Decoupled Illumination Priors for Spatially Controllable Multi-View Indoor Scene Relighting

Chenjian Gao[1], Linning Xu[1†], and Tianfan Xue[1,2,3†]

[1] Multimedia Laboratory, The Chinese University of Hong Kong
[2] Shanghai AI Laboratory
[3] CPII under InnoHK
{gc025, tfxue}@ie.cuhk.edu.hk linningxu@link.cuhk.edu.hk
[†] Corresponding Authors

Abstract. Indoor scene relighting demands photorealism, precise spatial control, and strict multi-view consistency. While diffusion-based image editing models enable semantic lighting manipulation via text prompts, enforcing exact 3D light placement often disrupts their generative priors. We propose Lume-Palette, a progressive framework that leverages semantic lighting priors for spatially controllable multi-view indoor relighting. The approach decouples relighting into two stages: (1) illumination distillation, which extracts canonical illumination palettes from a pretrained diffusion model to preserve realistic material–light interactions, and (2) illumination casting, which explicitly maps target spatial lighting conditions defined from coarse 3D geometry. To efficiently handle dense multi-view and multi-modal inputs, we introduce an asymmetric multi-view conditioning strategy that selectively injects essential spatial context. Experiments on diverse synthetic scenes and real-world scenes demonstrate that Lume-Palette produces photorealistic, spatially controllable, and multi-view consistent relighting results.

Keywords: Indoor Scene Relighting · Image-based Illumination

1 Introduction

Indoor scene relighting stands as a pivotal challenge in computer vision and computer graphics, with applications ranging from immersive augmented reality (AR) and 3D virtual photography to interior design [10, 21, 31, 63]. In these interactive environments, users explore the space from varying viewpoints, necessitating illumination that is not only realistic but also consistent across different viewpoints. Unlike object-centric relighting [22, 35, 66], scene-level relighting presents a unique set of complexities, as the geometry is no longer isolated and observed from limited angles. It requires maintaining coherent global illumination and shadow propagation amidst intricate occlusion relationships and diverse material properties. Due to these complexities, while traditional inverse rendering methods [2, 4] strive for physical interpretability, they often struggle to accurately invert material interactions, yielding results that lack photorealism.

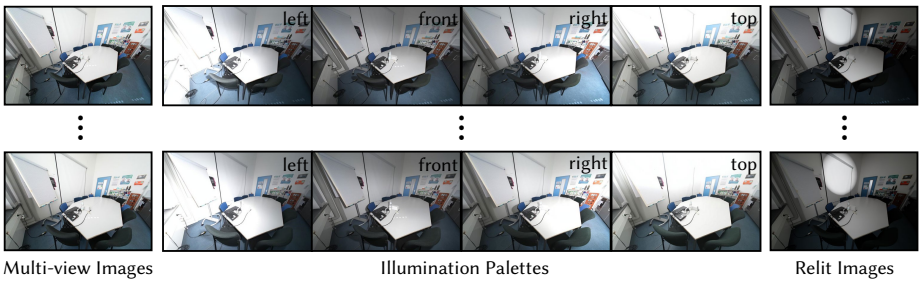

Fig. 1: Multi-view indoor scene relighting with Lume-Palette. Our progressive framework first extracts canonical "illumination palettes" from multi-view source images to capture material-light interactions. These palettes subsequently guide the synthesis of photorealistic, multi-view consistent relit images under new target illuminations.

Recently, diffusion-based image editing models [3, 26, 62] have excelled at manipulating lighting effects via text prompts, yielding photorealistic results. However, text-driven control lacks explicit spatial control over light sources and struggles to preserve consistency when the viewpoint shifts. To achieve spatial lighting control and multi-view consistency, recent approaches [30, 35] attempt to fine-tune diffusion models to accept spatial condition maps. Forcing models to accommodate these invasive spatial constraints from scratch by relying on synthetic data can disturb their generative priors. Consequently, while these approaches achieve spatial alignment, they tend to produce synthetic artifacts, failing to balance precise illumination controllability with natural realism.

Our goal is to achieve explicit spatial control for multi-view consistent relighting without losing the diffusion model's native lighting prior. Drawing inspiration from photometric stereo [11, 16], which reveals material properties by capturing a scene under a set of canonical lighting directions, we propose **Lume-Palette**, a progressive framework that decouples the relighting process into illumination distillation and illumination casting. Since pre-trained diffusion models are not natively trained to condition on explicit spatial lighting maps, forcing them to condition on this foreign modality requires learning a complex geometry-to-appearance mapping from scratch. This heavy adaptation inevitably disrupts the model's well-tuned generative priors, degrading photorealism. To bypass this, we first distill canonical "illumination palettes" via text prompts—a modality the model is inherently aligned with—preserving authentic material responses. These high-quality references then allow the casting stage to focus solely on spatial distribution guided by the geometric conditions. Directionally relit images have also been used by Poirier-Ginter *et al.* [48] as priors for per-scene relightable radiance-field optimization. In Lume-Palette, these priors are instead distilled into illumination palettes and cast through a feed-forward diffusion relighting engine, enabling spatially controllable 3D lighting across multiple views.

In the illumination distillation stage, we distill the generative prior of a pre-trained diffusion model into a photorealistic **illumination palette**—a set of

reference images capturing the scene under canonical lighting directions (e.g., left, right, front, and top), as illustrated in Fig. 1. By fine-tuning the model on real-world photographs to align with fixed lighting descriptive prompts, we induce the synthesis of these canonical lighting patterns. This semantic-driven distillation avoids invasive conditioning, preserving the rich generative prior. By revealing how the scene's surfaces respond to canonical illumination, they serve as a rich generative prior to guide the synthesis of highly photorealistic results.

In the subsequent illumination casting stage, we leverage the illumination palette to synthesize target illumination. Specifically, we allow users to freely place virtual light sources within a reconstructed coarse 3D mesh, and then render the scene's response to represent the target lighting conditions. Unlike explicit source parameterization [40], our adopted receiver-centric lighting condition naturally accommodates off-screen lights while maintaining cross-view 3D consistency. To synthesize the final relit scene, the network must jointly process the multi-view source images alongside their corresponding lighting conditions and the previously distilled illumination palettes. However, simply concatenating all these dense, multi-modal conditions across all viewpoints is computationally intractable. To overcome this, we design an asymmetric multi-view conditioning scheme that avoids feeding dense conditions to every view. For each prediction, only the active view receives the full source image, lighting condition, and illumination palette, while the other views are represented by lightweight noisy latents as spatial anchors, enabling scalable joint processing for multi-view relighting.

Our approach achieves photorealistic, spatially controllable, and multi-view consistent relighting results by leveraging the generative priors of powerful image generative models via illumination distillation and illumination casting. Experiments on synthetic and real-world scenes show that Lume-Palette produces visually plausible relighting results that follow user-specified spatial lighting conditions while maintaining cross-view coherence.

