# LIGHTSHIFT: CONTROLLABLE INDOOR RELIGHTING VIA VIDEO DIFFUSION PRIORS

**Anonymous authors**
Paper under double-blind review

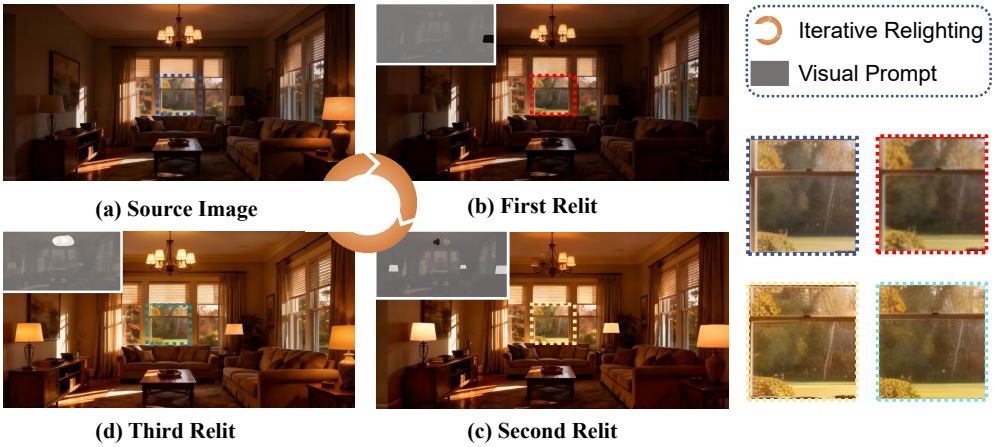


Figure 1: Iterative relighting results produced by **LightShift**. Starting from the (a) source image, LightShift is applied repeatedly, with each new relighting step taking the output of the previous step as its input. Using different user-provided visual prompts (shown in the gray insets), we obtain (b) the first relit, (c) the second relit, and (d) the third relit images. The highlighted box regions show that, despite sequential illumination changes, LightShift consistently preserves scene content and intrinsic properties while accurately adapting illumination according to each prompt.

## ABSTRACT

Controllable image-based indoor relighting aims to render a source image under user-specified illumination conditions while preserving intrinsic scene properties and structural consistency. Previous approaches typically leverage pretrained image diffusion models with auxiliary conditioning branches, such as ControlNet, to incorporate estimated geometry (e.g., depth and normals), intrinsic properties (e.g., albedo), and lighting instructions. However, errors in these estimates can propagate to the relit outputs, degrading visual quality and scene consistency. We present **LightShift**, a controllable indoor relighting framework that leverages pretrained video diffusion priors without requiring geometric or intrinsic estimates at inference time. Our key idea is in-sequence visual conditioning: we represent the source image, a scribble-based lighting instruction, and the relit output as a unified three-frame sequence, transferring pretrained video priors to modeling illumination changes without an auxiliary conditioning branch. To support training, we introduce a two-stage data synthesis pipeline that constructs relighting pairs from in-the-wild indoor photographs. The first stage generates coarse relighting pairs through intrinsic-based re-rendering, while the second uses the first-stage model to synthesize relit source images paired with real photographic targets. Quantitative and qualitative evaluations, together with a user study, demonstrate the effectiveness of LightShift in controllable indoor relighting, with improved lighting control and scene consistency across scenes.

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

# 1 INTRODUCTION

Illumination editing is a fundamental task that aims to modify an image's lighting while preserving its semantic content. As a core technique underpinning numerous applications in virtual reality (VR), film production, and computational photography, it has recently attracted growing attention in the computer vision community. Early classical 3D graphics methods (Debevec et al., 2000; Peers et al., 2007; Yu & Smith, 2019) model image appearance using physically based illumination models that rely on multiple input captures. While these approaches enable precise and explicit illumination control, their reliance on additional captures or recovered scene properties limits their applicability to relighting from a single in-the-wild image.

Recent advances in large-scale pretrained diffusion models have opened new opportunities for controllable image relighting. Building on their powerful generative priors, several methods (Zhang et al., 2025; Jin et al., 2024; Zeng et al., 2024; Liu et al., 2025; Chadebec et al., 2025) have explored object (Jin et al., 2024), portrait (Ren et al., 2024; Kim et al., 2024), and outdoor scene (Phongthawee et al., 2024) relighting from a single image, achieving diverse and realistic in-the-wild lighting effects. However, extending these capabilities to indoor environments remains highly challenging due to multiple interacting light sources and diverse material properties, which produce complex global illumination, soft shadows, and inter-reflections. Recent studies (Choi et al., 2025; Xing et al., 2025) address these challenges by conditioning pretrained image diffusion models on additional scene information. These approaches incorporate estimated scene properties through auxiliary conditioning branches, such as ControlNet (Zhang et al., 2023), using image-space cues such as albedo and geometry maps (Choi et al., 2025), or latent intrinsic representations (Xing et al., 2025; Zeng et al., 2024) to guide generation. However, their relighting results remain dependent on estimated geometric and intrinsic properties, which are often imperfect due to the limitations of current estimators (Li et al., 2025; Choi et al., 2025; Liang et al., 2025). Errors in these estimates can propagate to the relit outputs, degrading visual quality and scene consistency.

To address these limitations, we explore pretrained video generative models for controllable indoor relighting. Our core hypothesis is that video pretraining provides useful priors for preserving scene structure while modeling appearance changes across frames, extending beyond motion dynamics. Such capabilities are particularly relevant to relighting, where illumination changes while the underlying scene remains unchanged. Prior studies (Lin et al., 2025; Yu et al., 2025) have demonstrated the effectiveness of transferring video generative priors to downstream tasks. Building on this insight, we introduce **LightShift**, which reformulates controllable indoor relighting as a next-frame prediction task within a pretrained video diffusion model, without requiring estimated geometry or intrinsic properties at inference time.

To implement this formulation, we introduce *in-sequence visual conditioning*, which represents the source image, lighting instruction, and relit result as a unified frame sequence. Unlike GenLit (Bharadwaj et al., 2025), which generates relighting sequences under varying point-light conditions injected through a ControlNet branch, LightShift uses the sequence dimension to jointly model scene context, spatial lighting instructions, and the resulting illumination change. Specifically, the source image serves as the first frame. Following prior work (Mei et al., 2023; Choi et al., 2025), we adopt scribbles to specify regions to be brightened or darkened and represent this visual prompt as the second frame. The model then synthesizes the relit result as the third frame, integrating scene context and spatial lighting control through its temporal layers. By placing the lighting instruction within the sequence, this design transfers pretrained video priors to modeling illumination changes for controllable indoor relighting. As shown in Table 2, in-sequence visual conditioning outperforms the evaluated cross-attention and latent-concatenation conditioning strategies.

