts and relighted images that preserve the content and identity of the original images. To address those problems, we propose a novel dataset synthesis method consisting of three main components of text generation, text-driven lighting image generation, and image-based relighting as shown in Figure 3.

We first generate diverse light-aware texts using our crafted language hierarchy. We use the generated text as a prompt to generate a lighting image in the form of either RGB image or HDR panorama map. We transfer the lighting distribution from the generated lighting image to the foreground portrait and background scene, separately. By compositing them together, we complete the synthesis of the relighted portrait image. Using the synthesized data, we develop a lighting-specific foundational model for Text-to-Relight with auxiliary task and data augmentation. The details of each component is described in below.

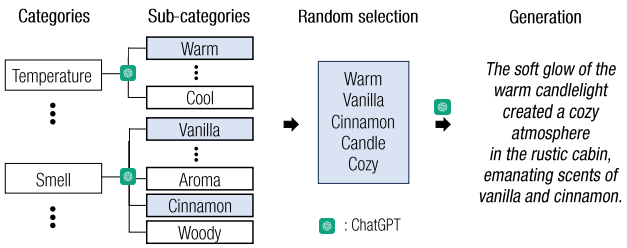


Figure 4: **Pipeline for text generation with hierarchy.** For sub-categories generation, we ask *'Generate words related to* {category}. *Write 30 or more words on a single line, separated by commas.'*; and for sentence generation, we ask *'could you describe the lighting property of a random scene using the words of* {selected words}.' to ChatGPT, respectively.

## Scalable Prompt Generation with Hierarchy

We create large-scale text prompts that describe a scene in the context of lighting distribution using a large language model (LLM), *e.g.,* ChatGPT (OpenAI 2022). However, we empirically found that existing LLMs utilize a limited range of words to generate the text prompts from a simple question *e.g., could you describe an arbitrary scene with its lighting distribution?*, which prevents us from generating diverse and creative text prompts. Our approach, instead, is to explicitly select a few words from a predefined large vocabulary pool and provide them as a constraint on LLMs *e.g., could you describe the lighting property of a random scene using the words of 'cozy' and 'warm'?*. To define such a large vocabulary pool, we formulate a categorical hierarchy. For example, humans define high-level categories related to lighting, and LLM generates various sub-categories for each category. The overall process for our hierarchical text generation is described in Figure 4.

## Lighting Image Generation

Given a text prompt, we generate two types of lighting images: RGB image and HDR panorama map as shown in Figure 5. For RGB image, we employ a latent consistency model (Luo et al. 2023) to generate the RGB image from a text prompt using four denoising steps. We use this image as the lighting image when we relight the foreground and background. For HDR panorama map, we develop a customized text-guided panorama generation model by fine-tuning a pre-trained stable diffusion model (Rombach et al. 2022) on publicly available HDR panorama maps with paired text prompts extracted from an existing vision language model. It generates an HDR panorama map from a text prompt.

## Foreground Portrait Relighting

We apply two separate methods for portrait relighting depending on the modality of the data: a single image and OLAT images. For a single-image-based portrait relighting, we develop an image-based relighting model that takes as input a single portrait image and a lighting image, and outputs a relighted image as shown in the upper row of Figure 5

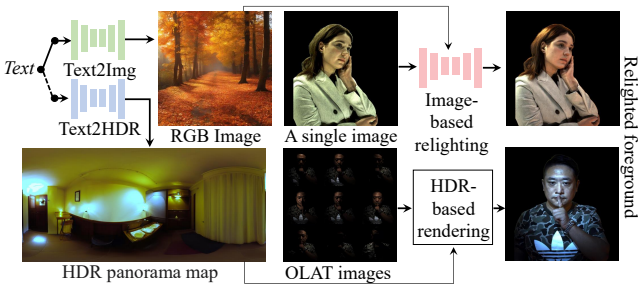


Figure 5: **Pipeline for foreground relighting**.

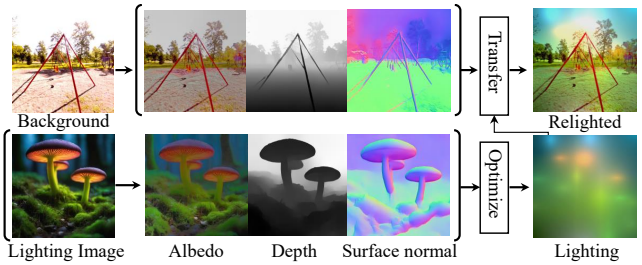


Figure 6: **Pipeline for background relighting**.

using internal lightstage dataset and many real-world data similar to existing method (Ren et al. 2023). For OLAT images captured from a lightstage as shown in the lower row of Figure 5, we incorporate a panorama map as a lighting image to pre-defined the illumination condition for each OLAT image and perform the relighting of the portrait by applying HDR rendering techniques (Zhang et al. 2021).

