LuxRemix: Lighting Decomposition and Remixing for Indoor Scenes

Ruofan Liang[1,2]   Norman Müller[1]   Ethan Weber[1]   Duncan Zauss[1]

Nandita Vijaykumar[2,*]   Peter Kontschieder[1,*]   Christian Richardt[1,*]

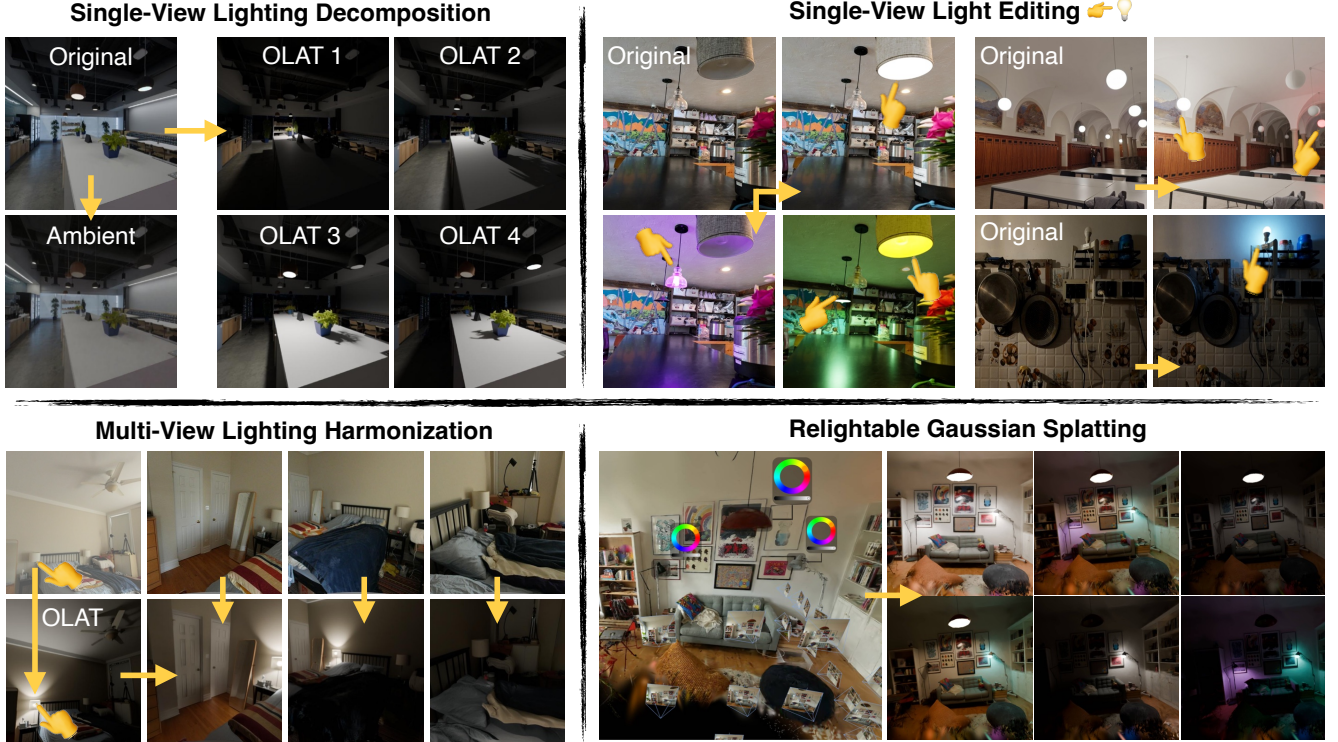[1] Meta Reality Labs   [2] University of Toronto

Figure 1. LuxRemix enables interactive light editing of indoor scenes. Our method decomposes complex scene lighting into one-light-at-a-time (OLAT) sources and ambient lighting, which can be remixed for relighting effects. In the top right, we apply our method to single images, where we can change lights and their colors. LuxRemix also enables multi-view-consistent harmonization of the decomposed lighting across multi-view images. By combining these capabilities, we enable real-time relighting of indoor scenes using 3D Gaussian splatting.

Abstract

We present a novel approach for interactive light editing in indoor scenes from a single multi-view scene capture. Our method leverages a generative image-based light decomposition model that factorizes complex indoor scene illumination into its constituent light sources. This factorization enables independent manipulation of individual light sources, specifically allowing control over their state (on/off), chromaticity, and intensity. We further introduce multi-view lighting harmonization to ensure consistent propagation of the lighting decomposition across all scene views. This is integrated into a relightable 3D Gaussian splatting representation, providing real-time interactive control over the individual light sources. Our results demonstrate highly photorealistic lighting decomposition and relighting outcomes across diverse indoor scenes. We evaluate our method on both synthetic and real-world datasets and provide a quantitative and qualitative comparison to state-of-the-art techniques. For video results and interactive demos, see luxremix.github.io.

1. Introduction

Controlling lighting in indoor scenes is fundamental to photography, cinematography, and virtual production workflows. Professional photographers carefully adjust individual light sources to achieve desired aesthetic and functional lighting

---

* Joint advising

conditions. However, this fine-grained control is typically lost after capture: images and 3D reconstructions bake lighting into the recorded appearance, making post-capture lighting adjustments difficult or impossible. Enabling flexible post-capture editing of individual light sources in multi-view captured scenes dramatically expands creative possibilities for photographers, filmmakers, and digital artists, allowing them to refine lighting decisions after the fact or explore alternative lighting configurations without physical recapture.

Existing approaches to scene relighting face significant limitations. Data-driven methods require dense multi-lighting captures under controlled conditions, which is impractical for real-world indoor scenes and offers limited generalization [6, 10]. Optimization-based inverse rendering can decompose scenes into geometry, materials, and lighting, but remains computationally intensive and often fails to produce plausible results when lighting conditions change significantly [65, 83]. Recent prior-driven approaches leverage pretrained diffusion models for relighting, but focus primarily on objects [40, 90], portraits [15], or simple scenes under distant illumination [51]. Indoor scenes present unique challenges: spatially varying illumination from multiple near-field sources creates complex lighting interactions that are difficult to decompose and edit, especially when selectively switching individual lights on or off. Moreover, existing single-image methods [58] cannot maintain 3D consistency across multiple viewpoints, limiting their applicability to multi-view reconstructions.

We present LuxRemix, a novel approach for interactive light editing in indoor scenes from a single multi-view capture. Our method decomposes the complex lighting of an indoor scene into individually controllable light sources, enabling users to interactively adjust the intensity and color of each light or switch it on or off entirely. The approach operates in three stages: first, we leverage our proposed generative image-based light decomposition model to separate the contribution of each light source; second, we propagate these decompositions consistently across all views using our multi-view lighting harmonization; third, we train a relightable 3D Gaussian splatting representation that enables real-time interactive manipulation of individual light sources from any viewpoint.

Unlike existing methods that treat lighting as a global property or focus on distant illumination, our approach models spatially varying global illumination from multiple near-field sources in complex indoor environments. While prior work on single-image lighting decomposition [58] provides individual light control, it cannot ensure consistency across multiple views. Conversely, multi-view relighting methods [1, 67] achieve 3D consistency but require controlled multi-lighting captures or produce only global relighting effects without per-light control. Our method bridges this gap by combining the fine-grained control of single-image decom-

position with the 3D consistency of multi-view methods.

The key insight enabling our approach is that modern diffusion models encode rich priors about indoor lighting that can be leveraged for decomposition, while multi-view geometric constraints provide the necessary consistency to propagate these decompositions across views. By formulating lighting decomposition as a multi-view harmonization problem and encoding the results in a fast, differentiable 3D representation (3D Gaussians), we achieve both high-quality per-light control and real-time interactive performance.

Our main contributions are:

1. A single-image lighting decomposition model to factorize complex indoor scene lighting into individually controllable light sources.
2. A multi-view lighting harmonization method to ensure 3D consistency of the individual light source decompositions across all captured viewpoints.
3. Encoding the decomposed, consistent lighting in a relightable 3D Gaussian splatting representation, enabling real-time interactive manipulation of individual near-field light sources from novel viewpoints.
4. A large-scale synthetic dataset of 12,000 generated indoor scenes with ground-truth per-light decompositions.

2. Related Work

There are three major approaches for editing scene lighting: (1) data-driven methods that require dense multi-lighting capture with limited generalization; (2) optimization-based inverse rendering that can be fragile for complex scenes; and (3) prior-driven methods that fine-tune pretrained models on task-specific data. We focus on prior-driven approaches for robustness and generalization. This research direction has gained momentum over the last year, with most relighting work focusing on humans [15, 95], objects [40, 90], or simple scenes [51, 55] under distant illumination. Scene relighting is more complex due to spatially varying illumination that is challenging to edit, e.g., to selectively switch lights on or off after the fact.

2.1. Inverse Rendering and Image Decompositions

Inverse rendering recovers scene geometry, materials, and lighting from images, enabling relighting and image editing. Traditional optimization-based approaches [54, 65, 83] leverage differentiable rendering but are computationally intensive and may lack plausibility when lighting changes. Recent neural and diffusion-based approaches have shown promising results, particularly for single images.

Single-image inverse rendering. Early work on intrinsic decomposition [13, 30] separated images into shading and albedo components, which was later extended by diffusion-based approaches [57, 78, 92] that jointly estimate intrinsic layers and enable material-aware synthesis. For complex

indoor scenes, methods have progressively improved from recovering shape, lighting, and SVBRDF [47, 48, 64] to using vision transformers [102] and diffusion models [46, 51] for more robust decomposition.

Multi-view inverse rendering. Multi-view methods can reconstruct more accurate geometry and appearance than single-view inverse rendering, but require multi-view input. Early approaches learned volumetric reflectance representations [6] and pre-integrated lighting networks [10, 17], while differentiable rendering enabled extraction of meshes with physically-based materials [34, 64, 65, 83]. Recent work has expanded to unconstrained captures [11, 25, 50], large-scale scenes [49, 101], and practical scenarios including near- and far-field sources [27], low dynamic range inputs [54], and physics-based indirect illumination [18, 21].

Single-view lighting decomposition and editing. Lighting decomposition and editing enables flexible workflows where lighting is manipulated after capture [12, 62]. Einabadi et al. [24] provide a comprehensive survey of deep neural models for illumination estimation and relighting. While latent-based methods [5, 84] achieve complex relighting effects, they lack detailed control over individual light sources. LightLab provides parametric control over individual light source intensity and color by fine-tuning diffusion models on photograph pairs [58]. Our aim is to extend this capability to multi-view captured scenes.

2.2. Image- and Video-based Relighting

Prior-driven relighting approaches involve training or fine-tuning a pretrained model on large datasets with (pseudo-)ground-truth relighting pairs. This allows the model to learn lighting priors and appearance changes under varying illumination, often generalizing well to unseen scenes, especially when fine-tuned on real-world data.