Building relighting training data for real-world indoor scenes is costly, requiring controlled illumination changes across scenes with diverse geometry and materials. To address this, we propose a two-stage data synthesis pipeline that constructs relighting pairs from existing indoor photographs. In Stage I, we generate coarse relighting pairs through re-rendering with estimated intrinsic properties, and automatically construct scribble prompts from the target shading maps following (Choi et al., 2025). These pairs provide initial supervision for learning illumination changes, although their quality is limited by errors in the estimated scene properties. In Stage II, we use the Stage I model to relight additional indoor photographs under inverse scribble prompts. The generated images serve as source inputs, while the original photographs are retained as targets, providing real photographic

supervision for the final model. We then train LightShift on the combined data from both stages, bringing together coarse synthetic illumination transitions and real photographic targets.

Our main contributions are as follows: (1) We introduce *LightShift*, a controllable indoor relighting framework based on *in-sequence visual conditioning*, which jointly models the source image, spatial lighting instruction, and relit result within a pretrained video diffusion model, without requiring geometric or intrinsic estimates at inference time. (2) We propose a two-stage data synthesis pipeline that combines intrinsic-based re-rendering with model-generated source images paired with real photographic targets, enabling training without manually captured relighting pairs. (3) Quantitative and qualitative evaluations, together with a user study, demonstrate improved lighting control and scene consistency, while ablation studies validate the effectiveness of video pretraining, in-sequence visual conditioning, and the two-stage data synthesis pipeline.

## 2 Related Work

Image-based relighting modifies scene illumination in photographs after capture. Existing work covers portraits, objects, outdoor scenes, and indoor environments under complex illumination (Chen et al., 2024; Pandey et al., 2021; Ren et al., 2024; Jin et al., 2024; Kim et al., 2024; Sun et al., 2019; Yu et al., 2020; Xing et al., 2025). Indoor scenes are especially challenging: cluttered geometry, occlusions, and mixed natural/artificial light produce hard shadows, specular highlights, and inter-reflections.

Many methods follow an *inverse–then–re-rendering* pipeline (Yu & Smith, 2019; Careaga & Aksoy, 2023; Li & Snavely, 2018; Thies et al., 2019; Yi et al., 2023; Kocsis et al., 2025): a network predicts intrinsic properties (albedo, normals, depth, lighting), and a second stage renders the image under new illumination. This design exposes physically meaningful controls, but single-view inverse rendering is ill-posed and errors in intrinsics are easily amplified by simplified light-transport models, often hurting realism in indoor scenes. Recent approaches instead adopt a *generative* formulation based on diffusion and other deep generative models (Bhattad et al., 2024; Chaturvedi et al., 2025; Chadebec et al., 2025). Some still expose an intrinsic-like interface and let a powerful generator act as a learned "renderer" (Careaga & Aksoy, 2024; Zeng et al., 2024; Magar et al., 2025; Lyu et al., 2025; Kocsis et al., 2025; Liang et al., 2026), while others treat relighting as image editing or inpainting and directly map an input image plus a high-level control signal to a relit result (Chadebec et al., 2025; Liu et al., 2025; Zhang et al., 2025; Li et al., 2025; Erel et al., 2025; Chaturvedi et al., 2026). These intrinsic-free methods improve realism, but precise spatial control and content preservation under large lighting changes remain challenging in complex indoor environments.

Scribble-based control offers a simple way to specify local lighting edits. Scribbles have been used in segmentation, editing, inpainting, and colorization (Lin et al., 2016; Zeng et al., 2022; Sharma et al., 2024; Chen et al., 2023; Wong et al., 2024), and sparse strokes can naturally indicate regions to be brightened or darkened without pixel-accurate shading (Mei et al., 2023). Closest to our setting, ScribbleLight (Choi et al., 2025) uses a ControlNet-conditioned diffusion model that takes an RGB image, predicted intrinsic maps, and a scribble map to perform indoor relighting. However, its reliance on estimated intrinsic maps at inference time makes relighting quality sensitive to decomposition errors. We retain the intuitive scribble interface but remove the need for intrinsic estimates at inference time, incorporating the lighting instruction directly into the frame sequence through in-sequence visual conditioning.

Recently, video diffusion models denoise whole frame sequences and learn temporal priors that keep scene structure stable while appearance evolves smoothly across frames (Ho et al., 2022; Blattmann et al., 2023b; Singer et al., 2022). Stable Video Diffusion (SVD) (Blattmann et al., 2023a) provides a pretrained image-to-video diffusion backbone with spatio-temporal layers for modeling interactions across frames. GenLit (Bharadwaj et al., 2025) introduces a video diffusion model for image relighting, trained on ray-traced relighting sequences, and is capable of generating short near-field relighting videos from a single image using parametric point-light controls. It encodes lighting conditions through a ControlNet branch and uses the video sequence to represent illumination changes along a light trajectory. In contrast, LightShift adopts *in-sequence visual conditioning*, representing the source image, a spatial lighting instruction, and the relit result as consecutive frames. This formulation repurposes the sequence dimension of the pretrained video model to integrate scene context and scribble-based lighting control through its temporal layers.

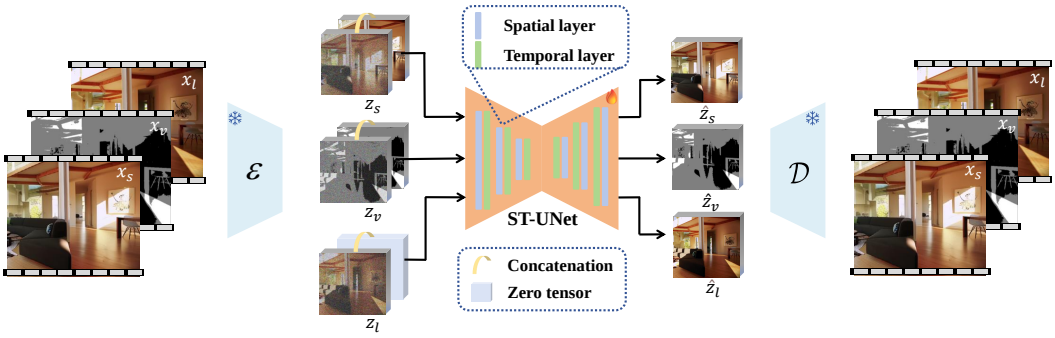


Figure 2: **Overview of *LightShift*.** We formulate relighting through in-sequence visual conditioning, modeling $(x_s, x_v, x_l)$ as a short sequence composed of the source image, visual prompt, and relit image. The pretrained VAE encoder $\mathcal{E}$ maps these frames into the latent space, after which the Spatio-Temporal U-Net (ST-UNet) is fine-tuned on our two-stage synthesized dataset to transfer pretrained video priors to modeling illumination changes while preserving scene consistency. During inference, given a single input image and a user-provided visual prompt, *LightShift* generates the relit result by iteratively denoising an initial Gaussian noise sample.

