ScribbleLight: Single Image Indoor Relighting with Scribbles

Jun Myeong Choi[1] Annie Wang[1] Pieter Peers[2] Anand Bhattad[3] Roni Sengupta[1]

[1]University of North Carolina at Chapel Hill [2]College of William & Mary
[3]Toyota Technological Institute at Chicago

https://chedgekorea.github.io/ScribbleLight

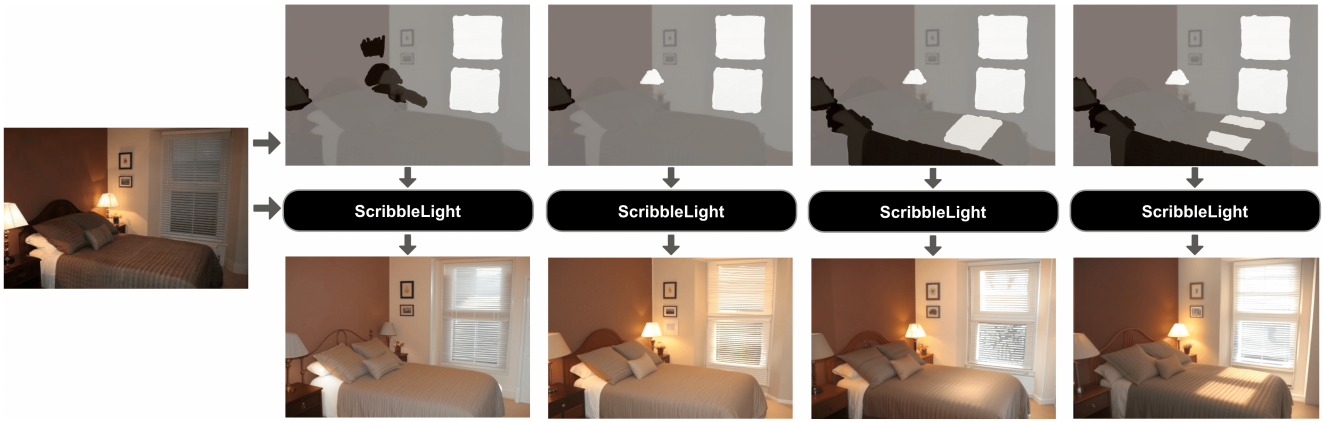


Figure 1. We introduce ScribbleLight, a generative model designed for indoor scene relighting from a single RGB image, that allows users to iteratively refine lighting effects in a photo using simple scribble annotations. For the first scribble example, the natural lighting coming from the window changes from nighttime to daytime by turning off both lamps and brightening the area near the window (left side) and casting a soft glow on the bed. In the second scribble example, only the right lamp is turned on. In the third scribble example, the natural lighting intensifies, thereby increasing the gloss on the bed's surface. For the final scribble example, adding the split in the light cast from the window causes the angle of the incoming light to change, creating a strong contrast with the warm glow of the bedside lamp.

Abstract

Image-based relighting of indoor rooms creates an immersive virtual understanding of the space, which is useful for interior design, virtual staging, and real estate. Relighting indoor rooms from a single image is especially challenging due to complex illumination interactions between multiple lights and cluttered objects featuring a large variety in geometrical and material complexity. Recently, generative models have been successfully applied to image-based relighting conditioned on a target image or a latent code, albeit without detailed local lighting control. In this paper, we introduce ScribbleLight, a generative model that supports local fine-grained control of lighting effects through scribbles that describe changes *in lighting. Our key technical novelty is an Albedo-conditioned Stable Image Diffusion model that preserves the intrinsic color and texture of the original image after relighting and an encoder-decoder-based ControlNet architecture that enables geometry-preserving lighting effects with normal map and scribble annotations. We demonstrate ScribbleLight's ability to create different lighting effects* (e.g.*, turning lights on/off, adding highlights, cast shadows, or indirect lighting from unseen lights) from sparse scribble annotations.*

1. Introduction

In today's digital age, the ability to control and visualize lighting in indoor spaces is crucial, especially in downstream applications such as real estate, virtual staging, and interior design. Traditional static images of indoor scenes often fail to show how a room would look under different lighting conditions, missing key elements that affect the space's aesthetics and mood. Scene relighting techniques offer a solution by enabling dynamic lighting adjustments within an image, letting users envision how a room transforms under various lighting scenarios, such as natural sunlight or with different lamps in the room turned on. The ability to relight provides a deeper and immersive understanding of the space, without needing a physical visit.

Indoor relighting presents unique challenges compared to outdoor settings, where a single, predictable light source (*i.e*., Sun) yields consistent and directional shadows; the resulting limited model space has been successfully exploited in prior outdoor relighting methods [29, 38, 61, 73, 77]. In contrast, indoor environments involve multiple light sources, such as ceiling lights, lamps, window-filtered daylight, as well as invisible light sources, each with unique characteristics in intensity, direction, and diffusion. These overlapping light sources create intricate, soft, layered shadows which create a challenging relighting environment. Indoor scenes are also composed of multiple objects made of different materials (*e.g*., furniture and doors), making indoor relighting more difficult than single-object relighting.

Despite recent advances, existing state-of-the-art indoor 3D relighting methods still require significant effort in dense scene capture [42, 48, 49, 81, 89]. Conversely, implicit relighting methods that use a latent space [8] or a reference image [85] to control lighting can only induce coarse global lighting changes and cannot control local details. Motivated by the success of using scribbles to guide various image manipulation tasks [41, 52, 53, 65], we propose ScribbleLight, where a user can intuitively and directly control relative lighting adjustments in an image with scribbles. ScribbleLight enables the user to iteratively indicate areas in the image they wish to brighten or darken with scribbles, from coarse to fine-grained, based on their preferences. Our generative model can relight an image from scribble input to create diverse illumination effects - turning lights on and off, adding cast shadows, highlights, inter-reflections, etc. Our method bridges the gap between technical complexity and creative flexibility, offering a streamlined, user-friendly way to achieve professional-quality lighting adjustments in intricate indoor scenes.

Scribbles only offer high-level guidance. Consequently, to resolve the inherent guidance-ambiguities, we exploit the general image priors embedded in large pretrained generative diffusion model (*i.e*., Stable Diffusion v2 [60]) and control the lighting effects with ControlNet [82]. However, a naive implementation fails to preserve the color and texture, *i.e*., intrinsic albedo, of the original image in the relit image. We therefore introduce an Albedo-conditioned Stable Image Diffusion model that generates realistic images conditioned on the intrinsic albedo of the scene. To support large lighting changes and to improve robustness with respect to an imperfect albedo (predicted by an Intrinsic Image Decomposition [11]) we inject uncertainty in the training process by adding noise to the (albedo) condition, thereby reducing dependency on the exact content of the albedo and forcing the diffusion model to rely more on the embedded image priors. We add lighting control during the albedo-conditioned diffusion process via a ScribbleLight ControlNet, where the control signal is the latent embedding of the scribbles and normals obtained from an encoder-decoder network that reconstructs the normal and *shading* map from the input. The decoder's ability to predict the intended shading (from scribles) and reconstruct the normals improves the likelihood that the latent code includes all the necessary information for relighting.

ScribbleLight enables flexible lighting control while retaining the intricate color and texture details of the original scene, overcoming challenges that arise due to the sparsity of scribbles. As no other prior indoor relighting method performs scribble-driven single-image relighting of indoor rooms, we compare our approach to baselines derived from existing approaches [38, 80] that also use Stable Diffusion and ControlNet for relighting. Quantitative and qualitative evaluations show that our method significantly outperforms the baseline methods, which often fail to preserve the albedo of the input image and control local lighting details. We also perform an extensive ablation study to demonstrate the effectiveness of ScribbleLight.