Single-image relighting. Recent diffusion-based methods enable detailed lighting control for object relighting, including DiLightNet [90], Neural Gaffer [40], and IC-Light [95]. For portrait relighting with environment maps, methods range from light stage training [66, 93] and 3D-aware representations [60, 72] to physics-driven architectures [42] and diffusion models trained on synthetic faces [15]. However, environment maps assume distant illumination, which is unrealistic for indoor scenes. Relighting without environment maps can leverage the background image as a reference [73, 80]. For outdoor scenes, methods range from self-supervised inverse rendering [89] to shadow prediction via learned ray-marching [31] or conditional diffusion models [45]. For indoor scenes, diffusion-based methods enable explicit control via light source parameters [58] or user scribbles [19], while Careaga and Aksoy [14] combine path-tracing with neural rendering for physically-based relighting.

Video relighting. Extending image relighting to video introduces temporal consistency challenges, as frame-by-frame application produces lighting and appearance flicker. Training-free approaches [99] adapt image models [95] with cross-frame attention, while end-to-end methods [61, 91] leverage video diffusion models trained on hybrid synthetic-real datasets. To avoid error accumulation in two-stage pipelines, DiffusionRenderer [51] uses G-buffers to decouple inverse and forward rendering, while UniRelight [35] jointly estimates albedo and synthesizes outputs in one pass.

Multi-view relighting. Multi-view relighting methods leverage geometric information to achieve 3D consistency and novel view synthesis. Early methods used proxy geometry from multi-view stereo with geometry-aware networks [23, 67]. NeRF-in-the-Wild [59] and GaRe [3] use appearance embeddings for varying illumination, which lacks explicit relighting control. Recent diffusion-based approaches synthesize multi-illumination data from single-illumination captures [69] or relight input images before reconstruction [1, 55, 97], avoiding brittle inverse rendering optimization.

2.3. 3D-Consistent and Real-time Relighting

Gaussian splatting for inverse rendering. 3D Gaussian splatting [41] has been extended to inverse rendering, addressing challenges in normal estimation and occlusion via regularization [52], ray tracing [28], or hybrid representations [87]. Methods also tackle global illumination [16], diffusion-guided material decomposition [22], and specialized effects including inter-reflections [32, 53] and radiance transfer [56, 77, 100], for high-quality relighting results.

Multi-view harmonization and consistency. Multi-view harmonization addresses the challenge of reconciling inconsistent illumination, appearance, and motion across captured views to enable accurate 3D reconstruction and relighting. For extreme illumination variation, Alzayer et al. [1] use diffusion models to harmonize inputs to a reference illumination before reconstruction, overcoming limitations of prior methods on specular objects. LightSwitch [55] incorporates material guidance for consistent multi-view relighting. CAT3D [29] and SEVA [98] generate consistent novel views for harmonization and reconstruction, while SimVS [79] handles real-world inconsistencies by training on synthetic data that simulates illumination variation and scene motion.

Real-time relightable representations. Neural rendering approaches combine multi-view stereo geometry with learned material and illumination representations for interactive relighting [68]. 3D Gaussians have recently emerged as a powerful representation for real-time rendering and relighting of 3D objects and scenes, including precomputed radiance transfer [94], BRDF decomposition [7, 28, 39], diffusion-guided estimation [22], and neural features [26, 56]. Recent advances enable real-time global illumination

Multi-View Capture **Single-View Light Decomp.** **MV Lighting Harmonization** **3DGS with Real-Time Light Edit**

Ambient

One light at a time (OLAT)

LuxRemix-MV

OLAT-Splatting

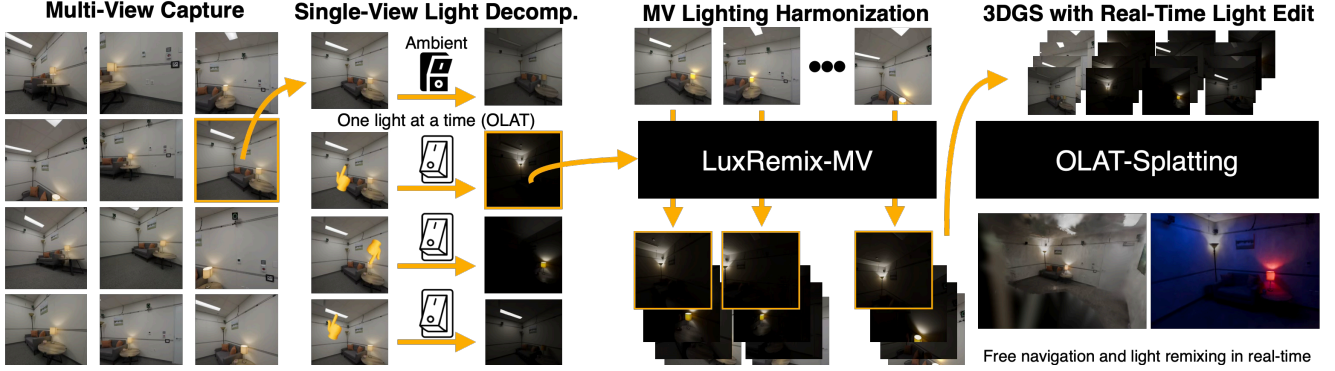Free navigation and light remixing in real-time

Figure 2. **Overview.** Given a multi-view capture of an indoor scene, we apply our single-image lighting decomposition model to generate one-light-at-a-time (OLAT) decompositions. We propagate these decompositions across all views using multi-view lighting harmonization with geometric constraints. Repeating this for each light source yields a complete multi-view OLAT dataset. We train a relightable 3D Gaussian splatting representation on these decompositions, enabling real-time interactive control over each light from any viewpoint.

with multi-bounce lighting [36]. Avatar-specific methods [75, 76, 81] and outdoor scene methods [3] leverage domain knowledge for specialized relighting. However, most methods require controlled multi-view OLAT capture rather than casual captures.

3. Multi-view Light Editing

Our goal is to decompose the complex lighting of an indoor scene into individually controllable light sources given a single multi-view capture. This enables users to interactively adjust the intensity and color of each light or switch it on or off, with view-consistent results. Figure 2 provides an overview of our three-stage approach. First, we train our single-image lighting decomposition model for per-view generative light decomposition (Section 3.2). To achieve this, we generate a large-scale synthetic dataset with ground-truth per-light decompositions (Section 3.1), enabling us to fine-tune pretrained diffusion models that encode rich priors about indoor lighting. Second, we propagate these decompositions consistently across all views using multi-view geometric constraints through a lighting harmonization framework (Section 3.3). Third, we encode the decomposed lighting in a relightable 3D Gaussian splatting representation that enables real-time rendering and interactive manipulation of individual light sources from any viewpoint (Section 3.4).

3.1. Synthetic Multi-Light Data Generation

A key component of our method is the synthetic multi-light dataset used to train our single-image lighting decomposition and multi-view harmonization models. We start with 12,000 procedurally generated 3D models of indoor scenes [2], augmenting each scene with procedurally generated light sources using Infinigen [71] with up to six controllable lights in total. This includes ceiling, wall, floor, and table lamps, as well as environment lighting [44, 70]. The emitted light

Figure 3. **Synthetic multi-light data example.** We generate 10,000 synthetic 3D scenes with procedurally generated light sources. For each scene, we render equirectangular images for multiple lighting conditions, including fully lit, randomly lit, and separate ambient and one-light-at-a-time (OLAT) configurations. This dataset allows us to train high-quality models to decompose scene lighting into individual light sources. Our models train on perspective views sampled from these equirectangular images.

source colors are sampled from black-body color temperatures to cover a spectrum of white light and augmented with 10% HSV variations to cover a wider range of colors.

We render each scene using Blender's Cycles renderer [9] with all lights on and in multiple one-light-at-a-time (OLAT) configurations, where all lights except one are switched off.
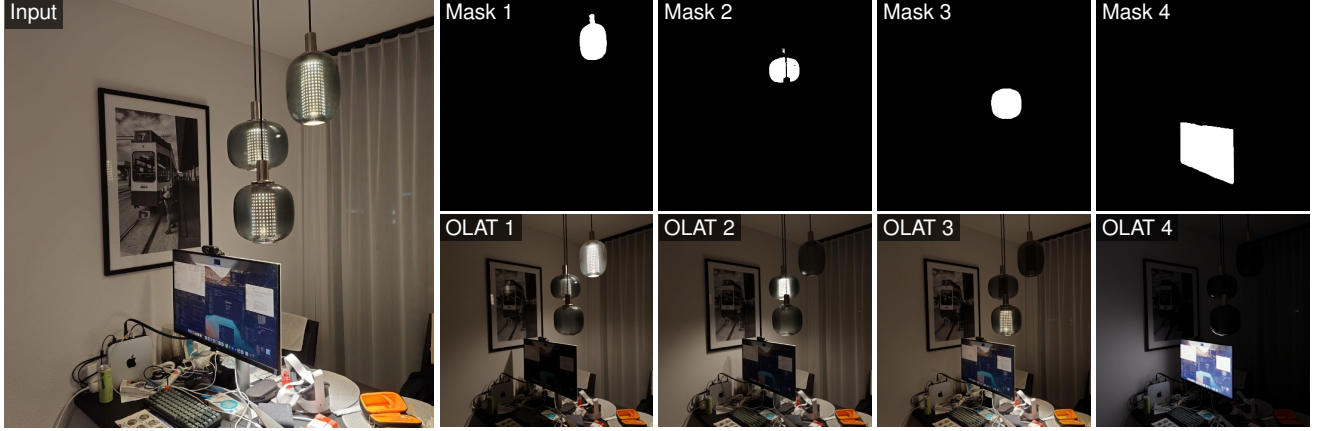
4

Figure 4. **Single-image Lighting Decomposition.** Given the light masks in the top row, our single-image lighting decomposition model realistically decomposes the scene lighting in the input image (on the left) into four individual one-light-at-a-time (OLAT) light sources.

To enable flexible viewpoint sampling during training, we render each scene as four equirectangular HDR images at roughly eye height rather than pre-rendering all possible perspective views, allowing us to sample diverse camera views on the fly while avoiding the high upfront rendering cost. For each light source, we generate three types of masks to guide the light editing process: (1) the light source area mask, computed from pixels with emissive materials only (dilated if too small); (2) the mask of the complete light fixture; and (3) the convex hull of the light fixture mask. For further details, please see the supplement. Figure 3 shows examples of our synthetic multi-light data with various light configurations.

3.2. Single-image Lighting Decomposition

Given a single input image of an indoor scene, I_{input}, we decompose the lighting into ambient lighting, I_{ambient}, and multiple one-light-at-a-time (OLAT) light sources, I_i, that can be edited individually:

$$I_{\text{input}} = \text{tonemap}\Big(I_{\text{ambient}} + \sum_{i=1}^{N} \boldsymbol{c}_i \cdot I_i\Big). \qquad (1)$$

The OLAT images are determined up to scale, requiring RGB scale factors $\boldsymbol{c}_i$ to recreate the input lighting.

