PIXLRelight: Controllable Relighting via Intrinsic Conditioning

Miguel Farinha **Ronald Clark**
Department of Computer Science
University of Oxford
{miguel.farinha,ronald.clark}@cs.ox.ac.uk

Abstract

We present PIXLRELIGHT, a feed-forward approach for physically controllable single-image relighting. Existing methods either provide limited lighting control (e.g. through text or environment maps), accumulate errors when chaining inverse and forward rendering, or require costly per-image optimization. Our key idea is to bridge physically based rendering (PBR) and learned image synthesis through a shared intrinsic conditioning that can be obtained from either real photographs or PBR renders. At training time, paired multi-illumination photographs are decomposed into albedo, diffuse shading, and non-diffuse residuals, which condition the model. At inference time, the same conditioning is computed from a path-traced render of a coarse 3D reconstruction of the input under user-specified PBR lights. A transformer-based neural renderer then applies the target illumination to the source photograph, preserving fine image detail through a per-pixel affine modulation. PIXLRELIGHT enables arbitrary PBR-style lighting control, achieves state-of-the-art relighting quality, and runs in under a tenth of a second per image. Code and models are available at https://mlfarinha.github.io/pixl-relight/.

Source Image

Our Relit Predictions

Figure 1: **PIXLRELIGHT** is a feed-forward transformer that produces photorealistic relightings of in-the-wild photographs in a single pass. The user authors the target illumination in a physically based renderer, with full control over light type, position, color, and intensity. The model is conditioned on the resulting intrinsic decomposition of the target appearance – albedo, diffuse shading, and a non-diffuse residual. Above, the same source image is relit under six different illuminations.

Preprint.

1 Introduction

Illumination is a primary component of image formation. In computer graphics, physically based rendering (PBR) engines such as Blender [3] and Unreal Engine [10] expose lighting controls directly: an artist places point, area, directional, environmental, or emissive light sources in a 3D scene, and a path tracer simulates their interaction with geometry and materials. Bringing this same physical control to in-the-wild photographs would unlock applications across computational photography, content creation, and visual effects – but is fundamentally harder, since recovering the geometry, materials, and light transport from an image is challenging.

Recent approaches address this gap in three ways, none of which jointly enables physical control, photorealism, and speed. A first line conditions relighting on HDR environment maps [18, 49], reference images [47], or screen-space scribbles [7] and masks [33]; these interfaces struggle to express spatially localized, multi-source illumination. A second line treats relighting as an inverse-then-forward rendering pipeline that estimates G-buffers and re-renders them under a target illumination [11, 28, 40, 50]; the intermediate buffers cannot encode every cue the renderer needs (e.g. transparency, subsurface scattering), and errors compound across the two stages [14]. The third, and closest, line combines PBR with a neural renderer: Careaga and Aksoy [5] reconstruct an approximate textured mesh from a photograph, ray-trace it under user-specified illumination, and pass the CG render to a neural renderer that produces the final image. Two limitations follow. First, because the mesh is reconstructed with diffuse reflectance only, the ray-traced render is diffuse-only, and the network must guess every specular highlight, transparent surface, or refractive cue in the output. Second, training each input pair requires a per-image differentiable-rendering optimization that fits a 3D lighting environment to the source image; this optimization is stable only for a narrow lighting parameterization, bounding the lighting distribution the network sees at training.

The natural alternative is to decompose the photograph into geometry and materials, place new lights in PBR, and re-render. This is exactly what computer graphics does well – given high-quality assets. For in-the-wild photographs, those assets are precisely what we cannot recover reliably: single-image geometry is coarse and material decomposition is under-constrained, so a path-traced render of such a reconstruction is riddled with artifacts. Our key insight is to harness both the advantages of PBR (controllability) and neural rendering (detailed photorealism). Our insight is to use each tool for what it does well: PBR for specifying *what* the target lighting should be, and a feed-forward neural model trained on real photographs for *how* to apply that lighting to the photograph photorealistically – absorbing the imperfections of the underlying scene reconstruction in the process. The two are bridged by an intrinsic decomposition of the target appearance into albedo, diffuse shading, and a non-diffuse residual [20]. The same input is produced from a real photograph at training and from a PBR render at inference. At training, paired captures from existing multi-illumination datasets [15, 25, 34] pass through a frozen intrinsic decomposition model [20] to directly supervise relighting which means there is no inverse-then-forward chain to train or per-image rendering optimization to run. At inference, the conditioning is produced by a path-traced render of a coarse 3D reconstruction in which the user freely controls lighting using arbitrary combinations of physically-based lights. Crucially, the model never sees the rendered RGB image; it sees only the intrinsic buffers, which carry a precise cue for the desired output lighting.

We call this approach PIXLRELIGHT, a feed-forward transformer that consumes a source image and the target intrinsics and predicts a relit RGB image with per-pixel detail. Following recent feed-forward dense-prediction architectures [16, 17, 29, 42], we tokenize the two inputs with asymmetric encoders – a ViT [9] for the source image and a ConvNeXt [31] for the smoother, lower-frequency intrinsic stack – and fuse them in a shared transformer trunk read out by a DPT head [35]. Rather than regressing RGB directly, the head produces an identity-initialized per-pixel affine modulation of the source: at initialization the network reproduces the input exactly, and during training it learns only the residual lighting transformation, preserving photorealistic detail by construction.

In summary, our contributions are: (1) a target-appearance intrinsic decomposition as a single conditioning interface for single-image relighting, computed from a real image at training and from a PBR render at inference; (2) a direct training strategy supervised end-to-end on real multi-illumination photographs, without an inverse-then-forward rendering pipeline or per-image rendering optimization; and (3) PIXLRELIGHT, which achieves state-of-the-art relighting quality while running in under a tenth of a second per image.