2. Related work

Image-based relighting aims to alter the lighting in photographs post-capture. Specialized methods have been proposed for relighting isolated objects [6, 22, 25, 27, 36, 56, 67, 69, 71, 78, 86, 87], human portraits [17, 18, 32, 37, 52, 54, 57, 59, 64], human bodies [9, 15, 68], outdoor scenes [29, 38, 61, 73, 77], and indoor scenes [8, 39, 51, 74, 80, 85]. Indoor scene relighting is especially challenging due to mixed natural and artificial light sources, occlusions, and intricate light interactions in a cluttered scene creating cast shadows, strong highlights, and inter-reflections.

Image-based relighting research has explored different lighting representations to control illumination in the rendered image. Explicit lighting representations such as shadow-maps [51], spherical Gaussians [39], or irradiance fields [80] directly specify the lighting, thereby offering the user only indirect control on the *effects* of lighting in the scene (*i.e*., the goal of the user). Alternatively, Xing *et al*. [74], Zhang *et al*. [85] and Bhattad *et al*. [8] use implicit lighting representations instead and navigate the latent space to control lighting effects. However, latent space editing only offers coarse global control and cannot control local details, making it difficult for the user to achieve their exact goal. In this paper, we use user-friendly scribbles to enable more fine-grained control of the lighting effects.

Image manipulation using scribbles offers an intuitive interface for specifying a user's intent. Scribbles have been used as a guide in a wide variety of tasks such as: segmentation [16, 33, 50, 66], image generation [13, 14, 26, 30, 34], image editing [21, 55, 58, 72, 75, 79], inpainting [53, 65], retrieval[19, 70] , and colorization [41]. Similar to us, Mei *et al*. [52] use scribbles to control relighting of human portraits. However, the geometrical complexity and mate-

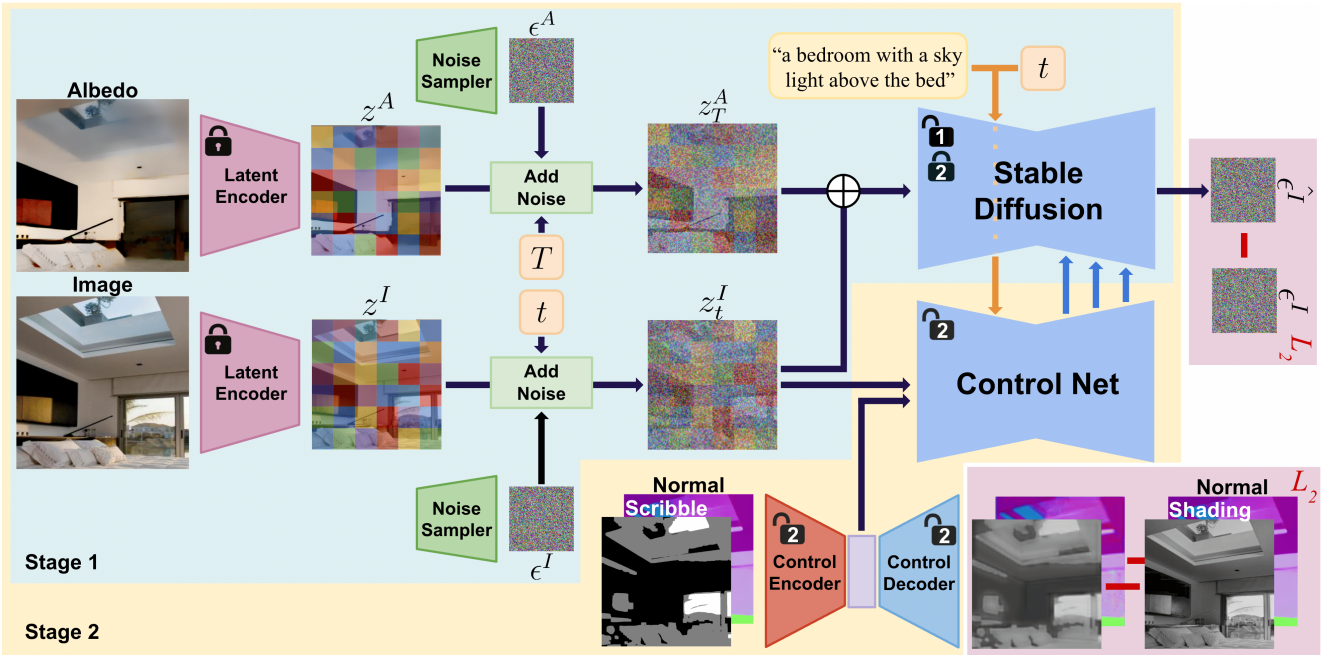


Figure 2. ScribbleLight consists of an Albedo-conditioned Stable image Diffusion model (trained in Stage 1), and a ControlNet (trained in Stage 2) that guides the albedo-conditioned diffusion model for relighting through a latent encoding of the scribbles and normals. To regularize the latent encoding, we jointly train a decoder that predicts the target shading (and normals) from the scribbles (and normals).

rial and lighting variations in indoor scenes make scribble-based relighting guidance more challenging.

Intrinsic image decomposition (IID) is a fundamental problem in computer vision that aims to separate an image into an illumination-dependent component (*i.e.*, shading) and an illumination-independent component (*i.e.*, reflectance or albedo). Early IID methods rely on heuristics based on physical properties or empirical observations [2, 3, 12, 28, 31, 40], often limited to Lambertian or simple scenes. Recent IID methods leverage machine learning trained on synthetic data [5, 10, 20, 35, 39, 44–46, 63, 88] to support more complex scenes and non-Lambertian reflectance. Other studies have shown that intrinsic images emerge within generative models and can be easily recovered [7, 24]. Relighting, a common downstream task of IID, is achieved by changing the illumination-dependent component and recompositing the intrinsic components. However, to achieve a plausible relit result, the modified illumination-dependent component has to contain all the details, making it a cumbersome and error-prone interface for relighting. In contrast, scribbles do not require the user to provide pixel-precise shading details, providing a more convenient and user-friendly control interface.

3. Method

We aim to generate plausible relit images of indoor scenes from a single photograph and guided by user-provided scribbles. Instead of viewing the guidance as target pixel or shading values, we instead view scribbles as a way for the user to indicate which areas need to be brightened (*e.g.*, turning on a light) and which areas need to be darkened (*e.g.*, adding a cast shadow). Therefore, we employ a binary scribble with '1' indicating brightening, and '0' indicating darkening. Unlabeled areas are left to the relighting model to determine the most plausible action.

Inspired by recent successes in using generative diffusion models for relighting tasks [4, 23, 38, 57, 78], we introduce ScribbleLight, a ControlNet-based single-image relighting solution for guiding an albedo-conditioned diffusion model (Section 3.1) with scribbles and normals (Section 3.2). Our training pipeline involves two distinct stages: first, fine-tuning an albedo-conditioned Stable Diffusion model, followed by separate training of the scribble-guided ControlNet. Figure 2 summarizes our pipeline.

3.1. Albedo-conditioned Image Diffusion

A key observation underpinning ScribbleLight is that relighting should preserve the underlying albedo intrinsic, *i.e.*, color and texture, of the input photograph. While Stable Diffusion v2 [60] trained on LAION-5B [62] provides a strong image prior, it lacks a strong constraint on the underlying albedo intrinsic. Therefore, we refine Stable Diffusion to produce an image **I** conditioned on an additional albedo image **A**. This will help the diffusion model to maintain the input image's color and texture information during

relighting. Concretely, we employ the pre-trained Latent VAE Encoder $\mathcal{E}^L$ to encode both the image $\mathbf{I}$ and the corresponding albedo $\mathbf{A}$ into the latent space, *i.e.*, Image Latent ($z^I = \mathcal{E}^L(I)$) and Albedo Latent ($z^A = \mathcal{E}^L(A)$).