## 3 METHOD

In this work, we propose LightShift, a framework that repurposes video diffusion models for controllable indoor relighting, leveraging priors learned through large-scale video pretraining. Given a single input image, our goal is to achieve controllable indoor relighting guided by user-provided visual prompts. We first provide a brief overview of diffusion model notation in Section 3.1. Section 3.2 then introduces the LightShift generative formulation and model architecture, including our in-sequence visual conditioning design. In Section 3.3, we further describe our two-stage data synthesis pipeline. An overview of LightShift is illustrated in Fig. 2.

### 3.1 PRELIMINARIES OF DIFFUSION MODELS

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Rombach et al., 2022; Karras et al., 2022) learn data distributions by adding and then removing noise. The *forward process* corrupts real data into a simple reference distribution (e.g., Gaussian noise), while the *reverse process* learns to denoise and reconstruct realistic samples.

In *latent diffusion* (Rombach et al., 2022), an image $\mathbf{x}_0$ is encoded into a compact latent representation $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$ using a pretrained autoencoder. This latent space greatly reduces computational cost while retaining the semantic and perceptual content of the image. During training, Gaussian noise is progressively added to the latent and a denoiser $\mathcal{Z}_\theta$ learns to reverse this process by iteratively refining noisy samples back to the clean latent space.

ControlNet (Zhang et al., 2023) introduces an auxiliary control branch initialized from parts of the pretrained diffusion model to encode additional conditioning signals. Recent works (Choi et al., 2025; Xing et al., 2025) incorporate scene properties and lighting controls into indoor relighting models through auxiliary branches, latent concatenation, or combinations of these strategies.

Our method builds upon the video diffusion model *Stable Video Diffusion (SVD)* (Blattmann et al., 2023a), which extends diffusion to temporal sequences and follows the *Elucidated Diffusion Model (EDM)* formulation (Karras et al., 2022). EDM defines the forward process by directly adding Gaussian noise with variance $\sigma_t^2$ to a clean latent sample $\mathbf{z}_0 \sim p(\mathbf{z})$:

$$\mathbf{z}_t = \mathbf{z}_0 + \sigma_t \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

where $\mathbf{z}_t \sim p(\mathbf{z}; \sigma_t)$ denotes the noised latent at noise level $\sigma_t$. To stabilize training across a wide range of noise levels, EDM employs a *preconditioned denoiser*, which combines the noisy input and network output using noise-dependent scaling functions:

$$\begin{aligned} \mathcal{Z}_\theta(\mathbf{z}_t, \sigma_t, \mathbf{c}) = c_{\text{skip}}(\sigma_t)\, \mathbf{z}_t \\ + c_{\text{out}}(\sigma_t)\, F_\theta(c_{\text{in}}(\sigma_t)\, \mathbf{z}_t,\ c_{\text{noise}}(\sigma_t),\ \mathbf{c})\,, \end{aligned} \tag{2}$$




Figure 3: Qualitative comparison with other methods using automatically generated scribbles given a target (GT) image.

where $F_\theta$ denotes the learnable UNet backbone and $\{c_{\text{in}}, c_{\text{out}}, c_{\text{skip}}, c_{\text{noise}}\}$ are scalar functions controlling input scaling, output weighting, the skip connection, and the noise embedding. Here, $\mathbf{c}$ denotes the conditioning information. This preconditioning stabilizes denoising across noise levels, forming the basis of our controllable video-based relighting framework.

### 3.2 Relighting with Video Generative Models

**Generative Formulation** We formulate controllable indoor relighting as a *next-frame prediction* task within a video generation framework, modeling the conditional distribution $D(\mathbf{x}_l \mid \mathbf{x}_s, \mathbf{x}_v)$, where $\mathbf{x}_s$ denotes the source image, $\mathbf{x}_v$ represents the visual control prompt, and $\mathbf{x}_l$ is the relit output. Through in-sequence visual conditioning, we organize each training sample as an ordered triplet:

$$\mathcal{X} = \big(\mathbf{x}_s, \mathbf{x}_v, \mathbf{x}_l\big).$$

By treating these images as consecutive frames, LightShift transfers pretrained video priors to modeling illumination changes while preserving scene consistency. Gaussian noise is added to the latent representation of $\mathbf{x}_l$ according to Eq. (1). In the reverse process, the conditional denoising model $\mathcal{Z}_\theta$ iteratively denoises the latent sequence, conditioned on $\mathbf{x}_s$ and $\mathbf{x}_v$. During this process, $\mathbf{x}_s$ and $\mathbf{x}_v$ are simultaneously denoised to provide scene context and lighting guidance. For conditioning, we follow *Stable Video Diffusion* (Blattmann et al., 2023a), which concatenates the conditioning latent with the noisy latent during denoising. In our formulation, a zero tensor is concatenated with the target latent to match the input dimensionality during conditioning.

At training time, the model parameters $\theta$ are updated by sampling data from the training set, adding Gaussian noise to the target latent, and optimizing the denoising objective. Following EDM (Karras et al., 2022), we use the preconditioned parameterization defined in Eq. (2) for denoising training.

At inference time, we initialize the relit latent with Gaussian noise and iteratively apply the learned denoiser $\mathcal{Z}_\theta$. Using $\mathbf{x}_s$ and $\mathbf{x}_v$ as the first and second frames of the sequence, respectively, the model generates the relit result at the third frame position.

**Encoder and Decoder** Rather than operating in pixel space, LightShift performs diffusion in a low-dimensional latent space, offering computational efficiency and suitability for high-resolution image generation (Rombach et al., 2022). We employ a frozen VAE, trained independently of the denoiser, to encode images into latents and decode the generated result. The VAE encodes the triplet $\mathcal{X} = (\mathbf{x}_s, \mathbf{x}_v, \mathbf{x}_l)$, corresponding to the source image, visual prompt, and relit image, respectively, as $\mathbf{z}^{(\mathcal{X})} = \mathcal{E}(\mathcal{X})$. After denoising, we retain only the relit latent and reconstruct it using the decoder. Given the relit latent code $\mathbf{z}^{(\mathbf{x}_l)}$, the relit image is obtained as $\hat{\mathbf{x}}_l = \mathcal{D}\big(\mathbf{z}^{(\mathbf{x}_l)}\big)$.

**Denoising U-Net** We implement in-sequence visual conditioning within the pretrained spatio-temporal U-Net. Rather than concatenating both the source image and visual prompt with the target latent or injecting the prompt through a ControlNet branch, we represent them at separate frame positions. The U-Net is then fine-tuned to generate the relit frame conditioned on the source and prompt frames. To achieve this, we generate the source image and visual prompt along a shared denoising trajectory, concatenating their conditioning latents with the corresponding noisy latents. A zero tensor is concatenated with the relit target latent to match the input dimensionality during conditioning. We disable the cross-attention layers that inject conditional image embeddings into the denoising process and use the source and visual-prompt latents to provide conditioning.