2 Related Works

2.1 Inverse Rendering and Physics-Based Relighting

Inverse rendering traditionally decomposes images into intrinsic components via optimization with hand-crafted priors [2, 4, 17, 37]. Deep learning advanced this by directly predicting SVBRDFs and lighting [12, 27, 50, 54], often utilizing vision transformers for dense estimation [61, 74]. To ensure physical plausibility, differentiable rendering [1,9,28,43] and neural implicit representations [6,7,56,68] are widely adopted to jointly optimize scene parameters and complex reflectance fields. However, explicit inversion remains computationally expensive and highly sensitive to geometric inaccuracies, often producing unnatural artifacts under complex global illumination. Instead of relying on explicit decomposition, our approach uses illumination palettes as intermediate appearance references that provide material-dependent lighting cues for spatially controllable relighting.

2.2 Generative and Diffusion-Based Relighting

While GANs laid the foundation for neural relighting [44, 57, 60, 73], diffusion models dominate image editing [51, 53, 69]. Text prompts manipulate lighting semantics [8, 19, 41], but lack spatial precision; recent methods therefore add spatial maps [23,67], light-transport constraints [66,70], or user guidance [10,30, 40]. Other systems explore lighting-aware editing or control from complementary perspectives, including intrinsic-space editing [39], unified geometry–illumination diffusion for controllable video [33], and practical light-control interfaces [13]. These approaches highlight explicit control, but adapting diffusion models to unfamiliar spatial modalities can disturb the pre-trained generative prior.

To preserve generative relighting priors, recent work distills lighting knowledge by isolating illumination semantics [25], extracting environment maps via LoRA [47], or distilling ray tracing into video models [5]. Directionally relit images have also been used as intermediate priors: Poirier-Ginter *et al.* [48] fine-tune a 2D diffusion model on MIT Multi-Illumination with global light-direction conditioning, and use the generated relit images for per-scene radiance-field optimization; ROGR [58] uses generative relighting for relightable 3D objects. LumePalette follows this line of using generated relighting results as illumination priors, but casts canonical relit images through receiver-centric spatial lighting conditions for feed-forward multi-view indoor relighting with user-placed 3D lights, without reconstructing a relightable radiance field per scene.

Rendered cues are another established control signal: DiffusionRenderer [29] uses G-buffers, LightIt [25] uses shading maps, and DiFaReli [49], Careaga and Aksoy [9], and Hybrelighter [72] use face, mesh, or reconstructed-scene renderings. We also render a receiver-centric lighting cue, but use it as a spatial layout rather than the sole appearance driver: it provides 3D-consistent guidance for the target illumination distribution, while the palette supplies view-specific material-response cues for the casting network.

2.3 Multi-View Consistent Relighting

Multi-view relighting must preserve consistency despite occlusions and viewpoint changes. Warping [45, 46] is brittle, and neural fields [7, 15, 52, 71] require per-scene optimization. Generative methods use video attention [18,63], joint denoising [14, 36, 55], propagation, including GS-Light [64], or harmonization [30, 59]. LightSwitch [35] avoids all-pair attention via mini-batch view shuffling for object relighting. While it samples different view subsets with similar per-view conditioning, our setting involves much denser conditions for each scene view. We therefore use asymmetric conditioning: only the active view receives these dense conditions, while inactive views are kept as lightweight latent anchors to provide cross-view cues through the shared latent state.

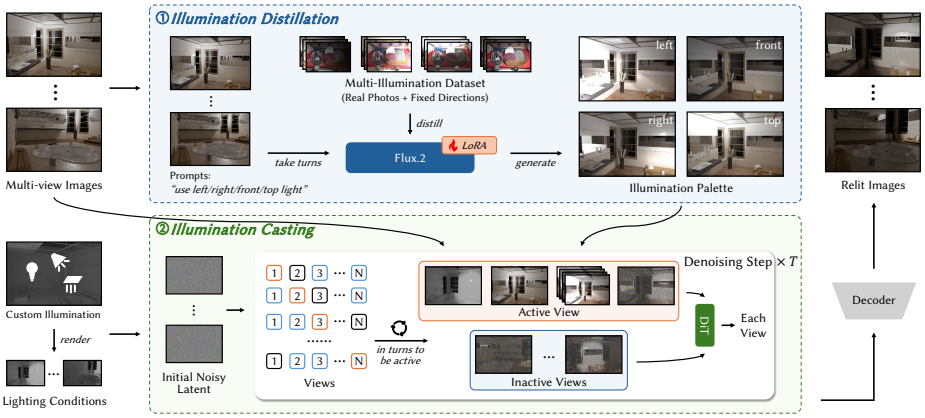

Fig. 2: Overview of Lume-Palette. Our framework decouples indoor relighting into two sequential stages. *Illumination Distillation* fine-tunes a diffusion model to extract canonical illumination palettes, capturing intrinsic material responses. *Illumination Casting* synthesizes the target illumination by combining these palettes with source images and user-defined spatial lighting conditions. To ensure scalable multi-view consistency, this stage employs an asymmetric multi-view conditioning scheme.

3 Method

3.1 Overview

We present **Lume-Palette**, a framework designed to synthesize photorealistic and spatially consistent indoor relighting results. Formally, given a set of N multi-view source images $\mathcal{I} = \{I_1, \ldots, I_N\}$ captured in the same scene under consistent lighting conditions, and a globally consistent user-defined spatial lighting condition represented as view-specific spatial lighting maps $\mathcal{L} = \{L_1, \ldots, L_N\}$, our framework predicts multi-view consistent relit images $\mathcal{R} = \{R_1, \ldots, R_N\}$. Here, each R_i is the relit version of the corresponding source image I_i, faithfully reflecting the target illumination. A naive approach to solve this problem is a single-stage pipeline that directly conditions an image editing model on explicit spatial lighting conditions. However, this end-to-end paradigm faces two critical challenges. First, in the real world, it is very hard to capture multi-view consistent and spatially varying relight pairs with precise spatial lighting annotations. Consequently, models must rely heavily on synthetic datasets, which inevitably compromises the photorealistic generative priors of the pretrained model. Second, it is hard to train a single network to simultaneously erase complex lighting in the original input and render complex target illumination, as it requires the network to learn intrinsic image decomposition, lighting estimation, and geometry estimation jointly without intermediate supervision.

To address both challenges, our approach decouples the relighting process into two sequential stages: **illumination distillation** and **illumination casting** (Fig. 2). The distillation stage learns a generative prior for canonical lighting

directions from real-world photographs and their corresponding directional annotations, which are simple textual descriptions indicating the light source positions (e.g., 'left", 'right"). By distilling this prior into specialized Low-Rank Adaptation (LoRA) modules, the model can generate images with canonical directional illumination for a given source image, serving as an intermediate illumination palette for the following lighting stage. The casting stage then synthesizes relighting results under specific lighting conditions. This is achieved by combining the original multi-view images and these reference palettes, guided by spatial lighting conditions. The lighting condition includes multi-view shading images generated by a rendering engine; it encourages multi-view consistent relighting through shared 3D-derived conditions. Consequently, our approach preserves the photorealistic prior by decoupling the native generative lighting prior from the enforcement of specific lighting control. Furthermore, by normalizing the complex source illumination with basic canonical lighting, the intermediate palettes alleviate the burden of erasing the original lighting during the casting stage.