We fine-tune a pretrained image editing DiT model using LoRA [37] to enable flexible light editing given the additional spatial prompt of the light selection mask, to decompose the input image into I_{ambient} and OLAT passes I_i. Our LoRA fine-tuning enables the DiT to focus on two light editing tasks: 1) OLAT decomposition using text instructions like "switch off all lights except the selected one" to generate an edited image with isolated individual light contribution; and 2) Turning off the light by prompting the model to "switch off only the selected light" to show the scene illuminated by all remaining sources, which we treat as ambient

lighting. To condition the model on which light to edit, we patchify the light mask and tokenize it via a single-layer MLP, then channel-wise add these tokens to the reference tokens alongside the input image features. To train these models, we use light composition to augment the training data by dynamically combining multiple OLAT images. Figure 4 shows single-image lighting decomposition results. We also prompt OLAT decomposition with "high/medium/low brightness" (matching EV0, EV-2, EV-4 of the target HDR OLAT) to help the model better learn light transport and capture a wider dynamic range.

3.3. Multi-view Lighting Harmonization

While existing methods can decompose lighting in single images or generate consistent multi-view imagery from scratch, no prior work addresses the challenge of propagating lighting decompositions across multi-view captures. This novel task requires both understanding the 3D scene geometry and maintaining photometric consistency while transferring complex lighting from conditional views to input views.

Given multi-view input images where lighting has been decomposed in one or more views, we aim to propagate the lighting consistently across all remaining views. Our approach takes as input the multi-view images with partial lighting decompositions along with their corresponding Plücker ray embeddings, and produces multi-view images with harmonized lighting across all views. Inspired by recent multi-view diffusion methods like CAT3D [29], SimVS [79], and SEVA [98], we follow the latter's approach to propagate lighting consistently while preserving geometric consistency. Similar to SimVS, we concatenate the original input views and sparse light-decomposed views along with the corresponding Plücker ray embeddings and reference view masks to the pretrained multi-view diffusion U-Net, and perform full-parameter fine-tuning of the U-Net. Please refer to the supplement for more details.

Figure 5. **Multi-view Lighting Harmonization.** We use the first input image to decompose the scene lighting into multiple one-light-at-a-time (OLAT) light components and the ambient lighting (highlighted in blue), which includes all unseen light sources and environment lighting. LuxRemix can turn on all three lamps for the top scene and turn off either lamp for the bottom scene. We then propagate these lighting components consistently across all views using the multi-view input images (top row). Sources: Zip-NeRF [4], VR-NeRF [86].

To generate high dynamic range (HDR) outputs, we run the lighting harmonization process three times with different exposure levels for each lighting condition ("high, medium, low" from Section 3.2). For each view and lighting condition, we then merge the three exposure-bracketed images to create the final HDR output following Debevec and Malik [20]. This per-light HDR reconstruction allows us to maintain the full dynamic range of the scene while ensuring consistent lighting across all viewpoints. Figure 5 shows our multi-view lighting harmonization results.

3.4. Real-time Remixable Lighting in 3D

We encode the decomposed lighting into a relightable 3D Gaussian splatting representation that enables real-time rendering and interactive manipulation of individual light sources from any viewpoint (see Figure 6). Building upon 3D Gaussian splatting [41], we extend each Gaussian with per-light HDR RGB coefficients, storing the appearance contribution from each light source (including ambient lighting) separately. At render time, we linearly combine these per-light contributions with user-controlled light intensity or color to produce the final appearance under arbitrary lighting configurations. This representation preserves the real-time

Figure 6. **Real-time Remixable Lighting.** The top left image shows the original scene lighting. All other images show new lighting conditions created interactively under varying viewpoints.

rendering capabilities of standard 3D Gaussian splatting while enabling independent control over each light source.

We optimize this representation in two stages. First, we pretrain a standard 3D Gaussian splatting model [88] on the original multi-view input images to establish the geometric structure and spatial distribution of Gaussians. Second, we freeze all geometric and appearance parameters, then

6

introduce and optimize per-light RGB coefficients for each Gaussian using the decomposed multi-view lighting images. We train these coefficients in linear HDR space with differentiable tone mapping, jointly solving for per-Gaussian and per-light color scaling factors that ensure the recombined lighting matches the original input images. Our training uses an L1 loss for the individual per-light images and an L1 composition loss for consistency with the original input views.

4. Evaluation

To quantitatively evaluate our single-view lighting decomposition and multi-view lighting harmonization models, we use 30 synthetic test scenes held out from training. We evaluate our models on this test set using PSNR, SSIM, and LPIPS [96] after channel-wise color rescaling w.r.t. the ground truth.

4.1. Single-Image Lighting Decomposition

We compare our LuxRemix-SV model for single-image lighting decomposition against two baseline methods: ScribbleLight [19], a diffusion-based approach for scribble-guided lighting editing, and Qwen-Image [82], a general-purpose image editing foundation model. We also evaluate several ablations of our model. The 'FLUX token' variant is fine-tuned as an in-context LoRA [38] by concatenating the input image and mask side-by-side as tokens. The 'SD' variant uses a U-Net-based latent diffusion model [74]. Our final LuxRemix-SV model uses channel-wise token addition where the mask is processed by a single-layer MLP to match the FLUX VAE latent dimensions, then added to the input condition image latents. Similar to the 'FLUX token' variant, we use LoRA for parameter-efficient fine-tuning.

Table 1 and Figure 7 show results for our single-image lighting decomposition model, which successfully isolates individual light contributions by switching off all lights except the selected one. Existing image editing models lack precise per-light control capability while our final method achieves the best performance across all metrics. Figure 8 shows a qualitative comparison of our LuxRemix-SV model with the baseline methods for light editing.

4.2. Multi-View Editing/Harmonization

Figure 5 shows qualitative results of our multi-view lighting harmonization model, which successfully propagates lighting consistently across all views. We also quantitatively compare our LuxRemix-MV model against two ablations in Table 2: 1) LuxRemix-SV processes each view independently without multi-view context, and 2) LuxRemix-MV-Edit, which extends LuxRemix-MV with additional light masks to perform mask-guided multi-view editing, instead of lighting harmonization from sparse reference views. The metrics demonstrate the necessity of our LuxRemix-MV model.

Table 1. **Evaluation of Single-Image Lighting Decomposition.** We compare our method and variants against baseline approaches on 30 synthetic test scenes.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
ScribbleLight [19]	14.39	0.395	0.688
Qwen-Image [82]	18.23	0.714	0.237
Our variants			
FLUX token [8]	25.20	0.865	0.101
SD [74]	27.13	0.857	0.099
LuxRemix-SV	**27.68**	**0.898**	**0.082**

Figure 7. **Qualitative Single-Image Lighting Decomposition.** Given an input image and light mask, we compare various methods for isolating individual light sources. Ground truth is shown at the bottom right. Existing methods fail to accurately decompose lighting, while ours performs best overall.

Table 2. **Evaluation of Multi-View Lighting Harmonization.** We compare our method against ablations on 30 synthetic test scenes.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
LuxRemix-SV	25.14	0.807	0.149
LuxRemix-MV-Edit	26.37	0.794	0.136
LuxRemix-MV	**30.76**	**0.867**	**0.091**

4.3. Real-time Remixable Lighting

Figure 6 shows our real-time remixable lighting results. Prior works do not provide the necessary level of fine-grained lighting control for interactive 3D scenes. Approaches like NeRF-W [59] and Splatfacto-W [85] can relight scenes at a per-image level when trained on multi-illumination captures, but lack per-light control. Text-based editing methods like Instruct-NeRF2NeRF [33] offer general scene editing but are also too imprecise. In contrast, our method enables interactive control over individual light sources with real-time feedback. Please see our supplementary material for our video results.
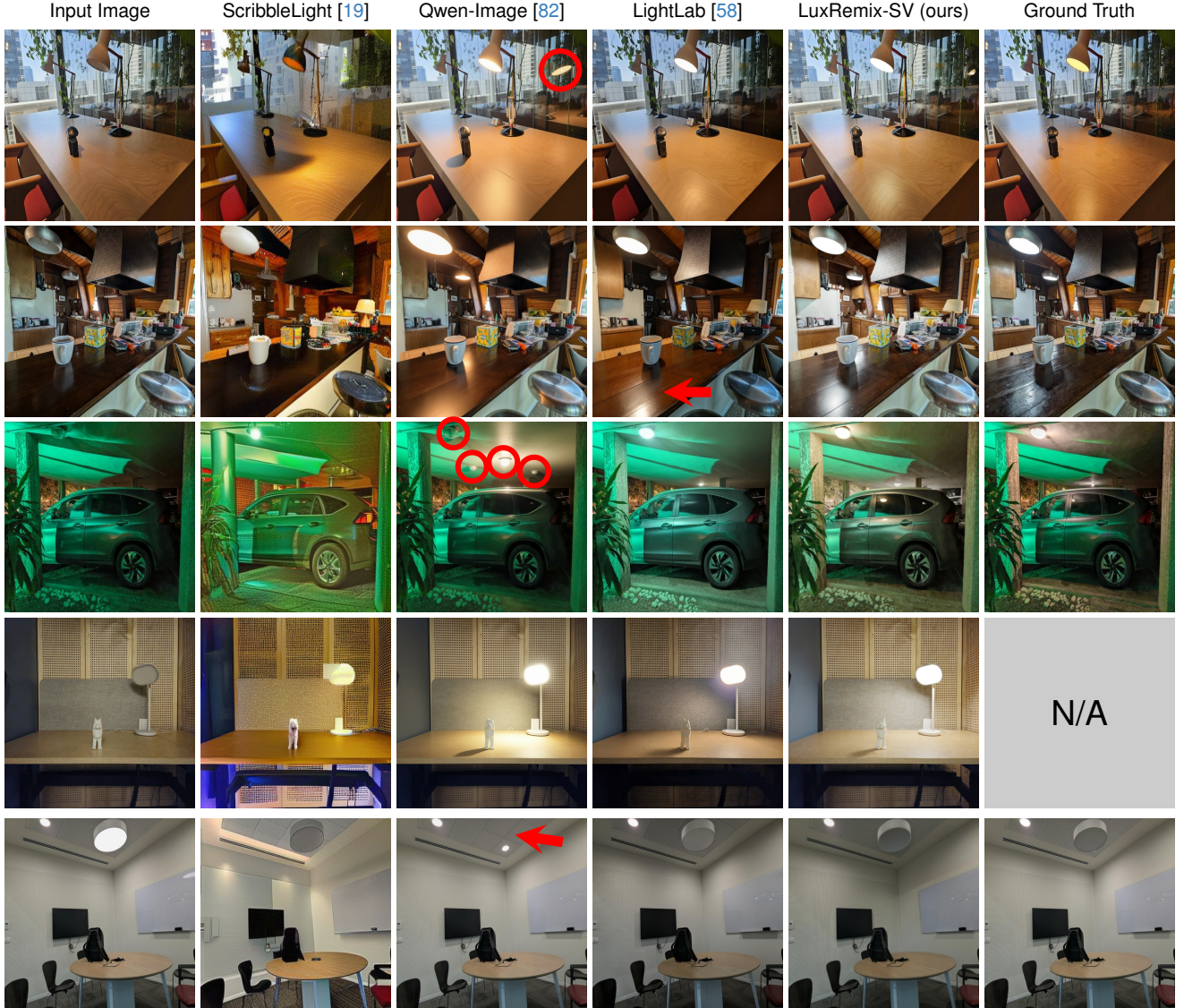
Figure 8. **Comparison of Single-Image Light Editing.** Given an input image (left) and a light mask, we compare our results with baseline methods for switching lights on and off. ScribbleLight [19] struggles to produce plausible lighting. Qwen-Image [82] shows great zero-shot image editing capabilities, but sometimes adds or removes lights or reflections as there is no mask conditioning. LightLab [58] sometimes struggles with light and object reflections. Our LuxRemix-SV model produces convincing results throughout. Image source: LightLab [58].