We follow Stable Diffusion's training process by adding ϵ_t^I noise to the image latents z^I for a randomly sampled time step $t \in \{1, ..., T\}$ and learning to denoise to z_{t-1}^I. However, directly conditioning the diffusion process on the albedo image poses two problems. First, because the scribbles do not encode any spatially varying intensity changes, the diffusion model tends to produce relit regions with little variation. Second, any errors present in the albedo map are included in the diffusion process yielding visually noticeable artifacts. We resolve both problems by also adding a fixed amount of noise ϵ_T^A to the albedo latents z^A (*i.e.*, introduce uncertainty) to preserve the fundamental color and structure of the scene while providing enough uncertainty to render different lighting conditions. In contrast to the image noise which varies per time step t, ϵ_T^A remains fixed at the level of $T = 200$, an optimal value we observed empirically. Next, we concatenate z_t^I and z_T^A into a single input, z_t, along the feature dimension. This results in a doubling of the input channels of the latent denoising U-net in Stable Diffusion, and we zero-initialize the additional convolution weights. Finally, we train the albedo-conditioned Stable Diffusion model (θ^S) with text prompt p using the following modified loss function:

$$\mathcal{L} = \mathbb{E}_{z_t^I, z_T^A, \epsilon^I \sim \mathcal{N}(0,1), t, p} \left[\left\| \epsilon - \epsilon_{\theta^S}(z_t^I, z_T^A, t, p) \right\|_2^2 \right]. \quad (1)$$

3.2. ScribbleLight ControlNet

We employ ControlNet [82] to guide the albedo-conditioned image diffusion using a user-provided scribble map $\mathbf{M}$ and normal map $\mathbf{N}$ to generate a relit image. We first concatenate and encode the scribble map $\mathbf{M}$ and normals $\mathbf{N}$ into a lighting feature map ($f = \mathcal{E}^C([\mathbf{M}, \mathbf{N}])$) using a learnable control encoder $\mathcal{E}^C$. To regularize the control encoder, we introduce an additional control decoder $\mathcal{D}^C$ that recovers the normal map $\mathbf{N}$ and predicts a monochromatic intrinsic shading component $\mathbf{S}_{mono}$ from the lighting feature map:

$$\mathcal{L}_D = \left\| \mathcal{D}^C(\mathcal{E}^C(\mathbf{M}, \mathbf{N})) - (\mathbf{S}_{mono}, \mathbf{N}) \right\|_2^2. \quad (2)$$

The Control Encoder-Decoder architecture ensures that the latent lighting features contain the scene geometry and shading information necessary for relighting.

The ControlNet takes as input the lighting feature map f, image latent code z_t^I at time-step t, and the text prompt p. Because our ControlNet is not conditioned on the albedo $\mathbf{A}$, we instead initialize it with the original prompt-conditioned Stable Diffusion v2 weights, and train it jointly with the Control Encoder-Decoder using the following loss:

$$\mathcal{L} = \mathcal{L}_D + \mathbb{E}_{z_t, f, \epsilon^I \sim \mathcal{N}(0,1), t, p} \left[\left\| \epsilon - \epsilon_{\theta^C}(z_t, f, t, p) \right\|_2^2 \right]. \quad (3)$$

3.3. Training Data and Scribble Generation

Existing large-scale indoor relighting datasets such as InteriorVerse [88] and OpenRooms [47] consist of synthetic scenes rendered from different views under two or more lighting conditions. There is a significant domain gap between these datasets and real photographs. To reduce the domain gap we opt to train ScribbleLight on the real indoor images from LSUN Bedrooms [76].

To train the albedo-conditioned image diffusion model (Section 3.1), we require an albedo map $\mathbf{A}$ for each image $\mathbf{I}$ from the training set. We compute $\mathbf{A}$ using a state-of-the-art IID [11]. We also require a corresponding text prompt $\mathbf{p}$ that we generate using BLIP-2 [43] from the image $\mathbf{I}$.

Training the ControlNet (Section 3.2) requires the normals $\mathbf{N}$ and scribbles $\mathbf{M}$. Additionally, to train the Control-decoder, we also require a corresponding monochromatic shading image $\mathbf{S}_{mono}$. The normal $\mathbf{N}$ are computed using DSINE [1] and the shading $\mathbf{S}_{mono}$ are provided by an IID method [11]. We generate the scribbles $\mathbf{M}$ automatically from the shading $\mathbf{S}_{mono}$ by setting $\mathbf{M}(x) = 1$ when $\mathbf{I}(x) > \mu + \sigma$, $\mathbf{M}$(x) = 0, when $\mathbf{I}(x) < \mu - \sigma$, and $\mathbf{M}(x) = 0.5$ otherwise, where μ and σ are the mean and standard deviation of the pixel intensity distribution in the training data. However, the thresholded scribbles exhibit edges that strongly align with the content in the input image $\mathbf{I}$, something unlikely to happen with user-drawn scribbles. Therefore, we perform an additional dilation and erosion with a kernel size randomly sampled between 3 and 19.

4. Experiment

4.1. Evaluation Framework

Dataset. We use a subset of 100K images from the LSUN Bedrooms dataset [76] as our training data, and we select 206 pairs of indoor room images, each pair with the same environment but two different lighting conditions, from the BigTime time-lapse dataset [44] for testing. We generate automatic scribble annotations from the IID generated shading map following Section 3.3. We also manually create hand-drawn scribble annotations for a small selection of images from LSUN Bedrooms and publicly available internet images. To differentiate between both, we denote the former as *'auto-generated scribbles'* and the latter as *'user scribbles'*. For illustration purposes, we display the scribble annotations overlaid over the source image; we use clean scribbles for computations.

Baselines. To our knowledge no prior method can guide indoor room relighting with scribble annotations. Therefore, we adapt two existing image-based relighting algo-

Figure 3. Qualitative comparison of relighting quality between LightIt* [38], RGB↔X [80] and ScribbleLight (Ours) with auto-generated scribbles given a target (GT) image.

	RMSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
LightIt*	0.341(0.302)	9.61(10.65)	0.232(0.332)	0.564(0.518)
RGB↔X	0.269(0.251)	12.47(12.99)	0.416(0.437)	0.439(0.425)
Ours	**0.206(0.190)**	**14.29(15.01)**	**0.436(0.504)**	**0.394(0.370)**

Table 1. Quantitative comparison of relighting accuracy between our ScribbleLight, RGB↔X [80] and LightIt* [38]. We compute the mean(best) errors with respect to a target image.

rithms as a comparison baseline. As a first baseline method, we retrain *LightIt* [38] on our training dataset, and directly pass the auto-generated scribble annotations as input to the control-encoder instead of the monochromatic shading map; we denote our retrained version as *'LightIt*'*. We employ *RGB↔X* [80] without retraining as our second baseline method; we first generate the intrinsic components (normal, albedo, roughness, and metallicity) using RGB→X, and then recompose the image using X→RGB with the shading map replaced by the scribble annotations.

Metrics. We employ four error metrics to assess performance: RMSE, PSNR, SSIM, and LPIPS [84]. Lower RMSE and higher PSNR indicate better per-pixel similarity to the reference. SSIM (higher is better) assesses structural similarity, while LPIPS (lower is better) evaluates intrinsic feature similarity. We compute both the average and best values for each metric over 5 replicates (*i.e.*, different diffusion seeds). Note that raw metric values do not necessarily capture the full relighting quality of the different outputs, but rather only measure the error with respect to a single reference image. For example, a low PSNR does not necessarily indicate poor performance since a relit result can be realistic while at the same time differ from the reference.