3.2 Illumination Distillation

As previous diffusion-based relighting methods show, diffusion models have an implicit prior regarding light-material interactions. However, since this prior is only implicit, it does not support flexible control and multi-view consistent relighting. Therefore, we first distill this prior into an explicit representation of light-material interactions, named the **illumination palette**. This palette is defined as the four lighting results of an input image lit from 4 canonical lighting directions $d \in \{\text{'left'}, \text{'right'}, \text{'front'}, \text{'top'}\}$. To achieve this, we train specialized Low-Rank Adaptation (LoRA) [20] modules of a pretrained diffusion model to generate this palette. During inference, these trained LoRA modules are applied independently to each individual image I_i within the multi-view source set $\mathcal{I} = \{I_1, \ldots, I_N\}$. Guided by text prompts, this process generates a view-specific illumination palette $\mathcal{B}_i = \{B_i^{left}, B_i^{right}, B_i^{front}, B_i^{top}\}$ for every viewpoint i, consisting of reference images capturing the scene's authentic shading response to the distinct canonical directions d.

To supervise this distillation, we leverage the MIT Multi-Illumination [42] dataset, which features 1,000 indoor scenes. For each scene, the camera remains strictly fixed while capturing photographs under 25 distinct lighting directions, ensuring perfectly aligned multiple illumination image groups. To construct our canonical illumination palette, we train four independent LoRA modules corresponding to target directions $d \in \{\text{'left'}, \text{'right'}, \text{'front'}, \text{'top'}\}$. These four LoRA modules are shared across all scenes. During a training iteration of each LoRA (like $d =$ 'left'), we select that lighting direction as the ground-truth target I_d, and randomly sample one another lighting out of the 24 available lighting from the same scene in the dataset as the input image I_{src}. By forcing the model to map these diversely lit inputs to the same fixed target, we ensure the network learns to override arbitrary initial illumination with the desired canonical lighting. Formally, since our base model [26] operates on a flow matching framework [34], we optimize the LoRA module θ by minimizing the velocity matching

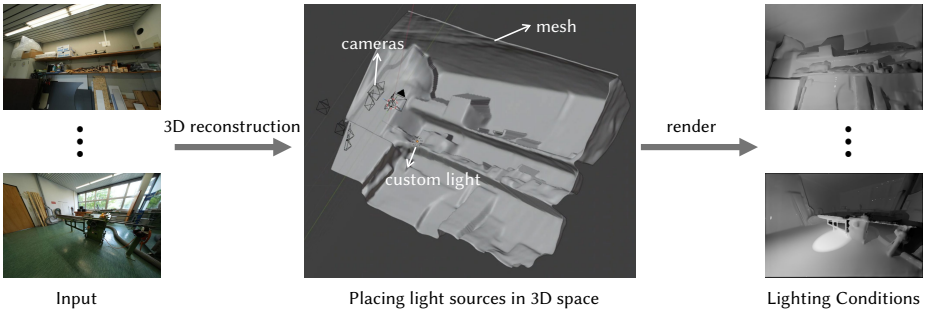

Fig. 3: Generation of spatial lighting conditions. First, a coarse 3D mesh is reconstructed from multi-view input images [32]. Users can then explicitly place custom light sources within this 3D space. Finally, the scene is rendered to produce view-specific lighting conditions for the illumination casting stage.

objective. Let $\mathbf{z}_0 = \mathcal{E}(I_d)$ be the latent representation of the target image encoded by a pretrained VAE, and $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ be sampled Gaussian noise. We construct the time-dependent noisy state as $\mathbf{z}_t = (1-t)\mathbf{z}_0 + t\mathbf{z}_1$. The optimization objective is defined as:

$$\min_{\theta} \left\| v_\theta(\mathbf{z}_t, t, \mathcal{E}(I_{src}), \text{“use } d \text{ light”}) - (\mathbf{z}_1 - \mathbf{z}_0) \right\|^2$$

where v_θ represents the velocity prediction network equipped with the LoRA module θ, and $t \in [0, 1]$ is the time step.

At inference, given a set of multi-view source images $\mathcal{I} = \{I_1, \ldots, I_N\}$, we apply the four fine-tuned LoRA modules to each viewpoint independently to predict the scene's appearance under the canonical directions. The resulting view-specific collection of illumination palettes, denoted as $\mathcal{B} = \{B_i^{left}, B_i^{right}, B_i^{front}, B_i^{top}\}_{i=1}^N$, serves as lighting references that capture how scene materials respond to canonical illumination, such as specularities and shading falloff, ready for the subsequent illumination casting stage.

3.3 Illumination Casting

While the illumination palettes provide robust material cues, they lack precise light control due to the ambiguity of text prompts. The goal of the illumination casting stage is to combine these palettes with explicit spatial lighting conditions to synthesize consistent multi-view relighting under user-defined light sources. Formally, the input to this stage includes the multi-view source images $\mathcal{I} = \{I_1, \ldots, I_N\}$, the distilled illumination palettes $\mathcal{B} = \{B_i^{left}, B_i^{right}, B_i^{front}, B_i^{top}\}_{i=1}^N$ estimated in the previous illumination distillation, and the target spatial lighting conditions $\mathcal{L} = \{L_1, \ldots, L_N\}$. By conditioning the model on these combined inputs, this stage outputs the final synthesized relit images $\mathcal{R} = \{R_1, \ldots, R_N\}$. As detailed below, $\mathcal{L}$ expresses the specific distribution of illumination by allowing users to freely place various light sources within the 3D space, providing more precise control than abstract text prompts.

Spatial Lighting Condition To explicitly define the target spatial illumination, we construct the lighting condition images in the rendering engine using the estimated scene 3D geometry. First, given the multi-view source images $\mathcal{I} = \{I_1, \ldots, I_N\}$, we utilize Depth Anything 3 [32] to extract their corresponding depth maps and reconstruct a coarse 3D mesh of the scene. Second, to support flexible control of illumination, we allow users to arbitrarily place and configure various light sources (such as point, spot, or area lights) within this reconstructed 3D space. Details of the user interface are provided in the supplementary material. Given the estimated geometry and customized lighting, we assign a uniform, untextured white material to the 3D mesh and render it from the N target viewpoints under these user-defined lights (this is sometimes called a "white 3D model"), yielding the view-specific spatial lighting conditions $\mathcal{L} = \{L_1, \ldots, L_N\}$.

The main idea of this design is to use multi-view 2D rendering of an untextured surface, rather than the explicit 3D location or types of lighting sources, as the lighting condition. This offers two advantages. First, it supports arbitrary choices of lighting, including off-screen or complex lighting setups without requiring visible light sources. Second, because these conditions $\mathcal{L}$ are derived from a shared 3D representation, this approach provides 3D-consistent lighting conditions across views.