## Background Relighting

Many existing works (Careaga and Aksoy 2023; Kocsis, Sitzmann, and Nießner 2023) represent a background as multiple layers of intrinsic values: $I = A * S$ where $I$, $A$, and $S$ represent the background image, its albedo, and its shading, respectively. Furthermore, the shading can be described as a function of geometry and lighting: $S = s(L, G)$ where $s$ is a rendering function that outputs the shading map $S$ as a function of input lighting information $L$ and geometry $G$ which is often composed of depth $D$ and surface normal $N$, *i.e.,* $G \rightarrow \{D, N\}$. Assuming $L$ is under the definition of a point lighting (Iwahori, Sugie, and Ishii 1990), we can expand this to the multiple light sources as follows: $S = \sum_{i=1}^{n} S_i = \sum_{i=1}^{n} s(L_i, \{D, N\})$, where each $S_i$ corresponds to the shading contribution from an individual light $L_i$ and $n$ represents the number of point lights.

Since lighting $L$ is completely decomposed from other intrinsic values, it is possible to transfer the lighting distribution to other background:

$$\tilde{I} = \hat{A} * \tilde{S} = \hat{A} * \sum_{i=1}^{n} s(L_i, \{\hat{D}, \hat{N}\}), \quad (1)$$

where $\hat{\cdot}$ denotes the properties of the source background, and

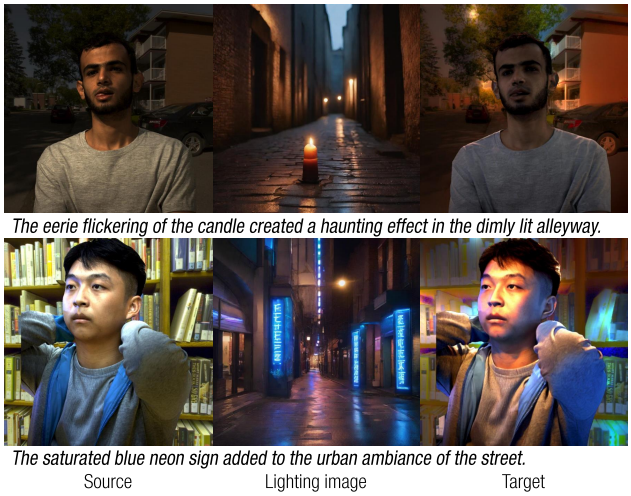


Figure 7: Examples of our simulated data. Source, lighting image, target and associated text prompt. Source and target have same content but different lighting environment.

$\tilde{I}$ and $\tilde{S}$ are the relighted background and associated shading, respectively.

To reconstruct the point lights from an image, we first initialize a set of 3D point lights (*e.g.,* 20) considering the intensity distribution of the gray-scaled image, and we optimize the parameters of each point light (*i.e.,* 3D position, lighting intensity, ellipsoid ratio, and a diffusion intensity) based on the photometric errors between the reconstructed and the original lighting image. We transfer the optimized point lights to other background images using Eq. 1.

### Data and Task Augmentation

We further push the visual and lighting diversity of the simulated dataset with various data and task augmentation strategies. For data augmentation, we perform text prompt augmentation, spatial image augmentation, real-data augmentation, and bi-direction relighting augmentation. For task augmentation, we jointly our Text2Relight model using two different tasks: portrait shadow removal (*i.e.,* remove portrait shadow or highlights while preserving original lighting environement) and light positioning (*e.g.,* position a colored light around the portrait prompted by a text).