### 3.3 Relighting Data Synthesis

To train LightShift, we require paired indoor images captured under varying lighting conditions. Although large-scale indoor relighting datasets such as OpenRooms (Li et al., 2021) exist, they are synthetic and exhibit a significant domain gap compared to real-world photographs. To mitigate this gap and maintain a fair comparison in data scale with existing methods, we follow Scribble-Light (Choi et al., 2025) and train on the real-world LSUN Bedrooms dataset (Yu et al., 2015). We construct relighting pairs from these photographs through a two-stage data synthesis pipeline.

**Stage I: Coarse Illumination Transitions via Re-rendering.** We synthesize coarse relighting pairs with a simplified inverse-rendering and re-rendering procedure. Given an indoor image $\mathbf{I}$, we estimate albedo and shading using an intrinsic decomposition network (Careaga & Aksoy, 2024), and surface normals using DSINE (Bae & Davison, 2024). We then sample random point lights and perturb shading while keeping the estimated albedo and geometry fixed.

Visual prompts are generated following ScribbleLight (Choi et al., 2025). We convert the target shading map to grayscale and threshold it into "brighten", "darken", and "neutral" regions using the global mean $\mu$ and standard deviation $\sigma_s$. Random morphological operations are applied to smooth and thicken strokes, producing a ternary scribble map $\{0, 0.5, 1\}$ rendered as a three-channel grayscale prompt.

We train LightShift on these synthesized pairs to obtain the Stage I model. Although the generated data contain noise, this stage provides initial supervision for learning illumination variations and scene consistency.

**Stage II: Fine Illumination Transitions via Self-Model Relighting.** Errors in the estimated intrinsic properties and geometry can limit the realism of Stage I re-renderings. To introduce real photographic targets into training, we use the Stage I model to synthesize alternative source images under modified illumination, as illustrated in Fig. 4.

In practice, we randomly sample an indoor image from LSUN that is disjoint from the Stage I data and treat it as the relit target image $\mathbf{I}_{\text{tar}}$. We then generate an inverse scribble map $M_{\text{inv}}$, where $M$ denotes the target illumination pattern of the sampled image. Next, we relight $\mathbf{I}_{\text{tar}}$ using LightShift trained in Stage I, conditioned on $M_{\text{inv}}$, obtaininga relit source image $\mathbf{I}_{\text{ref}}$. Finally, we form the



(a) Stage I Source Image (b) Stage II Source Image (c) Relit Image

Figure 4: Qualitative comparison between Stage I and Stage II synthesized data.

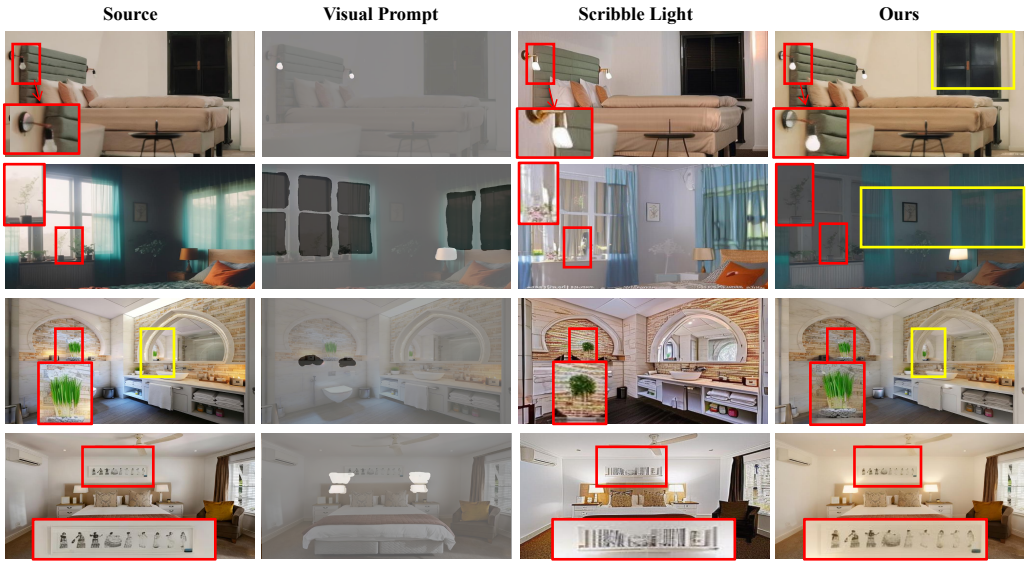


Figure 5: Qualitative comparison of relighting quality between ScribbleLight (Choi et al., 2025) and LightShift (ours) using user-provided visual prompts. The red boxes highlight regions used to assess scene-structure consistency, while the yellow boxes highlight regions illustrating the visual plausibility of lighting effects.

Stage II training triplets as $(\mathbf{I}_{\text{ref}}, M, \mathbf{I}_{\text{tar}})$. The generated image serves as the source input, while the original photograph is retained as the target.

We sample each training batch from both Stage I and Stage II to jointly retrain LightShift as the final model. Stage I provides synthetic illumination transitions with fixed estimated albedo and geometry, while Stage II contributes real photographic targets. We fine-tune the U-Net backbone on this mixed dataset to learn controllable illumination changes while preserving scene content. Although estimated scene properties are used during data synthesis, LightShift requires only a source image and a visual prompt at inference time.

## 4 Experiments

**Datasets.** Following previous work (Choi et al., 2025), we use the LSUN-Bedroom dataset (Yu et al., 2015) to construct synthetic relighting triplets using the two-stage data synthesis pipeline described in Section 3.3. We generate 50K samples for Stage I (Coarse Illumination Transition) and 50K samples for Stage II (Fine Illumination Transition). Unless otherwise specified, all variants of LightShift are trained on the combined set of these 100K triplets. We evaluate on BigTime (Li & Snavely, 2018) using 1,584 image pairs from 132 indoor scenes. Following the evaluation protocol of ScribbleLight (Choi et al., 2025), we generate scribble prompts from grayscale source–target differences and provide the same prompt to all evaluated methods. Details are provided in the supplementary material.

Table 1: Quantitative and user study results on indoor relighting.

| | Quantitative (BigTime) | | | User Study (%) | | |
|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | LPIPS↓ | Light | Consist. | Realism |
| IC-Light | 16.67 | 0.63 | **0.26** | 70% | 63% | 67% |
| RGB↔X | 10.79 | 0.44 | 0.50 | 86% | 95% | 95% |
| ScribbleLight | 12.14 | 0.44 | 0.45 | 83% | 85% | 86% |
| **LightShift** | **17.40** | **0.67** | 0.27 | – | – | – |

**Evaluation Metrics.** Building upon prior studies (Choi et al., 2025; Magar et al., 2025), we report peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and learned perceptual image patch similarity (LPIPS), all widely adopted in the literature. However, these reference-based metrics do not fully capture lighting controllability and perceived visual realism. To provide a more comprehensive assessment of relighting quality, we additionally conduct a user study comparing LightShift with the baseline methods.