Asymmetric Multi-View Conditioning Achieving multi-view consistency implies modeling the conditional joint probability distribution $p(\mathcal{R} \mid \mathcal{I}, \mathcal{L}, \mathcal{B})$. The most direct approach would be to simultaneously condition the generation on the complete collection of inputs across all N viewpoints within a single network pass, comprising all the condition images per view: the source image I_i, the spatial lighting condition L_i, and the four directional references $\{B_i^{left}, B_i^{right}, B_i^{front}, B_i^{top}\}$. However, concatenating such an extensive set of condition images leads to an unmanageable explosion in input dimensionality, making direct joint modeling computationally intractable.

To address this bottleneck while preserving spatial awareness, we formulate our approach as an asymmetric multi-view conditioning scheme. Since processing the full set of N views is intractable, we sample a group of V views ($V < N$, e.g., $N = 8, V = 4$) to serve as the local context window. Within this focused context window, we select one view index m as an active view and other views as inactive views. The model is trained to predict the velocity field exclusively for this active view. The model uses the complete set of that active view as dense guiding signals: the source image of that active view as $I_m \in \mathcal{I}$, the target spatial lighting condition $L_m \in \mathcal{L}$, and the illumination palettes $\{B_m^{left}, B_m^{right}, B_m^{front}, B_m^{top}\}$ distilled in the first stage. Since the model does not predict the velocity field for these inactive views, feeding them dense visual conditions is unnecessary. Instead, to maintain cross-view spatial awareness between the active and inactive views, we solely incorporate their concurrent noisy latents $\{\mathbf{z}_t^k\}_{k \neq m}$ to serve as lightweight spatial anchors. These inactive views are necessary here because they provide the global 3D context needed to ensure 3D consistency during the

denoising process. Unlike the dense conditions that require encoding multiple high-dimensional reference images, these noisy latents are merely the intermediate noisy states of the relit images, making them computationally trivial. By concatenating these asymmetric signals, the final input stream encompasses the active view's dense conditions and the noisy latents of inactive views. By omitting the dense condition encodings for the $V-1$ inactive views, this asymmetric split reduces the condition sequences from $6V$ to $V+5$. For example, when $V=4$, DiT inputs decrease from 24 to 9, substantially lowering the conditional memory footprint compared to a symmetric baseline.

To differentiate these varied inputs within the network, we leverage the 3D positional encoding native to the pre-trained image editing model, utilizing its first dimension to identify the condition type, thereby enabling the network to seamlessly integrate lighting references with global spatial awareness.

Training Objective At each training iteration, we first sample a scene alongside a local subset of V viewpoints. To formulate the training sample for relighting, we randomly select two distinct illuminations: a source illumination to provide the input images $\{I_i\}_{i=1}^{V}$, and a target illumination to render the ground-truth relit images $\{R_i\}_{i=1}^{V}$. Following our asymmetric conditioning scheme, we then uniformly sample an active view index $m \in \{1, \dots, V\}$.

The model is optimized via a flow matching [34] objective. During the forward process, the ground-truth relit images $\{R_i\}_{i=1}^{V}$ are mapped to their latent representations $\{\mathbf{z}_0^i\}_{i=1}^{V}$ via a pre-trained VAE [24]. We then sample Gaussian noise $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and construct the time-dependent noisy states via linear interpolation: $\mathbf{z}_t = (1-t)\mathbf{z}_0 + t\mathbf{z}_1$. Conditioned on the active view's dense guiding signals, specifically the source image I_m, the spatial lighting condition L_m, and the illumination palettes $\{B_m\}$ separately encoded via $\mathcal{E}(\cdot)$, the network v_θ predicts the constant velocity field $u_t = \mathbf{z}_1^m - \mathbf{z}_0^m$ for the active view. The optimization objective is defined as:

$$\mathcal{L} = \left\| v_\theta \left(\mathbf{z}_t^m, t, \underbrace{\mathcal{E}(I_m), \mathcal{E}(L_m), \{\mathcal{E}(B_m)\}}_{\text{active guiding signals}}, \underbrace{\{\mathbf{z}_t^k\}_{k \neq m}}_{\text{inactive views}} \right) - (\mathbf{z}_1^m - \mathbf{z}_0^m) \right\|^2 \tag{1}$$

Consistent with this design, gradients are back-propagated solely through the active target's prediction path, treating the inactive latents $\{\mathbf{z}_t^k\}_{k \neq m}$ purely as gradient-free spatial references.

Multi-View Aware Inference During inference, we concurrently synthesize the full set of N viewpoints. We initialize the process by drawing random Gaussian noise for the entire scene, denoted as $\mathcal{Z}_1 = \{\mathbf{z}_1^1, \dots, \mathbf{z}_1^N\}$. At each denoising step t, we compute the velocity fields for all N views. To achieve this, we formulate the prediction for each view $m \in \{1, \dots, N\}$ by setting it as the active view. For this active view m, we prepare its dense guiding signals, specifically the

Table 1: Quantitative comparison against baselines on synthetic scenes, including image fidelity and geometry-aware multi-view consistency.

	RGB↔X [67]	IC-Light [70]	LumiNet [63]	ScribbleLight [10]	Ours
PSNR ↑	10.497	13.982	16.593	12.591	**24.453**
SSIM ↑	0.525	0.622	0.607	0.504	**0.872**
LPIPS ↓	0.394	0.281	0.399	0.431	**0.126**
MV MSE ↓	0.0227	0.0184	0.0159	0.0218	**0.0090**
MV LPIPS ↓	0.176	0.143	0.195	0.208	**0.115**

encoded source image $\mathcal{E}(I_m)$, spatial lighting condition $\mathcal{E}(L_m)$, and illumination palettes $\{\mathcal{E}(B_m)\}$, alongside a context window of $V-1$ inactive spatial anchors sampled from the remaining noisy state $\mathcal{Z}_t \setminus \{\mathbf{z}_t^m\}$. Conditioned on these combined inputs, the network v_θ predicts the velocity field u_t^m exclusively for view m. After computing this independent forward pass for all N views, we aggregate the resulting velocity fields $\{u_t^1, \ldots, u_t^N\}$ to form the joint velocity for the entire scene. The ODE solver then uses this joint velocity to update the global scene state $\mathcal{Z}_t$ a step backward toward $t = 0$. Once the integration reaches $t = 0$, the latents are decoded via the VAE to produce the final multi-view relit output $\mathcal{R}$.

4 Experiments

4.1 Implementation Details

Both the illumination distillation and illumination casting stages are built upon Flux.2 Klein 9B [26] to leverage its powerful generative priors. The models in both stages are optimized using a flow matching objective [34] with AdamW [38], and are trained for 1 epoch with a learning rate of 1×10^{-4}. In the illumination distillation phase, we train LoRA modules with a rank of 32 and a batch size of 8. In the illumination casting stage, we use a higher LoRA rank of 128 and a batch size of 32 to handle the more complex spatial conditions. We train the casting stage on paired synthetic multi-view relighting data rendered under diverse lighting configurations, using 480 scenes for training and 10 held-out scenes for evaluation. During casting-stage training, we randomly sample a local subset of $V = 4$ viewpoints to form the context window.