2 Related work

Single-image relighting methods can be organized by the modality through which the user specifies the target lighting. Text and reference-image conditioning [47, 51] hallucinate plausible illumination but offer no physical control. Environment-map conditioning [14, 18, 28, 49] assumes lighting comes from infinity and cannot express the near-field sources common in indoor scenes. Parametric conditioning on a single light source placed in a reference view, as in GenLit [2] and SyncLight [38], restores spatial control but is not designed for multi-light, area-light, or emissive-geometry edits. Pixel-aligned intrinsic conditioning sidesteps both limitations: RGB↔X [50], Ouroboros [40], and V-RGBX [11] drive diffusion-based renderers from intrinsic buffers that include shading, but stop short of CG-style authoring. Closest to ours, Careaga and Aksoy [5] achieve Blender-authored physical control by routing a PBR render through a feed-forward neural renderer, but their pipeline requires a per-image differentiable-rendering optimization for training and a diffuse-only RGB render as conditioning. PIXLRELIGHT shares the Blender-as-authoring-interface design but conditions the network on rendered intrinsic fields rather than rendered RGB, and trains directly on paired multi-illumination captures.

Intrinsic image decomposition factors an RGB image into illumination-invariant material properties and illumination-dependent shading [1]. Recent diffusion-based decompositions [11, 20–22, 50] substantially improve over earlier hand-crafted [4, 12, 23] and supervised [26, 27, 45, 53] approaches. We adopt Marigold-IID-Lighting [20], which fine-tunes a pretrained diffusion backbone [37] on Hypersim [36] to produce albedo, diffuse shading, and a non-diffuse residual. We use it as a frozen extractor in both training and inference, and inherit its image-formation model as the bridge between real photographs and physically based renders.

Feed-forward dense prediction has steadily replaced iterative pipelines across vision. VGGT [42] estimates cameras, depth, point maps, and tracks for hundreds of views in a single pass, displacing the bundle-adjustment loop of classical SfM [13]; DUSt3R [43] and MASt3R [24] regress aligned pointmaps without explicit matching; RayZer [16] and LVSM [17] synthesize novel views without an intermediate 3D representation. A generic transformer trunk, scale, and task-specific supervision suffice to match or surpass much of the task-specific machinery they replace. Single-image relighting has predominantly remained in the diffusion paradigm [11, 14, 28, 50, 51], whose iterative sampling is slow and whose generative prior can drift from the input image – a serious problem in relighting, where most pixels should remain photometrically faithful to the source. Even within diffusion, recent work pursues single-step formulations explicitly motivated by inference latency [38, 40]. We adopt the feed-forward recipe from the outset, with a per-pixel modulation parameterization that preserves source identity by construction.

3 Method

We present PIXLRELIGHT, a feed-forward transformer that relights a single input image to match a user-specified target lighting condition. The target condition is supplied as an intrinsic decomposition of the desired appearance, which serves as a unified conditioning interface for both training and inference. We define the problem in Sec. 3.1, describe the architecture in Sec. 3.2, and detail training and inference in Secs. 3.3 and 3.4.

3.1 Problem definition

Let $I_S \in [0,1]^{3\times H\times W}$ be a source RGB image of a static scene under lighting L_S, and let $I_T \in [0,1]^{3\times H\times W}$ be the same scene under a different target lighting L_T. We adopt the intrinsic image-formation model followed by Ke et al. [20],

$$I \;=\; A \odot S \;+\; R, \tag{1}$$

where $A \in [0,1]^{3\times H\times W}$ is the diffuse albedo, $S \in \mathbb{R}^{3\times H\times W}_{\geq 0}$ is the diffuse shading, $R \in \mathbb{R}^{3\times H\times W}$ is the non-diffuse residual capturing specular highlights, transparency, and other non-Lambertian effects, and $\odot$ denotes element-wise multiplication. The triplet (A_T, S_T, R_T) thus fully encodes the target appearance under L_T: the albedo is shared with the source by construction, while S_T and R_T together carry every change induced by the target lighting.

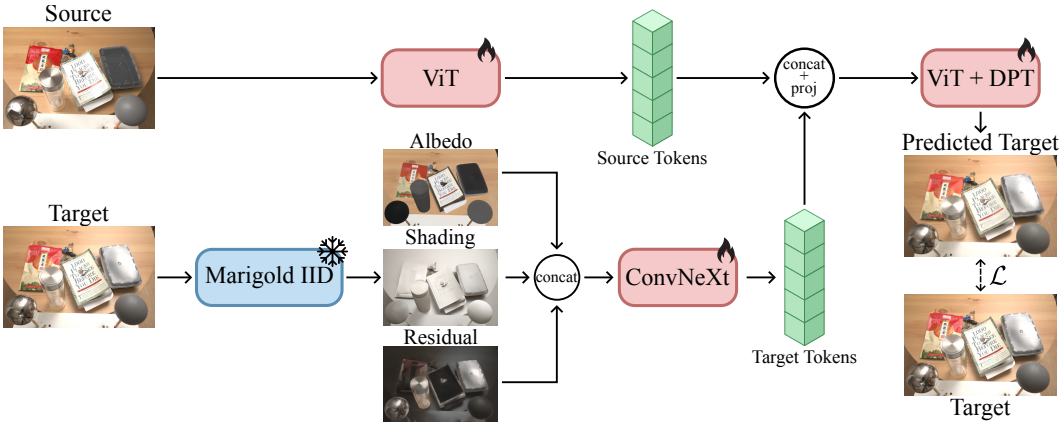

Figure 2: **Training pipeline.** The source image is patchified by a ViT branch and the channel-wise concatenated target intrinsics – extracted from the target image by a frozen Marigold-IID-Lighting model – are patchified by a ConvNeXt branch. The two token grids are fused per spatial location, projected to a common dimension, and processed by a self-attention transformer trunk. A DPT head reads out intermediate trunk features and predicts a per-pixel affine modulation of the source. Training is supervised end-to-end against the target image with pixel and perceptual losses.