5. Conclusions and Limitations

We introduced LuxRemix, the first method for multi-view lighting decomposition and remixing for indoor scenes. Our approach combines a single-image lighting decomposition model with multi-view harmonization to produce view-consistent per-light decompositions from arbitrary multi-view captures. By training on a large-scale synthetic dataset with ground-truth OLAT decompositions, our models learn to isolate individual light contributions and enable flexible relighting control. We further introduce a real-time remixable lighting representation based on 3D Gaussian splat-ting that allows users to interactively manipulate individual light sources, adjusting their intensities and colors from any viewpoint. Our method opens new possibilities for interactive scene editing, virtual production, and immersive experiences, where fine-grained control over lighting is essential for achieving photorealistic results. We believe this work establishes a foundation for future research in controllable multi-view scene lighting.

While introducing a significant step towards consistent, multi-view scene relighting, our approach has remaining limitations. First, our models are trained exclusively on static synthetic indoor scenes and may not generalize well to out-

door or dynamic scenes. Second, the limited diversity of light sources in our training data introduces some bias in the lighting decomposition, tending to favor light cones over more diffuse lighting configurations. Third, distant global illumination editing via HDRIs is not supported by our model, which is left for future work.

References

[1] Hadi Alzayer, Philipp Henzler, Jonathan T. Barron, Jia-Bin Huang, Pratul P. Srinivasan, and Dor Verbin. Generative multiview relighting for 3D reconstruction under extreme illumination variation. In *CVPR*, 2025. 2, 3

[2] Armen Avetisyan, Christopher Xie, Henry Howard-Jenkins, Tsun-Yi Yang, Samir Aroudj, Suvam Patra, Fuyang Zhang, Duncan Frost, Luke Holland, Campbell Orme, Jakob Engel, Edward Miller, Richard Newcombe, and Vasileios Balntas. SceneScript: Reconstructing scenes with an autoregressive structured language model. In *ECCV*, 2024. 4, 13

[3] Haiyang Bai, Jiaqi Zhu, Songru Jiang, Wei Huang, Tao Lu, Yuanqi Li, Jie Guo, Runze Fu, Yanwen Guo, and Lijun Chen. GaRe: Relightable 3D Gaussian splatting for outdoor scenes from unconstrained photo collections. In *ICCV*, 2025. 3, 4

[4] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-NeRF: Anti-aliased grid-based neural radiance fields. In *ICCV*, 2023. 6

[5] Anand Bhattad, James Soole, and D.A. Forsyth. StyLitGAN: Image-based relighting via latent control. In *CVPR*, 2024. 3

[6] Sai Bi, Zexiang Xu, Kalyan Sunkavalli, Miloš Hašan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. Deep reflectance volumes: Relightable reconstructions from multi-view photometric images. In *ECCV*, 2020. 2, 3

[7] Zoubin Bi, Yixin Zeng, Chong Zeng, Fan Pei, Xiang Feng, Kun Zhou, and Hongzhi Wu. GS^3: Efficient relighting with triple Gaussian splatting. In *SIGGRAPH Asia*, 2024. 3

[8] Black Forest Labs. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. arXiv:2506.15742, 2025. 7, 14, 16, 19

[9] Blender Online Community. *Blender – a 3D modelling and rendering package*. Blender Foundation, 2025. 4, 13

[10] Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan T. Barron, and Hendrik P.A. Lensch. Neural-PIL: Neural pre-integrated lighting for reflectance decomposition. In *NeurIPS*, 2021. 2, 3

[11] Mark Boss, Andreas Engelhardt, Abhishek Kar, Yuanzhen Li, Deqing Sun, Jonathan T. Barron, Hendrik P. A. Lensch, and Varun Jampani. SAMURAI: Shape and material from unconstrained real-world arbitrary image collections. In *NeurIPS*, 2022. 3

[12] Ivaylo Boyadzhiev, Sylvain Paris, and Kavita Bala. User-assisted image compositing for photographic lighting. *ACM Trans. Graph.*, 32(4):36:1–12, 2013. 3

[13] Chris Careaga and Yağız Aksoy. Intrinsic image decomposition via ordinal shading. *ACM Trans. Graph.*, 43(1):12:1–24, 2023. 2

[14] Chris Careaga and Yagiz Aksoy. Physically controllable relighting of photographs. In *SIGGRAPH*, 2025. 3

[15] Sumit Chaturvedi, Mengwei Ren, Yannick Hold-Geoffroy, Jingyuan Liu, Julie Dorsey, and Zhixin Shu. SynthLight: Portrait relighting with diffusion model by learning to re-render synthetic faces. In *CVPR*, 2025. 2, 3

[16] Hongze Chen, Zehong Lin, and Jun Zhang. GI-GS: Global illumination decomposition on Gaussian splatting for inverse rendering. In *ICLR*, 2025. 3

[17] Changwoon Choi, Juhyeon Kim, and Young Min Kim. IBL-NeRF: Image-based lighting formulation of neural radiance fields. *Comput. Graph. Forum*, 42(7):e14929, 2023. 3

[18] JunYong Choi, SeokYeong Lee, Haesol Park, Seung-Won Jung, Ig-Jae Kim, and Junghyun Cho. MAIR++: Improving multi-view attention inverse rendering with implicit lighting representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(6):5076–5093, 2025. 3

[19] Jun Myeong Choi, Annie Wang, Pieter Peers, Anand Bhattad, and Roni Sengupta. ScribbleLight: Single image indoor relighting with scribbles. In *CVPR*, 2025. 3, 7, 8

[20] Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *SIGGRAPH*, pages 369–378, 1997. 6, 14

[21] Youming Deng, Xueting Li, Sifei Liu, and Ming-Hsuan Yang. Physics-based indirect illumination for inverse rendering. In *3DV*, 2024. 3

[22] Kang Du, Zhihao Liang, and Zeyu Wang. GS-ID: Illumination decomposition on Gaussian splatting via diffusion prior and parametric light source optimization. In *ICCV*, 2025. 3

[23] Sylvain Duchêne, Clement Riant, Gaurav Chaurasia, Jorge Lopez Moreno, Pierre-Yves Laffont, Stefan Popov, Adrien Bousseau, and George Drettakis. Multiview intrinsic images of outdoors scenes with an application to relighting. *ACM Trans. Graph.*, 34(5):164:1–16, 2015. 3

[24] Farshad Einabadi, Jean-Yves Guillemaut, and Adrian Hilton. Deep neural models for illumination estimation and relighting: A survey. *Comput. Graph. Forum*, 40(6):315–331, 2021. 3

[25] Andreas Engelhardt, Amit Raj, Mark Boss, Yunzhi Zhang, Abhishek Kar, Yuanzhen Li, Deqing Sun, Ricardo Martin Brualla, Jonathan T. Barron, Hendrik P. A. Lensch, and Varun Jampani. SHINOBI: Shape and illumination using neural object decomposition via BRDF optimization in-the-wild. In *CVPR*, 2024. 3

[26] Jiahui Fan, Fujun Luan, Jian Yang, Miloš Hašan, and Beibei Wang. RNG: Relightable neural Gaussians. In *CVPR*, 2025. 3

[27] Fan Fei, Jiajun Tang, Ping Tan, and Boxin Shi. VMINer: Versatile multi-view inverse rendering with near- and far-field light sources. In *CVPR*, 2024. 3

[28] Jian Gao, Chun Gu, Youtian Lin, Hao Zhu, Xun Cao, Li Zhang, and Yao Yao. Relightable 3D Gaussians: Realistic point cloud relighting with BRDF decomposition and ray tracing. In *ECCV*, 2024. 3

[29] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T. Barron, and Ben Poole. CAT3D: Create anything in 3D with multi-view diffusion models. In *NeurIPS*, 2024. 3, 5, 15

[30] Elena Garces, Carlos Rodriguez-Pardo, Dan Casas, and Jorge Lopez-Moreno. A survey on intrinsic images: Delving deep into Lambert and beyond. *Int. J. Comput. Vis.*, 130: 836–868, 2022. 2

[31] David Griffiths, Tobias Ritschel, and Julien Philip. OutCast: Outdoor single-image relighting with cast shadows. *Comput. Graph. Forum*, 41(2):179–193, 2022. 3

[32] Chun Gu, Xiaofei Wei, Zixuan Zeng, Yuxuan Yao, and Li Zhang. IRGS: Inter-reflective Gaussian splatting with 2D Gaussian ray tracing. In *CVPR*, 2025. 3

[33] Ayaan Haque, Matthew Tancik, Alexei A. Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-NeRF2NeRF: Editing 3D scenes with instructions. In *ICCV*, 2023. 7

[34] Jon Hasselgren, Nikolai Hofmann, and Jacob Munkberg. Shape, light, and material decomposition from images using Monte Carlo rendering and denoising. In *NeurIPS*, 2022. 3

[35] Kai He, Ruofan Liang, Jacob Munkberg, Jon Hasselgren, Nandita Vijaykumar, Alexander Keller, Sanja Fidler, Igor Gilitschenski, Zan Gojcic, and Zian Wang. UniRelight: Learning joint decomposition and synthesis for video relighting. In *NeurIPS*, 2025. 3

[36] Chenxiao Hu, Meng Gai, Guoping Wang, and Sheng Li. Real-time global illumination for dynamic 3D Gaussian scenes. arXiv:2503.17897, 2025. 4

[37] 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 *ICLR*, 2022. 5, 14

[38] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context LoRA for diffusion transformers. arXiv:2410.23775, 2024. 7

[39] Yingwenqi Jiang, Jiadong Tu, Yuan Liu, Xifeng Gao, Xiaoxiao Long, Wenping Wang, and Yuexin Ma. GaussianShader: 3D Gaussian splatting with shading functions for reflective surfaces. In *CVPR*, 2024. 3

[40] 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 *NeurIPS*, 2024. 2, 3

[41] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139:1–14, 2023. 3, 6, 15

[42] 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 *CVPR*, 2024. 3

[43] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 15

[44] Jeremy Klotz and Shree K. Nayar. Minimal sensing for orienting a solar panel. *Solar Energy*, 300:113833, 2025. 4

[45] Peter Kocsis, Julien Philip, Kalyan Sunkavalli, Matthias Nießner, and Yannick Hold-Geoffroy. LightIt: Illumination modeling and control for diffusion models. In *CVPR*, 2024. 3

[46] Peter Kocsis, Vincent Sitzmann, and Matthias Nießner. Intrinsic image diffusion for single-view material estimation. In *CVPR*, 2024. 3

[47] 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 *CVPR*, 2020. 3

[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 *ECCV*, 2022. 3

[49] Zhen Li, Lingli Wang, Mofang Cheng, Cihui Pan, and Jiaqi Yang. Multi-view inverse rendering for large-scale real-world indoor scenes. In *CVPR*, 2023. 3

