SPATIAL–ILLUMINATION ALIGNMENT FOR CONTROLLABLE IMAGE ILLUMINATION EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Controllable image illumination editing modifies both the luminance intensity and distribution. Pixel-wise alignment measures reconstruction errors but does not explicitly organize illumination variation with regional brightness. Spatial–illumination alignment couples regional luminance levels with spatial arrangement to guide illumination redistribution. For spatial-illumination consistency, we introduce Illumination Intensity Ranking (IIR) loss, constructing luminance rankings over image patches. Combining luminance consistency and positional consistency, IIR supervises how bright each region should be and where regions of corresponding brightness should occur. We further integrate spatial-illumination alignment with latent bridge matching and enhance informative illumination injection by condition-confirmed initialization. Extensive experiments show that the proposed Illumination-Conditioned Latent Bridge Matching (IC-LBM) achieves state-of-the-art performance on indoor scene relighting, improving the controllability of image illumination editing. Ablations confirm that luminance and positional consistency contribute complementary improvements to illumination reconstruction. IIR provides a structured objective for predicting target illumination without intrinsic decomposition or 3D scene reconstruction. Spatial–illumination alignment introduces explicit illumination supervision for image illumination editing, providing a pioneering perspective on modeling illumination transformations.

1 INTRODUCTION

Image illumination editing modifies lighting attributes while preserving the overall realism of an image. Illumination editing of scenes, objects, and human faces is an important problem in computer vision. Early approaches (Li et al., 2020), (Boss et al., 2021) relied on physical modeling to manipulate illumination. Subsequently, a large body of work (Rao et al., 2024), (Mei et al., 2024), (Deng et al., 2024) incorporated 3D perception or scene reconstruction to simulate illumination variations. More recently, with the rapid development of controllable diffusion models, a number of methods (Zeng et al., 2024), (Kocsis et al., 2024a), (Zhang et al., 2025) have abandoned redundant dense constraints in high-dimensional spaces. With images, depth maps, and text prompts as conditions, controllable diffusion models precisely edit illumination from a single image with sparse physical definitions, demonstrating strong generative potential.

Existing methods adopt pixel-wise alignment, which preserves geometric details but provides insufficient constraints on illumination changes. Decomposition-based methods disentangle reflectance and illumination. However, decomposition based on sparse geometric definitions introduces deviation, while pixel-wise losses cannot independently constrain geometry and illumination. Latent-space learning discards physical modeling and relies on large-scale source–target pairs to drive training. Pixel-wise alignment lacks explicit supervision on the controllability of the transformation from the source to the target. During illumination editing, changes in illumination cause global variations in shadows, highlights, and color bleeding. Solely relying on pixel-wise alignment tends to focus on high-frequency information without explicit modeling of the illumination distribution, which is not conducive to overall illumination variation.

Effective condition injection and control mechanisms are significant for precise illumination editing. Diffusion models generate highly diverse results but struggle to preserve semantic details. Latent

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.

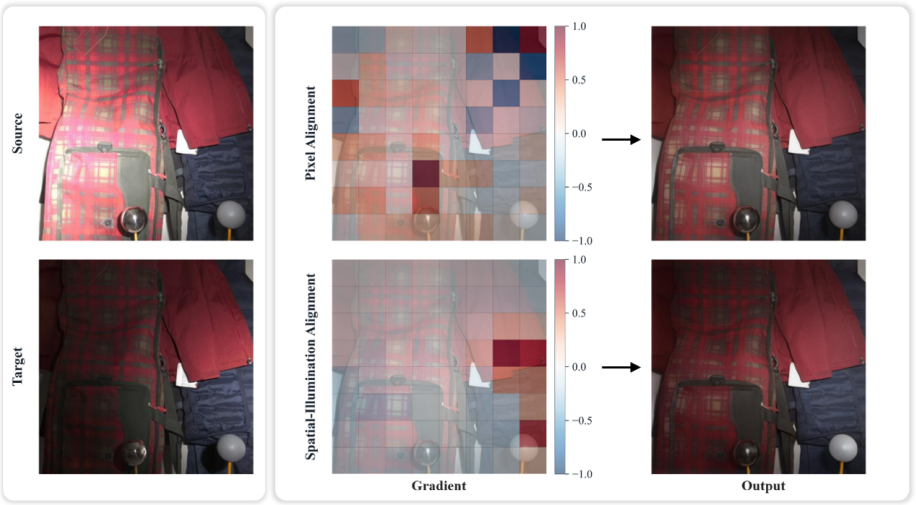

Figure 1: Visualization of gradient directions. The middle column shows the normalized gradient values for different regions. Pixel-wise alignment produces disparate local gradients without explicitly accounting for the spatial arrangement. Spatial–illumination alignment provides consistent regional gradients that guide the illumination toward the target locations.

bridge matching (Chadebec et al., 2025) enables direct condition injection for fidelity-preserving generation. However, latent bridge matching concatenates the condition with the noisy latent. As conditions overwhelmed by noise, the model tends to generate results close to the source image with poor conditional guidance. Effective incorporation of lighting conditions is therefore essential.

In this paper, we propose a new perspective that focuses image illumination editing on spatial–illumination alignment. Illumination variations manifest as changes in pixel values across different regions. As shown in Figure 1, merely aligning pixel values aims for global average optimization without clear spatial guidance. We divide the image into patches and compare luminance rankings between patches to characterize illumination variation. Specifically, we sort the patches by luminance and compare the value and distance between the output and target patches at the same rank. The spatial-illumination alignment constrains the consistency of both brightness levels and the spatial locations of bright regions. Unlike pixel-wise alignment, spatial–illumination alignment provides structured gradients that couple luminance levels with spatial locations, explicitly supervising the spatial redistribution of illumination. Without intrinsic disentanglement or 3D inverse rendering, spatial–illumination alignment successfully models illumination changes in the 2D image space. For sufficient condition control in latent bridge matching, we conduct condition-confirmed initialization to activate conditions.

Experimental results show that spatial–illumination alignment effectively captures global illumination variation. Experiments on the MIIW (Murmann et al., 2019) benchmark indicate consistent improvements in image quality and illumination controllability. With adherence to the target condition, spatial–illumination alignment explicitly constrains illumination variations and improves the consistency of regional illumination distributions.