4.2 Comparison with Baseline Methods

We evaluate Lume-Palette against recent open-source relighting methods, including RGB↔X [67], IC-Light [70], LumiNet [63], and ScribbleLight [10]. We first evaluate relighting accuracy on diverse synthetic scenes. We report image-fidelity metrics, including PSNR, SSIM, and LPIPS. In addition, we evaluate geometry-aware multi-view consistency by warping each generated relit view to its adjacent view using depth and camera poses, and computing MSE and LPIPS

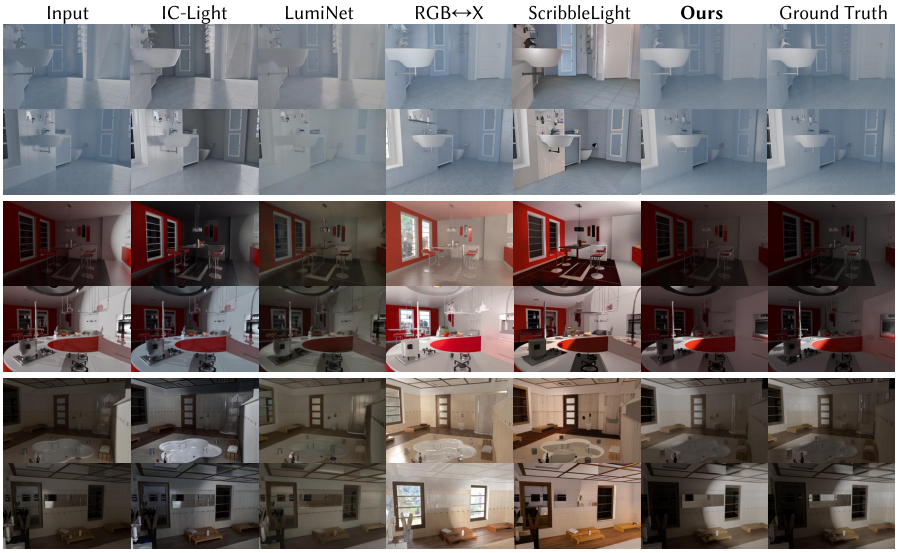

Fig. 4: Qualitative comparison against baselines on synthetic scenes. Our method achieves precise spatial control, realistic appearance, and multi-view consistency.

Table 2: Real-scene user study results using Bradley–Terry scores. Higher scores indicate stronger user preference.

Method	Realism ↑	Lighting adherence ↑	MV consistency ↑
RGB↔X	-1.498	0.807	-0.038
IC-Light	1.760	-1.031	-0.038
LumiNet	-0.571	-1.981	-0.542
ScribbleLight	-1.742	-0.596	-0.895
Ours	**2.051**	**2.801**	**1.514**

against the corresponding generated adjacent-view result in the overlapping region. As shown in Table 1, Lume-Palette achieves the best results among the evaluated baselines under the adopted metrics. Qualitative comparisons in Fig. 4 further illustrate the behavior of different methods under diverse spatial lighting configurations. IC-Light often struggles to disentangle and remove the baked-in source illumination, while LumiNet shows limited ability to impose complex target illumination. Methods relying on explicit intrinsic decomposition, such as RGB↔X and ScribbleLight, are sensitive to albedo prediction errors and may introduce noticeable color shifts. In comparison, Lume-Palette produces spatially aligned relighting results with visually plausible appearance and improved cross-view coherence. It captures a range of lighting effects, from high-frequency spotlight shadows to softer illumination from rectangular area lights, while keeping the resulting shading patterns coherent across viewpoints.

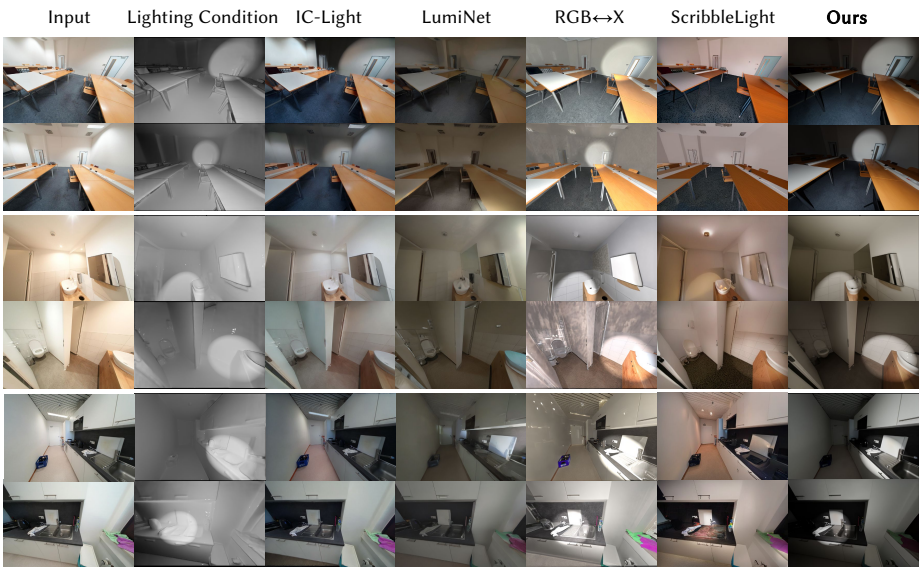

Fig. 5: Qualitative comparison against baselines on real scenes. Lume-Palette robustly synthesizes realistic shading that adheres to imposed spatial lighting conditions.

We further evaluate Lume-Palette on real scenes from ScanNet++ [65]. As shown in Fig. 5, Lume-Palette can synthesize plausible shading that follows the imposed spatial lighting conditions in complex real scenes. The results show spatially varying illumination effects rather than global tone changes, with brighter and darker regions appearing in accordance with the provided lighting maps. Since ground-truth relighting is unavailable for real scenes, we conduct a user study with 25 participants to assess perceptual quality. In each trial, participants compare two anonymized method outputs according to realism, lighting adherence to the reference diffuse lighting condition, and multi-view consistency. We collected 150 valid pairwise preference judgments spanning different scenes, method pairs, and evaluation criteria. The preferences are aggregated using a Bradley–Terry model, and the resulting scores are reported in Table 2. Our method is consistently preferred across all three dimensions.

4.3 Ablation Studies

We conduct ablation studies to validate the core components of our proposed framework.

Effectiveness of Illumination Distillation. To validate the effectiveness of illumination distillation, we train a variant (w/o Illumination Palette) that directly conditions the image editing model on explicit spatial maps without the intermediate illumination palettes. As shown in Table 3, this direct enforcement degrades the model's generative prior, leading to a noticeable drop in PSNR and

Fig. 6: Qualitative ablation of the illumination palette and multi-view awareness. Removing the illumination palette or multi-view awareness degrades performance.