We represent the target lighting condition through the channel-wise concatenation of the three target intrinsic maps:

$$C_T \;=\; [\,A_T\,;\,S_T\,;\,R_T\,] \;\in\; \mathbb{R}^{9\times H\times W}. \tag{2}$$

We seek a function f_θ such that

$$\hat{I}_T \;=\; f_\theta(I_S, C_T) \;\approx\; I_T. \tag{3}$$

C_T is the only carrier of lighting information seen by the model, and the same interface is used regardless of how C_T is produced: from a real photograph at training (Sec. 3.3), and from a physically based render at inference (Sec. 3.4).

3.2 Architecture

PIXLRELIGHT is a transformer that consumes the source image and the target intrinsics jointly and produces a relit RGB image. Its design follows recent feed-forward transformer architectures for dense prediction [16, 17, 29, 42]: a pair of asymmetric encoders feeds a shared transformer trunk, whose intermediate features are read out by a DPT head [35]. An overview is shown in Fig. 2.

Asymmetric Feature Encoders. The two inputs have very different spatial statistics, and we tokenize them accordingly. The source image I_S, dominated by high-frequency content, is patchified by a Vision Transformer (ViT) [9, 41]; the intrinsic stack C_T, dominated by smooth, lower-frequency structure, is patchified by a ConvNeXt [31, 46], whose convolutional inductive bias suits smoother inputs. Both branches use a patch size of p and produce token grids of size $(H/p) \times (W/p)$ with embedding dimension d. We then perform per-location fusion: at each spatial position, we concatenate the source and conditioning tokens channel-wise and project to dimension d with a small MLP, yielding a single fused token grid that is the input to the transformer trunk. Per-location fusion preserves the spatial layout of the conditioning and halves the sequence length compared with sequence-level concatenation.

Transformer Trunk. The fused token sequence is processed by a stack of L self-attention transformer blocks. We prepend a small set of learnable register tokens [8] to provide a global communication channel, and apply two-dimensional Rotary Positional Embeddings [39] to the patch tokens to encode their spatial layout. Following DPT-style dense prediction [35, 48], we expose the outputs of four intermediate blocks for multi-scale fusion in the head.

Modulation Head. The four intermediate token streams are first converted to a single dense feature map $F \in \mathbb{R}^{C' \times H \times W}$ with a DPT layer [35], then mapped with a 1×1 convolution to a six-channel output. Rather than regressing the relit RGB image directly, we parameterize the output as a per-pixel affine modulation of the source: most of the high-frequency content of $\hat{I}_T$ is already present in I_S, and asking the network to reproduce it from scratch wastes capacity that should be spent on transferring lighting. We split the six channels into a gain map $g \in \mathbb{R}^{3\times H\times W}$ and a bias map $b \in \mathbb{R}^{3\times H\times W}$, and form the prediction as

$$\hat{I}_T \;=\; \mathrm{clip}((1+g)\odot I_S + b,\; 0,\; 1)\,. \tag{4}$$

The output convolution is initialized to zero, so $g \equiv 0$ and $b \equiv 0$ at initialization and the network outputs $\hat{I}_T = I_S$ exactly. The model thus starts from an identity prior and learns only the residual lighting transformation. Gradients flow through the clip wherever the prediction lies inside $[0, 1]$, which holds at initialization and continues to hold during training.

3.3 Training

Training Data. We train on paired captures of static scenes under varying illumination, combining three datasets: the MIT Multi-Illumination Images in the Wild dataset (MIIW) [34], with 985 scenes captured under 25 different artificial flash conditions; BigTime [25], with 212 time-lapse scenes captured under varying natural illumination; and VIDIT [15], with 300 synthetic Unreal Engine scenes rendered under 40 lighting conditions formed by 5 color temperatures and 8 light directions. Together they span controlled artificial and uncontrolled natural lighting, real and synthetic captures, and indoor and outdoor scenes. Although smaller than the unpaired photo collections used by self-supervised relighting methods, these datasets provide paired multi-illumination supervision, which directly trains the photometric transfer we need without requiring per-image rendering optimizations [5]. For each batch, we randomly sample two images of the same scene and randomly assign them the roles of source and target, yielding dense supervision for arbitrary directional lighting changes between any two captured conditions. The target intrinsics C_T are produced on the fly by a frozen Marigold-IID-Lighting model [20] run for a single denoising step.

Training Loss. We supervise the predicted relit image directly against the ground-truth target image with a photometric loss,

$$\mathcal{L}(\hat{I}_T, I_T) \;=\; \|\hat{I}_T - I_T\|_1 \;+\; \lambda\, \mathcal{L}_{\mathrm{perc}}(\hat{I}_T, I_T), \tag{5}$$

where $\mathcal{L}_{\mathrm{perc}}$ is a VGG-based perceptual loss [19, 52] and λ balances the two terms. The gain and bias maps are not supervised directly; they emerge from end-to-end training with the only objective being to match the target image.

Implementation Details. The RGB encoder is a ViT-Large (24 blocks, d=1024, 16 heads); the intrinsics encoder is a ConvNeXt-Base whose final-stage features are projected to dimension d=1024. The transformer trunk consists of L=24 self-attention blocks with d=1024, 16 heads, and 8 register tokens; we feed the outputs of blocks $\{4, 11, 17, 23\}$ to the DPT head. All branches use patch size p=16. The model has approximately 640M parameters. We train with AdamW [32] for 200K iterations using a cosine learning-rate schedule peaking at 5×10^{-5} after 2.5K warmup steps, with λ=0.2 and gradients clipped at 1.0. Training uses bfloat16 mixed precision and gradient checkpointing on two H200 GPUs and takes approximately four days. Input images are resized to 512-pixel longer side with random aspect ratio in $[0.33, 1.0]$ and random horizontal flips; we deliberately avoid photometric augmentations on the source and target, which would corrupt the lighting signal the model is supervised to learn. We do, however, apply corruption augmentations to the conditioning C_T to simulate the artifacts produced by single-image geometry and material reconstruction at inference (Sec. C). Full hyperparameters are in Sec. A.