### Model Training

We repurpose a text-guided image editing model (*i.e.,* InstructPix2Pix (Brooks, Holynski, and Efros 2023)) as a lighting-specific foundational (Text2Relight) model by learning the following objectives:

$$\mathcal{L}_{T2R}(x) = ||\epsilon - f_\theta(\{z_t, I, M\}, t, T)(x)||_2^2. \quad (2)$$

$x$ is the latent pixel position, $z_t$ is the intermediate noisy latent at time $t$, $f_\theta$ is the denoiser (*i.e.,* UNet) which predicts the noise, $T$ is text, $I$ and $M$ are portrait image and foreground mask, respectively. $M$ guides the model to focus on the foreground, helping it learn effectively and prevent to generate lighting artifacts.

| Method | SSIM ↑ | LPIPS ↓ | CVS ↑ | FIS ↑ | LS ↑ |
|---|---|---|---|---|---|
| IP2P | 0.408 | 0.531 | 0.646 | 0.008 | 3.6 |
| GLIDE | 0.464 | 0.525 | 0.643 | -0.969 | 3.6 |
| MGIE | 0.415 | 0.464 | 0.802 | 0.465 | 3.3 |
| Ours | **0.546** | **0.401** | **0.886** | **0.584** | **3.8** |

Table 1: Comparison with other state-of-the-art methods. CVS, FIS, and LS denote CLIP vision score, face identity score, and LLaVA score, respectively.

**Dataset Summary.** Overall, our data has 1.5M pairs of relighting images and associated text prompts. We create 400K and 800K data from OLAT images and a single image, respectively. We create 100K data pair for shadow removal from a lightstage and 200K data pair for light positioning using a single image. The examples of our simulated data are shown in Figure 7.

## Experiments

**Datasets.** For quantitative evaluation, we use our data simulation pipeline to synthesize the ground-truth data for text-guided portrait relighting. We use real-world portrait images from (Kvanchiani et al. 2023) as testing sets. For qualitative evaluation, we use many real-world portrait images collected from license-free stock data.

**Metrics.** We utilize various metrics to measure the score between the generated images and ground truths. SSIM (Wang et al. 2004) measures structural similarity between two images. LPIPS (Zhang et al. 2018) evaluates perceptual similarity between two images by comparing their deep features from a neural network (Simonyan and Zisserman 2014). The CLIP vision score (CVS) calculates the cosine similarity between two images using CLIP vision encoder (Radford et al. 2021). To validate how effectively the identity of the portrait is preserved, we use the face identity score (FIS) that computes the cosine similarity between two face images in the feature embedding space constructed using ArcFace (Deng et al. 2019). To measure the score of how well the text prompt matches an image, we fine-tune the LLaVA (Liu et al. 2024) using the data obtained from ChatGPT (OpenAI 2022): a question template, a concatenated image, and an answer of ChatGPT. To this end, we create the concatenated image by merging a source image and an output image in parallel and input them into the ChatGPT to get the score of the lighting adjustment on a scale from 1 (very bad) to 5 (very good). The fine-tuned LLaVA measures the lighting adjustment score from 1 to 5.

**Baselines.** We compare our model with IP2P (Brooks, Holynski, and Efros 2023), GLIDE (Nichol et al. 2021), and MGIE (Fu et al. 2023) in Table 1. For the fair comparison, we use an instruction-based prompt template 'Change the portrait under lighting of {*Text prompt*}' for existing models since they tend to distort the content when using {*Text prompt*} as the text prompt.

### Comparison

IP2P (Brooks, Holynski, and Efros 2023) globally edits the source image based on the text prompt, and it some-

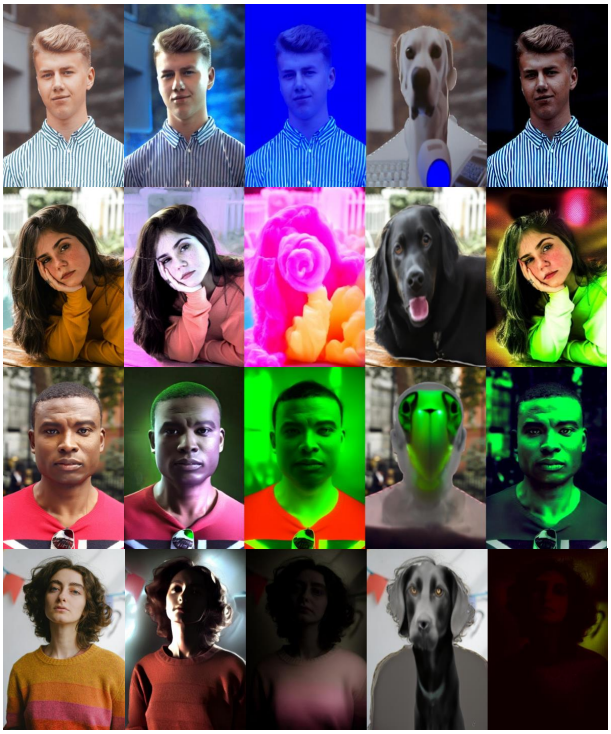


Figure 8: Qualitative comparison with text prompts : 'The blue light of the computer monitor', 'Sweet cotton candy', 'Eerie glow of green fluorescent lighting', and 'The chilling atmosphere of the fear-inducing dark room'. We compare ours with IP2P (Brooks, Holynski, and Efros 2023), GLIDE (Nichol et al. 2021), and MGIE (Fu et al. 2023).