Our contributions are summarized as follows: (1) We formulate illumination editing from the perspective of spatial–illumination alignment, effectively modeling illumination variation in the 2D image with the positional and value information. (2) We propose an Illumination Intensity Ranking (IIR) Loss for spatial–illumination alignment. Rather than matching pixel-by-pixel, IIR converts illumination into spatially aware luminance rankings and simultaneously supervises the brightness with spatial correspondence, explicitly matching luminance redistribution across different regions. (3) We propose Illumination-Conditioned Latent Bridge Matching (IC-LBM), injecting illumination cues into latent bridge matching for controllable image illumination editing. Extensive experiments demonstrate improved controllability in indoor scene image illumination editing.

2 Related Work

2.1 Image Illumination Editing with Decomposition

Explicit intermediate representations provide scene structure and material information for illumination editing. Decomposition-based methods recover editable factors to render the scene under target illumination. For indoor scenes, inverse rendering (Li et al., 2020) jointly estimates illumination and reflectance with spherical Gaussians. Neural-PIL (Boss et al., 2021) recovers geometry and illumination from multi-view observations. Accounting for the ambiguity of single-view decomposition, Intrinsic Image Diffusion (Kocsis et al., 2024b) models a distribution over material estimates rather than a single deterministic solution. Materialist (Wang et al., 2026a) and PI-Light (Liang et al., 2026) further combine intrinsic estimation with physical guidance and neural rendering. Other approaches use explicit scene representations or geometric cues without full intrinsic decomposition. For portrait relighting, 3D-aware generative representations (Rao et al., 2024), (Mei et al., 2024), (Deng et al., 2024) provide structural guidance. In low-light 3D restoration, Lumos3D Liu et al. (2025a) introduces geometric priors to support structure recovery. DiLightNet Zeng et al. (2024) and LightIt Kocsis et al. (2024a) perform conditional generation with radiance hints and shading maps. Light–geometry interaction maps Wang et al. (2026b) couple relighting with shadow generation. LightSwitch Litman et al. (2025) incorporates material guidance for multi-view consistency, and GeoRelight Xue et al. (2026) jointly models geometric reconstruction and relighting.

2.2 Image Illumination Editing with Latent Representation Learning

Latent representation methods learn representations of scene properties and illumination in latent space, reducing the reliance on explicit physical parameter estimation. Latent Intrinsics (Zhang et al., 2024) explores paired relighting supervision to induce latent representations of scene properties and illumination. Neural Gaffer (Jin et al., 2024) introduces diffusion-based object relighting conditioned on HDR environment maps. LumiNet (Xing et al., 2025) combines latent intrinsic features with the conditioning adapter for indoor scene relighting. Different conditioning interfaces further expand the range of controllability. ScribbleLight (Choi et al., 2025) defines sparse scribbles for indoor relighting, Text2Relight (Cha et al., 2025) specifies portrait illumination via text prompts, and CameraMaster (Yang et al., 2025) combines semantic and numerical camera controls for photographic retouching. UniLight (Zhang et al., 2026) introduces a shared lighting representation across modalities, while SyncLight (Serrano-Lozano et al., 2026) addresses consistency across multiple views. Physical guidance also informs latent representation learning. POLAR (Chen et al., 2026) constructs portrait relighting with one-light-at-a-time datasets, LightLab (Magar et al., 2025) exploits light additivity for parametric control, and IC-Light (Zhang et al., 2025) regularizes diffusion training with light-transport consistency. UniLumos (Liu et al., 2025b) incorporates geometric feedback, while the latent proxy method (Zheng et al., 2026b) encodes compact material and geometry cues.

2.3 Controllable Diffusion Models

Denoising diffusion probabilistic models (Ho et al., 2020) generate images through iterative denoising, and latent diffusion models (Rombach et al., 2022) operate in a compressed space. Flow matching (Hu et al., 2024) further explores image editing with latent representations. For controllable generation, ControlNet (Zhang et al., 2023) injects conditions through trainable branches parallel to the backbone. T2I-Adapter (Mou et al., 2024) encodes control signals by lightweight adapters, while Uni-ControlNet (Zhao et al., 2023) integrates local and global conditions. Attention-based methods (Hu et al., 2023), (Lee et al., 2024), (Kim et al., 2026), (Sun et al., 2024) designs different feature-fusion strategies for multi-modal control. For fidelity preserving illumination editing, diffusion bridge models (Zhou et al., 2024) connect source and target distributions, providing a formulation for fidelity-preserving image-to-image translation. Latent Bridge Matching (LBM) (Chadebec et al., 2025) constructs bridges in a compact latent space and supports conditional translation for image illumination editing.

3 METHOD

3.1 ILLUMINATION-CONDITIONED LATENT BRIDGE MATCHING

Given a source scene image x_0 and a condition image x_c of the target illumination, the model synthesizes $\hat{x}_1$. As illustrated in Figure 2, x_0 is mapped to the source latent $z_0 \in \mathbb{R}^{4\times64\times64}$ by a pretrained VAE encoder. x_c is encoded by the same VAE and compressed into a condition sequence c. Guided by the condition sequence, z_0 evolves toward the target latent $\hat{z}_1$, which is decoded into the predicted image $\hat{x}_1$. To explicitly guide the source toward the target, we introduce an objective that aligns the illumination distribution.

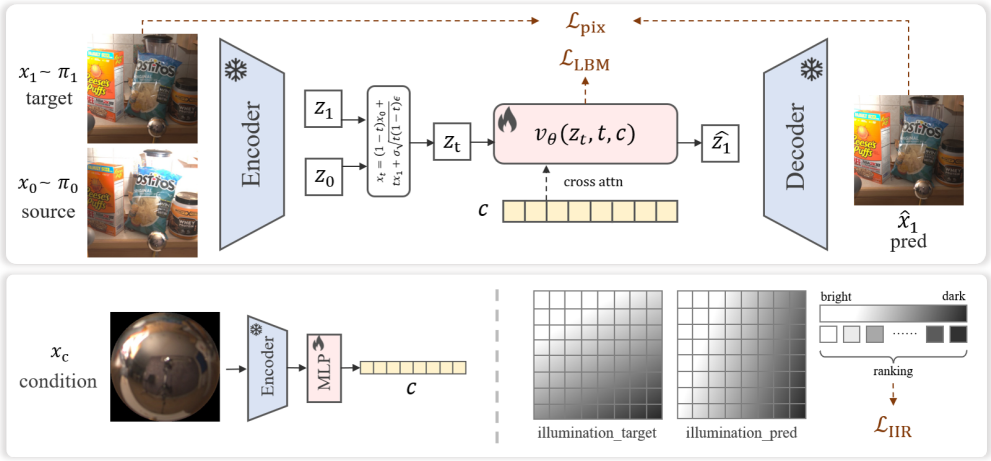

Figure 2: Overview of the proposed framework. Chrome probe images provide conditions that are injected into latent bridge matching through cross-attention. Spatial–illumination alignment provides an explicit constraint on illumination variations.