[50] Zhengqin Li, Dilin Wang, Ka Chen, Zhaoyang Lv, Thu Nguyen-Phuoc, Milim Lee, Jia-Bin Huang, Lei Xiao, Cheng Zhang, Yufeng Zhu, Carl S. Marshall, Yufeng Ren, Richard Newcombe, and Zhao Dong. LIRM: Large inverse rendering model for progressive reconstruction of shape, materials and view-dependent radiance fields. In *CVPR*, 2025. 3

[51] Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Zhi-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, and Zian Wang. DiffusionRenderer: Neural inverse and forward rendering with video diffusion models. In *CVPR*, 2025. 2, 3

[52] Zhihao Liang, Qi Zhang, Ying Feng, Ying Shan, and Kui Jia. GS-IR: 3D Gaussian splatting for inverse rendering. In *CVPR*, 2024. 3

[53] Zhihao Liang, Hongdong Li, Kui Jia, Kailing Guo, and Qi Zhang. GUS-IR: Gaussian splatting with unified shading for inverse rendering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(10):8364–8378, 2025. 3

[54] Zhi-Hao Lin, Jia-Bin Huang, Zhengqin Li, Zhao Dong, Christian Richardt, Tuotuo Li, Michael Zollhöfer, Johannes Kopf, Shenlong Wang, and Changil Kim. IRIS: Inverse rendering of indoor scenes from low dynamic range images. In *CVPR*, 2025. 2, 3

[55] Yehonathan Litman, Fernando De la Torre, and Shubham Tulsiani. LightSwitch: Multi-view relighting with material-guided diffusion. In *ICCV*, 2025. 2, 3

[56] Zhenyuan Liu, Yu Guo, Xinyuan Li, Bernd Bickel, and Ran Zhang. BiGS: Bidirectional Gaussian primitives for relightable 3D Gaussian splatting. In *3DV*, 2025. 3

[57] Jundan Luo, Duygu Ceylan, Jae Shin Yoon, Nanxuan Zhao, Julien Philip, Anna Frühstück, Wenbin Li, Christian Richardt, and Tuanfeng Y. Wang. IntrinsicDiffusion: Joint intrinsic layers from latent diffusion models. In *SIGGRAPH*, 2024. 2

[58] 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. In *SIGGRAPH*, 2025. 2, 3, 8

[59] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, 2021. 3, 7

[60] Yiqun Mei, Yu Zeng, He Zhang, Zhixin Shu, Xuaner Zhang, Sai Bi, Jianming Zhang, HyunJoon Jung, and Vishal M. Patel. Holo-relighting: Controllable volumetric portrait relighting from a single image. In *CVPR*, 2024. 3

[61] Yiqun Mei, Mingming He, Li Ma, Julien Philip, Wenqi Xian, David M George, Xueming Yu, Gabriel Dedic, Ahmet Levent Taşel, Ning Yu, Vishal M. Patel, and Paul Debevec. Lux post facto: Learning portrait performance relighting with conditional video diffusion and a hybrid dataset. In *CVPR*, 2025. 3

[62] Abhimitra Meka, Mohammad Shafiei, Michael Zollhöfer, Christian Richardt, and Christian Theobalt. Real-time global illumination decomposition of videos. *ACM Trans. Graph.*, 40(3):22:1–16, 2021. 3

[63] Konstantin Mishchenko and Aaron Defazio. Prodigy: An expeditiously adaptive parameter-free learner. In *ICML*, 2024. 15

[64] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting triangular 3D models, materials, and lighting from images. In *CVPR*, 2022. 3

[65] Merlin Nimier-David, Zhao Dong, Wenzel Jakob, and Anton Kaplanyan. Material and Lighting Reconstruction for Complex Indoor Scenes with Texture-space Differentiable Rendering. In *Eurographics Symposium on Rendering*, 2021. 2, 3

[66] Rohit Pandey, Sergio Orts-Escolano, Chloe LeGendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. Total relighting: Learning to relight portraits for background replacement. *ACM Trans. Graph.*, 40(4), 2021. 3

[67] Julien Philip, Michaël Gharbi, Tinghui Zhou, Alexei A. Efros, and George Drettakis. Multi-view relighting using a geometry-aware network. *ACM Trans. Graph.*, 38(4):78:1–14, 2019. 2, 3

[68] Julien Philip, Sébastien Morgenthaler, Michaël Gharbi, and George Drettakis. Free-viewpoint indoor neural relighting from multi-view stereo. *ACM Trans. Graph.*, 40(5):194:1–18, 2021. 3

[69] Yohan Poirier-Ginter, Alban Gauthier, Julien Philip, Jean-François Lalonde, and George Drettakis. A diffusion approach to radiance field relighting using multi-illumination synthesis. *Comput. Graph. Forum*, 43(4):e15147, 2024. 3

[70] Poly Haven. Poly Haven: The public 3D asset library. https://polyhaven.com/, 2025. 4

[71] Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, Zeyu Ma, and Jia Deng. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *CVPR*, 2024. 4, 13

[72] Pramod Rao, Gereon Fox, Abhimitra Meka, Mallikarjun B R, Fangneng Zhan, Tim Weyrich, Bernd Bickel, Hanspeter Pfister, Wojciech Matusik, Mohamed Elgharib, and Christian Theobalt. Lite2Relight: 3D-aware single image portrait relighting. In *SIGGRAPH*, 2024. 3

[73] 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 *CVPR*, 2024. 3

[74] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 7

[75] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable Gaussian codec avatars. In *CVPR*, 2024. 4

[76] Jonathan Schmidt, Simon Giebenhain, and Matthias Niessner. BecomingLit: Relightable Gaussian avatars with hybrid neural shading. In *NeurIPS*, 2025. 4

[77] Yahao Shi, Yanmin Wu, Chenming Wu, Xing Liu, Chen Zhao, Haocheng Feng, Jian Zhang, Bin Zhou, Errui Ding, and Jingdong Wang. GIR: 3D Gaussian inverse rendering for relightable scene factorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2025. 3

[78] Shanlin Sun, Yifan Wang, Hanwen Zhang, Yifeng Xiong, Qin Ren, Ruogu Fang, Xiaohui Xie, and Chenyu You. Ouroboros: Single-step diffusion models for cycle-consistent forward and inverse rendering. In *ICCV*, 2025. 2

[79] Alex Trevithick, Roni Paiss, Philipp Henzler, Dor Verbin, Rundi Wu, Hadi Alzayer, Ruiqi Gao, Ben Poole, Jonathan T. Barron, Aleksander Holynski, Ravi Ramamoorthi, and Pratul P. Srinivasan. SimVS: Simulating world inconsistencies for robust view synthesis. In *CVPR*, 2025. 3, 5

[80] Junying Wang, Jingyuan Liu, Xin Sun, Krishna Kumar Singh, Zhixin Shu, He Zhang, Jimei Yang, Nanxuan Zhao, Tuanfeng Y. Wang, Simon S. Chen, Ulrich Neumann, and Jae Shin Yoon. Comprehensive relighting: Generalizable and consistent monocular human relighting and harmonization. In *CVPR*, 2025. 3

[81] Shaofei Wang, Tomas Simon, Igor Santesteban, Timur Bagautdinov, Junxuan Li, Vasu Agrawal, Fabian Prada, Shoou-I Yu, Pace Nalbone, Matt Gramlich, Roman Lubachersky, Chenglei Wu, Javier Romero, Jason Saragih, Michael Zollhoefer, Andreas Geiger, Siyu Tang, and Shunsuke Saito. Relightable full-body Gaussian codec avatars. In *SIGGRAPH*, 2025. 4

[82] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report. arXiv:2508.02324, 2025. 7, 8, 16, 17, 19

[83] Liwen Wu, Rui Zhu, Mustafa B. Yaldiz, Yinhao Zhu, Hong Cai, Janarbek Matai, Fatih Porikli, Tzu-Mao Li, Manmohan

11

Chandraker, and Ravi Ramamoorthi. Factorized inverse path tracing for efficient and accurate material-lighting estimation. In *ICCV*, 2023. 2, 3

[84] Xiaoyan Xing, Konrad Groh, Sezer Karagolu, Theo Gevers, and Anand Bhattad. LumiNet: Latent intrinsics meets diffusion models for indoor scene relighting. In *CVPR*, 2025. 3

[85] Congrong Xu, Justin Kerr, and Angjoo Kanazawa. Splatfacto-W: A nerfstudio implementation of Gaussian splatting for unconstrained photo collections. arXiv:2407.12306, 2024. 7

[86] Linning Xu, Vasu Agrawal, William Laney, Tony Garcia, Aayush Bansal, Changil Kim, Samuel Rota Bulò, Lorenzo Porzi, Peter Kontschieder, Aljaž Božič, Dahua Lin, Michael Zollhöfer, and Christian Richardt. VR-NeRF: High-fidelity virtualized walkable spaces. In *SIGGRAPH Asia*, 2023. 6, 18

[87] Kai Ye, Chong Gao, Guanbin Li, Wenzheng Chen, and Baoquan Chen. GeoSplatting: Towards geometry guided Gaussian splatting for physically-based inverse rendering. In *ICCV*, 2025. 3

[88] Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: An open-source library for Gaussian splatting. *Journal of Machine Learning Research*, 26(34):1–17, 2025. 6, 15

[89] Ye Yu, Abhimitra Meka, Mohamed Elgharib, Hans-Peter Seidel, Christian Theobalt, and William A. P. Smith. Self-supervised outdoor scene relighting. In *ECCV*, 2020. 3

[90] Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. DiLightNet: Fine-grained lighting control for diffusion-based image generation. In *SIGGRAPH*, 2024. 2, 3

[91] Jianshu Zeng, Yuxuan Liu, Yutong Feng, Chenxuan Miao, Zixiang Gao, Jiwang Qu, Jianzhang Zhang, Bin Wang, and Kun Yuan. Lumen: Consistent video relighting and harmonious background replacement with video generative models. arXiv:2508.12945, 2025. 3

[92] 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 *SIGGRAPH*, 2024. 2

[93] Longwen Zhang, Qixuan Zhang, Minye Wu, Jingyi Yu, and Lan Xu. Neural video portrait relighting in real-time via consistency modeling. In *ICCV*, 2021. 3

[94] Libo Zhang, Yuxuan Han, Wenbin Lin, Jingwang Ling, and Feng Xu. PRTGaussian: Efficient relighting using 3D Gaussians with precomputed radiance transfer. In *Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2024. 3

[95] 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 *ICLR*, 2025. 2, 3

[96] 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. 7

[97] Xiaoming Zhao, Pratul P. Srinivasan, Dor Verbin, Keunhong Park, Ricardo Martin Brualla, and Philipp Henzler. IllumiNeRF: 3D relighting without inverse rendering. In *NeurIPS*, 2024. 3

[98] Jinghao (Jensen) Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. In *ICCV*, 2025. 3, 5, 15

[99] Yujie Zhou, Jiazi Bu, Pengyang Ling, Pan Zhang, Tong Wu, Qidong Huang, Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Anyi Rao, Jiaqi Wang, and Li Niu. Light-a-video: Training-free video relighting via progressive light fusion. In *ICCV*, 2025. 3

[100] Yongyang Zhou, Fang-Lue Zhang, Zichen Wang, and Lei Zhang. RTR-GS: 3D Gaussian splatting for inverse rendering with radiance transfer and reflection. In *ACM Multimedia*, 2025. 3

[101] Jingsen Zhu, Yuchi Huo, Qi Ye, Fujun Luan, Jifan Li, Dianbing Xi, Lisha Wang, Rui Tang, Wei Hua, Hujun Bao, and Rui Wang. I^2-SDF: Intrinsic indoor scene reconstruction and editing via raytracing in neural SDFs. In *CVPR*, 2023. 3

[102] Rui Zhu, Zhengqin Li, Janarbek Matai, Fatih Porikli, and Manmohan Chandraker. IRISformer: Dense vision transformers for single-image inverse rendering in indoor scenes. In *CVPR*, 2022. 3

LuxRemix: Lighting Decomposition and Remixing for Indoor Scenes
Supplementary Material

This supplemental document provides additional details for LuxRemix. We first describe the data preparation process in Section 6, then introduce additional details about our models in Section 7, and finally provide additional results and comparisons in Section 8.