3.4 Inference

At inference, the user provides a single image I_S and wishes to relight it under a freely chosen target lighting condition. Our conditioning interface requires a target intrinsic decomposition C_T that the user cannot directly author, so we bridge this gap with a physically based renderer (Fig. 3):

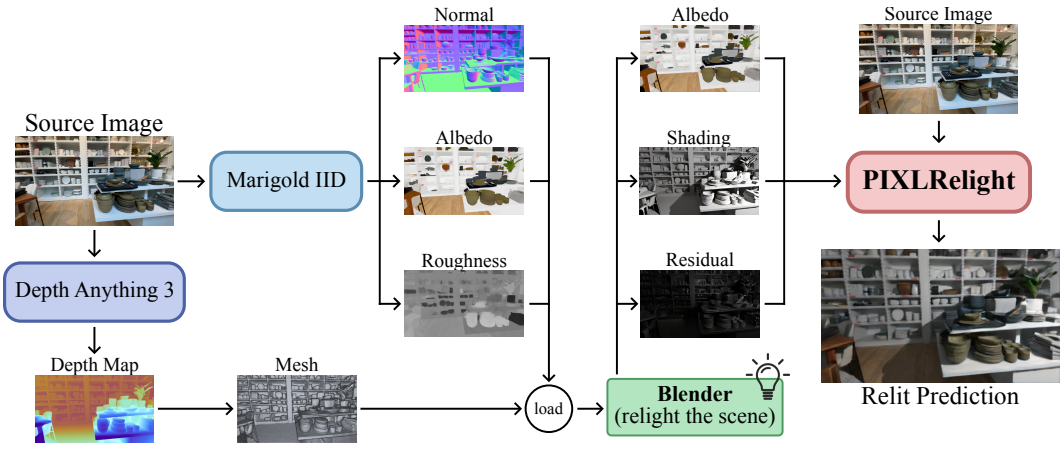

Figure 3: **Inference pipeline.** Given a single input image, geometry is recovered by Depth Anything 3 and unprojected to a triangle mesh, and materials are recovered by Marigold-IID-Appearance. The textured mesh is loaded into Blender, where the user authors the desired illumination; Blender Cycles then renders the scene and produces the target intrinsic maps C_T. PIXLRELIGHT takes as input the original image together with C_T and produces the final relit prediction.

1. **Geometry.** A metric depth map for I_S is estimated with Depth Anything 3 [29] and unprojected into a triangle mesh.
2. **Materials.** Marigold-IID-Appearance [20] extracts pixel-aligned albedo, surface normal, and roughness maps from I_S.
3. **User edit.** The mesh and materials are imported into Blender, where the user freely authors the desired illumination using arbitrary combinations of area lights, environment maps, sun lamps, and emissive geometry.
4. **Conditioning.** The three target intrinsic buffers $C_T = [A_T\,;\,S_T\,;\,R_T]$ are composed directly from Blender's Cycles render passes following the image-formation model of eq. (1); we provide the exact processing formulas in Sec. B.
5. **Relit prediction.** The original input image I_S together with C_T is passed to PIXLRELIGHT, which produces the final relit prediction $\hat{I}_T$ in a single forward pass.

The Blender render of the scene is never shown to the model. Single-image geometry and material estimates from off-the-shelf tools are coarse, and the rendered RGB inherits these errors. The intrinsic buffers C_T are still a valid lighting specification, because errors in geometry or materials corrupt C_T locally, at the affected pixels, rather than propagating globally through the rendered image. This locality, combined with the corruption augmentations of Sec. C, allows PIXLRELIGHT to be conditioned on coarse intrinsic buffers and still produce photorealistic relightings.

4 Experiments

4.1 Quantitative comparison

Baselines. We compare against five recent baselines, grouped by the lighting cue they consume. DiffusionRenderer [28] and UniRelight [14] consume an HDR environment map; since neither accepts an arbitrary target image, we estimate the target environment from the ground-truth target with DiffusionLight-Turbo [6]. RGBX [50], Ouroboros [40], and V-RGBX [11] consume a target diffuse-shading map alongside source G-buffers; we drive each with its own inverse-rendering stage, taking source G-buffers from the source image and target shading from the target image. All baselines use their official released checkpoints. We do not compare against Careaga and Aksoy [5], the closest prior work, because no code or model has been released.

Metrics. We report PSNR, SSIM [44], and LPIPS [52] at native target resolution. Following [14, 18, 21, 28], we apply a per-image, per-channel least-squares scale correction to absorb global exposure

ambiguity, uniformly across every method including ours. Inference times are forward-pass wall-clock on an NVIDIA RTX A6000, averaged over five runs after two warm-ups.

Datasets. We evaluate on the official test split of MIT Multi-Illumination Images in the Wild [34], comprising 30 indoor scenes. We additionally collect a small held-out set of six indoor scenes captured on a stationary tripod under two everyday lighting conditions each, evaluated in both directions for twelve source–target pairs. Neither benchmark overlaps with our training data, and the held-out set probes a lighting distribution distinct from MIIW's controlled flashes.

Results. Table 1 reports MIIW test-split results. PIXLRELIGHT outperforms every baseline by a wide margin, exceeding the next-best result on each metric by 9.8 dB PSNR and 0.130 SSIM (over Ouroboros) and 0.243 LPIPS (over RGBX), and runs in 0.09 s per image – at least an order of magnitude faster than every baseline. The margin holds on the held-out tripod set (Fig. 4), where PIXLRELIGHT is best on every metric in every scene, with per-scene PSNR gaps of 8–9 dB. The two baseline groups fail in distinct ways. UniRelight (environment-map cue) cannot represent the spatially varying near-field sources that dominate indoor scenes: in row 1 it misses the highlight cast by the desk lamp, and in row 3 the directional shadow behind the backpack is absent. V-RGBX (shading cue) inverse-renders source and target into G-buffers and shading, then re-renders both with a forward diffusion model; this chain fails on both ends – it neither preserves the source nor transfers the target lighting, missing the desk light in row 1, retaining the source's strong magenta cast in row 2, and washing out the lamp scene in row 3. PIXLRELIGHT avoids both failure modes: the full intrinsic stack carries the target lighting while the source RGB carries photographic detail. More comparisons in Secs. F and G.