3.1.1 LATENT BRIDGE MATCHING

π_0 and π_1 denote the data distributions of the source and target domains. Sample paired image pairs $(x_0, x_1) \sim \pi_0 \times \pi_1$ from the two distributions. The core objective of bridge matching is to learn a transport mapping from π_0 to π_1 (Albergo et al., 2025), (Liu et al., 2022), (Shi et al., 2023).

E and D are the encoder and decoder of a VAE. Given an image pair (x_0, x_1), encode the images into the latent space:

$$z_0 = E(x_0), \quad z_1 = E(x_1). \tag{1}$$

Then construct a Brownian bridge interpolation in the latent space:

$$z_t = (1-t)z_0 + tz_1 + \sigma\sqrt{t(1-t)}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{2}$$

where $\epsilon \sim \mathcal{N}(0, I)$, and $\sigma \geq 0$ is the noise magnitude. The marginal distributions of the interpolation satisfy $z_0 \sim \pi_0$ and $z_1 \sim \pi_1$ at the two endpoints. At an intermediate time t, the interpolation performs a stochastic linear transition between the source and target samples. The corresponding stochastic differential equation (SDE) is given by:

$$dz_t = \frac{z_1 - z_t}{1-t}\,dt + \sigma dB_t, \tag{3}$$

where B_t denotes standard Brownian motion. The drift term

$$v_\theta(z_t, t, c) = \frac{z_1 - z_t}{1-t} \tag{4}$$

determines the direction of sample evolution along the temporal dimension.

The drift estimation network $v_\theta(z_t, t, c)$ is employed to approximate the drift term. The training objective for latent bridge matching is defined as

$$\mathcal{L}_{\text{LBM}} = \mathbb{E}_{t,z_0,z_1,\epsilon}\left[\left\|\frac{z_1 - z_t}{1-t} - v_\theta(z_t, t, c)\right\|^2\right], \tag{5}$$

where $(z_0, z_1) = (E(x_0), E(x_1))$ and $t \sim \pi(t)$ denotes the timestep.

During inference, starting from the latent representation z_0 of the source image x_0, latent bridge matching iteratively solves the latent-space SDE to $\hat{z}_1$. The generated latent $\hat{z}_1$ is then mapped back to the pixel space with the decoder D, $\hat{x}_1 = D(\hat{z}_1)$.

To further constrain the texture and fine-grained details, we additionally introduce LPIPS loss on VGG16 as the pixel-wise loss:

$$\mathcal{L}_{\text{pix}} = LPIPS_{VGG16}(\hat{x}_1, x_1). \tag{6}$$

3.1.2 Condition Injection

Prior work (Chadebec et al., 2025) injects conditions by concatenating the illumination reference with the noisy feature. With direct concatenation, the conditioning signal tends to be overwhelmed by noise, causing the prediction to deviate from the condition. In the MIIW dataset (Murmann et al., 2019), each illumination direction is associated with a chrome probe. Encode the chrome probe image x_c to z_c and compress z_c into the token sequence c by a lightweight tokenizer. To prevent the tokenizer from producing arbitrary outputs that render the conditioning ineffective, we conduct condition-confirmed initialization. Following prior work (Zheng et al., 2026a), we initialize the tokenizer with a reconstruction objective, forcing the token sequence to retain the discriminative information. During joint training, the token sequence c is provided as the conditioning signal to the cross-attention modules of the U-Net. With condition-confirmed initialization, condition injection enables the U-Net to dynamically aggregate the illumination information, synthesizing the target illumination.

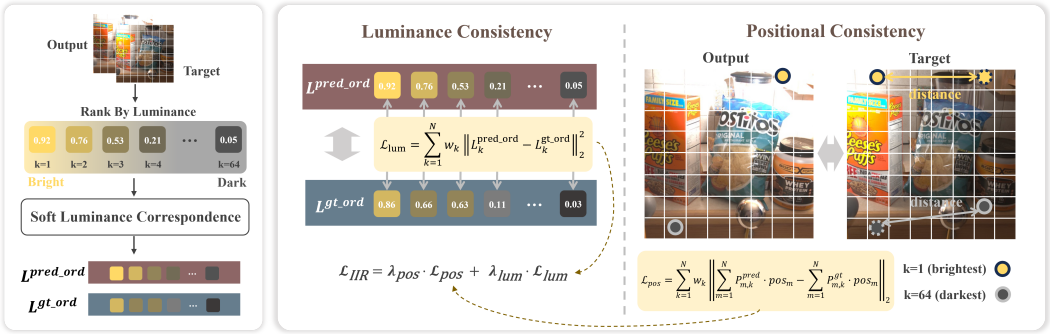

Figure 3: Illumination Intensity Ranking (IIR) loss. We divide each image into patches, rank by luminance and gain the luminance sequence via soft luminance correspondence. The IIR loss constrains illumination variations by comparing the luminance values and spatial positions of patches at corresponding ranks.

3.2 Spatial–Illumination Alignment

Pixel-wise alignment cannot explicitly model illumination variations, while spatial-illumination alignment constrains luminance variations across regions. As shown in Figure 3, we divide the image into patches, rank different regions by luminance, and locate illumination variations via the ranking order. To make the discrete ranking differentiable, we adopt a soft ranking strategy to generate soft luminance correspondence. To jointly optimize from both luminance and spatial arrangement, we constrain luminance consistency and positional consistency with regional luminance values and positions.

3.2.1 SOFT LUMINANCE CORRESPONDENCE

For the predicted image I^{pred} and the ground-truth image I^{gt}, we partition each image into a $G \times G$ grid, getting $N = G^2$ patches (with $G = 8$ and $N = 64$ in our experiments). After average pooling, each patch is converted to luminance by the ITU-R BT.601 standard. Sort the patches of the output and the target in descending order by luminance to obtain $L^{\text{pred}}, L^{\text{gt}} \in \mathbb{R}^N$.

