DRAWING THE LIGHT OF IMAGE VIA MATERIAL DECOUPLED ILLUMINATION REPRESENTATION

Anonymous authors
Paper under double-blind review



Figure 1: **Controlling illumination in images with hand-drawn strokes.** Our method enables users to directly specify illumination in an image through simple strokes. By adjusting the stroke values, users can achieve material-consistent changes in both illumination intensity and chromaticity.

ABSTRACT

Controllable image relighting is an important problem in image editing, and hand-drawn scribbles provide an intuitive interface for specifying the desired illumination. However, existing methods do not establish a consistent and effective mapping between scribble inputs and relighting results, limiting their ability to control illumination intensity, chromaticity, and complex spatial distributions. We address this limitation by introducing a material-decoupled illumination representation, termed the Lumi Map, which establishes an explicit mapping between user scribbles and the resulting illumination, thereby improving both relighting accuracy and controllability. Specifically, we use a renderer to synthesize source image-Lumi Map-relit image triplets and train the model to predict the target relighting result conditioned on the Lumi Map. To mitigate the domain gap introduced by synthetic data, we further perform reconstruction training on real relighting pairs, improving the model's generalization to real-world images. Finally, we present **Dior-Light**, an image relighting method controlled by hand-drawn strokes. Extensive experiments demonstrate that our method outperforms existing approaches in relighting accuracy and enables effective control over illumination intensity and chromaticity on in-the-wild images.

1 INTRODUCTION

Image relighting is a longstanding problem in image editing and has advanced rapidly with modern generative models. Given a photograph, the goal is to change its illumination while preserving

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the original viewpoint, geometry, surface reflectance, and subject identity. Recent diffusion-based relighting methods have substantially improved visual realism (Zhang et al., 2025; Jin et al., 2024; Zeng et al., 2024a). However, precise illumination control remains a key bottleneck. A straightforward approach to explicit control is to decompose an image into illumination-related components, such as an environment map or shading, and scene attributes, such as geometry and material properties (Zeng et al., 2024b; Careaga & Aksoy, 2024; Zhang et al., 2021b). The illumination components can then be modified independently before the image is recomposed. However, this approach relies heavily on the accuracy of image decomposition, while high-quality illumination representations are often difficult to obtain for real-world photographs.

Scribble-driven relighting instead investigates how to control target illumination using sparse lighting cues, enabling users to interactively adjust the relighting result through hand-drawn scribbles. Existing methods have demonstrated that the rich priors of generative models can achieve global relighting from local, sparse lighting conditions (Choi et al., 2025; Mei et al., 2023). However, such control is primarily expressed through the spatial distribution of illumination, while other lighting attributes, such as intensity and chromaticity, remain less explicitly controllable.

Achieving accurate and controllable scribble-driven relighting requires addressing two fundamental challenges. (a) Predicting a complete illumination distribution from a sparse illumination representation is inherently underdetermined. Without a well-defined mapping between the lighting scribble and the target illumination, precise control cannot be achieved. During inference, this ambiguity may manifest as drift in lighting attributes or as relighting results that remain constrained by the distribution of the training data. (b) Training data with physically consistent light transport are difficult to obtain. Triplets consisting of an input image, a lighting condition, and a relit image with a consistent mapping are typically synthesized using a renderer. However, the distributional gap between synthetic and real data can limit the model's performance on real-world images. In particular, models trained exclusively on synthetic data may produce results with overly smooth textures and an artificial appearance, compromising photorealism.

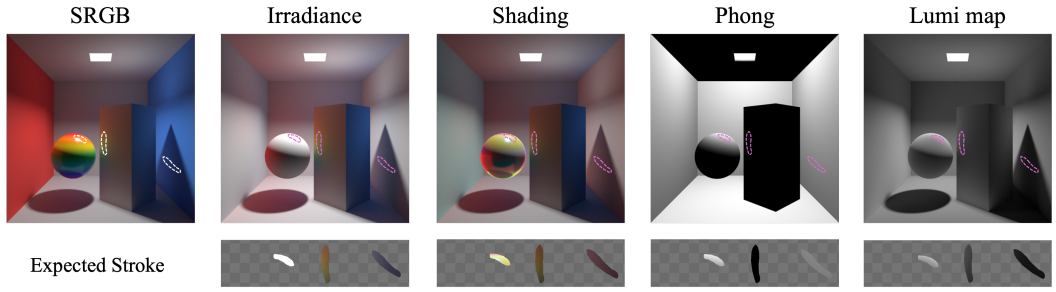

Figure 2: **Comparison of illumination representations.** In a Cornell box lit from above, irradiance includes material-dependent interreflections, while shading ignores indirect illumination and introduces additional errors through ideal Lambertian decomposition. Phong captures only the light—normal relationship, omitting occlusion and distance attenuation. In contrast, the Lumi Map directly represents the spatial distribution of source illumination incident on scene surfaces.

To address the first challenge, we define a material-decoupled illumination representation, termed the *Lumi Map*, which encodes not only the spatial distribution of illumination but also its intensity and chromaticity. Existing methods typically adopt shading maps or Phong-based representations as lighting conditions. The former entangle illumination with the material properties of object surfaces, whereas the latter fail to capture occlusion relationships, as illustrated in Fig. 2. We learn the mapping between the lumi map and the corresponding relighting result. At inference time, modifying the values of the lumi map enables controllable relighting while preserving material consistency. Specifically, we carefully design a data-synthesis pipeline to ensure a consistent correspondence between each lumi map and its relighting result. By strictly controlling exposure parameters, normalizing material properties, and varying camera viewpoints and light-source attributes, we increase data diversity while preserving this mapping. Finally, we directly export the lumi map in linear color space and use it as the illumination condition.