Table 1: **Quantitative evaluation on the MIIW test split [34].** Methods are grouped by the target-lighting cue they consume. All metrics use a per-image, per-channel least-squares scale correction applied uniformly to every method, including ours [14, 18, 21, 28]. Inference times are wall-clock times of one relighting forward pass on an NVIDIA RTX A6000, averaged over 5 runs after 2 warm-ups; intrinsic estimation, which differs across methods, is excluded. **Bold**: best; underline: second-best.

Method	**Lighting cue**	PSNR ↑	SSIM ↑	LPIPS ↓	Time (s) ↓
DiffusionRenderer [28]	Environment map	15.02	0.663	0.461	3.37
UniRelight [14]		16.71	0.672	0.444	58.58
RGBX [50]	Diffuse shading	16.20	0.690	0.438	13.98
Ouroboros [40]		19.38	0.763	0.463	0.62
V-RGBX [11]		17.79	0.725	0.467	841.13[†]
Ours	Diffuse shading	**29.18**	**0.893**	**0.195**	**0.09**

[†] V-RGBX produces noise on a single-frame input; we replicate the source 49 times to reach the model's minimum supported sequence length and report one full forward pass over that sequence.

4.2 Controllable relighting from authored illumination

The previous evaluations test how faithfully each method transfers a captured target lighting, but cannot test controllability. We now evaluate relighting under physically authored target illumination, supplied through the same intrinsic interface used at training but produced by a path tracer at inference.

Protocol. We automate the inference pipeline of Sec. 3.4 into a Blender script: given an input photograph, it recovers a textured mesh (Depth Anything 3 [29] for geometry, Marigold-IID-Appearance [20] for albedo/normals/roughness), inserts one of five preset lighting setups (cool side flash, warm overhead flash, dim overhead spot, soft frontal sun, warm interior sun), and renders the scene in linear HDR, exporting the Cycles passes needed to compose C_T via eq. (1) (formulas in Sec. B). PIXLRELIGHT consumes C_T alongside the source image; shading-conditioned baselines consume only the diffuse-shading channel of C_T together with their own inverse-rendered source G-buffers. We exclude DiffusionRenderer and UniRelight, which require an HDR environment map. We apply this pipeline to twenty in-the-wild images from DL3DV [30] under all five setups; with no ground-truth relit photograph available, the evaluation is qualitative. For display, sRGB method

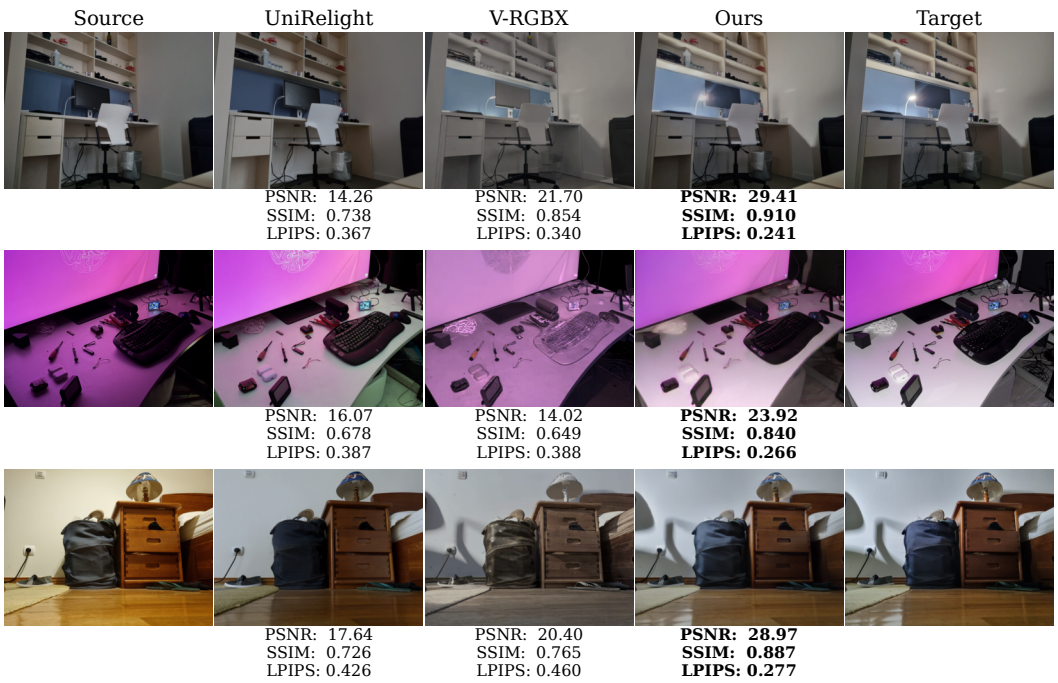

Figure 4: **Held-out tripod captures: paired source–target relighting.** Three representative pairs from a held-out set of six indoor scenes captured under two lighting conditions each. For visual clarity we display the most recent baseline per conditioning group: UniRelight (environment-map) and V-RGBX (shading). Per-image PSNR/SSIM/LPIPS are shown below each prediction; **bold** marks the best method per image. PIXLRELIGHT is best on every metric in every scene.

outputs are shown as produced; the linear-HDR path traced reference is tonemapped to sRGB via Reinhard auto-exposure (key=0.18).