For soft matching, we adopt the soft sorting method (Prillo & Eisenschlos, 2020). Given a luminance sequence $L \in \mathbb{R}^N$ and its k-th ranked luminance $\widetilde{L}_k$, we define a soft luminance correspondence $P \in \mathbb{R}^{N \times N}$, where $P_{m,k}$ denotes the soft assignment probability of the m-th spatial patch to the k-th luminance rank:

$$P_{m,k} = \frac{\exp\left(-|L_m - \widetilde{L}_k|/\tau\right)}{\sum_{j=1}^{N} \exp\left(-|L_j - \widetilde{L}_k|/\tau\right)}, \tag{7}$$

where $\tau > 0$ controls the assignment softness. As $\tau \to 0^+$, the probability mass concentrates on the best-matching patch. During training, τ linearly decays from 1.0 to 0.1 to progressively sharpen the correspondence. For each fixed rank k, $\sum_{m=1}^{N} P_{m,k} = 1$. $P_{m,k}$ defines a conditional soft-correspondence distribution over spatial patches.

Applying Eq. 7 to the predicted and ground-truth images gains the soft luminance correspondences P^{pred} and P^{gt}, respectively.

3.2.2 ILLUMINATION INTENSITY RANKING LOSS

To achieve spatial-illumination alignment, we propose an Illumination Intensity Ranking (IIR) loss with luminance consistency and positional consistency to constrain the illumination intensity and the spatial distribution.

For luminance consistency, we aggregate the target soft correspondence P^{gt} and the predicted luminance L^{pred} to get the predicted luminance values under target rankings. The k-th luminance in the target and the corresponding predicted luminance are defined as

$$L_k^{\text{gt_ord}} = \sum_{m=1}^{N} P_{m,k}^{\text{gt}} \cdot L_m^{\text{gt}}, \qquad L_k^{\text{pred_ord}} = \sum_{m=1}^{N} P_{m,k}^{\text{gt}} \cdot L_m^{\text{pred}}. \tag{8}$$

The luminance consistency loss measures the weighted difference between the target-aligned predicted luminance and the corresponding target luminance:

$$\mathcal{L}_{\text{lum}} = \sum_{k=1}^{N} w_k \left\| L_k^{\text{pred_ord}} - L_k^{\text{gt_ord}} \right\|_2^2, \tag{9}$$

where

$$w_k = \frac{e^{-k/N}}{\sum_{j=1}^{N} e^{-j/N}}. \tag{10}$$

The exponentially decaying weights assign greater importance to brighter ranks that are sensitive to illumination variation.

To further refine the spatial distribution of luminance, we define the positional consistency loss as

$$\mathcal{L}_{\text{pos}} = \sum_{k=1}^{N} w_k \left\| \sum_{m=1}^{N} P_{m,k}^{\text{pred}} \cdot pos_m - \sum_{m=1}^{N} P_{m,k}^{\text{gt}} \cdot pos_m \right\|_2, \tag{11}$$

where pos_m denotes the normalized spatial coordinate of the m-th patch. $\mathcal{L}_{\text{pos}}$ measures the difference between the spatial locations with the same luminance rank in the predicted and ground-truth images.

Together, the two terms align luminance ranks with spatial regions, providing structured supervision for both illumination intensity and spatial redistribution. The IIR loss is formulated as

$$\mathcal{L}_{\text{IIR}} = \lambda_{\text{lum}} \mathcal{L}_{\text{lum}} + \lambda_{\text{pos}} \mathcal{L}_{\text{pos}}, \tag{12}$$

where $\lambda_{\text{lum}} = 1.5$ and $\lambda_{\text{pos}} = 0.5$ in our experiments.

The overall training objective is

$$\mathcal{L} = \lambda_{\text{IIR}}\mathcal{L}_{\text{IIR}} + \lambda_{\text{LBM}}\mathcal{L}_{\text{LBM}} + \lambda_{\text{pix}}\mathcal{L}_{\text{pix}}, \tag{13}$$

where $\lambda_{\text{IIR}} = 5$, $\lambda_{\text{LBM}} = 1$, and $\lambda_{\text{pix}} = 5$ during experiments.

4 Experiments

4.1 Implementation Details

We conduct experiments on the MIIW dataset (Murmann et al., 2019) at a resolution of 512×512. Multi-Illumination provides 25 illumination conditions with 985 scenes for training and 30 scenes for testing, resulting in 591,000 training pairs and 18,000 testing pairs. Each pair consists of the source image, the target image, and the chrome probe as illumination reference. We adopt latent bridge matching with $\sigma = 0.2$, the timestep t uniformly sampled from $\{0, 0.25, 0.5, 0.75\}$ in training, and a single-step Euler update with $t = 0$ during inference. The denoiser is the SDXL-base-1.0 U-Net. We optimize the U-Net with AdamW with $\beta = (0.9, 0.999)$ and a weight decay of 0.01. The learning rates are 1×10^{-5} for the U-Net and 5×10^{-4} for the tokenizer, with per-step cosine annealing ($T_{\max} = 30{,}000$, $\eta_{\min} = 10^{-7}$). We initialize the model with pretrained weights from the image relighting method(Chadebec et al., 2025) and train for 30,000 steps with a batch size of 4 using bf16 mixed precision on a single A800 GPU.

4.2 Ablation Study

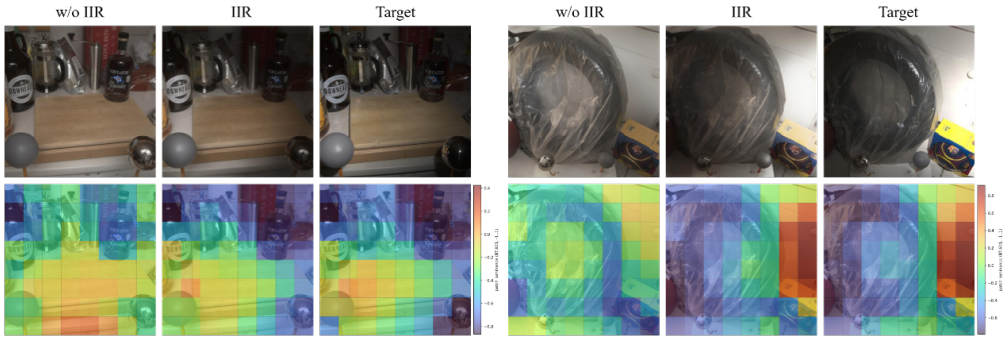

Figure 4: Effect of spatial–illumination alignment on regional luminance. The first row shows images, and the second row shows the corresponding luminance heatmaps. Spatial–illumination alignment improves controllability over both the location and intensity of luminance.