| Method | Ours | IP2P | MGIE | GLIDE |
|---|---|---|---|---|
| Preference | 66.17% | 14.83% | 15.50% | 1.17% |

Table 2: User study for preference. 2.33% select 'None'.

times dramatically changes the foreground of the portrait. GLIDE (Nichol et al. 2021) employs a mask to mark an edited region. It preserves the content in the region outside the mask, but the content inside the mask region is dramatically changed. MGIE (Fu et al. 2023) leverages a large multi-modal model to obtain an expressive text prompt. However, it does not ensure the preservation of the foreground contents. According to face identity score (FIS) values, IP2P and GLIDE tend to change the face of the portrait. In particular, GLIDE largely changes the human to other objects. MGIE has the larger FIS than IP2P and GLIDE, but it sometimes creates artifacts on the foreground. Although the LLaVA score (LS) does not show significant differences, the qualitative results indicate that only our model can create new shadows and lighting. Existing models merely perform style transfer rather than actual relighting. The detailed qualitative comparisons are shown in Figure 8.

We further compare the CLIP vision score performance according to 10 text prompts, as shown in Figure 9. Our

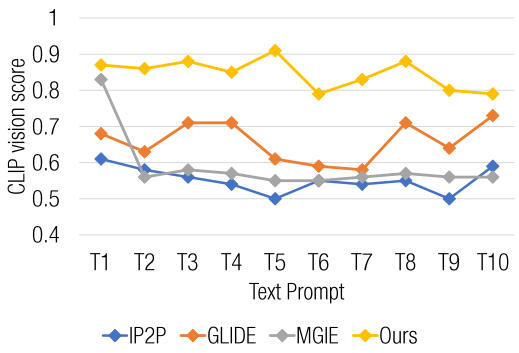


Figure 9: Performance comparison according to 10 text prompts (T1 $\sim$ T10).

| Method | SSIM ↑ | LPIPS ↓ | CVS ↑ |
|---|---|---|---|
| Ours w/o ATD | 0.545 | 0.419 | 0.866 |
| Ours w/o mask | 0.535 | 0.423 | 0.865 |
| Ours | **0.551** | **0.399** | **0.882** |

Table 3: Ablation study based on data and mask condition. ATD means the auxiliary task data.

model consistently performs well across all text prompts. MGIE's performance varies depending on the text, while other models show low performance for all text prompts.

## User Study

We conduct the user study to compare the user perceptual preference, involving 30 participants and 20 samples. We present users with randomly arranged images from various models: Ours, IP2P (Brooks, Holynski, and Efros 2023), MGIE (Fu et al. 2023), and GLIDE (Nichol et al. 2021), and ask the following question: which portrait relighting result best matches the 'text' while preserving the 'content'? Table 2 shows that our model's results are the most preferred, although some users occasionally choose images that contain the new contents described in the text.

## Ablation Study

We conduct an ablation study on data, mask condition, and crafted hierarchy.

**Effect of Auxiliary Task Augmentation.** We utilize auxiliary task data (*e.g.,* shadow-free and geometry lighting data). Table 3-(*Data*) summarizes the study of the effect of the auxiliary task augmentation. *Ours w/o ATD* means the model is trained without the auxiliary task data (ATD). Auxiliary task data helps with modeling the intrinsic appearance of humans under various shadows and highlights (with the delighting task), and the light positioning enables better geometric understanding, both of which lead to performance improvement. It is also presented in Figure 10.

**Mask Condition.** We use a mask condition for the Text2Relight model to distinguish the foreground and background. Table 3-(*Condition*) shows the comparison of the trained model with and without mask conditioning. When


Figure 10: Ablation study on mask condition and auxiliary task data (ATD). We use text prompts, 'The eerie and mysterious glow of the light added an enchanting touch to the dark forest' and 'Sunlight illuminates the her face', respectively.