4.2. Evaluation with Auto-Generated Scribbles

We first quantitatively compare the relighting quality with auto-generated scribble annotations over the test set. The results are summarized in Table 1 and show that ScribbleLight significantly outperforms the baseline methods across all metrics. Figure 3 qualitatively confirms the superior performance. Even though the rough scribble annotations do not include detailed shadow and shading information, ScribbleLight is able to produce a relighting result similar to the target image. In contrast, the RGB↔X results appear to overlay the coarse scribble annotation yielding an unrealistic result. LightIt* is able to interpret the scribble annota-

Figure 4. Qualitative comparison of relighting quality of LightIt* [38], RGB↔X [80] and ScribbleLight (Ours) with user-provided hand-drawn scribbles.

tions correctly, but fails to preserve the albedo of the source image, resulting in lighting that differs significantly from the target. In the supplementary material, we demonstrate that even with $\mathbf{S}_{mono}$ (instead of scribbles), ScribbleLight outperforms RGB↔X and LightIt*.

4.3. Evaluation with User Scribbles

We perform a number of qualitative tests to demonstrate that ScribbleLight also performs well on hand-drawn user scribbles. Figure 4 shows examples of turning lights on and off (rows 1 & 2), as well as to brighten or darken the incoming light from outside (rows 3 & 4) using scribble annotations. In addition, we also edit shadows by marking areas for darkening (rows 1-4). While ScribbleLight produces plausible results, RGB↔X and LightIt* fail to create realistic relighting results, producing similar artifacts as seen in Section 4.2. We observe that even when the scribbles are physically inconsistent (*e.g.*, the bright annotation on the side of the bed (row 4) is unlikely to be cast from a small light on the ceiling), ScribbleLight creates a physically plausible result by imagining a light source outside the image with appropriate shading effects such as the gradually decreasing lighting intensity on the side of the bed.

Despite the simplicity of the scribble annotations, ScribbleLight can still exert precise control over lighting conditions/effects. Figure 5 shows relit results of a source image with two lights turned on (row 1) or off (row 2), which can be separately toggled on or off with the appropriate scribbles. In both cases, ScribbleLight produces realistic soft highlights on the nearby walls from the lamps, even though this is not specified in the scribble annotations.

We also demonstrate in Figure 1 and Figure 6 how users

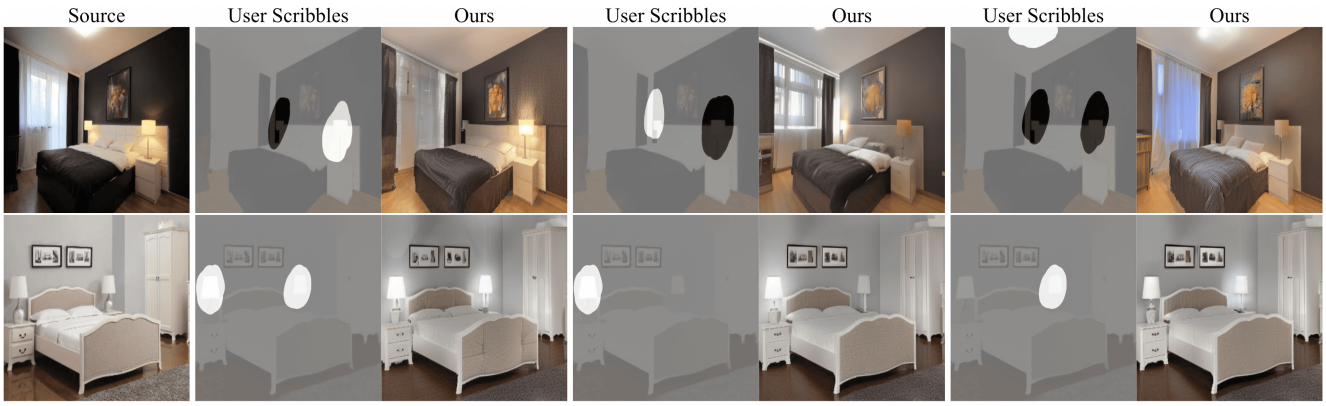


Figure 5. Demonstration of ScribbleLight's ability to generate different plausible relit images by turning on and off different lights while maintaining the intrinsics of the input photograph.

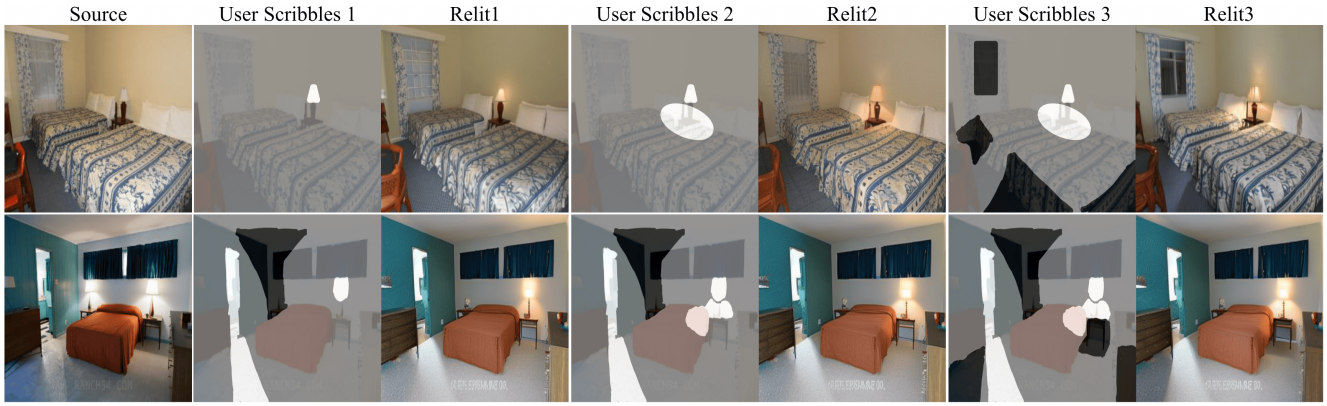


Figure 6. Minor changes to the scribbles yield proportional changes in the relit results, enabling a user to iteratively refine the scribbles to achieve the desired results.

can progressively refine the scribble annotations to improve the relit results, yielding a flexible and intuitive indoor scene relighting experience. For instance, in Figure 6 row 1, starting from a source image with the lamp turned off, we first turn on the lamp (user scribble 1), then add ambient lighting around the lit lamp (user scribble 2) creating a soft glow effect, and finally enhance realism by adding cast shadows near the bed and darkening the window.

Comparison with IC-Light. IC-Light [83] is a harmonization model that extracts the foreground from an image and edits the background lighting using a prompt or image. In contrast, our method does not require any foreground-background separation and can relight the image using scribbles. We provide a fair comparison between both techniques by carefully crafting prompts for lighting modifications. We observe (Table 2, Figure 8 in supplementary) that while IC-Light generates realistic images, it often fails to adhere to 'turn on/off'-prompts and changes geometry and texture of background elements.