We conduct ablation studies to evaluate the contribution of each component. As shown in Table 1, injection by attention without condition-confirmed initialization gains results comparable to results without conditions. Directly concatenating the condition with the noisy latent provides a modest improvement, while cross-attention-based injection with condition-confirmed initialization substantially improves performance. For spatial–illumination consistency, IIR loss further improves image quality and illumination similarity. As illustrated in Figure 4, omitting the IIR loss produces a luminance distribution that deviates considerably from the ground truth. Spatial–illumination alignment improves the localization of bright regions and aligns regional luminance with ground-truth levels.

Table 1: Ablation studies on condition injection and spatial–illumination alignment.

Condition Injection	IIR Loss	PSNR↑	RMSE↓	SSIM↑
×	×	16.47	0.1649	0.6106
Cross Attn	×	16.95	0.153	0.6287
Concatenation	×	17.58	0.1406	0.6288
Cross Attn + CCI	×	19.61	0.113	0.7167
Cross Attn +CCI	✓	**21.17**	**0.0932**	**0.7701**

We further ablate the components of the IIR loss. Figure 5 illustrates the different roles of the two consistencies. The luminance consistency adjusts brightness across all regions, while the positional consistency refines the spatial arrangement by the single sorting. In the first row showing concentrated illumination, luminance consistency establishes the overall distribution while positional consistency identifies the brightest region and suppresses excessive bright regions. In the second row showing dispersed illumination, luminance consistency adjusts the brightness of the broader illuminated area, while positional consistency refines the brightest region. Additional details of IIR components are provided in Appendix A.

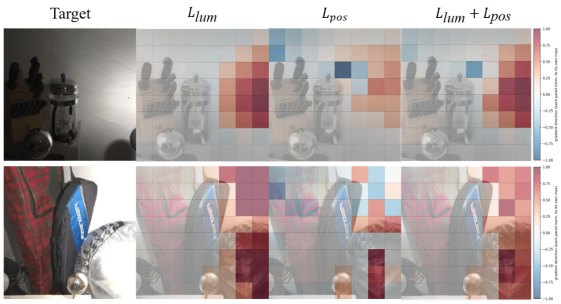

Figure 5: Gradient directions induced by the two IIR components. The luminance consistency loss adjusts brightness across all regions, while the positional consistency loss provides distinct gradients to refine the locations of bright regions.

4.3 Quantitative Evaluation

To evaluate the benefit of spatial–illumination alignment, we compare our method with pixel-aligned approaches. Table 2 reports results for latent representation learning (Rombach et al., 2022), latent bridge matching (Chadebec et al., 2025), cross-scene illumination editing (Xing et al., 2025), and light transport decomposition (Zhang et al., 2025). Compared with pixel-aligned methods, our approach improves PSNR by 3.6 dB and SSIM by 8.3%, while reducing RMSE by 33%.

Table 2: Quantitative comparison on MIIW. The spatial–illumination alignment method achieves the best results across all reported metrics.

Metrics	PSNR↑	RMSE↓	SSIM↑
Source	15.81	0.1901	0.7103
LDM	15.92	0.1792	0.6152
LBM	17.58	0.1406	0.6288
LumiNet	16.07	0.1682	0.6871
IC-Light	17.43	0.1482	0.6314
Ours	**21.17**	**0.0932**	**0.7701**

Table 3 reports the evaluation on illumination controllability. We assess brightness levels using mean luminance error. To measure localization accuracy of illumination, we define the brightest 20% of pixels as the bright region and compute the intersection over union (IoU). To evaluate fidelity to the lighting condition, we consider 24 images of the same scene under different lighting conditions and measure the probability that the target image is the closest match to the output in terms of luminance similarity. Removing IIR supervision reduces the target recall accuracy, indicating spatial–illumination alignment facilitates the controllability of condition. Compared with competing methods, our method accurately predicts both luminance levels and bright regions matching illumination conditions, substantially improving the controllability in indoor scene images relighting.

Table 3: Quantitative evaluation of illumination controllability.

	Mean Luminance Error ↓	Bright Region IoU ↑	Target Recall@1 ↑
Source	0.0968	43.12%	0%
LumiNet	0.0925	49.70%	20.87%
IC-Light	0.0637	53.80%	23.38%
LBM	0.0610	49.66%	6.47%
Ours w/o IIR	0.0326	58.98%	26.53%
Ours	**0.0146**	**67.93%**	**54.39%**

4.4 Qualitative Comparation

Figure 6 compares the outputs of different methods and the corresponding luminance heatmaps. Under both the uniform or point-source illumination, spatial–illumination alignment precisely cali-

brates regional brightness across all scenes. Our method predicts luminance distributions consistent with the targets. Figure 7 shows the results from the same source image under different target conditions. Our method produces distinct outputs for different conditions while consistently matching the target illumination distributions.

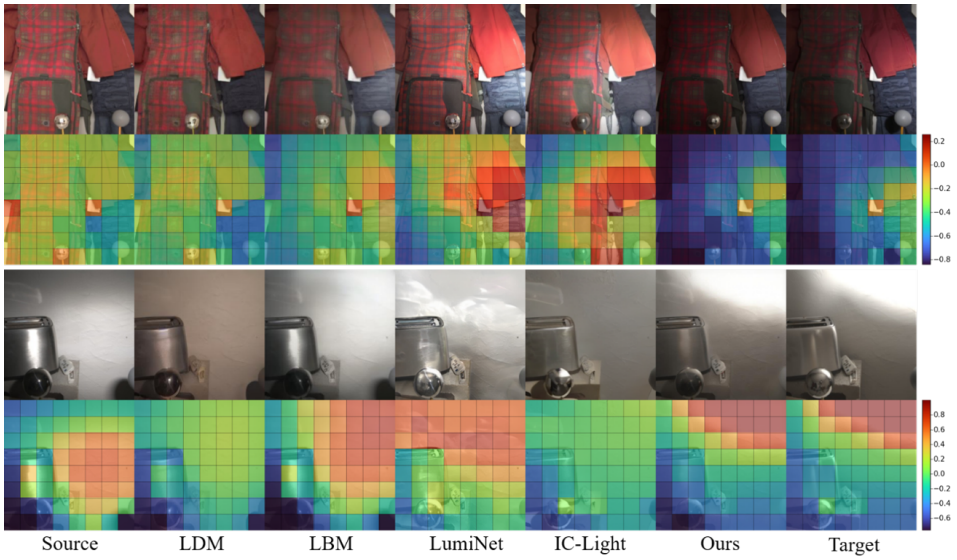

Figure 6: Visual comparison of relighting results and luminance heatmaps on MIIW.