Results. Figure 5 shows five DL3DV scenes relit under Blender-authored illumination, alongside the path-traced render of each reconstruction. Without ground truth, identity preservation – not plausibility alone – separates the methods. V-RGBX, the most recent shading-conditioned baseline, drifts from the source: it desaturates the scene in rows 1, 2, and 5, and overshoots the requested spotlight in row 4. PIXLRELIGHT reproduces the authored lighting in each row – side-lit shadows in the lamp store (row 1), warm interior glow in the bar (row 2) and empty room (row 5), warm overhead falloff in the gift shop (row 3), and a focused spotlight on the fruit stand (row 4) – while leaving the underlying scene unchanged. The Path Traced column reveals a limitation common to all single-image PBR pipelines, including the closest prior work [5]: the underlying 3D reconstruction is coarse and the recovered materials are imperfect, so the rendered RGB visibly drifts from the source. PIXLRELIGHT is robust to this drift because the source image and C_T enter the model through separate branches: the source carries photographic content and C_T carries the intrinsic components, so the model never has to disentangle the two from a single rendered RGB. Further results are in Sec. H.

4.3 Ablation studies

We ablate the two principal architectural choices of PIXLRELIGHT: the fusion of source and intrinsic features inside the transformer trunk, and the modulation head. Both variants are trained from scratch under the protocol of Sec. 3.3 and differ from the full model in exactly one component. Table 2 reports MIIW results. Removing the source ViT branch – so the source enters the network only through the modulation head – costs 3.36 dB PSNR, 0.062 SSIM, and 0.179 LPIPS; without scene structure available to self-attention, predictions hew to the source illumination rather than transferring the requested condition (see Sec. I). Replacing the modulation head with a direct-RGB regression costs 1.41 dB PSNR, 0.060 SSIM, and 0.164 LPIPS; the loss is consistent but visually subtle, since

Figure 5: **Relighting from path-traced illumination on DL3DV scenes [30].** Each row shows a source image, V-RGBX (the most recent shading-conditioned baseline), Blender's full RGB render of the reconstructed scene under the authored lighting (Path Traced), and PIXLRELIGHT. V-RGBX produces plausible relightings but drifts from the source. PIXLRELIGHT transfers the authored lighting while preserving the source's photographic detail.

direct regression still recovers the global lighting but must regenerate source-aligned texture from scratch. Both ablated variants still outperform every baseline in Tab. 1, indicating that the supervision regime – direct training on paired multi-illumination captures – drives most of our gains, with the architectural choices providing the rest.

Table 2: **Architectural ablations on the MIIW test split.** Both variants are trained from scratch and differ from the full model in exactly one component. Intrinsics-only trunk: the source ViT branch is removed; the source enters the network only through the modulation head. Direct regression head: the modulation of eq. (4) is replaced by a sigmoid-activated RGB regression. Both variants still beat every baseline in Tab. 1.

Variant	PSNR ↑	SSIM ↑	LPIPS ↓
Intrinsics-only trunk	25.82	0.831	0.374
Direct regression head	27.77	0.833	0.359
Ours	**29.18**	**0.893**	**0.195**

5 Conclusion

We present PIXLRELIGHT, a feed-forward transformer that brings the physical lighting control of computer graphics to in-the-wild photographs. By separating *what* the target lighting should be from *how* it is applied to a real photograph, and bridging the two through a single intrinsic-decomposition interface, we train directly on paired multi-illumination photographs and accept arbitrary path-traced illumination at inference. PIXLRELIGHT achieves state-of-the-art relighting quality in under a tenth of a second per image – an order of magnitude faster than prior approaches – enabling interactive lighting authoring on real photographs.

References

[1] Harry Barrow, J Tenenbaum, A Hanson, and E Riseman. Recovering intrinsic scene characteristics. *Comput. vis. syst*, 2(3-26):2, 1978.

[2] 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*, pages 1–12, 2025.

[3] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation.

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

[5] Chris Careaga and Yağız Aksoy. Physically controllable relighting of photographs. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–10, 2025.

[6] Worameth Chinchuthakun, Pakkapon Phongthawee, Amit Raj, Varun Jampani, Pramook Khungurn, and Supasorn Suwajanakorn. Diffusionlight-turbo: Accelerated light probes for free via single-pass chrome ball inpainting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[7] 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*, pages 5720–5731, 2025.

[8] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] Epic Games. Unreal engine.

[11] Ye Fang, Tong Wu, Valentin Deschaintre, Duygu Ceylan, Iliyan Georgiev, Chun-Hao Paul Huang, Yiwei Hu, Xuelin Chen, and Tuanfeng Yang Wang. V-rgbx: Video editing with accurate controls over intrinsic properties. *arXiv preprint arXiv:2512.11799*, 2025.

[12] Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2335–2342. Ieee, 2009.

[13] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[14] Kai He, Ruofan Liang, Jacob Munkberg, Jon Hasselgren, Nandita Vijaykumar, Alexander Keller, Sanja Fidler, Igor Gilitschenski, Zan Gojcic, and Zian Wang. Unirelight: Learning joint decomposition and synthesis for video relighting. *arXiv preprint arXiv:2506.15673*, 2025.

[15] Majed El Helou, Ruofan Zhou, Johan Barthas, and Sabine Süsstrunk. Vidit: Virtual image dataset for illumination transfer. *arXiv preprint arXiv:2005.05460*, 2020.

[16] Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, et al. Rayzer: A self-supervised large view synthesis model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4918–4929, 2025.

[17] Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias. *arXiv preprint arXiv:2410.17242*, 2024.

[18] 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. *Advances in Neural Information Processing Systems*, 37:141129–141152, 2024.

[19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[20] Bingxin Ke, Kevin Qu, Tianfu Wang, Nando Metzger, Shengyu Huang, Bo Li, Anton Obukhov, and Konrad Schindler. Marigold: Affordable adaptation of diffusion-based image generators for image analysis, 2025.

[21] 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*, pages 5198–5208, 2024.

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

[23] Edwin H Land and John J McCann. Lightness and retinex theory. *Journal of the Optical society of America*, 61(1):1–11, 1971.