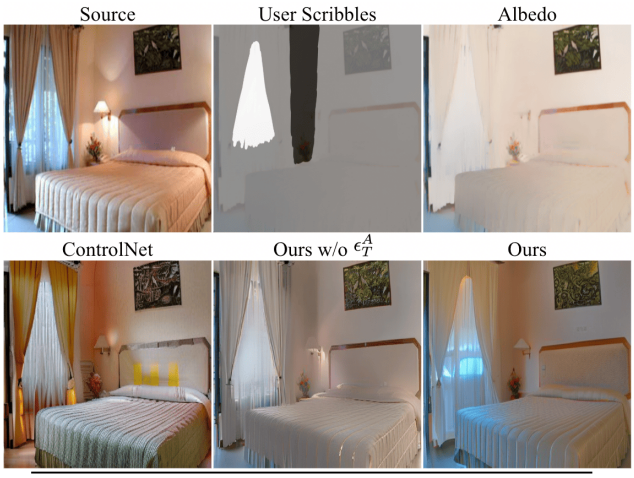
User Study. We conducted a user study in which each user was presented with the relighting results of RGB↔X [80] and ScribbleLight in randomized order. The user was asked to pick the best relit image conforming to the scribble, while preserving the geometry and texture of the original image. We evaluated 30 test examples, each judged by 30 participants. We note that 95.6% of the participants prefer ScribbleLight over RGB↔X [80], supporting the observations of quantitative evaluation with user scribbles.

	RMSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
IC-Light	0.278	11.46	0.357	0.505
ScribbleLight	**0.210**	**14.21**	**0.443**	**0.397**

Table 2. We compare ScribbleLight, which utilizes auto-generated scribbles, with IC-light, with carefully crafted prompts that reflect the scribbles, across 50 test images.

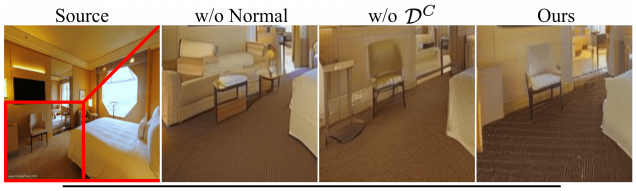
4.4. Ablation study

Design of Albedo-conditioned Stable Image Diffusion. A key contributor to the quality of the relighting is the albedo-conditioning of the diffusion model, ensuring better preservation of the intrinsic color and texture of the input. The robustness of the albedo-conditioned diffusion process is further improved by adding a noise ϵ_T^A to the latent albedo condition z^A (in addition to the noise ϵ_t^I added to the image



Albedo	Add ϵ_T^A	RMSE ↓	PSNR ↑	LPIPS ↓
ControlNet	–	0.2305	13.19	0.4839
StableDiff	–	0.2082	14.07	0.4193
StableDiff	✓	**0.2059**	**14.29**	**0.3942**

Table 3. Design of Albedo-conditioned Stable Image Diffusion. We show that adding noise to the albedo latent improves albedo preservation and generation of realistic lighting effects.



Normal	$\mathcal{D}^C$	RMSE ↓	PSNR ↑	LPIPS ↓
–	✓	0.2224	13.61	0.4251
✓	–	0.2098	14.06	0.4093
✓	✓	**0.2059**	**14.29**	**0.3942**

Table 4. Inclusion of both normals and the control-decoder improves geometry preservation in the relit images.

latent z^I). We ablate the efficacy of both components using auto-generated scribbles in two experiments: (i) we directly add the albedo as input to the ControlNet instead of conditioning Stable Diffusion on the albedo, and (ii) we train the Albedo-conditioned Stable Diffusion without adding noise to the albedo latent. From Table 3, we observe that albedo conditioning better preserves the intrinsic colors and texture than injecting the albedo via ControlNet (2nd row, 1st column in figure, 1st row in table vs 2nd row, 2nd column in figure, 2nd row in table). Furthermore, we observe that adding noise to the albedo latent (2nd row, 3rd column in the figure, 3rd row in the table) is more robust to inaccuracies in the predicted albedo (*e.g.*, the soft glow around the lamp) and provides larger variations in lighting effects (*e.g.*, blue sunlight peeking through the window).

Design of ScribbleLight ControlNet. Another key component in ScribbleLight is the control encoder-decoder for normal and scribble maps. We verify the importance of the normal map **N**, and the regularizing role of the control decoder $\mathcal{D}^C$. Table 4 indicates that omitting the normal map (2nd column in the figure, 1st row in table) results in the creation of more random objects, even in empty spaces, due to the absence of 3D geometric guidance. Furthermore, omitting the regularizing control-decoder $\mathcal{D}^C$ (3rd column in the figure, 2nd row in table) also creates hallucinations (4th column in the figure, 3rd row in table).

Limitations. ScribbleLight struggles to rectify strong physical inconsistencies in the user-defined scribbles and often generates realistic but physically implausible lighting effects (Figure 7). Furthermore, ScribbleLight does not support colored lighting adjustments, leading to relit results biased toward commonly seen colors like yellow and blue from its learned prior or source image colors.

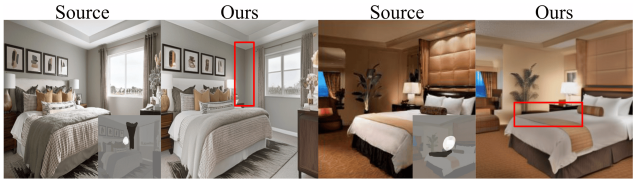


Figure 7. Given a physically incorrect scribble, *i.e.*, the location of shadow does not match the light source, ScribbleLight often creates implausible lighting effects that best match the user scribbles.

5. Conclusion

In this paper, we introduce ScribbleLight, a generative model for scribble-based single-image relighting of indoor scenes. We show that scribbles are a viable control signal for realistic and physically plausible relighting while significantly reducing user efforts and providing flexibility by enabling progressive coarse-to-fine editing. Our key technical contributions are the introduction of an Albedo-conditioned Stable Image Diffusion variant that better preserves the intrinsic color and texture of the input image during relighting, and ScribbleLight's ControlNet that better preserves geometrical shading information and guides the relighting based on a latent encoding of the surface normals and rough scribble annotations. Our method outperforms existing image-based relighting algorithms adapted for scribble-based relighting in both quantitative and qualitative evaluations. We demonstrate the effectiveness of ScribbleLight in generating various lighting effects, *e.g.*, turning a light source on or off or adding strong highlights and cast shadows. We also show that ScribbleLight can generate multiple replicates that match the scribbles. For future work we would like to enhance scribble generation to improve user control and precision in relighting. Additional avenues for future research include incorporating colored scribbles to allow users to control the color of the lighting.

Acknowledgement. This project was supported by a career start-up funding grant from the UNC CS Department and in part by the NIH project #1R21EB035832. Pieter Peers was supported in part by NSF grant IIS-1909028.

References

[1] Gwangbin Bae and Andrew J. Davison. Rethinking inductive biases for surface normal estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[2] Jonathan T Barron and Jitendra Malik. Intrinsic scene properties from a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 17–24, 2013.

[3] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1670–1687, 2014.

[4] Dina Bashkirova, Arijit Ray, Rupayan Mallick, Sarah Adel Bargal, Jianming Zhang, Ranjay Krishna, and Kate Saenko. Lasagna: Layered score distillation for disentangled object relighting. *arXiv preprint arXiv:2312.00833*, 2023.

[5] Anil S Baslamisli, Thomas T Groenestege, Partha Das, Hoang-An Le, Sezer Karaoglu, and Theo Gevers. Joint learning of intrinsic images and semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–302, 2018.

[6] Anand Bhattad and David A Forsyth. Cut-and-paste object insertion by enabling deep image prior for reshading. In *3DV*, 2022.

[7] Anand Bhattad, Daniel McKee, Derek Hoiem, and David Forsyth. Stylegan knows normal, depth, albedo, and more. *Advances in Neural Information Processing Systems*, 36: 73082–73103, 2023.