Please see our project website https://luxremix.github.io for the following additional supplemental material:

- **Interactive Lighting Blend Explorer:**
  A browser-based demo for interactive single-image light editing based on our lighting decompositions.
- **Single-Image Lighting Decomposition:**
  Shows the decomposition of seven light sources and their flexible recombination for various relighting effects.
- **Multi-View Lighting Harmonization:**
  Visualizes the decomposition and harmonization of multiple light sources across multiple views.
- **Real-time Remixable Lighting:**
  Screen recordings from our modified Splatfacto renderer in the Nerfstudio viewer demonstrating our interactive relighting with 3D Gaussian splatting. Note that temporarily lower rendering resolutions are due to the dynamic resolution rendering implemented in the Nerfstudio viewer.

6. Data

Here we provide additional information for how we created the training data for LuxRemix. We render scenes from Avetisyan et al. [2] with procedurally generated light sources using Infinigen [71] into many one-light-at-a-time (OLAT) equirectangular images. We choose equirectangular 360° images to save on the rendering cost and allow dynamic sampling of diverse perspective camera viewpoints on the fly while training the model. Please see Table 3 for a comparison of the rendering and storage costs.

Table 3. **Rendering and storage comparison.** The perspective views are rendered at resolution 512×512 pixels, and the equirectangular 360° images are rendered at resolution 2048×1024 pixels. 64 samples per pixel are used for Blender's Cycles path tracer [9].

Setup	Render time	Storage
20,000 perspective images	140,000 sec	17.2 GB
1,000 360° images	27,000 sec	6.8 GB
	= 19%	= 40%

Rendering scenes once. To reduce our dataset size while maintaining flexibility during training, we store equirectangular 360° images instead of pre-rendering a large collection of perspective views that our model ingests. Equirectangular representations compactly encode the full scene from a single viewpoint, allowing arbitrary camera perspectives to be sampled on the fly. In contrast, storing every possible perspective image incurs significant storage overhead and limits viewpoint diversity at training time. For example, training on 20K pre-rendered perspective images with resolution 512×512 pixels would require 17.2 GB of storage. Instead we could store 1K equirectangular images with a resolution of 2048×1024 pixels in just 6.8 GB and sample these from new viewpoints on the fly, assuming we can always sample at least 20 distinct views per equirectangular image. This design substantially reduces disk usage and data preparation cost while preserving the ability to generate diverse, view-consistent training inputs dynamically. We found it particularly useful to render the dataset once and use the stored data for the entire project.

Implementation details. As described in Section 6, we use Blender's Cycles renderer [9] to generate equirectangular HDR images for each synthetic room. For each room, we sample four distinct locations to capture equirectangular images at a resolution of 2048×1024 pixels. Rendering is performed with 64 samples per pixel (spp), using the OptiX denoiser to enhance path-traced results. We used approximately 2,800 GPU hours on NVIDIA A100s to render the complete dataset. In total, we rendered 49,600 different equirectangular views across 12,400 synthetic rooms. The additional depth map and light mask are rendered for each 360° image to assist effective perspective view sampling. Combining all different light passes, the total dataset size is about 4.5 TB.

On-the-fly sampling. As our model is designed for regular perspective views, during training, the perspective views are sampled from the rendered equirectangular 360° images. For single-image light editing, we sample perspective views such that the target light source is visible within the field of view (FOV). Specifically, we use the equirectangular light mask to select a visible light source, and then use its center to guide the sampling range of azimuth and elevation for the perspective projection.

For multi-view lighting, after selecting an initial perspective view that contains the target light source within the field of view, we sample additional perspective views to incrementally increase the overall coverage of visible regions. These views are drawn from multiple source equirectangular 360° images and are chosen to ensure that consecutive views share overlapping regions, promoting consistency across frames. To guide this sampling, we project depth maps from the various equirectangular images into a global coordinate system, allowing us to accurately identify co-visible areas within the room. Additionally, depth information is used to avoid
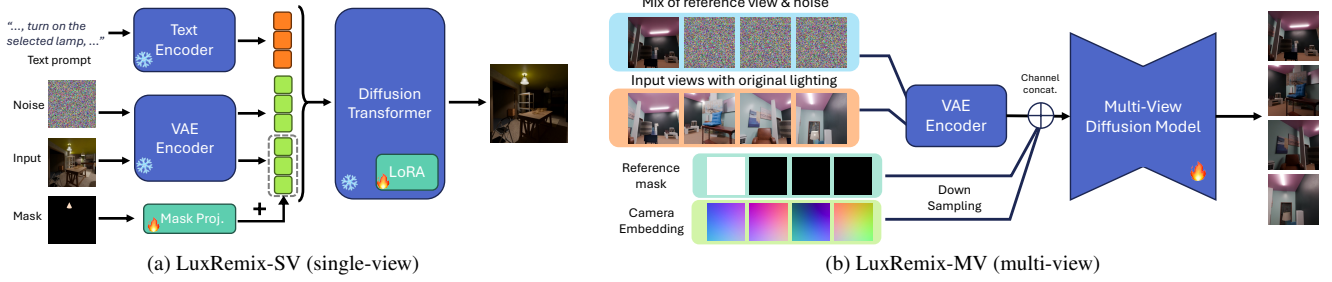
(a) LuxRemix-SV (single-view)   (b) LuxRemix-MV (multi-view)

Figure 9. Overview of our models for **(a)** single-view lighting decomposition, and **(b)** multi-view lighting harmonization.

sampling views that are too close to scene geometry or the camera, ensuring a diverse and meaningful set of training perspectives.

After sampling the perspective views, we generate training data pairs on-the-fly. For the OLAT decomposition task, we select a target OLAT image I_{target} and its corresponding full-light image I_{full}, forming the training pair $(I_{\text{full}}, I_{\text{target}})$. For the one-light-off task, we compute $I_{\text{one-off}} = I_{\text{full}} - c_{\text{target}} \cdot I_{\text{target}}$, where c_{target} is the scaling factor for the target OLAT image as indicated in the rendering metadata. The resulting pair $(I_{\text{full}}, I_{\text{one-off}})$ teaches the model to turn off the selected light source. To further augment the training set, we perform additional lighting composition by loading multiple OLAT images and adding them to the full-light image. This produces new, well-lit image variants such as $I'_{\text{full}} = I_{\text{full}} + \sum_{i=1}^{N} c'_i \cdot I_i$, enriching the diversity of the training data. To preserve the high dynamic range of the original images when converting to the sRGB color space for model training, we apply AgX tone mapping[1] to the sampled linear images, which has been shown to be effective in preserving more realistic highlight effects.

7. Models

7.1. Single-image Light Editing – LuxRemix-SV

Our single-image editing model builds upon a pretrained text-based image editing diffusion transformer (DiT) with an architecture similar to FLUX.1 Kontext [8]. Because full-parameter fine-tuning is prohibitively expensive, we adopt LoRA [37] for efficient lightweight adaptation. Importantly, the base DiT model already demonstrates strong capability for general instruction-based image editing (see Figure 11), which allows LoRA fine-tuning to specialize the model for the light editing task using only limited synthetic training data, and generalize to real-world images using the generative prior inherent in the base DiT model. Figure 9a illustrates the model design.

We use LoRA fine-tuning for two closely related light editing tasks: (1) OLAT decomposition and (2) turning off

the target light. Below are the instruction templates we used for the two tasks:

1. "{*trigger word:* OLAT}. *Darken the room to a night scene: eliminate all light sources and window light. The only light source should be the selected {light type}, on at {low,medium,high} brightness, according to the selection mask.*"
2. "{*trigger word:* LTOFF}. *Only turn off the selected {light type}, according to the selection mask. Keep the remaining light sources unchanged.*"

We use two distinct trigger words to differentiate between the two closely related tasks. For the OLAT decomposition task, we further specify the desired brightness level *"{low, medium, high} brightness"*, for the target light source. Specifically, we instruct the model to generate LDR OLAT images corresponding to approximate exposure values of EV–4 (low), EV–2 (medium), and EV0 (high) relative to the HDR target after tone mapping. This approach encourages the model to better learn light transport and enables reconstruction of HDR OLAT images through multi-exposure fusion [20]. Figure 10 shows two examples of our additional brightness level control.

Other than text prompts, we also provide the light mask as a spatial prompt for the model. The mask is processed by a single-layer MLP to match the DiT's input latent di-
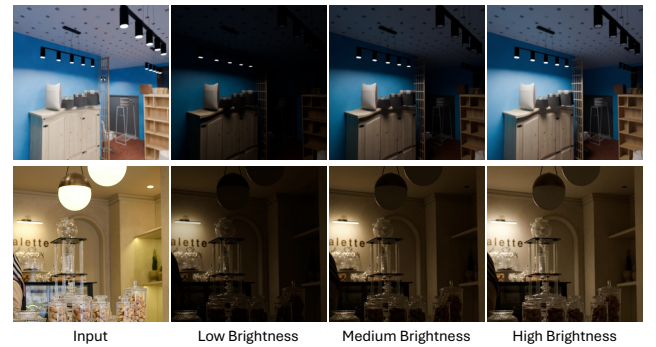

Figure 10. LuxRemix-SV's OLAT decomposition with the control of different brightness levels.

---

mensions, then added to the input condition image latents. This additional single-layer MLP is fine-tuned along with the LoRA parameters to further improve the model's ability to understand the light mask.

Training details. We insert LoRA adapters into all attention blocks of the DiT model, using a LoRA rank of 32. Fine-tuning is performed for 3,000 iterations with a batch size of 192, leveraging the Prodigy optimizer [63] for adaptive learning rate adjustment and bias correction. Training was conducted on 48 NVIDIA A100-40GB GPUs, completing in approximately 12 hours.

7.2. Multi-view Harmonization – LuxRemix-MV

Our multi-view harmonization model extends the lighting decomposition from single to multiple views. While it is possible to use the single-view LuxRemix-SV model for multi-view propagation (e.g., by processing diptychs or grids of images), this becomes computationally infeasible for large image sets due to high resource demands. To address this, we use a scalable multi-view diffusion model based on a pretrained U-Net architecture (similar to CAT3D [29] and SEVA [98]) for efficient multi-view lighting propagation.