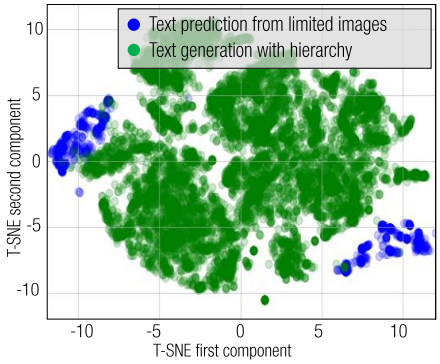


Figure 11: The T-SNE distribution of the texts predicted from limited sets of environment images using a vision language model (blue); and generated from a large-language model using our crafted language hierarchy (green).

the mask is not conditioned, the model sometimes generates the lighting artifacts, such as saturation and blur in the foreground, leading to poor performance. This issue is also confirmed in the qualitative results, as illustrated in Figure 10.

**Crafted Language Hierarchy.** As shown in Figure 11, our crafted language hierarchy significantly enhances the large language model's ability to produce a wider variety of text prompts, surpassing the distributional limitations typically observed in text generated by vision-language models. This strategic enhancement ensures that our simulated data achieves greater diversity.

## Application

**Shadow Removal.** Our model can support portrait shadow removal. For example, in the left two rows of Figure 12, we use text prompts, 'Eliminate the shadow from the portrait' and 'Remove the shadow', for the shadow removal.

**Light Positioning.** Our model can support the light positioning with a coarse text description about the light position. In Figure 12, we perform a series of text-guided relighting: 'Put a blue light on the right side', and 'Put a yellow light on the left side'.

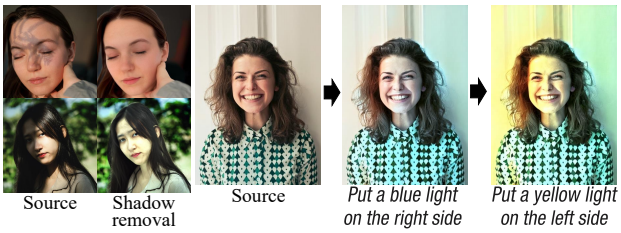


Figure 12: Examples of shadow removal and lighting position. First, shadow removal examples are demonstrated in the left two rows. Second, light positioning is shown in the right three images.


Figure 13: Examples of background harmonization.

**Background Harmonization.** We can use Text2Relight for background harmonization. For example, we apply the text prompt, 'Natural relighting', to the initially composited portrait image with a novel background, and re-composite the relighted foreground and the intact background. As shown in Figure 13, the model relights the foreground considering the lighting distribution of the background.

## Conclusion

We introduce Text2Relight, an end-to-end relighting model that can control the lighting distribution of a single image. We address the core dataset challenge by introducing a novel data simulation pipeline that can synthesize scalable ground-truth pairs of the relighting images and associated text prompt in three steps: generating diverse light-aware text prompts with our crafted language hierarchy, text-conditioned lighting image generation, and image-based foreground and background relighting. We perform repurposing of a pre-trained diffusion model as a lighting-specific foundational model by learning our large-scale simulation data. In our experiments, we demonstrate that Text2Relight effectively changes the lighting distribution that well-reflects the semantics of text prompts while maintaining the original contents from the input image, outperforming existing text-guided image editing models.

**Limitation.** Our model sometimes generates some strong point lights in the background, which appear unnatural. With weak spatial lighting context (accurate 3D position and color information) from the nature of text, our model is sometimes confused to localize the text-specified lighting.

## Acknowledgments

Junuk Cha and Seungryul Baek were supported by IITP grants (No. RS-2020-II201336 Artificial intelligence graduate school program (UNIST) 25%; No. RS-2021-II212068 AI innovation hub 25%; No. RS-2022-II220264 Comprehensive video understanding and generation with knowledge-based deep logic neural network 50%), funded by the Korean government (MSIT).

## References

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended diffusion for text-driven editing of natural images. In *CVPR*.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*.

Careaga, C.; and Aksoy, Y. 2023. Intrinsic Image Decomposition via Ordinal Shading. *ACM Transactions on Graphics*.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*.

Fu, T.-J.; Hu, W.; Du, X.; Wang, W. Y.; Yang, Y.; and Gan, Z. 2023. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *NIPS*.

Hou, A.; Sarkis, M.; Bi, N.; Tong, Y.; and Liu, X. 2022. Face relighting with geometrically consistent shadows. In *CVPR*.

Hou, A.; Zhang, Z.; Sarkis, M.; Bi, N.; Tong, Y.; and Liu, X. 2021. Towards high fidelity face relighting with realistic shadows. In *CVPR*.