For the second challenge, we train a lumi map estimator to infer lumi maps from real relighting image pairs, thereby constructing training triplets of the form $(I_{src}, L_{tar}, I_{tar})$, where I_{src} is the source image, L_{tar} is the target illumination condition, and I_t is the target image. Recent work,

such as Vision Banana (Gabeur et al., 2026), has shown that generative models can learn visual representations through fine-tuning with a relatively small amount of data. We realize this process through a two-stage training strategy. In the first stage, using synthetic data, we jointly train a general-purpose image editing model on two tasks: Lumi-driven image relighting and lumi map estimation from images. In the second stage, we incorporate real relighting pairs and estimate the illumination representation of the target image online during training, denoted as the target lighting L_{tar}. We then use L_{tar} as the conditioning signal, together with the source image I_{src}, to predict the target image I_{tar}. This design aligns the forward process used in training with that used at inference time, while also improving the model's ability to generate realistic relit images.

We evaluate the accuracy of our method in performing relighting from sparse illumination representations on paired data with ground-truth targets. Both quantitative and qualitative results demonstrate the effectiveness of our approach. In addition, we evaluate its scribble-driven relighting performance on manually annotated real-world images. The experimental results show that our method achieves state-of-the-art relighting accuracy.

Our contributions can be summarized as follows: (a) We introduce a material-decoupled illumination representation that facilitates learning a more consistent mapping between illumination conditions and relighting results, thereby enabling more accurate control. (b) Building upon this high-quality illumination representation, we propose a unified training framework for relighting and illumination estimation, which extends training from synthetic data to real-world data. (c) Finally, we present a state-of-the-art scribble-driven image relighting method that produces high-quality relighting results on real-world images.

2 Related Work

Diffusion-based relighting. Recent methods leverage the image priors of diffusion models to improve relighting realism (Jin et al., 2024; Xing et al., 2025) and combine them with diverse conditioning signals to provide different degrees of control (Zeng et al., 2024a; Kocsis et al., 2024; Magar et al., 2025; Yen et al., 2026). IC-Light (Zhang et al., 2025) scales diffusion-based illumination editing to more than ten million training examples under a consistent-light-transport constraint; its released models relight a supplied foreground using either text or a background image as the condition. TokenLight (Chaturvedi et al., 2026) encodes lighting attributes—including source position, chromaticity, intensity, ambient illumination, and diffuse level—as tokens, enabling parameter-based control of the resulting illumination. Lume-Palette (Gao et al., 2026) focuses on spatial consistency in multi-view indoor relighting: it derives dense spatial lighting conditions from coarse reconstructed geometry and uses them to cast canonical illumination priors. General-purpose multimodal generators, such as Qwen-Image-Edit (Wu et al., 2025) and Nano Banana 2 (Google DeepMind, 2026), also support relighting as an image-editing task conditioned on input images and textual instructions. Although these methods explore how additional conditions can improve relighting control, practical controllability remains limited by the difficulty of providing conditioning signals that are both easy to author and sufficiently precise.

Scribble-guided relighting. Digital art shows that drawing can specify illumination directly in image space (Okabe et al., 2007). SmartShadow (Zhang et al., 2021a) is an early data-driven system that converts sparse 2D marks into artistic shadows. Although limited to line drawings rather than photorealistic relighting, it establishes such marks as an intuitive illumination interface. LightPainter (Mei et al., 2023) enables interactive portrait relighting from freehand shading strokes. It represents surface illumination with Phong shading and samples training strokes from quantized Phong maps. Because Phong omits visibility and distance attenuation, these strokes cannot describe complex spatial light transport; the method's portrait-specific design further limits its use on unconstrained photographs. ScribbleLight (Choi et al., 2025) extends scribble guidance to indoor scenes using albedo-conditioned diffusion and a ControlNet (Zhang et al., 2023) with predicted normals and binary brighten–darken annotations. Its strokes serve as binary labels for bright and shadowed regions, encoding neither continuous relative magnitude nor chromaticity. Unlike these methods, our material-decoupled representation retains spatial support, relative magnitude, and chromaticity per sparse sample, enabling fine-grained brightness and color control.

Illumination representations for image decomposition. Prior work exposes illumination at different levels of abstraction. The Phong model (Phong, 1975) expresses appearance as a combination

of ambient, diffuse, and specular terms, but its coefficients conflate illumination with material response. Intrinsic-image decomposition (Barrow & Tenenbaum, 1978; Careaga & Aksoy, 2024) factors an observed image into reflectance and shading. Although shading provides an editable image-space layer, it remains coupled with scene geometry and depends on decomposition accuracy. Irradiance offers a more physically grounded representation, defined as the cosine-weighted integral of incident radiance over the hemisphere at a surface. Under distant illumination, diffuse irradiance is dominated by low-frequency components (Ramamoorthi & Hanrahan, 2001). Although independent of the local receiving material, irradiance may include indirect illumination and therefore depend on materials elsewhere in the scene. Recent diffusion models expose such decomposed quantities as editing controls. RGB↔X (Zeng et al., 2024b) estimates albedo, normals, roughness, metallicity, and diffuse irradiance from an RGB image and synthesizes images from all or a subset of these channels. Diffusion Renderer (Liang et al., 2025) estimates editable G-buffers, including albedo, normals, depth, roughness, and metallicity, from video and conditions a learned renderer on these buffers and an environment map. These factorized interfaces support relighting and material editing but require dense intrinsic estimates and explicit lighting inputs, rather than sparse image-space observations.