The base model is pretrained to generate novel views given a few reference images. However, our task does not require synthesizing entirely new views; instead, we have access to multi-view input images prior to lighting decomposition, and our goal is to propagate the lighting decomposition from sparse reference views to all other output views. To accomplish this, we concatenate the original input views and the sparse light-decomposed views, together with the corresponding Plücker ray embeddings and reference view masks, as inputs to the pretrained multi-view diffusion U-Net, by extending the model's channels of the input projection layer, as illustrated in Figure 9b. To distinguish between OLAT decomposition and one-light-off editing, we also include a binary mask in the input condition: all zeros for OLAT decomposition, and all ones for one-light-off editing.

Training details. We conduct full-parameter fine-tuning of the U-Net for 30,000 iterations. The training progresses in three stages: For the first 15,000 iterations, the model is trained on 4-view input batches of size 192. In the next 10,000 iterations, the number of input views is increased to 8, with a batch size of 144. Finally, during the last 5,000 iterations, the model is further trained with 15-view input batches of size 96. We utilize the AdamW optimizer [43] with a learning rate of 5×10^{-5}. Training is performed on 48 NVIDIA A100-40GB GPUs and completes in approximately 28 hours.

Inference details. Our model is fine-tuned to handle up to 15 views per forward pass. To accommodate larger multi-view datasets, we employ a sequential multi-pass strategy. The core idea is to process images in batches while maintaining consistency by conditioning on both the original reference views and previously generated frames. Let $\mathcal{U}$ denote the set of unprocessed target frames and $\mathcal{P}$ the set of processed frames. We use a distance metric $d(I_i, I_j)$ based on camera pose similarity to guide the selection process:

- **Pass** 1: We select target frames from $\mathcal{U}$ that are spatially closest to the original source reference views $\mathcal{R}^*$.
- **Pass** $k > 1$: We iteratively select remaining frames from $\mathcal{U}$ that are closest to any frame in the processed set $\mathcal{P}$.

For each pass, we construct a dynamic reference set $\mathcal{R}_k$ that always includes the original source references $\mathcal{R}^*$. Additionally, if a target frame is selected due to its proximity to a previously generated frame $I_{\text{prev}} \in \mathcal{P}$, we optionally include I_{prev} as a secondary reference in $\mathcal{R}_k$. This approach allows the model to "chain" visual information from source views to distant targets, ensuring that subsequent passes remain consistent with the lighting decomposition established in earlier steps.

7.3. Gaussian Splatting with Lighting Control

Following the two-stage training pipeline described in Section 3.4, we extend the 3D Gaussian splatting model [41] to enable precise lighting control. We achieve this by fitting per-light RGB parameters for each Gaussian, utilizing the light-decomposed multi-view images.

Stage 1: Pretraining a standard 3DGS. We first establish the scene's geometric structure by training a standard 3D Gaussian splatting model on the original multi-view images using gsplat [88]. Upon convergence, these pretrained Gaussians serve as the basis for incorporating additional parameters for lighting control.

Stage 2: Fitting per-light RGB parameters. We augment each Gaussian from the previous stage with a set of learnable lighting parameters $\mathbf{L}_i \in \mathbb{R}^{M \times 3}$, where M denotes the number of decomposed light sources (including ambient lighting). Modeled in a linear HDR space to represent the physical accumulation of light, these parameters allow us to render the image for a specific light source m, denoted as $\hat{I}_m$, via standard rasterization using the Gaussians' existing geometry and the new learnable lighting parameters. We optimize these per-light RGB parameters as follows:

1. **Joint Optimization:** We jointly optimize the per-light RGB parameters $\mathbf{L}$ alongside the shared geometry parameters. This step refines the pretrained Gaussians' geometry and appearance to better align with the light-decomposed multi-view images. We train the Gaussians in this step for 4,000 iterations.

2. **Light Fitting (Frozen Geometry):** To prevent the model from explaining lighting residuals by altering geometry, we freeze all geometric parameters and focus solely on optimizing $\mathbf{L}$ to fit the light-decomposed images for the remaining 2,000 iterations.

Training Objectives. The optimization in Stage 2 incorporates three losses: one for the fidelity of individual light
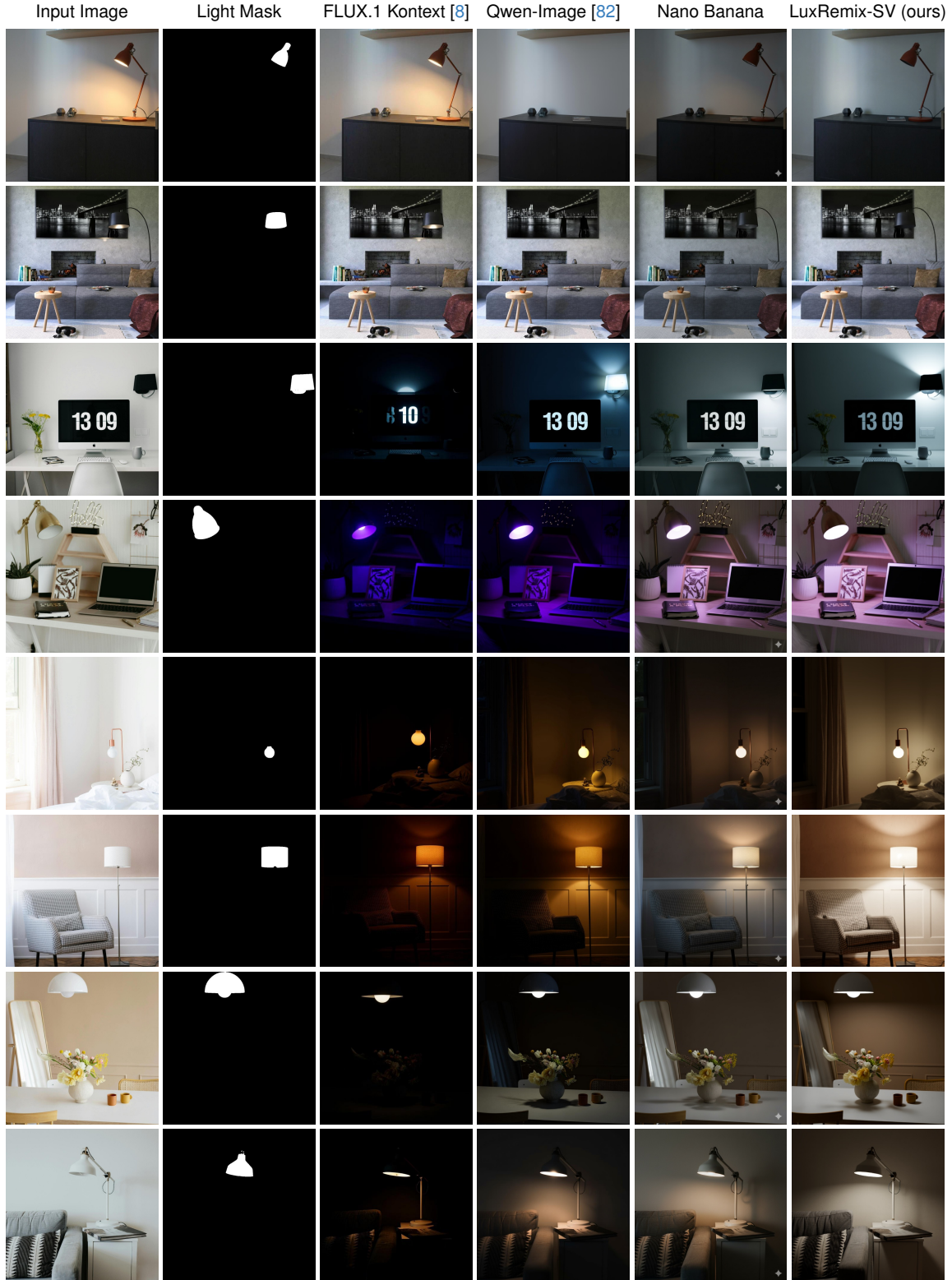
Figure 11. **Additional Comparisons of Single-Image Light Editing.** Given an input image and a light mask (left), we compare our light editing results with baseline methods for switching lights on and off. FLUX.1 Kontext [8] sometimes removes or modifies image details. Qwen-Image [82] sometimes removes lights entirely. Nano Banana produces plausible light edits. Our LuxRemix-SV model produces convincing results throughout with fine-grained controllability over individual light intensities and colors. Image source: Pexels.
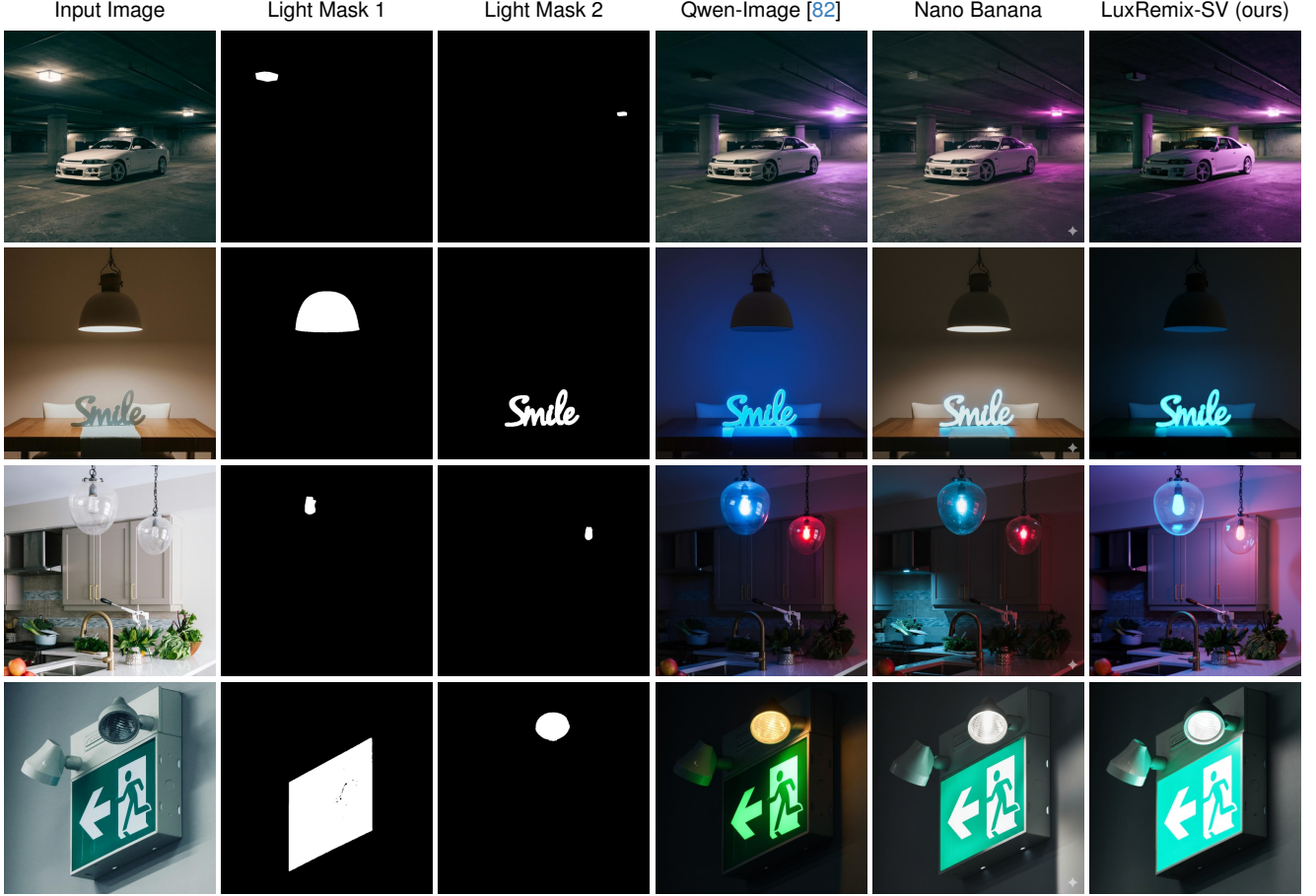
16

Figure 12. **Comparisons of Single-Image Light Editing with Multiple Masks.** Given an input image and two light masks (left), we compare our light editing results (right) with baseline image editing models. While Qwen-Image [82] and Nano Banana enable basic prompt-based light editing, only our LuxRemix-SV approach enables realistic and fine-grained light editing control via light mask conditioning. Image sources: Pexels and Unsplash.

renderings, one for the consistency of the recombined lighting, and one for the spatial smoothness of the Gaussian parameters.