### 4.1 Main Results

We compare with representative image relighting methods, including IC-Light (Zhang et al., 2025), RGB↔X (Zeng et al., 2024), and ScribbleLight (Choi et al., 2025), with ScribbleLight as the closest baseline in interface protocol and training data. We implement ScribbleLight by predicting albedo and surface normals with IID (Careaga & Aksoy, 2023) and DSINE (Bae & Davison, 2024), then conditioning its diffusion model on these intrinsic maps and the scribble input. Since IC-Light and RGB↔X lack native scribble conditioning, we adapt them by using the scribble as an environment-lighting cue for IC-Light and as a new lighting configuration for RGB↔X.

**Quantitative Comparison.** The quantitative results (PSNR, SSIM, and LPIPS) are presented in Table 1. Our approach achieves the highest PSNR and SSIM among the evaluated methods, while obtaining LPIPS comparable to IC-Light. Despite IC-Light relying on large-scale paired training data, LightShift synthesizes its training pairs from single images. Additionally, LightShift substantially outperforms ScribbleLight (Choi et al., 2025), which also uses LSUN-Bedroom for training. These results demonstrate the effectiveness of our method in producing high-quality and structurally consistent indoor relighting.

**User Study.** To provide a more comprehensive evaluation, we conducted a user study comparing LightShift with three baselines: IC-Light (Zhang et al., 2025), RGB↔X (Zeng et al., 2024), and ScribbleLight (Choi et al., 2025). We invited 30 participants to evaluate 30 randomly selected cases from each dataset (15 with freehand user scribbles to represent interactive use, and 15 with automatically generated visual prompts). We adopted a Two-Alternative Forced Choice (2AFC) protocol. Each comparison presented the output of LightShift alongside that of one baseline. For each case, participants evaluated the results under three separate criteria. For each criterion, they selected the better result between the two options: (1) *Light: Light Controllability* (how well the output follows the brighten/darken instructions), (2) *Consist.: Scene Consistency* (preservation of object identity and geometry), and (3) *Realism* (overall visual plausibility). Table 1 reports the percentage of comparisons in which participants preferred LightShift over each baseline. Participants prefer LightShift across all three criteria, indicating improved lighting control, scene consistency, and visual realism.

**Qualitative Comparison with Automatic Scribbles.** In Figure 3, we present a qualitative comparison using automatically generated visual prompts against RGB↔X, IC-Light, and ScribbleLight on four challenging scenes with diverse illumination conditions, including both artificial and natural light sources. Under the adapted scribble-conditioning setup, RGB↔X introduces artifacts resembling the visual prompt. IC-Light can alter global illumination, but it struggles to follow some fine-grained lighting instructions and retains lighting patterns close to the input in the illustrated examples. ScribbleLight can respond to prompts in some cases, yet it introduces unnatural lighting patterns and irregular textures that degrade visual quality. In contrast, *LightShift* follows the lighting instructions while better preserving scene content and structure, producing visually plausible results under both local edits and global illumination changes.

Table 2: Analysis of pretrained video diffusion and scribble injection.

| | PSNR↑ | SSIM↑ | LPIPS↓ |
|---|---|---|---|
| **Pretrained Video Diffusion Model** | | | |
| (a) w/o full pretraining | 9.80 | 0.42 | 0.54 |
| (b) w/o temporal weights | 15.10 | 0.58 | 0.33 |
| (c) Image diffusion concat. | 12.60 | 0.45 | 0.52 |
| **Scribble Injection Mechanism** | | | |
| (d) Cross-Attention Inject | 13.97 | 0.61 | 0.31 |
| (e) Scribble–noise concat. | 13.90 | 0.56 | 0.37 |
| (f) ControlNet-style | 13.50 | 0.60 | 0.32 |
| (g) **Ours** | **17.40** | **0.67** | **0.27** |

**Qualitative Comparison with User Scribbles.** We conduct several qualitative evaluations to further demonstrate the advantages of our method. We select ScribbleLight as the primary baseline for comparison under user-provided visual prompts, as it is the closest to our setting in both interface and task formulation. As shown in Figure 5, challenging regions are highlighted with red and yellow boxes. These examples illustrate that LightShift not only performs the desired lighting edits more accurately, but also produces more visually plausible lighting effects. Notably, LightShift better captures light transmission through frosted glass (row 1) and handles translucent materials more realistically (row 2). In contrast, ScribbleLight exhibits lighting and texture artifacts in these regions. These examples show that LightShift can produce plausible illumination changes while preserving scene structure, without requiring explicit intrinsic estimates at inference time.

## 5 Ablation Studies

**Effectiveness of Video Pretrained Priors.** Our core hypothesis is that priors learned through video pretraining can be transferred to modeling illumination changes while preserving scene consistency. To examine the contribution of video pretraining, we compare three variants: training LightShift from scratch, initializing only its spatial layers with pretrained weights, and replacing the video diffusion backbone with an image diffusion model using spatial concatenation as input. As shown in Table 2, initializing the full video diffusion model achieves the best performance. These results support the benefit of transferring pretrained video priors to controllable indoor relighting.

**Effectiveness of In-Sequence Visual Conditioning.** We compare three alternative scribble conditioning strategies in Table 2: encoding the scribble using CLIP and injecting it through cross-attention, concatenating the scribble latent with the noisy target latent, and using an auxiliary ControlNet-style branch to guide the same pretrained video diffusion backbone. In contrast, in-sequence visual conditioning represents the source image, scribble, and relit result as consecutive frames, providing a sequence-based formulation for transferring video pretrained priors to illumination changes. Our formulation outperforms all three alternatives across PSNR, SSIM, and LPIPS. In particular, it improves PSNR by 3.90 dB over ControlNet-style conditioning, supporting the effectiveness of incorporating lighting instructions directly into the frame sequence.

## 6 Conclusion

We introduced *LightShift*, a framework for controllable indoor relighting that transfers pretrained video priors to modeling illumination changes without requiring geometric or intrinsic estimates at inference time. Our key design, in-sequence visual conditioning, represents the source image, visual prompt, and relit result as consecutive frames, formulating relighting as a next-frame prediction task while preserving scene consistency. To support *LightShift* training, we propose a two-stage data synthesis pipeline that combines coarse re-rendered pairs with model-generated source images paired with real photographic targets. Quantitative and qualitative evaluations, together with a user study, demonstrate the effectiveness of LightShift in lighting control, scene preservation, and visual realism. Ablation studies further support the benefits of video pretraining and in-sequence visual conditioning for controllable indoor relighting.

## AI USE STATEMENT

We used generative AI for language polishing and assistance with implementing and editing code. We reviewed and tested the code and take responsibility for the final implementation.