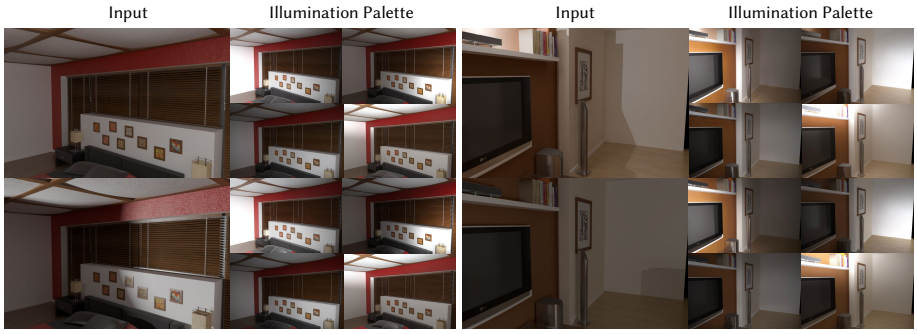

Fig. 7: Robustness of the illumination palette to varying input illumination. Changes in input light sources and shadows have almost no effect on the extracted palette.

a worse LPIPS score compared to the full model. Qualitatively, Fig. 6 shows that removing this component leads to noticeable synthetic artifacts and a loss of intrinsic material properties. Furthermore, Fig. 7 demonstrates the robustness of the Illumination Palette, showing that it can consistently extract canonical illumination regardless of varying initial input lighting. This stable representation provides reliable appearance priors for the subsequent casting stage.

Impact of Asymmetric Multi-View Conditioning. By removing the inactive spatial anchors, we train a single-view variant (w/o Multi-View Aware). As demonstrated quantitatively and qualitatively, the overall relighting quality of this baseline is inferior to our full Lume-Palette. Fundamentally, multi-view inputs provide cross-view spatial anchors and parallax cues, helping the model maintain consistent light-surface interactions across viewpoints. Deprived of this multi-view spatial awareness, the model generates noticeably flatter shading and less accurate light-surface interactions. This confirms that our asymmetric conditioning scheme is essential for deeply understanding scene materials, thereby enhancing the overall photorealism of the final relighting results more faithfully.

Table 3: Quantitative ablation of illumination palettes and multi-view awareness.

Model Variant	PSNR ↑	SSIM ↑	LPIPS ↓
w/o Multi-View Aware & Illumination Palette	21.973	0.834	0.140
w/o Multi-View Aware	23.618	0.864	0.131
w/o Illumination Palette	22.748	0.854	0.128
Full Lume-Palette	**24.453**	**0.872**	**0.126**

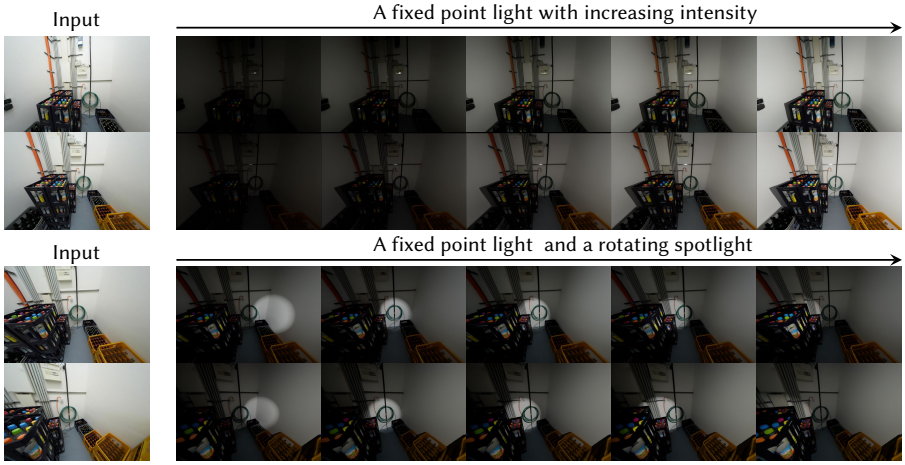

Fig. 8: Continuous spatial and dynamic controllability of light intensity and direction. The generated illumination, including specular highlights and cast shadows, updates naturally and remains firmly attached to the underlying geometry.

4.4 Spatial and Dynamic Controllability

A core advantage of Lume-Palette is its fine-grained lighting control. As illustrated in Fig. 8, our framework supports continuous spatial and dynamic controllability beyond discrete lighting presets. Users can adjust the intensity of a fixed point light or rotate a spotlight within the 3D space, and the relit appearance changes progressively with the edited light. The generated illumination, including specular highlights, cast shadows, and local brightness falloff, updates naturally and remains attached to the underlying geometry.

Beyond single-light editing, Lume-Palette can handle more complex lighting layouts. Fig. 9 shows results under diverse setups, including point lights at different positions, multiple lights, and a rectangular area light. These examples demonstrate that the receiver-centric lighting representation can express different source types and spatial arrangements within the same framework. Multiple lights jointly affect the scene with spatially varying contributions, while area lighting produces broader and softer shading patterns than point lighting.

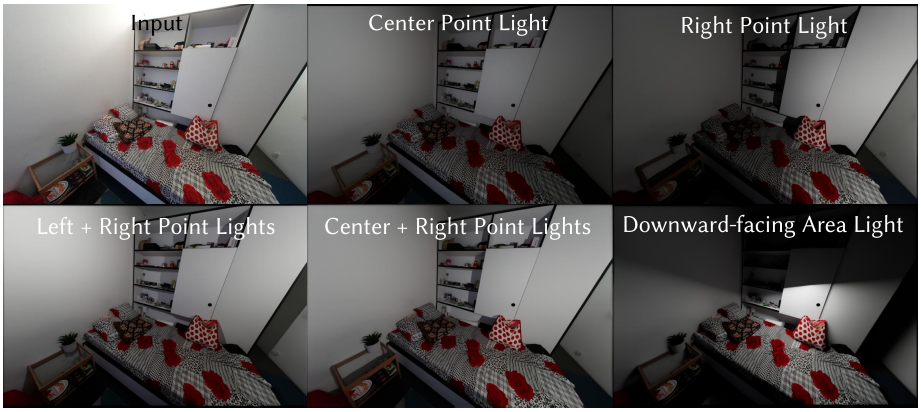

Fig. 9: Lume-Palette supports flexible lighting configurations, including point lights at different positions, multiple simultaneous lights, and rectangular area lights.

5 Conclusion and Limitations

We presented Lume-Palette, a progressive framework for spatially controllable multi-view indoor scene relighting. By decoupling relighting into illumination distillation and illumination casting, our method leverages canonical illumination palettes to preserve realistic material–light interactions, while using receiver-centric spatial lighting conditions to support explicit 3D light placement. To enable efficient joint multi-view generation, we further introduced an asymmetric conditioning strategy that provides dense guidance only to the active view while maintaining cross-view spatial anchors through inactive latents. Experiments on synthetic and real-world scenes show that Lume-Palette produces photorealistic, spatially aligned, and multi-view consistent relighting results, comparing favorably with baseline methods under the adopted metrics.

Despite these encouraging results, our method still has several limitations. Since receiver-centric lighting conditions are rendered from coarse reconstructed geometry, inaccurate geometry may degrade relighting quality, especially around thin structures, reflective or transparent materials, and scenes with complex indirect illumination. We hope this limitation motivates future work on more robust geometry-aware conditioning for controllable multi-view relighting.

Acknowledgements

This study was supported in part by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd., a CUHK-led InnoCentre under the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government. This study was supported by CUHK-CUHK(SZ)-GDSTC Joint Collaboration Fund No. 2025A0505000053. We would like to thank the reviewers for their insightful comments.