3 Illumination Representation with Material Decoupling

A sparse lighting stroke is useful only if its retained value has a predictable effect on the completed image. Table 1 summarizes the symbols used below. For fixed geometry $\mathcal{G}$, material field $\mathcal{M}$, light source parameters Λ, and camera $\mathcal{C}$, we seek a dense signal that can later be sparsified while retaining image-space support, relative magnitude, and chromaticity. The signal should exclude the original material response but retain the geometric factors that place direct illumination and cast shadows. We derive it under the standard steady-state, surface-only light-transport model: radiance is conserved along unobstructed rays, no participating medium is present, and scattering vertices are described by opaque BRDFs. These assumptions match the reflected component used by our canonical direct-diffuse pass. All radiance quantities below may be evaluated spectrally or per RGB channel. We follow the standard convention that both incident and outgoing directions point away from their local surface.

Table 1: **Notation for light transport.**

Symbol	Meaning
$\mathcal{G}, \mathcal{M}, \Lambda, \mathcal{C}$	Geometry, materials, light source parameters, and camera.
x, n_x	Surface point and its geometric normal.
$\omega_i, \omega_o, \mathcal{H}_x$	Incident/outgoing directions and the upper hemisphere at x.
L_i, L_o, L_e	Incident, outgoing, and emitted radiance.
$L_{i,\mathrm{src}}, L_{i,\mathrm{ind}}$	Direct-source and indirectly reflected incident radiance.
$f_r^{\mathcal{M}}$	BRDF determined by the original materials.
p, x_p	Pixel and its first surface intersection.
I_s, I_t, L_s, L_t	Source/target images and their lumi-map conditions.

Full light transport. At a non-emissive surface point x, energy balance gives the classical rendering equation (Kajiya, 1986; Pharr et al., 2023):

$$L_o^{\mathcal{G},\mathcal{M},\Lambda}(x,\omega_o) = \int_{\mathcal{H}_x} \underbrace{f_r^{\mathcal{M}}(x,\omega_i,\omega_o)}_{\text{material coupling}} L_{i,\mathrm{src}}^{\mathcal{G},\Lambda}(x,\omega_i)(n_x\cdot\omega_i)_+ \,\mathrm{d}\omega_i + \underbrace{\int_{\mathcal{H}_x} f_r^{\mathcal{M}}(x,\omega_i,\omega_o) L_{i,\mathrm{ind}}^{\mathcal{G},\mathcal{M},\Lambda}(x,\omega_i)(n_x\cdot\omega_i)_+ \,\mathrm{d}\omega_i}_{\text{indirect reflection}}. \tag{1}$$

Here $(a)_+ = \max\{a, 0\}$. Equation (1) is complete within the stated surface model: L_i includes source emission as well as radiance that has undergone any number of earlier reflections. To expose the remaining material dependence, we partition the incident field into radiance arriving directly from a source and radiance arriving after at least one surface reflection:$L_i^{\mathcal{G},\mathcal{M},\Lambda} = L_{i,\mathrm{src}}^{\mathcal{G},\Lambda} + L_{i,\mathrm{ind}}^{\mathcal{G},\mathcal{M},\Lambda}$.

Canonicalizing the receiver material. We replace the BRDF at the receiver by a white, unit-albedo Lambertian constant, $f_r^{\mathcal{M}} \to f_r^{\mathrm{can}} = \rho^{\mathrm{can}}/\pi = 1/\pi$ with $\rho^{\mathrm{can}} = 1$. Substitution into Eq. (1) gives

$$L_{o,\mathrm{can}}^{\mathcal{G},\mathcal{M},\Lambda}(x,\omega_o) = \frac{1}{\pi}\int_{\mathcal{H}_x}\Big[L_{i,\mathrm{src}}^{\mathcal{G},\Lambda}(x,\omega_i) + L_{i,\mathrm{ind}}^{\mathcal{G},\mathcal{M},\Lambda}(x,\omega_i)\Big](n_x\cdot\omega_i)_+ \,\mathrm{d}\omega_i. \tag{2}$$

This canonicalization removes only the receiver's original BRDF. The indirect field $L_{i,\mathrm{ind}}$ still depends on the materials of earlier scattering surfaces, so canonicalizing the receiver alone does not yet produce a material-independent signal.