[8] Anand Bhattad, James Soole, and David A Forsyth. Stylitgan: Image-based relighting via latent control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4231–4240, 2024.

[9] Diogo Carbonera Luvizon, Vladislav Golyanik, Adam Kortylewski, Marc Habermann, and Christian Theobalt. Relightable neural actor with intrinsic decomposition and pose control. In *European Conference on Computer Vision*, pages 465–483. Springer, 2024.

[10] Chris Careaga and Yağız Aksoy. Intrinsic image decomposition via ordinal shading. *ACM Transactions on Graphics*, 43 (1):1–24, 2023.

[11] Chris Careaga and Yağız Aksoy. Colorful diffuse intrinsic image decomposition in the wild. *ACM Transactions on Graphics (TOG)*, 43(6):1–12, 2024.

[12] Qifeng Chen and Vladlen Koltun. A simple model for intrinsic image decomposition with depth cues. In *Proceedings of the IEEE international conference on computer vision*, pages 241–248, 2013.

[13] Tao Chen, Ming-Ming Cheng, Ping Tan, Ariel Shamir, and Shi-Min Hu. Sketch2photo: Internet image montage. *ACM transactions on graphics (TOG)*, 28(5):1–10, 2009.

[14] Wengling Chen and James Hays. Sketchygan: Towards diverse and realistic sketch to image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9416–9425, 2018.

[15] Zhaoxi Chen and Ziwei Liu. Relighting4d: Neural relightable human from videos. In *European conference on computer vision*, pages 606–623. Springer, 2022.

[16] Sungduk Cho, Hyungjoon Jang, Jing Wei Tan, and Won-Ki Jeong. Deepscribble: interactive pathology image segmentation using deep neural networks with scribbles. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 761–765. IEEE, 2021.

[17] Jun Myeong Choi, Max Christman, and Roni Sengupta. Personalized video relighting with an at-home light stage. In *European Conference on Computer Vision*, pages 394–410. Springer, 2024.

[18] Jun Myeong Choi, Johnathan Leung, Noah Frahm, Max Christman, Gedas Bertasius, and Roni Sengupta. Building secure and engaging video communication by using monitor illumination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4377–4386, 2024.

[19] Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Aneeshan Sain, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. Scenetrilogy: On human scene-sketch and its complementarity with photo and text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10972–10983, 2023.

[20] Partha Das, Sezer Karaoglu, and Theo Gevers. Pie-net: Photometric invariant edge guided network for intrinsic image decomposition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19790–19799, 2022.

[21] Tali Dekel, Chuang Gan, Dilip Krishnan, Ce Liu, and William T Freeman. Sparse, smart contours to represent and edit images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3511–3520, 2018.

[22] Kangle Deng, Timothy Omernick, Alexander Weiss, Deva Ramanan, Jun-Yan Zhu, Tinghui Zhou, and Maneesh Agrawala. Flashtex: Fast relightable mesh texturing with lightcontrolnet. In *European Conference on Computer Vision*, pages 90–107. Springer, 2024.

[23] Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12736–12746, 2023.

[24] Xiaodan Du, Nicholas Kolkin, Greg Shakhnarovich, and Anand Bhattad. Generative models: What do they know? do they know things? let's find out! *arXiv preprint arXiv:2311.17137*, 2023.

[25] Frédéric Fortier-Chouinard, Zitian Zhang, Louis-Etienne Messier, Mathieu Garon, Anand Bhattad, and Jean-François Lalonde. Spotlight: Shadow-guided object relighting via diffusion. *arXiv preprint arXiv:2411.18665*, 2024.

[26] Chengying Gao, Qi Liu, Qi Xu, Limin Wang, Jianzhuang Liu, and Changqing Zou. Sketchycoco: Image generation from freehand scene sketches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5174–5183, 2020.

[27] Jian Gao, Chun Gu, Youtian Lin, Zhihao Li, Hao Zhu, Xun Cao, Li Zhang, and Yao Yao. Relightable 3d gaussians: Realistic point cloud relighting with brdf decomposition and

ray tracing. In *European Conference on Computer Vision*, pages 73–89. Springer, 2024.

[28] Elena Garces, Adolfo Munoz, Jorge Lopez-Moreno, and Diego Gutierrez. Intrinsic images by clustering. In *Computer graphics forum*, pages 1415–1424. Wiley Online Library, 2012.

[29] James AD Gardner, Evgenii Kashin, Bernhard Egger, and William AP Smith. The sky's the limit: Relightable outdoor scenes via a sky-pixel constrained illumination prior and outside-in visibility. In *European Conference on Computer Vision*, pages 126–143. Springer, 2024.

[30] Arnab Ghosh, Richard Zhang, Puneet K Dokania, Oliver Wang, Alexei A Efros, Philip HS Torr, and Eli Shechtman. Interactive sketch & fill: Multiclass sketch-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1171–1180, 2019.

[31] Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2335–2342. IEEE, 2009.

[32] Andrew Hou, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. Face relighting with geometrically consistent shadows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4217–4226, 2022.

[33] Peiliang Huang, Junwei Han, Nian Liu, Jun Ren, and Dingwen Zhang. Scribble-supervised video object segmentation. *IEEE/CAA Journal of Automatica Sinica*, 9(2):339–353, 2021.

[34] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[35] Michael Janner, Jiajun Wu, Tejas D Kulkarni, Ilker Yildirim, and Josh Tenenbaum. Self-supervised intrinsic image decomposition. *Advances in neural information processing systems*, 30, 2017.

[36] 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 *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[37] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25096–25106, 2024.

[38] Peter Kocsis, Julien Philip, Kalyan Sunkavalli, Matthias Nießner, and Yannick Hold-Geoffroy. Lightit: Illumination modeling and control for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9359–9369, 2024.

[39] Peter Kocsis, Vincent Sitzmann, and Matthias Nießner. Intrinsic image diffusion for indoor single-view material estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5198–5208, 2024.

[40] Edwin H Land and John J McCann. Lightness and retinex theory. *Josa*, 61(1):1–11, 1971.

[41] Anat Levin, Dani Lischinski, and Yair Weiss. Colorization using optimization. In *ACM SIGGRAPH 2004 Papers*, pages 689–694. 2004.

[42] Junxuan Li, Hongdong Li, and Yasuyuki Matsushita. Lighting, reflectance and geometry estimation from 360 panoramic stereo. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10586–10595. IEEE, 2021.

[43] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[44] Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[45] Zhengqi Li and Noah Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *Proceedings of the European conference on computer vision (ECCV)*, pages 371–387, 2018.

[46] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2475–2484, 2020.

[47] Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Meng Song, Yuhan Liu, Yu-Ying Yeh, Rui Zhu, Nitesh Gundavarapu, Jia Shi, et al. Openrooms: An end-to-end open framework for photorealistic indoor scene datasets. *arXiv preprint arXiv:2007.12868*, 2020.

[48] Zhengqin Li, Jia Shi, Sai Bi, Rui Zhu, Kalyan Sunkavalli, Miloš Hašan, Zexiang Xu, Ravi Ramamoorthi, and Manmohan Chandraker. Physically-based editing of indoor scene lighting from a single image. In *European Conference on Computer Vision*, pages 555–572. Springer, 2022.

[49] Zhen Li, Lingli Wang, Mofang Cheng, Cihui Pan, and Jiaqi Yang. Multi-view inverse rendering for large-scale real-world indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12499–12509, 2023.

[50] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3159–3167, 2016.