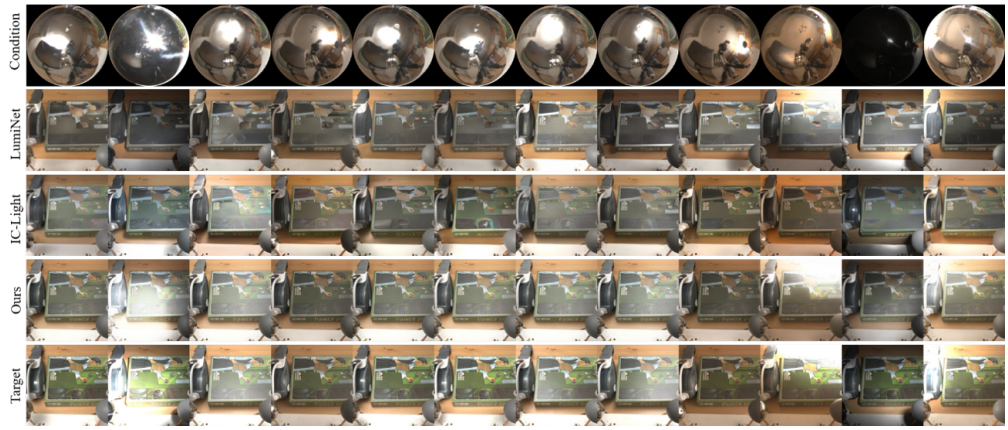

Figure 7: Visual comparison under different illumination conditions.

5 Conclusion

We presented spatial–illumination alignment for controllable image illumination editing. We introduce regional brightness and spatial arrangement to jointly characterize illumination variations. The proposed IIR loss captures spatial-illumination alignment through luminance ranking, coupling luminance consistency with positional consistency to guide illumination redistribution. Experiments on MIIW demonstrate improved performance over state-of-the-art methods while ablation studies support the effectiveness of components. Spatial–illumination alignment introduces the explicit illumination supervision for image illumination editing, providing a pioneering perspective on modeling illumination transformations. Although the IIR loss improves the prediction of illumination redistribution, IC-LBM still exhibits local color shifts and limited recovery of fine details. We will further incorporate geometric constraints to improve structural fidelity.

AI USE STATEMENT

In this work, we used generative AI tools for literature searches, language polishing, implementation assistance, and feasibility analysis. We did not use AI to generate research ideas. We reviewed all AI-assisted work and manually verified the literature references provided by large language models (LLMs). Both authors reviewed the manuscript text polished by LLMs and validated and tested LLM-generated code to ensure its correctness. We take responsibility for the final content of this study, including any text, claims, or results produced with the assistance of generative AI.

REFERENCES

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan T. Barron, and Hendrik P. A. Lensch. Neural-pil: Neural pre-integrated lighting for reflectance decomposition. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Junuk Cha, Mengwei Ren, Krishna Kumar Singh, He Zhang, Yannick Hold-Geoffroy, Seunghyun Yoon, HyunJoon Jung, Jae Shin Yoon, and Seungryul Baek. Text2relight: Creative portrait relighting with text guidance. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2): 1980–1988, 2025. doi: 10.1609/aaai.v39i2.32194.

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.

Zhuo Chen, Chengqun Yang, Zhuo Su, Zheng Lv, Jingnan Gao, Xiaoyuan Zhang, Xiaokang Yang, and Yichao Yan. Polar: A portrait olat dataset and generative framework for illumination-aware face modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28871–28881, 2026.

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

Boyang Deng, Yifan Wang, and Gordon Wetzstein. Lumigan: Unconditional generation of relightable 3d human faces. In *2024 International Conference on 3D Vision (3DV)*, pp. 302–312. IEEE, 2024. doi: 10.1109/3DV62453.2024.00081.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Minghui Hu, Jianbin Zheng, Daqing Liu, Chuanxia Zheng, Chaoyue Wang, Dacheng Tao, and Tat-Jen Cham. Cocktail: Mixing multi-modality controls for text-conditional image generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Vincent Tao Hu, Wei Zhang, Meng Tang, Pascal Mettes, Deli Zhao, and Cees Snoek. Latent space editing in transformer-based flow matching. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(3):2247–2255, 2024. doi: 10.1609/aaai.v38i3.27998.

Haian Jin, Yuan Li, Fujun Luan, Yuanbo Xiangli, Sai Bi, Kai Zhang, Zexiang Xu, Jin Sun, and Noah Snavely. Neural gaffer: Relighting any object via diffusion. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Sungnyun Kim, Junsoo Lee, Kibeom Hong, Daesik Kim, and Namhyuk Ahn. Diffblender: Composable and versatile multimodal text-to-image diffusion models. *Expert Systems with Applications*, 297:129345, 2026. doi: 10.1016/j.eswa.2025.129345.

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

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

Jonghyun Lee, Hansam Cho, YoungJoon Yoo, Seoung Bum Kim, and Yonghyun Jeong. Compose and conquer: Diffusion-based 3d depth aware composable image synthesis. In *International Conference on Learning Representations*, 2024.

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

Zhexin Liang, Zhaoxi Chen, Yongwei Chen, Tianyi Wei, Tengfei Wang, and Xingang Pan. Pi-light: Physics-inspired diffusion for full-image relighting. In *International Conference on Learning Representations*, 2026.

Yehonathan Litman, Fernando De la Torre, and Shubham Tulsiani. Lightswitch: Multi-view relighting with material-guided diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27750–27759, 2025. doi: 10.1109/ICCV51701.2025.02576.

Hanzhou Liu, Peng Jiang, Jia Huang, and Mi Lu. Lumos3d: A single-forward framework for low-light 3d scene restoration. *arXiv preprint arXiv:2511.09818*, 2025a.