Iwahori, Y.; Sugie, H.; and Ishii, N. 1990. Reconstructing shape from shading images under point light source illumination. In *[1990] Proceedings. 10th International Conference on Pattern Recognition*.

Ji, C.; Yu, T.; Guo, K.; Liu, J.; and Liu, Y. 2022. Geometry-aware single-image full-body human relighting. In *ECCV*.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-based real image editing with diffusion models. In *CVPR*.

Kim, H.; Jang, M.; Yoon, W.; Lee, J.; Na, D.; and Woo, S. 2024. SwitchLight: Co-design of Physics-driven Architecture and Pre-training Framework for Human Portrait Relighting. *arXiv preprint arXiv:2402.18848*.

Kocsis, P.; Sitzmann, V.; and Nießner, M. 2023. Intrinsic Image Diffusion for Single-view Material Estimation. *arXiv preprint arXiv:2312.12274*.

Kvanchiani, K.; Petrova, E.; Efremyan, K.; Sautin, A.; and Kapitanov, A. 2023. EasyPortrait–Face Parsing and Portrait Segmentation Dataset. *arXiv preprint arXiv:2304.13509*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *NIPS*.

Luo, S.; Tan, Y.; Huang, L.; Li, J.; and Zhao, H. 2023. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*.

Mei, Y.; Zeng, Y.; Zhang, H.; Shu, Z.; Zhang, X.; Bi, S.; Zhang, J.; Jung, H.; and Patel, V. M. 2024. Holo-Relighting: Controllable Volumetric Portrait Relighting from a Single Image. *arXiv preprint arXiv:2403.09632*.

Mei, Y.; Zhang, H.; Zhang, X.; Zhang, J.; Shu, Z.; Wang, Y.; Wei, Z.; Yan, S.; Jung, H.; and Patel, V. M. 2023. Light-Painter: interactive portrait relighting with freehand scribble. In *CVPR*.

Nestmeyer, T.; Lalonde, J.-F.; Matthews, I.; and Lehrmann, A. 2020. Learning physics-guided face relighting under directional light. In *CVPR*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

OpenAI. 2022. Introducing chatgpt.

Pandey, R.; Orts-Escolano, S.; Legendre, C.; Haene, C.; Bouaziz, S.; Rhemann, C.; Debevec, P. E.; and Fanello, S. R. 2021. Total relighting: learning to relight portraits for background replacement. *ACM Trans. Graph.*

Ponglertnapakorn, P.; Tritrong, N.; and Suwajanakorn, S. 2023. DiFaReli: Diffusion face relighting. In *ICCV*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Ren, M.; Xiong, W.; Yoon, J. S.; Shu, Z.; Zhang, J.; Jung, H.; Gerig, G.; and Zhang, H. 2023. Relightful Harmonization: Lighting-aware Portrait Background Replacement. *arXiv preprint arXiv:2312.06886*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Saito, S.; Schwartz, G.; Simon, T.; Li, J.; and Nam, G. 2023. Relightable gaussian codec avatars. *arXiv preprint arXiv:2312.03704*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Song, G.; Cham, T.-J.; Cai, J.; and Zheng, J. 2021. Half-body Portrait Relighting with Overcomplete Lighting Representation. In *Computer Graphics Forum*.

Sun, T.; Barron, J. T.; Tsai, Y.-T.; Xu, Z.; Yu, X.; Fyffe, G.; Rhemann, C.; Busch, J.; Debevec, P.; and Ramamoorthi, R. 2019. Single image portrait relighting. *ACM Transactions on Graphics (TOG)*.

Tajima, D.; Kanamori, Y.; and Endo, Y. 2021. Relighting Humans in the Wild: Monocular Full-Body Human Relighting with Domain Adaptation. In *Computer Graphics Forum*.

Wang, Y.; Holynski, A.; Zhang, X.; and Zhang, X. 2023. Sunstage: Portrait reconstruction and relighting using the sun as a light stage. In *CVPR*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*.

Yeh, Y.-Y.; Nagano, K.; Khamis, S.; Kautz, J.; Liu, M.-Y.; and Wang, T.-C. 2022. Learning to relight portrait images via a virtual light stage and synthetic-to-real adaptation. *ACM Transactions on Graphics (TOG)*.

Zhang, L.; Zhang, Q.; Wu, M.; Yu, J.; and Xu, L. 2021. Neural video portrait relighting in real-time via consistency modeling. In *ICCV*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

Zhou, H.; Hadap, S.; Sunkavalli, K.; and Jacobs, D. W. 2019. Deep single-image portrait relighting. In *ICCV*.