[24] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pages 71–91. Springer, 2024.

[25] Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[26] 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*, pages 2475–2484, 2020.

[27] 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*, pages 7190–7199, 2021.

[28] 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*, pages 26069–26080, 2025.

[29] Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

[30] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024.

[31] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[33] 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*, pages 1–11, 2025.

[34] 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*, pages 4080–4089, 2019.

[35] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021.

[36] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10912–10922, 2021.

[37] 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*, pages 10684–10695, 2022.

[38] 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.

[39] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[40] Shanlin Sun, Yifan Wang, Hanwen Zhang, Yifeng Xiong, Qin Ren, Ruogu Fang, Xiaohui Xie, and Chenyu You. Ouroboros: Single-step diffusion models for cycle-consistent forward and inverse rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10386–10397, 2025.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[42] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025.

[43] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20697–20709, 2024.

[44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[45] Zian Wang, Jonah Philion, Sanja Fidler, and Jan Kautz. Learning indoor inverse rendering with 3d spatially-varying lighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12538–12547, 2021.

[46] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16133–16142, 2023.

[47] 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*, pages 442–452, 2025.

[48] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

[49] 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*, pages 1–12, 2024.

[50] 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*, New York, NY, USA, 2024. Association for Computing Machinery.

[51] 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.

[52] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[53] Rui Zhu, Zhengqin Li, Janarbek Matai, Fatih Porikli, and Manmohan Chandraker. Irisformer: Dense vision transformers for single-image inverse rendering in indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2822–2831, 2022.

Appendix

This appendix collects implementation details and additional results. Section A lists the full training hyperparameters; Sec. B describes how the target intrinsic conditioning C_T is composed from Blender's Cycles render passes at inference; Sec. C details the corruption augmentations applied to C_T during training; and Sec. D discusses the limitations and failure modes of our pipeline. The remaining sections present additional qualitative results: an extended version of the banner figure (Sec. E), additional scenes from the MIIW test split (Sec. F), further scenes from the held-out tripod set (Sec. G), additional comparisons on relit DL3DV scenes (Sec. H), and qualitative ablations (Sec. I).

A Training details

We collect here the hyperparameters omitted from Sec. 3.3. Table 3 lists the optimization, data, architecture, and compute settings used to produce the results reported in the main paper. All hyperparameters were selected through small-scale runs and held fixed for the final training; we did not perform extensive sweeps.

Table 3: **Training hyperparameters for PIXLRELIGHT.**

Group	Hyperparameter	Value
Optimization	Optimizer	AdamW [32]
	(β_1, β_2)	$(0.9, 0.95)$
	Weight decay	0.05
	Peak learning rate	5×10^{-5}
	Final learning rate	1×10^{-5}
	Schedule	cosine, 2,500 warmup steps
	Iterations	200,000
	Batch size (per GPU)	42
	Effective batch size	84 (2 GPUs)
	Gradient clipping (max norm)	1.0
	Mixed precision	bfloat16
Data	Datasets	MIIW [34], BigTime [25], VIDIT [15]
	Longer-side resolution	512
	Random aspect ratio	$[0.33, 1.0]$
	Random horizontal flip	yes
	Photometric augmentations	none on the source/target (would corrupt the lighting signal)
	Conditioning augmentations	corruption pipeline on C_T (Sec. C)
Architecture	Source encoder	ViT-Large (24 blocks, d=1024, 16 heads)
	Intrinsics encoder	ConvNeXt-Base, projected to d=1024
	Trunk depth	$L = 24$ self-attention blocks
	Trunk width	$d = 1024$, 16 heads
	Register tokens	8
	DPT readout blocks	$\{4, 11, 17, 23\}$
	Patch size	$p = 16$
	RoPE base frequency	100.0
	Total parameters	$\approx$ 640M
Loss	Pixel loss	ℓ_1
	Perceptual loss weight λ	0.2
Compute	Hardware	2×NVIDIA H200
	Wall-clock time	$\approx$ 4 days

B Composing target intrinsics from Blender

To produce the target intrinsic conditioning C_T at inference time, we read Blender's Cycles render passes for the user-lit scene and compose them following the image-formation model of Marigold-IID-Lighting [20]. The albedo and diffuse shading are obtained directly from the diffuse passes:

$$A_T = \mathrm{clip}(\textit{diffuse\_color},\, 0,\, 1), \tag{6}$$

$$S_T = \max(\textit{diffuse\_direct} + \textit{diffuse\_indirect},\, 0)\,. \tag{7}$$

The non-diffuse residual R_T aggregates every non-Lambertian light-transport contribution that Cycles exposes as a render pass – glossy reflection, transmission (refraction and transparency), participating media (volume scattering), and self-emission:

$$\begin{aligned} R_T \;=\; \max\Big(\; & \textit{glossy\_color} \odot (\textit{glossy\_direct} + \textit{glossy\_indirect}) \\ & + \textit{transmission\_color} \odot (\textit{transmission\_direct} + \textit{transmission\_indirect}) \\ & + (\textit{volume\_direct} + \textit{volume\_indirect}) \\ & + \textit{emission}\;,\, 0\Big), \end{aligned} \tag{8}$$

where the $*\_\textit{color}$ terms are the per-material reflectance/transmittance passes, the $*\_\{\textit{direct}, \textit{indirect}\}$ terms are the corresponding direct- and indirect-lighting passes, $\textit{volume}\_\{\textit{direct}, \textit{indirect}\}$ are the in-scattered radiance from participating media, $\textit{emission}$ is the self-emission pass, and $\odot$ is element-wise multiplication. This decomposition exhausts the non-diffuse light-transport channels Cycles makes available, so any path-traced lighting effect not encoded in the diffuse passes – specular highlights, refraction through transparent objects, glow from emissive geometry, or scattering through fog – is captured by R_T.