## REPRODUCIBILITY STATEMENT

Implementation and training details are provided in the appendix. The main paper describes the model architecture, training objective, data-synthesis pipeline, datasets, baselines, and evaluation metrics used in the experiments. The appendix further reports the backbone configuration, optimization settings, evaluation-set construction, and ablation protocols.

## REFERENCES


Gwangbin Bae and Andrew J Davison. Rethinking inductive biases for surface normal estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9535–9545, 2024.

Shrisha Bharadwaj, Haiwen Feng, Giorgio Becherini, Victoria Fernandez Abrevaya, and Michael J Black. Genlit: Reformulating single-image relighting as video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025.

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

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22563–22575, 2023b.

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

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

Clément Chadebec, Onur Tasar, Sanjeev Sreetharan, and Benjamin Aubin. Lbm: Latent bridge matching for fast image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 29086–29098, 2025.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 369–379, 2025.

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

Dave Zhenyu Chen, Haoxuan Li, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Scenetex: High-quality texture synthesis for indoor scenes via diffusion priors. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 21081–21091, 2024.

Xi Chen, Yau Shing Jonathan Cheung, Ser-Nam Lim, and Hengshuang Zhao. Scribbleseg: Scribble-based interactive image segmentation. *arXiv preprint arXiv:2303.11320*, 2023.

Jun Myeong Choi, Annie Wang, Pieter Peers, Anand Bhattad, and Roni Sengupta. Scribblelight: Single image indoor relighting with scribbles. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5720–5731, 2025.

Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. 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.

Yotam Erel, Rishabh Dabral, Vladislav Golyanik, Amit H. Bermano, and Christian Theobalt. Practilight: Practical light control using foundational diffusion models. *arXiv preprint arXiv:2509.01837*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Haian Jin, Yuan Li, Fujun Luan, Yuanbo Xiangli, Sai Bi, Kai Zhang, Zexiang Xu, Jin Sun, and Noah Snavely. Neural gaffer: Relighting any object via diffusion. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 141129–141152. Curran Associates, Inc., 2024. doi: 10.52202/079017-4481. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ff7373914a96956f2a7cacbdf3b0b8d8-Paper-Conference.pdf.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Hoon Kim, Minje Jang, Wonjun Yoon, Jisoo Lee, Donghyun Na, and Sanghyun Woo. Switchlight: Co-design of physics-driven architecture and pre-training framework for human portrait relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25096–25106, 2024.

Peter Kocsis, Lukas Höllein, and Matthias Nießner. Intrinsix: High-quality pbr generation using image priors. *arXiv preprint arXiv:2504.01008*, 2025.

Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9039–9048, 2018.

Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Meng Song, Yuhan Liu, Yu-Ying Yeh, Rui Zhu, Nitesh Gundavarapu, Jia Shi, et al. Openrooms: An open framework for photorealistic indoor scene datasets. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7190–7199, 2021.

Zongming Li, Lianghui Zhu, Haocheng Shen, Longjin Ran, Wenyu Liu, and Xinggang Wang. Translight: Image-guided customized lighting control with generative decoupling. *arXiv preprint arXiv:2508.14814*, 2025.

Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Chih-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, 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.

Ruofan Liang, Norman Müller, Ethan Weber, Duncan Zauss, Nandita Vijaykumar, Peter Kontschieder, and Christian Richardt. Luxremix: Lighting decomposition and remixing for indoor scenes. *arXiv preprint arXiv:2601.15283*, 2026.

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

Yijing Lin, Mengqi Huang, Shuhan Zhuang, and Zhendong Mao. Realgeneral: Unifying visual generation via temporal in-context learning with video models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14994–15004, 2025.

Yong Liu, Wenpeng Xiao, Qianqian Wang, Junlin Chen, Shiyin Wang, Yitong Wang, Xinglong Wu, and Yansong Tang. Dreamlight: Towards harmonious and consistent image relighting. *arXiv preprint arXiv:2506.14549*, 2025.

Linjie Lyu, Valentin Deschaintre, Yannick Hold-Geoffroy, Miloš Hašan, Jae Shin Yoon, Thomas Leimkühler, Christian Theobalt, and Iliyan Georgiev. Intrinsicedit: Precise generative image manipulation in intrinsic space. *ACM Transactions on Graphics (TOG)*, 44(4):1–13, 2025.

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 *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2025.

Yiqun Mei, He Zhang, Xuaner Zhang, Jianming Zhang, Zhixin Shu, Yilin Wang, Zijun Wei, Shi Yan, HyunJoon Jung, and Vishal M Patel. Lightpainter: Interactive portrait relighting with freehand scribble. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 195–205, 2023.

Rohit Pandey, Sergio Orts-Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul E Debevec, and Sean Ryan Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Trans. Graph.*, 40(4):43–1, 2021.

Pieter Peers, Naoki Tamura, Wojciech Matusik, and Paul Debevec. Post-production facial performance relighting using reflectance transfer. *ACM Transactions on Graphics (TOG)*, 26(3):52–es, 2007.

Pakkapon Phongthawee, Worameth Chinchuthakun, Nontaphat Sinsunthithet, Varun Jampani, Amit Raj, Pramook Khungurn, and Supasorn Suwajanakorn. 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.

Mengwei Ren, Wei Xiong, Jae Shin Yoon, Zhixin Shu, Jianming Zhang, HyunJoon Jung, Guido Gerig, and He Zhang. Relightful harmonization: Lighting-aware portrait background replacement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6452–6462, 2024.

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

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

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Trans. Graph*, 38(4):1–12, 2019.

Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

Hallee E Wong, Marianne Rakic, John Guttag, and Adrian V Dalca. Scribbleprompt: fast and flexible interactive segmentation for any biomedical image. In *European Conference on Computer Vision*, pp. 207–229. Springer, 2024.

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

Renjiao Yi, Chenyang Zhu, and Kai Xu. Weakly-supervised single-view image relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8402–8411, 2023.

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

Xin Yu, Tianyu Wang, Soo Ye Kim, Paul Guerrero, Xi Chen, Qing Liu, Zhe Lin, and Xiaojuan Qi. Objectmover: Generative object movement with video prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17682–17691, 2025.

Ye Yu and William AP Smith. Inverserendernet: Learning single image inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3155–3164, 2019.

Ye Yu, Abhimitra Meka, Mohamed Elgharib, Hans-Peter Seidel, Christian Theobalt, and William AP Smith. Self-supervised outdoor scene relighting. In *European Conference on Computer Vision*, pp. 84–101. Springer, 2020.

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

Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. Rgb↔x: Image decomposition and synthesis using material- and lighting-aware diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, SIGGRAPH '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705250. doi: 10.1145/3641519.3657445. URL https://doi.org/10.1145/3641519.3657445.

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

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

To complement the main content of the paper, we provide additional experimental details in Appendix A, followed by additional ablation studies in Appendix B, further quantitative and qualitative comparisons with LumiNet (Xing et al., 2025) in Appendix C, and additional visual comparisons in Appendix D.