[51] Jundan Luo, Duygu Ceylan, Jae Shin Yoon, Nanxuan Zhao, Julien Philip, Anna Frühstück, Wenbin Li, Christian Richardt, and Tuanfeng Wang. Intrinsicdiffusion: Joint intrinsic layers from latent diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

[52] Yiqun Mei, He Zhang, Xuaner Zhang, Jianming Zhang, Zhixin Shu, Yilin Wang, Zijun Wei, Shi Yan, HyunJoon

Jung, and Vishal M Patel. Lightpainter: Interactive portrait relighting with freehand scribble. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 195–205, 2023.

[53] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

[54] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas Lehrmann. Learning physics-guided face relighting under directional light. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5124–5133, 2020.

[55] Kyle Olszewski, Duygu Ceylan, Jun Xing, Jose Echevarria, Zhili Chen, Weikai Chen, and Hao Li. Intuitive, interactive beard and hair synthesis with generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7446–7456, 2020.

[56] Yohan Poirier-Ginter, Alban Gauthier, Julien Phillip, J-F Lalonde, and George Drettakis. A diffusion approach to radiance field relighting using multi-illumination synthesis. In *Computer Graphics Forum*, page e15147. Wiley Online Library, 2024.

[57] Puntawat Ponglertnapakorn, Nontawat Tritrong, and Supasorn Suwajanakorn. Difareli: Diffusion face relighting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22646–22657, 2023.

[58] Tiziano Portenier, Qiyang Hu, Attila Szabo, Siavash Arjomand Bigdeli, Paolo Favaro, and Matthias Zwicker. Faceshop: Deep sketch-based face image editing. *arXiv preprint arXiv:1804.08972*, 2018.

[59] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6452–6462, 2024.

[60] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[61] Viktor Rudnev, Mohamed Elgharib, William Smith, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Nerf for outdoor scene relighting. In *European Conference on Computer Vision*, pages 615–631. Springer, 2022.

[62] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

[63] Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David W Jacobs, and Jan Kautz. Neural inverse rendering of an indoor scene from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8598–8607, 2019.

[64] Soumyadip Sengupta, Brian Curless, Ira Kemelmacher-Shlizerman, and Steven M Seitz. A light stage on every desk. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2420–2429, 2021.

[65] Nakul Sharma, Aditay Tripathi, Anirban Chakraborty, and Anand Mishra. Sketch-guided image inpainting with partial discrete diffusion process. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6024–6034, 2024.

[66] Zhenyu Shu, Xiaoyong Shen, Shiqing Xin, Qingjun Chang, Jieqing Feng, Ladislav Kavan, and Ligang Liu. Scribble-based 3d shape segmentation via weakly-supervised learning. *IEEE transactions on visualization and computer graphics*, 26(8):2671–2682, 2019.

[67] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7495–7504, 2021.

[68] Daichi Tajima, Yoshihiro Kanamori, and Yuki Endo. Relighting humans in the wild: Monocular full-body human relighting with domain adaptation. In *Computer Graphics Forum*, pages 205–216. Wiley Online Library, 2021.

[69] Ashish Tiwari, Satoshi Ikehata, and Shanmuganathan Raman. Merlin: Single-shot material estimation and relighting for photometric stereo. In *European Conference on Computer Vision*, pages 251–269. Springer, 2024.

[70] Fang Wang, Le Kang, and Yi Li. Sketch-based 3d shape retrieval using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1875–1883, 2015.

[71] Tong Wu, Jia-Mu Sun, Yu-Kun Lai, and Lin Gao. De-nerf: Decoupled neural radiance fields for view-consistent appearance editing and high-frequency environmental relighting. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–11, 2023.

[72] C Xiao, D Yu, X Han, Y Zheng, and H Fu. Sketchhairsalon: deep sketch-based hair image synthesis.(2021). *arXiv preprint arXiv:2109.07874*.

[73] Zixuan Xie, Rengan Xie, Rong Li, Kai Huang, Pengju Qiao, Jingsen Zhu, Xu Yin, Qi Ye, Wei Hua, Yuchi Huo, et al. Holistic inverse rendering of complex facade via aerial 3d scanning. *arXiv preprint arXiv:2311.11825*, 2023.

[74] Xiaoyan Xing, Konrad Groh, Sezer Karagolu, Theo Gevers, and Anand Bhattad. Luminet: Latent intrinsics meets diffusion models for indoor scene relighting. In *CVPR*, 2025.

[75] Shuai Yang, Zhangyang Wang, Jiaying Liu, and Zongming Guo. Deep plastic surgery: Robust and controllable image editing with human-drawn sketches. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 601–617. Springer, 2020.

[76] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[77] Ye Yu, Abhimitra Meka, Mohamed Elgharib, Hans-Peter Seidel, Christian Theobalt, and William AP Smith. Self-supervised outdoor scene relighting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 84–101. Springer, 2020.

[78] 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*, pages 1–12, 2024.

[79] Yu Zeng, Zhe Lin, and Vishal M Patel. Sketchedit: Mask-free local image manipulation with partial sketches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5951–5961, 2022.

[80] Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. Rgb↔x: Image decomposition and synthesis using material- and lighting-aware diffusion models. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24*, page 1–11. ACM, 2024.

[81] Edward Zhang, Michael F Cohen, and Brian Curless. Emptying, refurnishing, and relighting indoor spaces. *ACM Transactions on Graphics (TOG)*, 35(6):1–14, 2016.

[82] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

[83] 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 *The Thirteenth International Conference on Learning Representations*, 2025.

[84] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[85] Xiao Zhang, William Gao, Seemandhar Jain, Michael Maire, David Forsyth, and Anand Bhattad. Latent intrinsics emerge from training to relight. *Advances in Neural Information Processing Systems*, 37:96775–96796, 2024.

[86] Zitian Zhang, Frédéric Fortier-Chouinard, Mathieu Garon, Anand Bhattad, and Jean-François Lalonde. Zerocomp: Zero-shot object compositing from image intrinsics via diffusion. In *WACV*, 2025.

[87] Xiaoming Zhao, Pratul Srinivasan, Dor Verbin, Keunhong Park, Ricardo Martin Brualla, and Philipp Henzler. Illuminerf: 3d relighting without inverse rendering. *Advances in Neural Information Processing Systems*, 37:42593–42617, 2024.

[88] Jingsen Zhu, Fujun Luan, Yuchi Huo, Zihao Lin, Zhihua Zhong, Dianbing Xi, Rui Wang, Hujun Bao, Jiaxiang Zheng, and Rui Tang. Learning-based inverse rendering of complex indoor scenes with differentiable monte carlo raytracing. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–8, 2022.

[89] Jingsen Zhu, Yuchi Huo, Qi Ye, Fujun Luan, Jifan Li, Dianbing Xi, Lisha Wang, Rui Tang, Wei Hua, Hujun Bao, et al. I2-sdf: Intrinsic indoor scene reconstruction and editing via raytracing in neural sdfs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12489–12498, 2023.

ScribbleLight: Single Image Indoor Relighting with Scribbles

Supplementary Material

Along with this supplementary Material, we provide additional visual materials (e.g., images and videos) in an websit, accessible via https://chedgekorea.github.io/ScribbleLight.

A. Implementation Details

We fine-tune a pre-trained Stable Diffusion v2 model [60] for Albedo-conditioned Image Diffusion and ScribbleLight ControlNet. To reconstruct the monochromatic shading map $\mathbf{S}_{mono}$ and the normal map $\mathbf{N}$ from the lighting feature map f, we utilize a control decoder $\mathcal{D}^C$. This control decoder $\mathcal{D}^C$ is structured similarly to the control encoder $\mathcal{E}^C$, consisting of 4 residual blocks, but with a transposed architecture. For training, we use a batch size of 16 and the AdamW optimizer with a learning rate of 1e–5. All inputs are resized to 512×512. Training each model takes approximately 48 hours on 4 A6000 GPUs. We employ the DDPM noise scheduler with 1000 diffusion steps during training. For inference, we apply the DDIM scheduler and sample only 20 steps.