Removing material-dependent interreflection. The lumi map deliberately retains only the direct-source component of Eq. (2); it is not asserted to approximate the complete image. For pixel p, let $x_p = x_{\mathcal{G},\mathcal{C}}(p)$ be the first surface intersection under camera $\mathcal{C}$. We define

$$\begin{aligned} L_t(p) &\equiv L^{\mathrm{dir}}_{o,\mathrm{can}}(x_p, \omega_o) \\ &= \frac{1}{\pi} \int_{\mathcal{H}_{x_p}} L^{\mathcal{G},\Lambda_t}_{i,\mathrm{src}}(x_p, \omega_i)(n_{x_p} \cdot \omega_i)_+ \, \mathrm{d}\omega_i. \end{aligned} \tag{3}$$

Together, the source field, its directional support, and the receiver cosine encode source magnitude and chromaticity as well as geometry-dependent visibility, orientation, and distance coupling. Consequently, Eq. (3) depends on $\mathcal{G}$, Λ_t, and $\mathcal{C}$, but contains no factor from the original material field $\mathcal{M}$. It preserves the spatial support of direct illumination and cast shadows while excluding both the receiver response and material-colored interreflection.

In practice, we replace every non-emitting surface material by the same white, unit-albedo matte and extract the direct-diffuse component. The resulting RGB condition L_t is pixel-aligned with the target image I_t: visibility and attenuation define the support of light and shadow, while per-channel values encode relative magnitude and chromaticity. These values are relative control signals rather than calibrated radiometric measurements. Equation (3) shows that this representation provides a more direct interface for illumination control and simplifies the construction of conditioning inputs at inference time.

4 Method

In this section, we describe how to train a hand-drawn relighting model using the lumi map defined in Section 3. Due to space constraints, we briefly describe Lumi Map synthesis and sparsification in Supplementary C.2 and our two-stage training strategy in Section 4.2.

4.1 From Dense Rendering Supervision to Sparse Strokes

Paired variation in source position and chromaticity. The rendering pipeline is organized as scene, lighting condition, and camera (Figure 3). Its scenes include isolated subjects, interacting objects, and interiors so that illumination produces self-shadowing, cast shadows, and foreground–background interactions. For each fixed scene and view, we vary the source position and chromaticity. Monochromatic basis renders are combined in linear-radiance space to cover colored illumination without changing geometry or material.

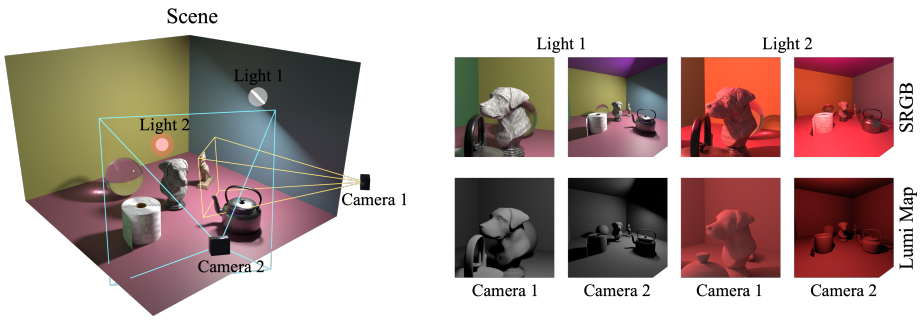

Figure 3: **Construction of paired rendered supervision.** A scene is sampled under multiple cameras and varying light source parameters (*left*). Each configuration yields a rendered sRGB image (*top*) and a pixel-aligned, canonical-material lumi map (*bottom*); both share the same geometry, viewpoint, and source.

Pixel-aligned supervision. Camera and source directions cover the scene, and configurations with a visible emitter are filtered out. For each configuration, one pass renders the image with its

original materials and a second pass renders Eq. (3) with the canonical material. The shared scene, camera, and source establish pixelwise correspondence between the target image and its lumi map. The synthetic training and validation sets are split by scene, excluding validation geometry from synthetic training. During data synthesis, we keep the exposure settings fixed and dynamically vary the light-source intensity to generate diverse relighting results.

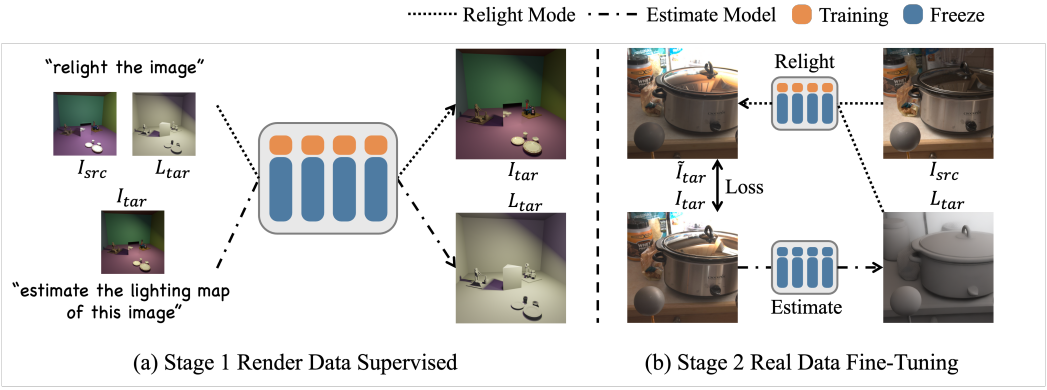

Figure 4: **Two-stage training strategy.** *Stage 1* optimizes a shared network for relighting from a rendered lumi map and for estimating that map from an image. *Stage 2* adds paired real-image reconstruction using a target-derived pseudo-control with stopped gradients, while rendered supervision remains active. The figure illustrates the case of a dense lumi map, whereas training uses lumi maps with randomly sampled sparsity levels.