## A Experimental Details

**Implementation Details.** We implement LightShift on top of the image-to-video variant of Stable Video Diffusion (Blattmann et al., 2023a). We follow the EDM noise schedule and loss weighting used in SVD and use 25 inference steps by default. During fine-tuning, all parameters of the spatio-temporal U-Net are updated, while the VAE remains frozen. Training is performed with mixed precision (fp16), gradient checkpointing, and an 8-bit AdamW optimizer with a learning rate of $1 \times 10^{-5}$. We train on two RTX 3090 GPUs with a per-GPU batch size of 1 for a total of 200K optimization steps. The training resolution is $512 \times 256$.

**Evaluation Datasets.** To evaluate controllable indoor relighting, we follow prior work (Choi et al., 2025) and adopt the BigTime (Li & Snavely, 2018) dataset, which consists of time-lapse videos of the same scene captured under varying illumination. This dataset allows us to construct complex and diverse relighting evaluation samples.

We focus on six types of illumination transitions: Dark→Bright, Dark→Mid, Bright→Dark, Bright→Mid, Mid→Dark, and Mid→Bright. Specifically, we select 132 indoor sequences and process each sequence independently. For every frame, we convert the image to grayscale and compute a scalar *brightness score* as the mean gray value. Frames within each sequence are then sorted by brightness and divided into three pools: the darkest $25\%$ (Dark), the brightest $25\%$ (Bright), and the middle $50\%$ (Mid). We then randomly sample image pairs from these pools to construct the six illumination-transition types described above.

For each pair $(\mathbf{I}_{\text{src}}, \mathbf{I}_{\text{tgt}})$, we automatically generate a scribble indicating the desired lighting changes. We first convert both images to grayscale and compute the signed per-pixel intensity difference $\Delta = \text{gray}(\mathbf{I}_{\text{tgt}}) - \text{gray}(\mathbf{I}_{\text{src}})$. Pixels with $|\Delta|$ exceeding a fixed threshold (50 in the 0–255 range) are marked as "edited" and then dilated to produce thick strokes. The sign of $\Delta$ determines whether each stroke indicates brightening or darkening. This process yields a paired evaluation set with target-derived scribble instructions for assessing relighting against reference images.

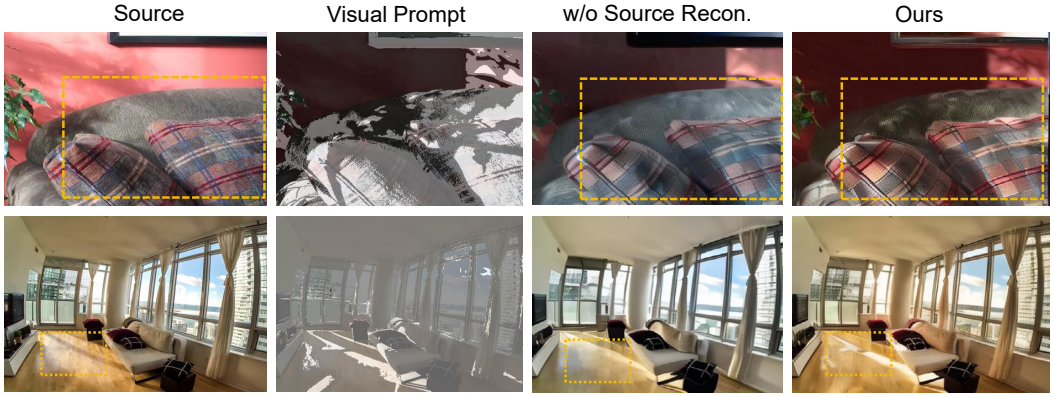


Figure 6: Comparison between different generation formulations.

## B Additional Ablation Studies

**Effectiveness of the Generation Formulation.** Our formulation represents the source image together with the visual prompt and relit result in the generation sequence. To examine the role of source-image reconstruction, we remove it and instead provide the source latent through channel-wise concatenation with the noisy relit latent. This variant retains the pretrained video backbone

Table 3: Ablation on generation formulation and training data.

| | PSNR↑ | SSIM↑ | LPIPS↓ |
|---|---|---|---|
| **Generation Formulation** | | | |
| (a) w/o source reconstruction | 15.30 | 0.59 | 0.33 |
| (b) **Ours** | **17.40** | **0.67** | **0.27** |
| **Training Data** | | | |
| (c) Stage I only | 17.23 | 0.64 | 0.31 |
| (d) Stage II only | 16.63 | 0.64 | 0.30 |

and the visual prompt as a frame. As shown in Table 3(a), removing source-image reconstruction reduces performance, supporting the benefit of including the source image in the generation sequence. Qualitative results in Figure 6 further show color bleeding, texture degradation, and less precise lighting control in this variant.

**Effectiveness of the Two-Stage Data Synthesis.** We examine the effect of the two-stage data synthesis pipeline in Table 3(c) and Table 3(d). We first train LightShift using only Stage I coarse illumination transitions generated via re-rendering. While these pairs provide supervision for learning illumination changes, their realism is limited by intrinsic estimation errors and simplified re-rendering. The model trained on Stage I alone produces less realistic lighting effects in Figure 7. We also train a variant exclusively on Stage II pairs, which consist of model-generated source images and real photographic targets. This variant improves LPIPS over Stage I-only training but yields lower PSNR, and neither variant matches the combined dataset across the reported metrics. Combining both stages achieves the best performance, supporting the use of coarse synthetic illumination transitions together with real photographic targets.

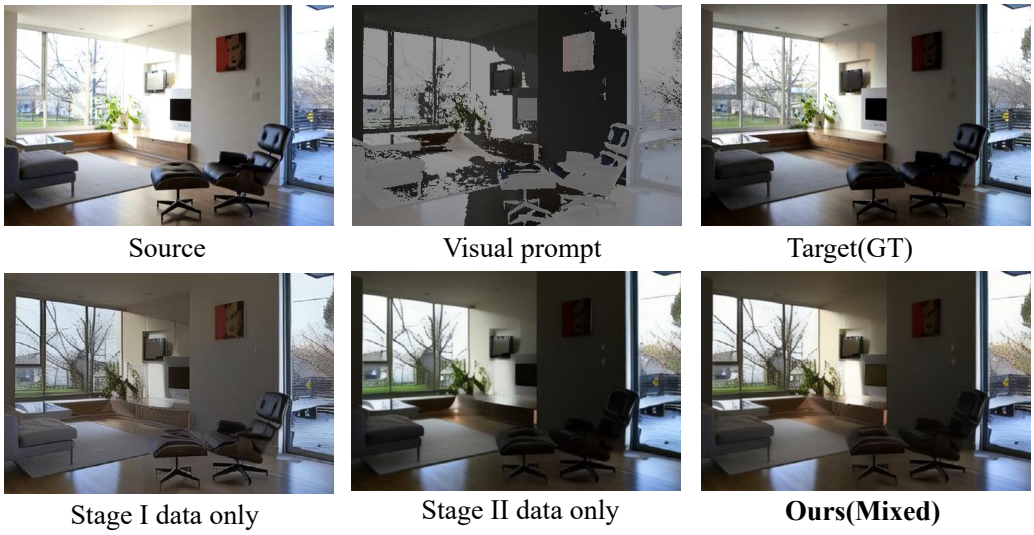


Figure 7: Comparison of models trained on different data combinations.