B. Evaluation with Monochromatic Shading

This section is analogous to Section 4, but instead of using scribbles, we evaluate indoor scene relighting performance using the monochromatic shading map $\mathbf{S}_{mono}$.

Dataset and Baselines. As detailed in Section 4.1, we trained both LightIt* [38] and our method using the LSUN bedrooms dataset [76]. Instead of using auto-generated scribbles as the ControlNet input, however, here we utilize the monochromatic shading map instead. For our second baseline, we employed RGB↔X [80] without retraining: intrinsic components (normal, albedo, roughness, and metallicity) were extracted from the source image, and the irradiance field was derived from the target image using RGB→X. The source image was then relit by recomposing it with its intrinsic components and the target image's irradiance field through X→RGB. Since IIDiffusion [39] employs spherical Gaussians as its lighting representation, it was not feasible to perform a comparison based on scribbles. However, we extend the comparison with IIDiffusion by extracting intrinsic components from the source image and the spherical Gaussians from the target image, and recomposing the source image under the target spherical Gaussians, similar to RGB↔X.

Evaluation. We quantitatively compare the relighting quality using the monochromatic shading $\mathbf{S}_{mono}$ over the test set. The results, summarized in Table 5, show that ScribbleLight outperforms the baseline methods across all metrics.

	RMSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
IIDiffusion	0.137	18.02	0.690	0.367
LightIt*	0.227	13.17	0.390	0.447
RGB↔X	0.261	13.36	0.570	0.364
Ours	**0.132**	**18.22**	**0.697**	**0.275**

Table 5. Quantitative comparison of relighting accuracy between our ScribbleLight, RGB↔X [80], LightIt* [38], and IIDiffusion [39] with monochromatic shading map created from the target image. We compute the errors with respect to a target image.

Additionally, we present the qualitative results in website (see the 'Monochromatic Shading Map' section).

C. Comparison with IC-Light

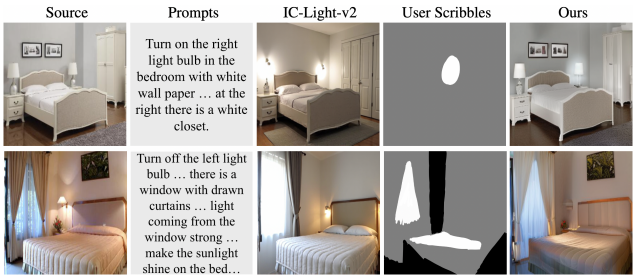


Figure 8. IC-Light uses prompt-driven relighting but lacks precise control, e.g., turn on only 1 light source (row 1) or add direct lighting on the bed (row 2). IC-Light also often changes the composition of the scene by removing or adding any objects, e.g., removes picture frames on the wall (row 1 & 2) and reshapes the curtains (row 2). In contrast, ScribbleLight provides precise lighting control while preserving the composition of the scene.

D. Visual Materials

The structure of website is as follows. As demonstrated in Figure 1 and Figure 6, the top section presents a demo video and three iterative examples showcasing how ScribbleLight iteratively refines lighting effects. The 'User Scribble' section provides additional examples of user scribble comparisons, as illustrated in Figure 4. The 'Monochromatic Shading Map' section features qualitative comparisons referenced in Appendix B. We also present the target shading utilized by each method during relighting in Figure 9. Finally, the 'Turning On/Off the Light' section includes additional examples from Figure 5.

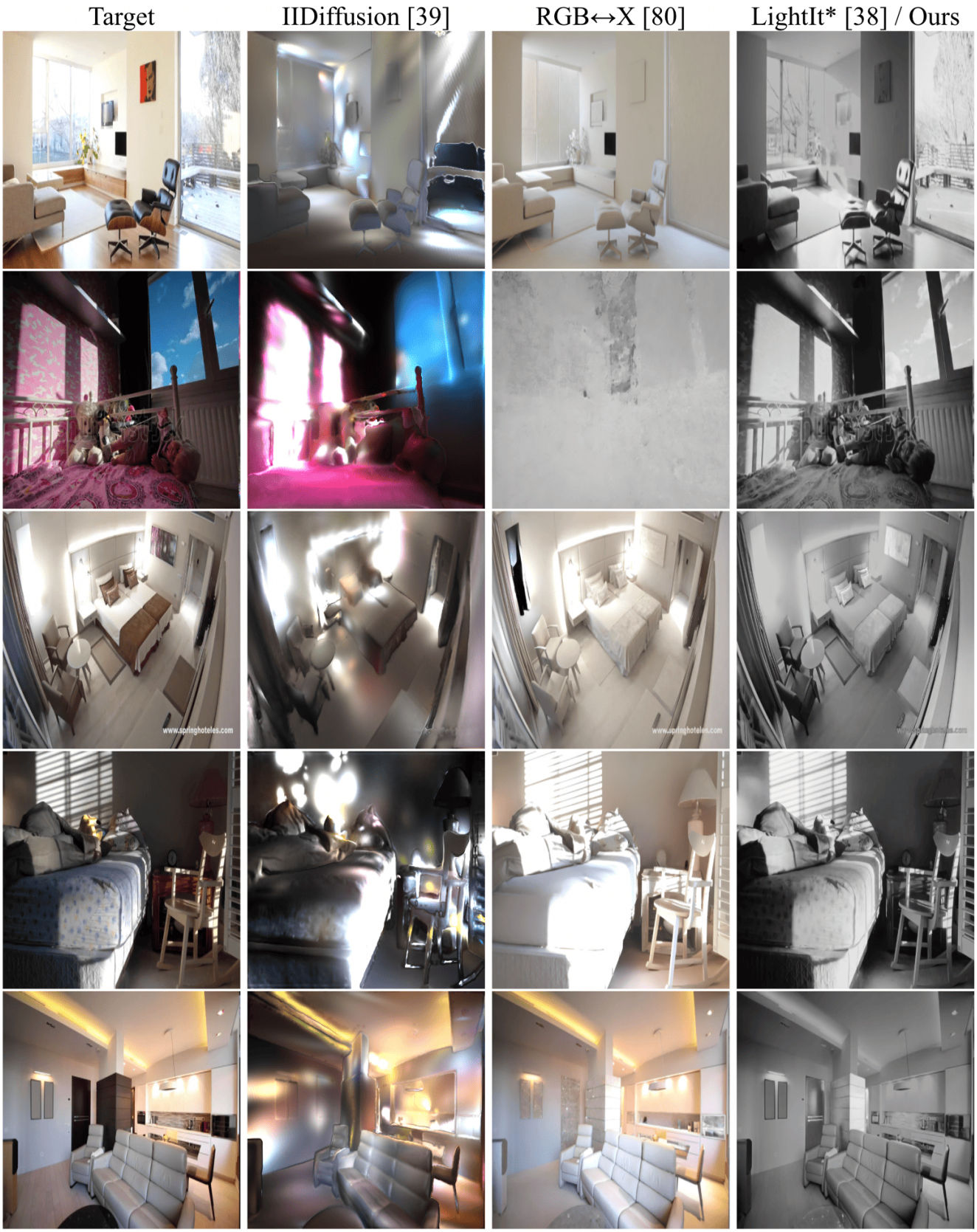


Figure 9. We demonstrate the target lighting representation used by each method—IIDiffusion [39], RGB↔X [80], LightIt* [38], and Ours—when performing relighting with a monochromatic shading map.