4.2 Two-Stage Training with Rendered Supervision and Real Reconstruction

Stage 1: learning dense illumination and sparse completion. We denote the source and target images by I_{src} and I_{tar}, and their corresponding lumi-map illumination conditions by L_{src} and L_{tar}. We train the DiT (Peebles & Xie, 2023) model in latent space, where each image $I$is encoded into a latent $z = \mathcal{E}(I)$. On the dense and sparsified rendered pairs of Section 4.1, we train with the standard flow-matching objective (Lipman et al., 2023; Liu et al., 2023). Let the clean latent–condition pairs for mode $m \in \{G, E\}$ be $(x_0^G, c_G) = (z_{tar}, (z_{src}, z_{L_{tar}}))$ for relighting and $(x_0^E, c_E) = (z_{L_{src}}, z_{src})$ for estimation. The mode m is controlled by varying the editing prompt, as shown in Fig. 4.With timestep $t \sim \mathcal{U}(0,1)$ and $\epsilon \sim \mathcal{N}(0, I)$, let $x_t^m = (1-t)x_0^m + t\epsilon$ and $v^m = \mathrm{d}z_t^m/\mathrm{d}t = \epsilon - x_0^m$. Stage 1 optimizes

$$\mathcal{L}_{\text{flow}}(\theta) = \mathbb{E}_{m,\,c,\,\epsilon,\,t} \left\| v_\theta(x_t^m, t, c_m) - v^m \right\|^2, \tag{4}$$

so the same parameters learn relighting $(I_{src}, L_{tar} \to I_{tar})$ and lumi estimation $(I_{tar} \to L_{tar})$ under their respective mode labels.

Stage 2: estimator-guided reconstruction on real photos. Training only on rendered data can bias the model toward synthetic appearances and degrade photorealistic generation. We therefore incorporate real relighting pairs (Murmann et al., 2019; Toschi et al., 2023), which lack ground-truth target illumination conditions L_{tar}. For each pair (I_{src}, I_{tar}), the target image supplies a pseudo-control that conditions relighting of the source. Each real-data optimization step first estimates this condition using the estimation mode:

$$\widetilde{z}_{L_{tar}} = \mathrm{sg}[E_\theta(z_{I_{tar}})], \tag{5}$$

where E_θ denotes the model's forward pass in estimation mode and sg denotes the stop-gradient operator. The reconstruction loss is then defined as:

$$\mathcal{L}_{\text{rec}}(\theta) = \mathbb{E}_{z_{src},\,\widetilde{z}_{L_{tar}},\,\epsilon,\,t} \left\| v_{\theta,G}(x_t^G, t, z_{src}, \widetilde{z}_{L_{tar}}) - v^G \right\|^2. \tag{6}$$

Using only the reconstruction loss conditioned on estimated illumination may expose the model to erroneous inputs when the estimated condition is inaccurate, potentially degrading its performance. Therefore, in Stage 2, we additionally use synthetic data to jointly train the model in both relighting and estimation modes. The overall training objective is defined as:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{flow}}(\theta) + \mathcal{L}_{\text{rec}}(\theta) \tag{7}$$

5 EXPERIMENTS

Implementation details. We use Qwen-Edit-2511Wu et al. (2025) as the backbone and fine-tune it with LoRA (Hu et al., 2022) using a rank of 64. For render data, we construct 4.7K scenes in Blender and render 115K images under diverse lighting conditions. For real-world relighting data, we collect 36K images from MultiIllumMurmann et al. (2019) and ReNeToschi et al. (2023). In Stage 1, the model is trained for 15K steps at a resolution of 512×512 with a learning rate of 1×10^{-4}. In Stage 2, it is trained for an additional 5K steps at a resolution of 1024×1024 with a learning rate of 5×10^{-5}. We use the AdamW (Loshchilov & Hutter, 2019) optimizer. During inference, the classifier-free guidance (Ho & Salimans, 2022) scale is set to 4, and sampling steps is set to 40.

Benchmarks and metrics. We construct three benchmarks—ReNe (Toschi et al., 2023), MultiIllum (Murmann et al., 2019), and SynthBench—by selecting held-out image pairs with pronounced illumination changes, with 600 pairs in each benchmark. The three benchmarks exhibit progressively increasing lighting complexity. ReNe contains fixed scenes with a single subject relit by varying the position of a white light source. MultiIllum includes more diverse content but still uses a single white light source. SynthBench offers the greatest diversity, covering various relighting scenarios, including portraits and complex scenes. We additionally collect 64 photographic images from the Internet as an in-the-wild test set to evaluate generalization. Following prior work, we measure similarity to the ground truth using RMSE, PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018).

Baselines. We select ScribbleLight (Choi et al., 2025), a scribble-driven relighting method, as a baseline, and additionally include IC-Light (Zhang et al., 2025) and multimodal image editing models, including Banana (Google DeepMind, 2026), Qwen-Edit (Wu et al., 2025). To ensure a fair comparison, each baseline is evaluated using the conditioning format best suited to its design. Due to space constraints, we limit comparisons in the main paper to the most relevant methods or those with the best overall performance. Additional comparisons are provided in the Supplementary D.2.