## C Additional Comparisons with LumiNet

We compare LightShift with **LumiNet** (Xing et al., 2025), a state-of-the-art indoor relighting method designed for reference-based illumination transfer. Since LumiNet requires a *target lighting image* rather than a scribble, we adapt the input by converting our scribble map into a grayscale image and using it as the reference.

Table 4: Quantitative comparison of relighting quality on the paired evaluation set from BigTime Li & Snavely (2018).

| Method | PSNR↑ | SSIM↑ | LPIPS↓ |
|---|---|---|---|
| LumiNet Xing et al. (2025) | 15.41 | 0.61 | 0.29 |
| **LightShift** | **17.40** | **0.67** | **0.27** |

Under this adapted input setting, LightShift achieves better quantitative results (Table 4) and produces fewer visible artifacts (Fig. 8). LumiNet's outputs exhibit texture artifacts resembling the scribble strokes in the illustrated examples. This comparison highlights the difficulty of directly using sparse lighting instructions with a reference-based model. Since scribble maps differ from LumiNet's intended photographic references, these results should be interpreted within this adapted scribble-guided setting.

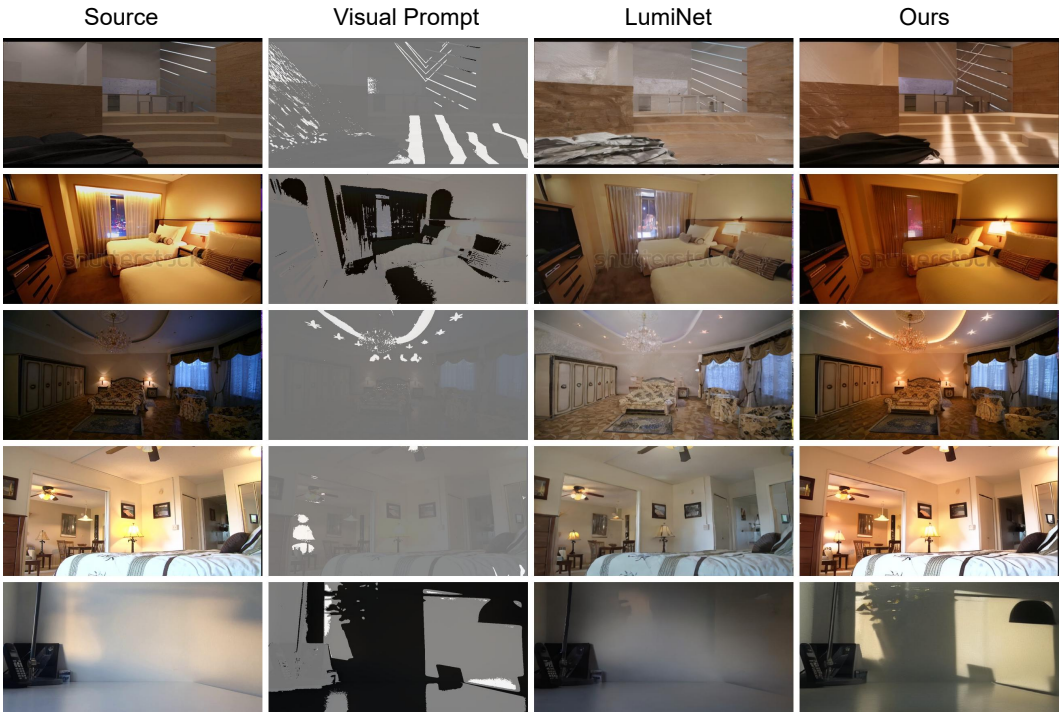


Figure 8: Qualitative comparison with LumiNet (Xing et al., 2025) and LightShift (ours) using auto-generated scribbles. LumiNet receives the grayscale scribble as its reference image.

## D More Qualitative Comparisons with ScribbleLight

Figure 9 presents additional qualitative examples comparing our method with ScribbleLight.

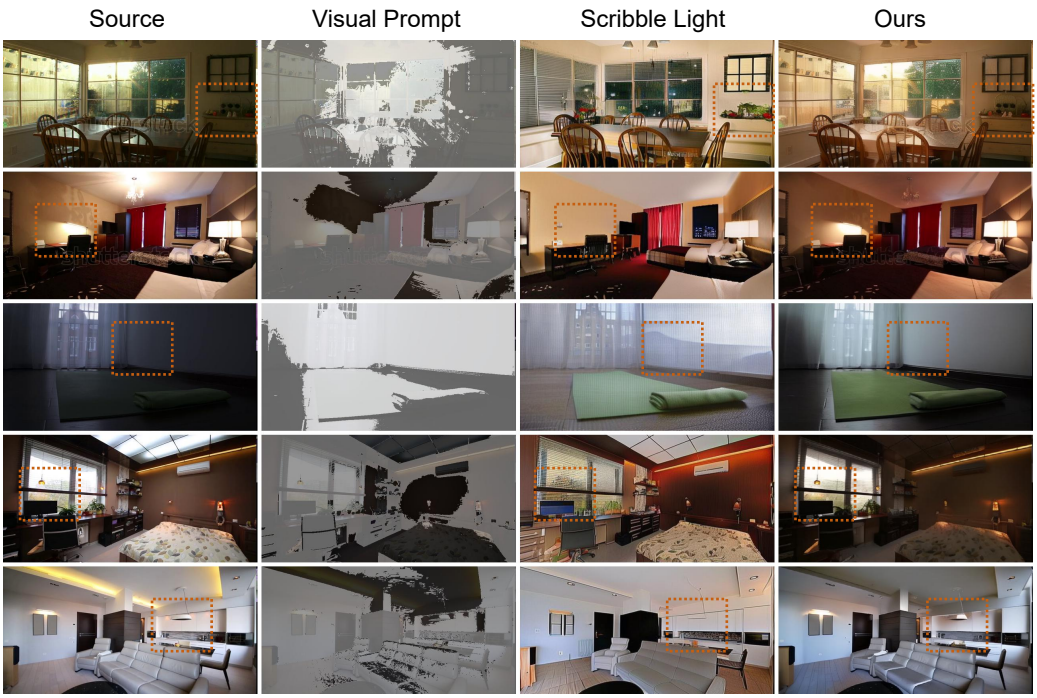


Figure 9: Additional qualitative comparisons between ScribbleLight (Choi et al., 2025) and LightShift (ours).