Pengwei Liu, Hangjie Yuan, Bo Dong, Jiazheng Xing, Jinwang Wang, Rui Zhao, Weihua Chen, and Fan Wang. Unilumos: Fast and unified image and video relighting with physics-plausible feedback. In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

Xingchao Liu, Lemeng Wu, Mao Ye, and Qiang Liu. Let us build bridges: Understanding and extending diffusion generative models. *arXiv preprint arXiv:2208.14699*, 2022.

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 *ACM SIGGRAPH 2025 Conference Papers*, 2025. doi: 10.1145/3721238.3730696.

Yiqun Mei, Yu Zeng, He Zhang, Zhixin Shu, Xuaner Zhang, Sai Bi, Jianming Zhang, HyunJoon Jung, and Vishal M. Patel. Holo-relighting: Controllable volumetric portrait relighting from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4263–4273, 2024.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5):4296–4304, 2024. doi: 10.1609/aaai.v38i5.28226.

Lukas Murmann, Michael Gharbi, Miika Aittala, and Fredo Durand. A dataset of multi-illumination images in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4080–4089, 2019.

Sebastian Prillo and Julian Eisenschlos. Softsort: A continuous relaxation for the argsort operator. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7793–7802. PMLR, 2020.

Pramod Rao, Gereon Fox, Abhimitra Meka, Mallikarjun B R, Fangneng Zhan, Tim Weyrich, Bernd Bickel, Hanspeter Pfister, Wojciech Matusik, Mohamed Elgharib, and Christian Theobalt. Lite2relight: 3d-aware single image portrait relighting. In *ACM SIGGRAPH 2024 Conference Papers*, 2024. doi: 10.1145/3641519.3657470.

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.

David Serrano-Lozano, Anand Bhattad, Luis Herranz, Jean-François Lalonde, and Javier Vazquez-Corral. Synclight: Controllable and consistent multi-view relighting. *arXiv preprint arXiv:2601.16981*, 2026.

Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion schrödinger bridge matching. In *Advances in Neural Information Processing Systems*, volume 36, pp. 62183–62223, 2023. doi: 10.52202/075280-2717.

Yanan Sun, Yanchen Liu, Yinhao Tang, Wenjie Pei, and Kai Chen. Anycontrol: Create your artwork with versatile control on text-to-image generation. In *European Conference on Computer Vision*, pp. 92–109. Springer, 2024.

Lezhong Wang, Duc Minh Tran, Ruiqi Cui, Thomson TG, Anders Bjorholm Dahl, Siavash Arjomand Bigdeli, Jeppe Revall Frisvad, and Manmohan Chandraker. Materialist: Physically based editing using single-image inverse rendering. *International Journal of Computer Vision*, 134:267, 2026a. doi: 10.1007/s11263-026-02833-z.

Shan Wang, Peixia Li, Chenchen Xu, Ziang Cheng, Jiayu Yang, Hongdong Li, and Pulak Purkait. Joint shadow generation and relighting via light-geometry interaction maps. In *International Conference on Learning Representations*, 2026b.

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

Yuxuan Xue, Ruofan Liang, Egor Zakharov, Timur Bagautdinov, Chen Cao, Giljoo Nam, Shunsuke Saito, Gerard Pons-Moll, and Javier Romero. Georelight: Learning joint geometrical relighting and reconstruction with flexible multi-modal diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29771–29780, 2026.

Qirui Yang, Yang Yang, Ying Zeng, Xiaobin Hu, Bo Li, Huanjing Yue, Jingyu Yang, and Peng-Tao Jiang. Cameramaster: Unified camera semantic-parameter control for photography retouching. *arXiv preprint arXiv:2511.21024*, 2025.

Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. Dilightnet: Fine-grained lighting control for diffusion-based image generation. In *ACM SIGGRAPH 2024 Conference Papers*, 2024.

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 *International Conference on Learning Representations*, 2025.

Xiao Zhang, William Gao, Seemandhar Jain, Michael Maire, David A. Forsyth, and Anand Bhattad. Latent intrinsics emerge from training to relight. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Zitian Zhang, Iliyan Georgiev, Michael Fischer, Yannick Hold-Geoffroy, Jean-François Lalonde, and Valentin Deschaintre. Unilight: A unified representation for lighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29684–29694, 2026.

Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K. Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Guangting Zheng, Yiyuan Zhang, Tao Yang, Yunpeng Chen, Rui Zhu, Jiajun Deng, and Yanyong Zhang. Genfirst: Generation before reconstruction for stable end-to-end latent generative modeling. *arXiv preprint arXiv:2608.29335*, 2026a.

Haoze Zheng, Zihao Wang, Xianfeng Wu, Yajing Bai, Yexin Liu, Yun Li, Xiaogang Xu, and Harry Yang. Learning latent proxies for controllable single-image relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b.

Linqi Zhou, Aaron Lou, Samar Khanna, and Stefano Ermon. Denoising diffusion bridge models. In *International Conference on Learning Representations*, 2024.

A Details of the IIR Loss

To ensure the reproducibility of the method, we have uploaded the core code to an anonymous repository: https://anonymous.4open.science/r/anonymous-code-E3C4. We further conduct ablation studies on the number of patches and the relative weighting of the two components of the IIR loss. Figure 8 shows the effects of varying the number of patches and the relative weights of the IIR components. When the number of patches is small, the performance remains close to the results without IIR. As the number of patches increases, the model captures illumination variations more accurately. Removing either the luminance consistency or the positional consistency results in degraded performance.

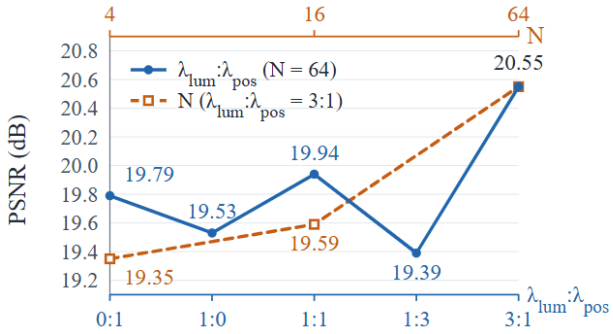

Figure 8: Ablation of loss weights and patch count. We compare different weight ratios of $\mathcal{L}_{\text{lum}}$ to $\mathcal{L}_{\text{pos}}$ and different numbers of patches. All results are obtained at a resolution of 512×512, with a batch size of 4, after $10{,}000$ training steps.