References

1. Azinovic, D., Li, T.M., Kaplanyan, A., Nießner, M.: Inverse path tracing for joint material and lighting estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2447–2456 (2019)
2. Barron, J.T., Malik, J.: Shape, illumination, and reflectance from shading. IEEE Trans. Pattern Anal. Mach. Intell. **37**(8), 1670–1687 (2015). https://doi.org/10.1109/TPAMI.2014.2377712, https://doi.org/10.1109/TPAMI.2014.2377712
3. Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv e-prints pp. arXiv–2506 (2025)
4. Bell, M., Freeman, W.T.: Learning local evidence for shading and reflectance. In: Proceedings of the Eighth International Conference On Computer Vision (ICCV-01), Vancouver, British Columbia, Canada, July 7-14, 2001 - Volume 1. pp. 670–677. IEEE Computer Society (2001). https://doi.org/10.1109/ICCV.2001.10095, https://doi.ieeecomputersociety.org/10.1109/ICCV.2001.10095
5. Bharadwaj, S., Feng, H., Becherini, G., Fernandez Abrevaya, V., Black, M.J.: Genlit: Reformulating single-image relighting as video generation. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–12 (2025)
6. Bi, S., Xu, Z., Sunkavalli, K., Hašan, M., Hold-Geoffroy, Y., Kriegman, D., Ramamoorthi, R.: Deep reflectance volumes: Relightable reconstructions from multi-view photometric images. In: European Conference on Computer Vision. pp. 294–311. Springer (2020)
7. Boss, M., Braun, R., Jampani, V., Barron, J.T., Liu, C., Lensch, H.: Nerd: Neural reflectance decomposition from image collections. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12684–12694 (2021)
8. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)
9. Careaga, C., Aksoy, Y.: Physically controllable relighting of photographs. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–10 (2025)
10. Choi, J.M., Wang, A., Peers, P., Bhattad, A., Sengupta, R.: Scribblelight: Single image indoor relighting with scribbles. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5720–5731 (2025)
11. Debevec, P., Hawkins, T., Tchou, C., Duiker, H.P., Sarokin, W., Sagar, M.: Acquiring the reflectance field of a human face. In: Proceedings of the 27th annual conference on Computer graphics and interactive techniques. pp. 145–156 (2000)
12. Deschaintre, V., Aittala, M., Durand, F., Drettakis, G., Bousseau, A.: Single-image svbrdf capture with a rendering-aware deep network. ACM Transactions on Graphics (ToG) **37**(4), 1–15 (2018)
13. Erel, Y., Dabral, R., Golyanik, V., H. Bermano, A., Theobalt, C.: Practilight: Practical light control using foundational diffusion models. ACM Transactions on Graphics (TOG) **44**(6), 1–11 (2025)
14. Gao, C., Jiang, B., Li, X., Zhang, Y., Yu, Q.: Genesistex: Adapting image denoising diffusion to texture space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4620–4629 (2024)
15. Gao, J., Gu, C., Lin, Y., Li, Z., Zhu, H., Cao, X., Zhang, L., Yao, Y.: Relightable 3d gaussians: Realistic point cloud relighting with brdf decomposition and ray tracing. In: European Conference on Computer Vision. pp. 73–89. Springer (2024)

16. Goldman, D.B., Curless, B., Hertzmann, A., Seitz, S.M.: Shape and spatially-varying brdfs from photometric stereo. IEEE Trans. Pattern Anal. Mach. Intell. **32**(6), 1060–1071 (2010). https://doi.org/10.1109/TPAMI.2009.102, https://doi.org/10.1109/TPAMI.2009.102
17. Grosse, R., Johnson, M.K., Adelson, E.H., Freeman, W.T.: Ground truth dataset and baseline evaluations for intrinsic image algorithms. In: 2009 IEEE 12th International Conference on Computer Vision. pp. 2335–2342. Ieee (2009)
18. He, K., Liang, R., Munkberg, J., Hasselgren, J., Vijaykumar, N., Keller, A., Fidler, S., Gilitschenski, I., Gojcic, Z., Wang, Z.: Unirelight: Learning joint decomposition and synthesis for video relighting. arXiv preprint arXiv:2506.15673 (2025)
19. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross-attention control. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=_CDixzkzeyb
20. Hu, E.J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=nZeVKeeFYf9
21. Ji, G., Sawyer, A.O., Narasimhan, S.G.: Digital kitchen remodeling: editing and relighting intricate indoor scenes from a single panorama. arXiv preprint arXiv:2504.16086 (2025)
22. Jin, H., Li, Y., Luan, F., Xiangli, Y., Bi, S., Zhang, K., Xu, Z., Sun, J., Snavely, N.: Neural gaffer: Relighting any object via diffusion. Advances in Neural Information Processing Systems **37**, 141129–141152 (2024)
23. Kim, H., Jang, M., Yoon, W., Lee, J., Na, D., Woo, S.: Switchlight: Co-design of physics-driven architecture and pre-training framework for human portrait relighting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25096–25106 (2024)
24. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: Bengio, Y., LeCun, Y. (eds.) 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings (2014), http://arxiv.org/abs/1312.6114
25. Kocsis, P., Philip, J., Sunkavalli, K., Nießner, M., Hold-Geoffroy, Y.: Lightit: Illumination modeling and control for diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9359–9369 (2024)
26. Labs, B.F.: FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2 (2025)
27. Li, Z., Shafiei, M., Ramamoorthi, R., Sunkavalli, K., Chandraker, M.: Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2475–2484 (2020)
28. Li, Z., Shi, J., Bi, S., Zhu, R., Sunkavalli, K., Hašan, M., Xu, Z., Ramamoorthi, R., Chandraker, M.: Physically-based editing of indoor scene lighting from a single image. In: European Conference on Computer Vision. pp. 555–572. Springer (2022)
29. Liang, R., Gojcic, Z., Ling, H., Munkberg, J., Hasselgren, J., Lin, C.H., Gao, J., Keller, A., Vijaykumar, N., Fidler, S., et al.: Diffusion renderer: Neural inverse and forward rendering with video diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26069–26080 (2025)
30. Liang, R., Müller, N., Weber, E., Zauss, D., Vijaykumar, N., Kontschieder, P., Richardt, C.: Luxremix: Lighting decomposition and remixing for indoor scenes. arXiv preprint arXiv:2601.15283 (2026)