5.1 COMPARISON WITH EXISTING METHODS

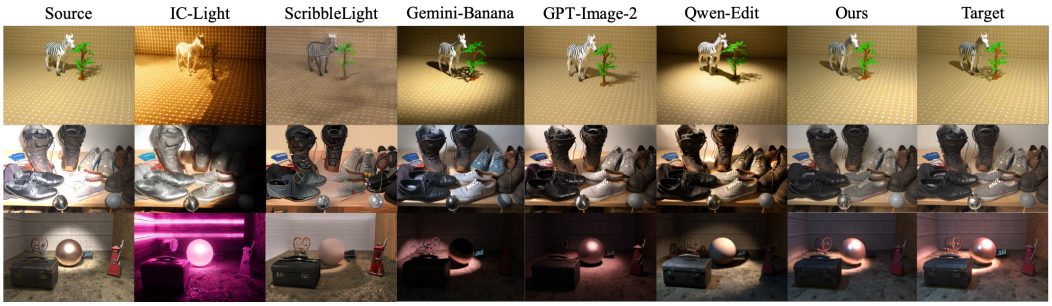

Figure 5: Qualitative comparison of relighting accuracy with existing methods across 3 benchmarks.



Figure 6: Visual analysis of relighting results on in-the-wild images.

Qualitative analysis. Figure 5 presents a visual comparison between the relighting results of different methods and the corresponding ground truth. The first row shows results on the ReNe benchmark, where our method more accurately reproduces the locations of shadows and highlights, as well as the target illumination intensity. The second row presents results on MultiIllum, demonstrating that our method better preserves object materials while accurately matching the illumination chromaticity. The third row evaluates relighting in complex synthetic scenes, where competing methods fail to reproduce the target illumination due to insufficient control, whereas dior achieves accurate relighting.

We present qualitative results on the in-the-wild dataset in Fig. 6. The first row shows that our method remains responsive to scribble guidance and produces realistic relighting results even in challenging cases, such as underwater scenes. In contrast, existing methods either fail to generate illumination with the desired direction and color or produce low-quality images. The second row further highlights the advantage of our method in terms of physical consistency, for example, in reproducing shadows cast by strong illumination and distinct specular highlights on surfaces of different materials..

Table 2: Quantitative comparison and ablation across three benchmarks.

	ReNe				**MultiIllum**				**SynthBench**			
Method	RMSE↓	PSNR↑	SSIM↑	LPIPS↓	RMSE↓	PSNR↑	SSIM↑	LPIPS↓	RMSE↓	PSNR↑	SSIM↑	LPIPS↓
IC-Light	86.52	9.47	0.3450	0.4783	96.83	8.70	0.3753	0.5742	85.02	9.74	0.3425	0.5850
ScribbleLight	47.46	14.64	0.2819	0.5632	56.29	13.39	0.5355	0.5834	76.10	10.90	0.3945	0.5356
Gemini-Banana	59.37	13.17	0.4576	0.3319	61.98	13.08	0.5798	0.3554	57.42	14.11	0.5581	0.3192
GPT-Image-2	47.89	14.64	0.2320	0.4256	66.53	12.38	0.5663	0.3496	62.37	12.94	0.4902	0.3420
Qwen-Image-Edit	50.16	14.45	0.4350	0.3483	64.64	12.50	0.5968	0.3677	57.48	13.86	0.5771	0.3256
w Irradiance	24.16	20.93	0.7333	0.1942	36.41	17.54	0.7602	0.2410	32.62	19.10	0.7974	0.1920
w Shading	37.95	16.74	0.7089	0.2537	38.76	16.78	0.7483	0.2554	35.89	18.29	0.7563	0.2134
w Phong	39.66	16.64	0.6858	0.2402	51.51	14.59	0.7026	0.2964	49.94	14.67	0.7778	0.2441
Dior w/o Stage 2	25.09	20.34	0.7321	0.2195	45.86	15.65	0.6947	0.2871	29.82	19.69	0.8293	0.1656
DIOR (ours)	**20.03**	**22.38**	**0.7542**	**0.1740**	**33.49**	**18.21**	**0.7659**	**0.2275**	**24.89**	**21.12**	**0.8637**	**0.1166**

Quantitative analysis. Table 2 compares the relighting accuracy of our method with existing approaches. Existing methods provide limited illumination control, resulting in substantial discrepancies from the target relit images. In contrast, our method enables effective illumination control and achieves the best performance across all evaluation settings.

Table 3: Elo ratings on itw images.

Method / variant	Elo↑	W/L/T (%)
IC-Light	877.99	41.1/58.9/0.0
ScribbleLight	409.56	4.2/95.8/0.0
Gemini-Banana	1169.55	63.4/36.6/0.0
GPT-Image-2	1113.67	53.6/46.1/0.3
Qwen-Image-Edit	661.25	16.3/83.7/0.0
w Irradiance	1320.85	82.2/17.8/0.0
w Shading	1109.59	68.3/31.7/0.0
w Phong	953.00	54.9/45.1/0.0
w/o stage2	1181.85	70.6/29.4/0.0
DIOR (ours)	**1410.65**	**87.2**/12.8/0.0

To quantitatively evaluate relighting on in-the-wild images without ground truth, we conduct randomized pairwise comparisons of outputs from different methods, using GPT-5.2 as a vision–language model (VLM) judge. The VLM selects the preferred result based on three criteria: adherence to the input conditions, light-transport consistency, and photorealism. We compute Elo ratings from approximately $2,000$ comparisons. Our method achieves the best performance, with a win rate of 87.2% and an Elo rating of 1410.65. We emphasize that this evaluation is not a rigorous assessment of physical simulation accuracy but serves as a proxy for user experience in practical applications. Further details are provided in Supplementary D.1.