We also conduct an ablation study on the weighting scheme w_k. As shown in Table 4, removing the rank-based weighting applied to the top-k brightest patches leads to a slight decrease in performance.

Table 4: Ablation study of the weighting strategy. All results are obtained at a resolution of 512×512, with a batch size of 4, after $10{,}000$ training steps.

	PSNR↑	RMSE↓	SSIM↑
w/o w_k weighting	19.87	0.1159	0.6440
with w_k weighting	20.55	0.1072	0.7133

We evaluate the computational overhead introduced by IIR. Since IIR reuses the computational graph of the pixel-wise loss, it incurs minimal additional overhead. At a resolution of 512×512, with $N = 64$ patches and a batch size of 4, the IIR loss introduces an additional 9.76 MFLOPs, accounting for only 0.0005% of the total loss computation.

Finally, we discuss how IIR operates in several special cases. When both the target and the output are under uniform illumination, the luminance values across patches are nearly identical, making the ranking constraint less informative. In this case, the soft luminance correspondence approaches

$$P_{m,k} \approx \frac{1}{N}, \qquad \forall\, m, k, \tag{14}$$

and IIR accordingly reduces to evaluating luminance differences without introducing artificial distinctions caused by ranking.

When an image contains multiple bright regions, the soft sorting computes the luminance correspondence probabilistically, allowing $\mathcal{L}_{\mathrm{lum}}$ to capture the luminance distribution with multiple light sources. Meanwhile, $\mathcal{L}_{\mathrm{pos}}$ consistently constrains the spatial centroid consistency, further refining the estimated locations of the bright regions.

B Additional Luminance Visualizations

To further demonstrate the effectiveness of spatial–luminance alignment, we provide additional qualitative comparisons with competing methods. Figure 9 compares the generated images and the corresponding luminance heatmaps. Our method more accurately captures luminance variations across diverse scenes and illumination conditions.

C Additional Visualizations within the Same Scene

To demonstrate the improved controllability of our method for image relighting, we present additional predictions for the same scene under different lighting conditions. Figure 10 compares results of the image captured under 14th-direction (central, uniform illumination) in MIIW as the source image and different chrome probes as conditions.

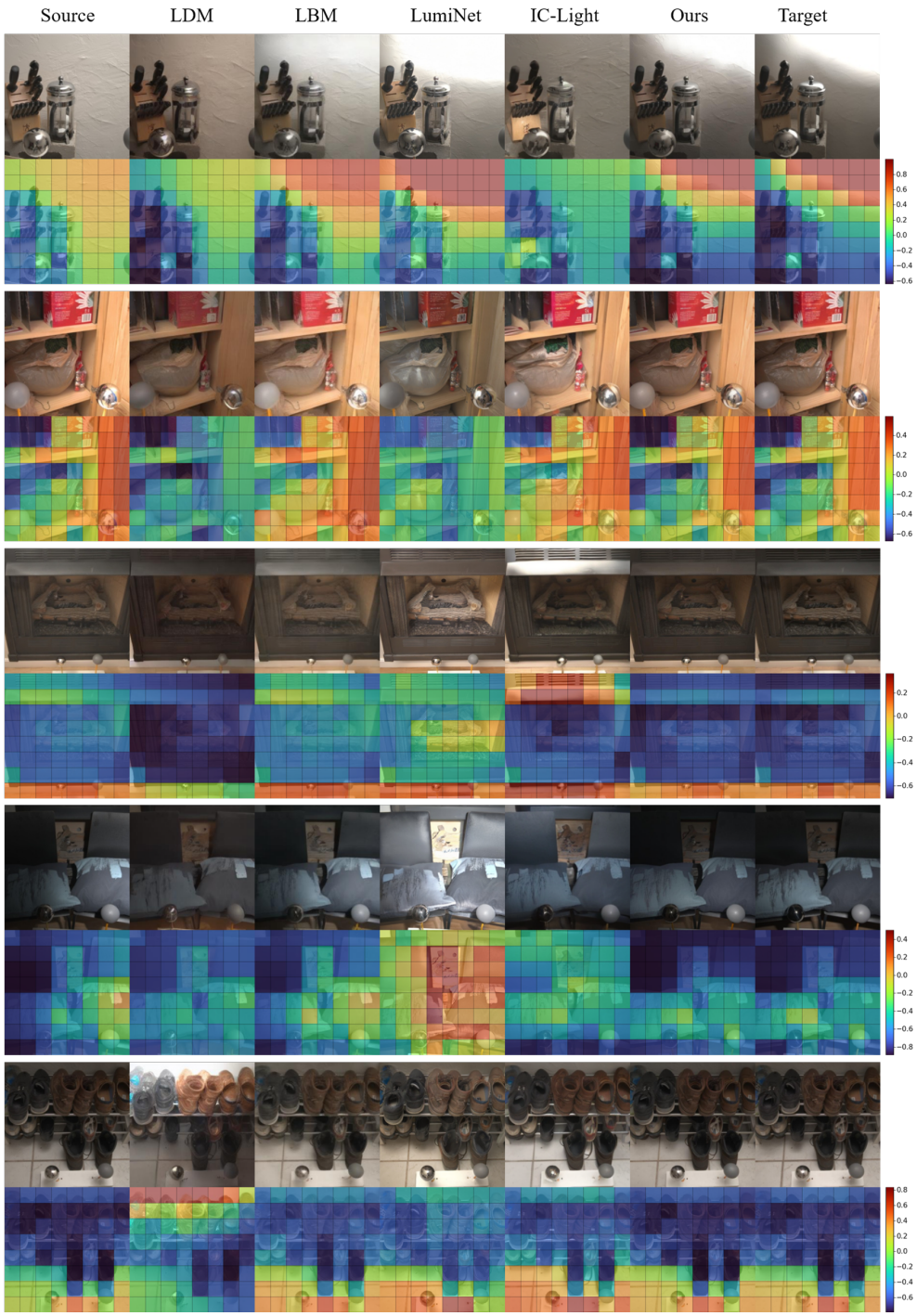

Figure 9: Qualitative comparisons with competing methods. Our method consistently produces results matching the targets across diverse scenes and varying illumination intensities.



Figure 10: Comparisons with competing methods within the same scene. Our method accurately predicts bright region locations and illumination distributions across lighting conditions in the same scene.



Figure 10: Comparisons with competing methods within the same scene. Our method accurately predicts bright region locations and illumination distributions across lighting conditions in the same scene. (continued)