31. Lin, C.H., Huang, J.B., Li, Z., Dong, Z., Richardt, C., Li, T., Zollhöfer, M., Kopf, J., Wang, S., Kim, C.: IRIS: Inverse rendering of indoor scenes from low dynamic range images. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
32. Lin, H., Chen, S., Liew, J.H., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
33. Lin, Y., Chen, Y.W., Tsai, Y.H., Clark, R., Yang, M.H.: Illumicraft: Unified geometry and illumination diffusion for controllable video generation. Advances in Neural Information Processing Systems **38**, 27798–27829 (2026)
34. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
35. Litman, Y., De la Torre, F., Tulsiani, S.: Lightswitch: Multi-view relighting with material-guided diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27750–27759 (2025)
36. Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W.: Syncdreamer: Generating multiview-consistent images from a single-view image. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=MN3yH2ovHb
37. Lombardi, S., Simon, T., Saragih, J., Schwartz, G., Lehrmann, A., Sheikh, Y.: Neural volumes: Learning dynamic renderable volumes from images. arXiv preprint arXiv:1906.07751 (2019)
38. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=Bkg6RiCqY7
39. Lyu, L., Deschaintre, V., Hold-Geoffroy, Y., Hašan, M., Yoon, J.S., Leimkühler, T., Theobalt, C., Georgiev, I.: Intrinsicedit: Precise generative image manipulation in intrinsic space. ACM Transactions on Graphics (TOG) **44**(4), 1–13 (2025)
40. Magar, N., Hertz, A., Tabellion, E., Pritch, Y., Rav-Acha, A., Shamir, A., Hoshen, Y.: Lightlab: Controlling light sources in images with diffusion models. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–11 (2025)
41. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 4296–4304 (2024)
42. Murmann, L., Gharbi, M., Aittala, M., Durand, F.: A multi-illumination dataset of indoor object appearance. In: 2019 IEEE International Conference on Computer Vision (ICCV) (Oct 2019)
43. Nimier-David, M., Vicini, D., Zeltner, T., Jakob, W.: Mitsuba 2: A retargetable forward and inverse renderer. ACM Transactions on Graphics (ToG) **38**(6), 1–17 (2019)
44. Pandey, R., Orts-Escolano, S., Legendre, C., Haene, C., Bouaziz, S., Rhemann, C., Debevec, P.E., Fanello, S.R.: Total relighting: learning to relight portraits for background replacement. ACM Trans. Graph. **40**(4), 43–1 (2021)
45. Philip, J., Gharbi, M., Zhou, T., Efros, A.A., Drettakis, G.: Multi-view relighting using a geometry-aware network. ACM Trans. Graph. **38**(4), 78–1 (2019)
46. Philip, J., Morgenthaler, S., Gharbi, M., Drettakis, G.: Free-viewpoint indoor neural relighting from multi-view stereo. ACM Transactions on Graphics (TOG) **40**(5), 1–18 (2021)

47. Phongthawee, P., Chinchuthakun, W., Sinsunthithet, N., Jampani, V., Raj, A., Khungurn, P., Suwajanakorn, S.: Diffusionlight: Light probes for free by painting a chrome ball. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 98–108 (2024)
48. Poirier-Ginter, Y., Gauthier, A., Phillip, J., Lalonde, J.F., Drettakis, G.: A diffusion approach to radiance field relighting using multi-illumination synthesis. In: Computer Graphics Forum. vol. 43, p. e15147. Wiley Online Library (2024)
49. Ponglertnapakorn, P., Tritrong, N., Suwajanakorn, S.: Difareli: Diffusion face relighting. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22646–22657 (2023)
50. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10912–10922 (2021)
51. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
52. Rudnev, V., Elgharib, M., Smith, W., Liu, L., Golyanik, V., Theobalt, C.: Nerf for outdoor scene relighting. In: European Conference on Computer Vision. pp. 615–631. Springer (2022)
53. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 conference proceedings. pp. 1–10 (2022)
54. Sengupta, S., Gu, J., Kim, K., Liu, G., Jacobs, D.W., Kautz, J.: Neural inverse rendering of an indoor scene from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8598–8607 (2019)
55. Shi, Y., Wang, P., Ye, J., Mai, L., Li, K., Yang, X.: MVDream: Multi-view diffusion for 3d generation. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=FUgrjq2pbB
56. Srinivasan, P.P., Deng, B., Zhang, X., Tancik, M., Mildenhall, B., Barron, J.T.: Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7495–7504 (2021)
57. Sun, T., Barron, J.T., Tsai, Y.T., Xu, Z., Yu, X., Fyffe, G., Rhemann, C., Busch, J., Debevec, P., Ramamoorthi, R.: Single image portrait relighting. ACM Trans. Graph **38**(4), 1–12 (2019)
58. Tang, J., Levine, M., Verbin, D., Garbin, S.J., Niessner, M., Martin-Brualla, R., Srinivasan, P.P., Henzler, P.: ROGR: Relightable 3D Objects using Generative Relighting. In: Advances in Neural Information Processing Systems (2025)
59. Trevithick, A., Paiss, R., Henzler, P., Verbin, D., Wu, R., Alzayer, H., Gao, R., Poole, B., Barron, J.T., Holynski, A., et al.: Simvs: Simulating world inconsistencies for robust view synthesis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16464–16474 (2025)
60. Wang, R., Zhang, Q., Fu, C.W., Shen, X., Zheng, W.S., Jia, J.: Underexposed photo enhancement using deep illumination estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6849–6857 (2019)
61. Wang, Z., Philion, J., Fidler, S., Kautz, J.: Learning indoor inverse rendering with 3d spatially-varying lighting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12538–12547 (2021)

62. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
63. Xing, X., Groh, K., Karaoglu, S., Gevers, T., Bhattad, A.: Luminet: Latent intrinsics meets diffusion models for indoor scene relighting. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 442–452 (2025)
64. Ye, J., Zhuang, J., Mu, L., Zheng, W., Hu, J., Zou, X., Wang, J., Hu, H.: Training-free multi-view extension of ic-light for textual position-aware scene relighting. arXiv preprint arXiv:2511.13684 (2025)
65. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12–22 (2023)
66. Zeng, C., Dong, Y., Peers, P., Kong, Y., Wu, H., Tong, X.: Dilightnet: Fine-grained lighting control for diffusion-based image generation. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–12 (2024)
67. Zeng, Z., Deschaintre, V., Georgiev, I., Hold-Geoffroy, Y., Hu, Y., Luan, F., Yan, L.Q., Hašan, M.: Rgbx: Image decomposition and synthesis using material- and lighting-aware diffusion models. In: ACM SIGGRAPH 2024 Conference Papers. SIGGRAPH '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3641519.3657445, https://doi.org/10.1145/3641519.3657445
68. Zhang, K., Luan, F., Wang, Q., Bala, K., Snavely, N.: Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5453–5462 (2021)
69. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
70. Zhang, L., Rao, A., Agrawala, M.: Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=u1cQYxRI1H
71. Zhang, X., Srinivasan, P.P., Deng, B., Debevec, P., Freeman, W.T., Barron, J.T.: Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. ACM Transactions on Graphics (ToG) **40**(6), 1–18 (2021)
72. Zhao, H., Akers, J., Elmieh, B., Kemelmacher-Shlizerman, I.: Hybrelighter: Combining deep anisotropic diffusion and scene reconstruction for on-device real-time relighting in mixed reality. arXiv preprint arXiv:2508.14930 (2025)
73. Zhou, H., Hadap, S., Sunkavalli, K., Jacobs, D.W.: Deep single-image portrait relighting. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7194–7202 (2019)
74. Zhu, R., Li, Z., Matai, J., Porikli, F., Chandraker, M.: Irisformer: Dense vision transformers for single-image inverse rendering in indoor scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2822–2831 (2022)