5.2 Ablation Study

Table 2 reports their impact on relighting accuracy. We repeat training with different illumination representations and evaluate relighting performance using the corresponding sparse inputs at inference time. Replacing the Lumi Map with alternative illumination representations consistently degrades performance. The gap from the full model is relatively small on ReNe and MultiIllum, where the light-source attributes are comparatively simple, but becomes substantially larger on SynthBench, which contains more complex illumination. This result highlights the advantage of the material-decoupled design of the Lumi Map at inference time. Omitting the Stage 2 training strategy while keeping the total training budget unchanged also leads to a clear performance drop on real-image benchmarks, demonstrating the effectiveness of reconstruction training with real relighting pairs.

Figure 7 presents the corresponding qualitative ablations. When trained with illumination representations other than the Lumi Map, the model follows the lighting scribbles less reliably, as shown in the first row. Shading additionally causes noticeable color shifts, such as the oversaturated facial appearance in the second row. Without Stage 2 training, the relighting results appear less natural, with abrupt illumination transitions across the face and inconsistent blending between the foreground and the black background.

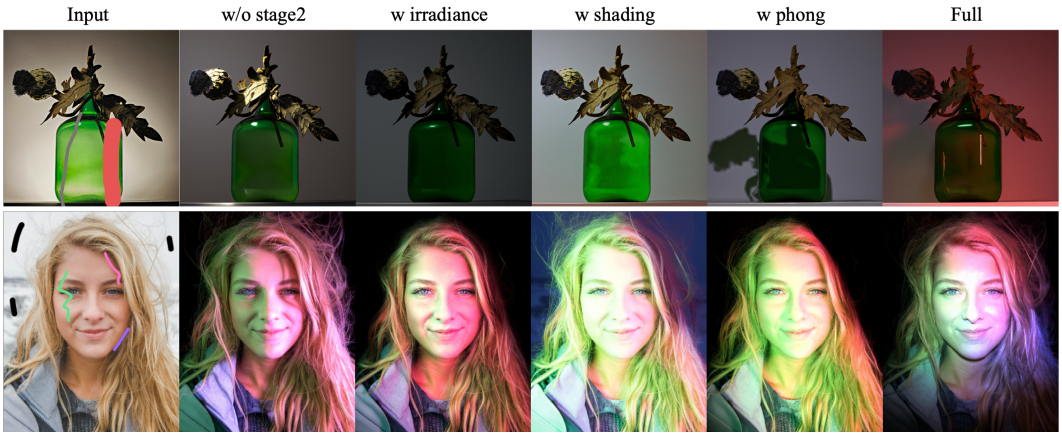

Figure 7: **Qualitative ablation of illumination representations and Stage 2.** Each row uses the same source photograph and sparse lumi drawing for the three alternative representations, Dior w/o Stage 2, and the final Dior model. The cases compare chromatic and spatial responses, respectively. No relit reference is available, so the figure is not used for quantitative ranking.

5.3 Additional analysis for application

In this section, we demonstrate the applicability of our method to practical relighting scenarios. As shown in Figure 8, users can adjust the chromaticity and intensity of the illumination to achieve customized lighting effects. They can also progressively add strokes to obtain more refined relighting results. Overall, our method enables users to create complex relighting effects through intuitive hand-drawn strokes.



Figure 8: Visualization of illumination intensity and chromaticity control with our method.

6 Conclusion

In this work, we presented **Dior-Light**, a hand-drawn image relighting method built upon the **Lumi Map**, a material-decoupled representation that encodes the spatial distribution, relative intensity, and chromaticity of direct illumination. We constructed pixel-aligned synthetic training triplets and adopted a two-stage training strategy that combines rendered supervision with estimator-guided reconstruction on real relighting pairs. Experiments on three paired benchmarks show that the proposed representation and training strategy improve relighting accuracy over the evaluated baselines, while qualitative results indicate that the model can respond consistently to sparse hand-drawn controls on in-the-wild images. Overall, Dior-Light provides a practical approach to controlling illumination distribution, intensity, and chromaticity without explicit geometry or material estimation or per-image optimization.

AI USE STATEMENT

We used AI tools for translation, language polishing, and grammar correction. The authors reviewed and take responsibility for all content.

REFERENCES

Harry G. Barrow and Jay M. Tenenbaum. Recovering intrinsic scene characteristics from images. In Allen R. Hanson and Edward M. Riseman (eds.), *Computer Vision Systems*, pp. 3–26. Academic Press, 1978.

Chris Careaga and Yağız Aksoy. Colorful diffuse intrinsic image decomposition in the wild. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 43(6), 2024.

Sumit Chaturvedi, Yannick Hold-Geoffroy, Mengwei Ren, Jingyuan Liu, He Zhang, Yiqun Mei, Julie Dorsey, and Zhixin Shu. Tokenlight: Precise lighting control in images using attribute tokens. *arXiv preprint arXiv:2604.15310*, 2026.