The first term is the photometric loss between the rendered light image $\hat{I}_m$ and the ground-truth light image I_m:

$$\mathcal{L}_{\text{olat}} = \frac{1}{M} \sum_{m=1}^{M} \left(\|\hat{I}_m - I_m\|_1 + \lambda \mathcal{L}_{\text{D-SSIM}}(\hat{I}_m, I_m) \right). \tag{2}$$

The second term is a composition consistency loss between the recombined lighting $\hat{I}_{\text{comp}} = \sum_m w_m \hat{I}_m$ and the original input image I_{ori}, where w_m is a learnable per-light scaling factor for light recombination. This ensures that the learned lighting coefficients, when recombined, accurately reconstruct the original appearance:

$$\mathcal{L}_{\text{comp}} = \|\mathcal{T}(\hat{I}_{\text{comp}}) - I_{\text{ori}}\|_1, \tag{3}$$

where $\mathcal{T}(x) = (x + \beta)^{\frac{1}{\gamma}}$ is a differentiable tone mapping function with a learnable gamma γ and offset β.

The third component is a spatial smoothness loss. To mitigate high-frequency noise where adjacent Gaussians learn divergent light responses, we impose a spatial smoothness regularizer based on K-Nearest Neighbors (KNN). For each Gaussian i, we penalize the deviation of its lighting coefficients from its K nearest spatial neighbors $\mathcal{N}(i)$:

$$\mathcal{L}_{\text{smooth}} = \frac{1}{NK} \sum_{i=1}^{N} \sum_{j \in \mathcal{N}(i)} \|\mathbf{L}_i - \mathbf{L}_j\|_2^2. \tag{4}$$

This encourages local consistency in light reflectance, helping to reduce sparkling artifacts during relighting. We apply this loss every 100 optimization iterations after the initial 4,000 iterations.

The final objective is the weighted sum of three losses:

$$\mathcal{L} = \mathcal{L}_{\text{olat}} + \lambda_{\text{comp}} \mathcal{L}_{\text{comp}} + \lambda_{\text{smooth}} \mathcal{L}_{\text{smooth}}. \tag{5}$$

After training, we can randomly blend multiple per-light renderings from the same set of Gaussians to get consistent
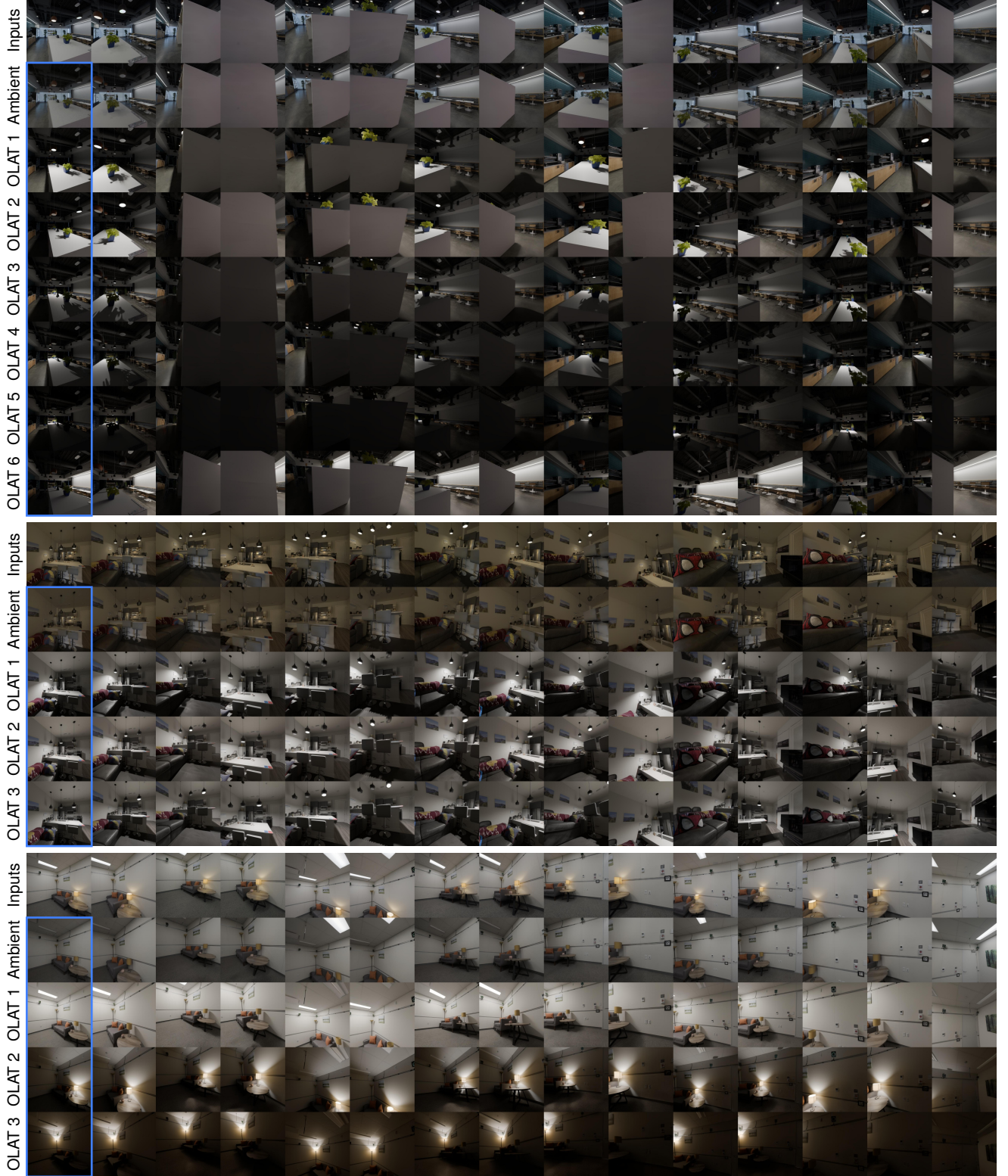
Figure 13. **Additional Multi-view Lighting Harmonization Results.** We use the first input image to decompose the scene lighting into the ambient lighting and multiple one-light-at-a-time (OLAT) light components (highlighted in blue for each scene). We then propagate these lighting components consistently across all views using the multi-view input images (top row). Source: VR-NeRF [86].

relighting results. We include screen recording videos in the supplement to show our interactive relighting with 3DGS.

8. Additional Results and Comparisons

8.1. Single-Image Lighting Editing

We show additional editing results in Figures 11 and 12. In Figure 11, we show results of switching lights off and on, with various light colors. The baseline comparisons to FLUX.1 Kontext [8], Qwen-Image [82] and Gemini 2.5 Flash Image a.k.a. Nano Banana show prompt-based editing results trying to match our results. Figure 12 shows more advanced editing results with two editable light sources and a mixture of out-of-distribution light shapes and novel colors. In the supplemental material, we include a video demonstrating the decomposition of seven distinct light sources, and recombining them in many different ways.

8.2. Multi-view Lighting Harmonization

Figure 13 shows additional multi-view lighting harmonization results. Note how our single-view model LuxRemix-SV can recover individual OLAT lights with plausible shadows for the plant in the top scene. Our multi-view harmonization model LuxRemix-MV then propagates these single-view lighting decompositions consistently across views. For the middle scene, our model can cleanly disentangle the three hanging lights. And for the bottom scene, our model can switch on the standing lamp in the center of the scene (see "OLAT 3"). Note also that the ambient lighting in this scene correctly includes the out-of-view ceiling lights.

8.3. Failure cases

Figure 14 shows some failure cases of our single-image light decomposition and editing method. In our synthetic training data, the shape of the light spread is biased towards conical shapes, which does not always match real-world lights. In multi-light images, our lighting decomposition sometimes fails to adhere to the provided light mask. Using a different random seed can result in different OLAT decompositions.
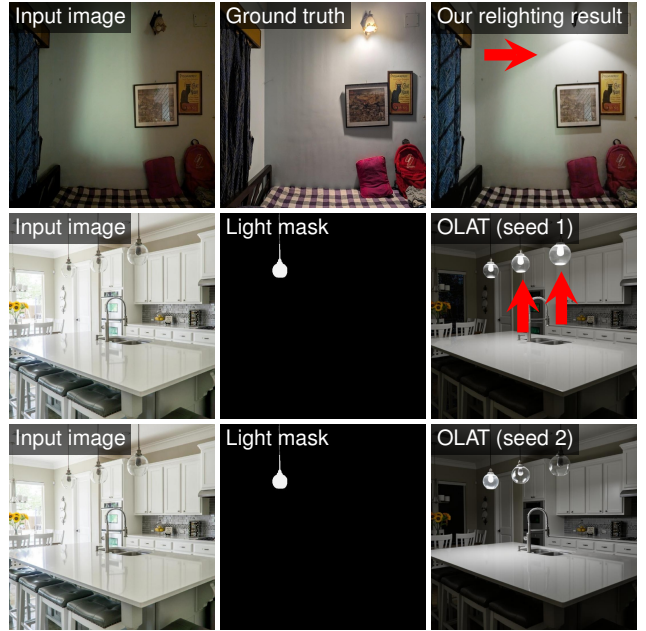

Figure 14. **Failure cases.** We show some failure cases of our light decomposition and editing method. **Top:** the shape of the light spread is biased towards conical shapes, which does not align with the ground truth. **Middle:** in multi-light images, our lighting decomposition sometimes fails to adhere to the provided light mask. **Bottom:** starting from a different random seed can result in different OLAT decompositions.