Jun Myeong Choi, Annie Wang, Pieter Peers, Anand Bhattad, and Roni Sengupta. Scribblelight: Single image indoor relighting with scribbles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Valentin Gabeur, Shangbang Long, Songyou Peng, Paul Voigtlaender, Shuyang Sun, Yanan Bao, Karen Truong, Zhicheng Wang, Wenlei Zhou, Jonathan T. Barron, et al. Image generators are generalist vision learners. *arXiv preprint arXiv:2604.20329*, 2026.

Chenjian Gao, Linning Xu, and Tianfan Xue. Decoupled illumination priors for spatially controllable multi-view indoor scene relighting. *arXiv preprint arXiv:2607.08879*, 2026.

Google DeepMind. Nano banana 2: Combining pro capabilities with lightning-fast speed. Official Google product announcement, 2026. URL `https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/`. Gemini 3.1 Flash Image.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Haian Jin, Yuan Li, Fujun Luan, Yuanbo Xiangli, Sai Bi, Kai Zhang, Zexiang Xu, Jin Sun, and Noah Snavely. Neural gaffer: Relighting any object via diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

James T. Kajiya. The rendering equation. In *Proceedings of the 13th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pp. 143–150, 1986.

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 (CVPR)*, pp. 9359–9369, 2024.

Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Chih-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, and Zian Wang. Diffusion renderer: Neural inverse and forward rendering with video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26069–26080, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Nadav Magar, Amir Hertz, Eric Tabellion, Yael Pritch, Alex Rav-Acha, Ariel Shamir, and Yedid Hoshen. Lightlab: Controlling light sources in images with diffusion models. *arXiv preprint arXiv:2505.09608*, 2025.

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 (CVPR)*, 2023.

Lukas Murmann, Michaël Gharbi, Miika Aittala, and Frédo Durand. A dataset of multi-illumination images in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Makoto Okabe, Yasuyuki Matsushita, Li Shen, and Takeo Igarashi. Illumination brush: Interactive design of all-frequency lighting. In *Pacific Conference on Computer Graphics and Applications (Pacific Graphics)*, pp. 171–180, 2007. doi: 10.1109/PG.2007.9.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Matt Pharr, Wenzel Jakob, and Greg Humphreys. *Physically Based Rendering: From Theory to Implementation*. MIT Press, 4 edition, 2023. URL https://pbr-book.org/4ed/contents.

Bui Tuong Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18(6): 311–317, 1975. doi: 10.1145/360825.360839.

Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pp. 497–500, 2001. doi: 10.1145/383259.383317.

Marco Toschi, Riccardo De Matteo, Riccardo Spezialetti, Daniele De Gregorio, Luigi Di Stefano, and Samuele Salti. Relight my nerf: A dataset for novel view synthesis and relighting of real world objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Eric Veach. *Robust Monte Carlo Methods for Light Transport Simulation*. PhD thesis, Stanford University, 1997.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Xiaoyan Xing, Konrad Groh, Sezer Karaoglu, Theo Gevers, and Anand Bhattad. LumiNet: Latent intrinsics meets diffusion models for indoor scene relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 442–452, 2025.

Chi-En Yen, Duy-Khanh Ngo, Wen-Wei Tang, Huu-Phu Do, Wen-Hsiao Peng, and Ching-Chun Huang. FreeLit: Paired-free indoor relighting via physics-guided diffusion. *arXiv preprint arXiv:2607.13656*, 2026.

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*, 2024a.

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 *ACM SIGGRAPH 2024 Conference Papers*, 2024b.

Lvmin Zhang, Jinyue Jiang, Yi Ji, and Chunping Liu. SmartShadow: Artistic shadow drawing tool for line drawings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5391–5400, 2021a.

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 (ICCV)*, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *International Conference on Learning Representations (ICLR)*, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

Xiuming Zhang, Pratul P. Srinivasan, Boyang Deng, Paul Debevec, William T. Freeman, and Jonathan T. Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 40(6):1–18, 2021b.

APPENDIX

Discussion of closely related work. LightIt (Kocsis et al., 2024), DiLightNet (Zeng et al., 2024a), and Lume-Palette (Gao et al., 2026) share our use of explicit illumination representations. LightIt uses direct shading that includes cast shadows as a diffusion conditioning signal and trains a shading estimator to construct lighting conditions for real-image training data. DiLightNet renders diffuse and specular radiance hints with homogeneous proxy materials, accounting for shadows and indirect illumination. Lume-Palette similarly renders scenes with a uniform, untextured white material to obtain spatial lighting conditions for multi-view indoor relighting. These methods provide closely related formulations, but their representations and interaction mechanisms are not designed specifically for sparse, hand-drawn lighting control.

Our work instead investigates how the choice of illumination representation affects relighting accuracy and controllability under sparse hand-drawn guidance at inference time. We compare several existing representations, formulate the material-decoupled Lumi Map based on the rendering equation, and validate its effectiveness through ablation studies. The Lumi Map is designed as a user-editable lighting interface, with stroke locations and values specifying the spatial distribution, intensity, and chromaticity of illumination. We further develop a unified relighting and estimation training strategy that enables the model to respond to these sparse conditions. Thus, although these works share the use of illumination representations, they differ in their design objectives and application settings. Our contribution lies in an illumination representation tailored to sparse, user-specified control at inference time, together with a training strategy that supports this interaction.