The albedo A_T lies in $[0, 1]$ by construction, but S_T and R_T are HDR. To match the distribution that Marigold-IID-Lighting was trained on, we apply a joint 98^{th}-percentile rescaling: we compute $\tau = \max(\mathrm{p98}(S_T),\, \mathrm{p98}(R_T),\, \varepsilon)$ and replace $S_T \leftarrow \mathrm{clip}(S_T, 0, \tau)/\tau$ and $R_T \leftarrow \mathrm{clip}(R_T, 0, \tau)/\tau$. Sharing the cutoff τ across both channels preserves their relative magnitudes, which encodes the diffuse-to-specular balance of the target lighting.

C Conditioning augmentations

The intrinsic conditioning C_T seen at inference is composed from Blender render passes of a coarse, single-image reconstruction. It carries artifacts that never appear in C_T at training, where it is extracted from a real photograph: missing-geometry holes, silhouette cracks at depth discontinuities, render speckle, denoiser blur, posterized banding, and per-channel exposure or color shifts inherited from the upstream estimators. Training only on clean, photograph-derived C_T would let the model overfit to its smooth statistics and degrade at inference. We therefore apply a stochastic corruption pipeline to C_T during training, designed to resemble these artifacts.

The pipeline is applied per sample, on GPU, after Marigold-IID-Lighting and before the conditioning encoder. It consists of eight independently gated augmentations grouped into four families:

- **Photometric.** Per-channel multiplicative *color cast* and additive bias; per-channel *gamma*.
- **Structural.** *Holes*: a low-resolution bilinearly-upsampled noise mask, optionally biased toward Sobel edges, replaces a sample-specific top fraction of pixels with the channel minimum plus a small noise floor. *Edge cracks*: a quantile threshold on the Sobel edge map produces a narrow silhouette mask which is dilated and used to multiplicatively darken those pixels.
- **Noise.** Per-pixel *salt-and-pepper* replacement and additive *Gaussian* noise.
- **Frequency.** Separable *Gaussian blur* (denoiser-style smoothing) and *posterization* into a sample-specific number of levels.

Augmentations are applied in the order photometric → structural → noise → blur, so that downstream noise stacks on top of the perturbed tonal range and the introduced structural defects. A global Bernoulli gate $p_{\text{apply}} = 0.7$ wraps the entire pipeline, so 30% of samples remain strictly clean;

among the rest, each augmentation fires independently with its own per-sample probability, and the output is finally clipped to the input value range. The full set of probabilities and strength ranges is listed in Tab. 4; values are deliberately mild and designed to mimic the renderer's failure modes rather than destroy the lighting signal.

Table 4: **Conditioning augmentation parameters.** The pipeline as a whole fires with probability p_{apply}; conditional on firing, each augmentation fires independently per sample with the listed probability and its strength is sampled uniformly from the range. Photometric augmentations operate per intrinsic group (albedo, shading, residual).

Augmentation	**Probability**	**Strength range**
Pipeline applied (p_{apply})	0.70	—
Color cast (per-channel scale)	0.50	scale $\in [0.85, 1.15]$, bias $\in [-0.06, 0.06]$
Gamma	0.30	$\gamma \in [0.75, 1.35]$
Holes (edge-biased)	0.50	affected fraction $\in [0.005, 0.040]$
Edge cracks (silhouette darkening)	0.50	quantile $\in [0.92, 0.99]$, strength $\in [0.4, 1.0]$
Salt-and-pepper	0.25	per-pixel fraction $\in [0, 0.003]$
Gaussian noise	0.50	$\sigma \in [0, 0.04]$
Gaussian blur	0.25	$\sigma \in [0.4, 1.2]$
Posterization	0.25	levels $\in [12, 64]$

D Limitations and failure cases

PIXLRELIGHT inherits the failure modes of its frozen dependencies. At training, the target intrinsics are produced by Marigold-IID-Lighting [20]; systematic errors in its decomposition – such as baking a cast shadow into albedo, or attributing a colored highlight to diffuse rather than non-diffuse shading – bias the supervisory signal. At inference, the same decomposer is applied to a Blender render, and the upstream geometry (Depth Anything 3 [29]) and material (Marigold-IID-Appearance [20]) estimators add their own errors. The corruption augmentations of Sec. C make the model robust to local artifacts, but global reconstruction failures still propagate: when entire objects are mis-localized or fused with the background, the resulting C_T no longer specifies the user's intended lighting on the original geometry. Figure 6 shows one such case on a DL3DV scene where the depth estimator collapses a bicycle into the floor; the path-traced render carries this error into C_T, and PIXLRELIGHT – which has no direct view of the original geometry beyond the source RGB – inherits it in the relit output. We expect future improvements in feed-forward geometry, materials, and intrinsic decomposition to translate directly into better authoring fidelity, without retraining the relighting network.

A second limitation is the relatively small training corpus: $985 + 212 + 300 \approx 1{,}500$ scenes from MIIW, BigTime, and VIDIT is still two orders of magnitude smaller than the unpaired photo collections used by self-supervised relighting methods. Although our quantitative margin and the held-out tripod evaluation indicate that the intrinsic-conditioning interface generalizes beyond the training distribution, scaling paired multi-illumination supervision – whether through new captures, simulated multi-illumination renders, or synthetic-to-real adaptation – is a natural avenue for further gains.

Figure 6: **Failure case from upstream reconstruction errors.** A DL3DV scene where the single-image depth estimator collapses the bicycle into the floor. The path-traced render carries this error into C_T, and PIXLRELIGHT inherits it in the relit output.

E Multi-light authored relighting on a single scene

Figure 7: **Banner figure expansion.** A single source image (left) is relit by PIXLRELIGHT (right) under six different Blender-authored illuminations. The middle column shows the corresponding path-traced render of the reconstructed scene. PIXLRELIGHT consumes only the intrinsic buffers derived from these renders, together with the source image, and produces a sharper and more photorealistic relighting that retains the source's photographic detail while transferring the authored lighting.

F Additional qualitative results on MIIW

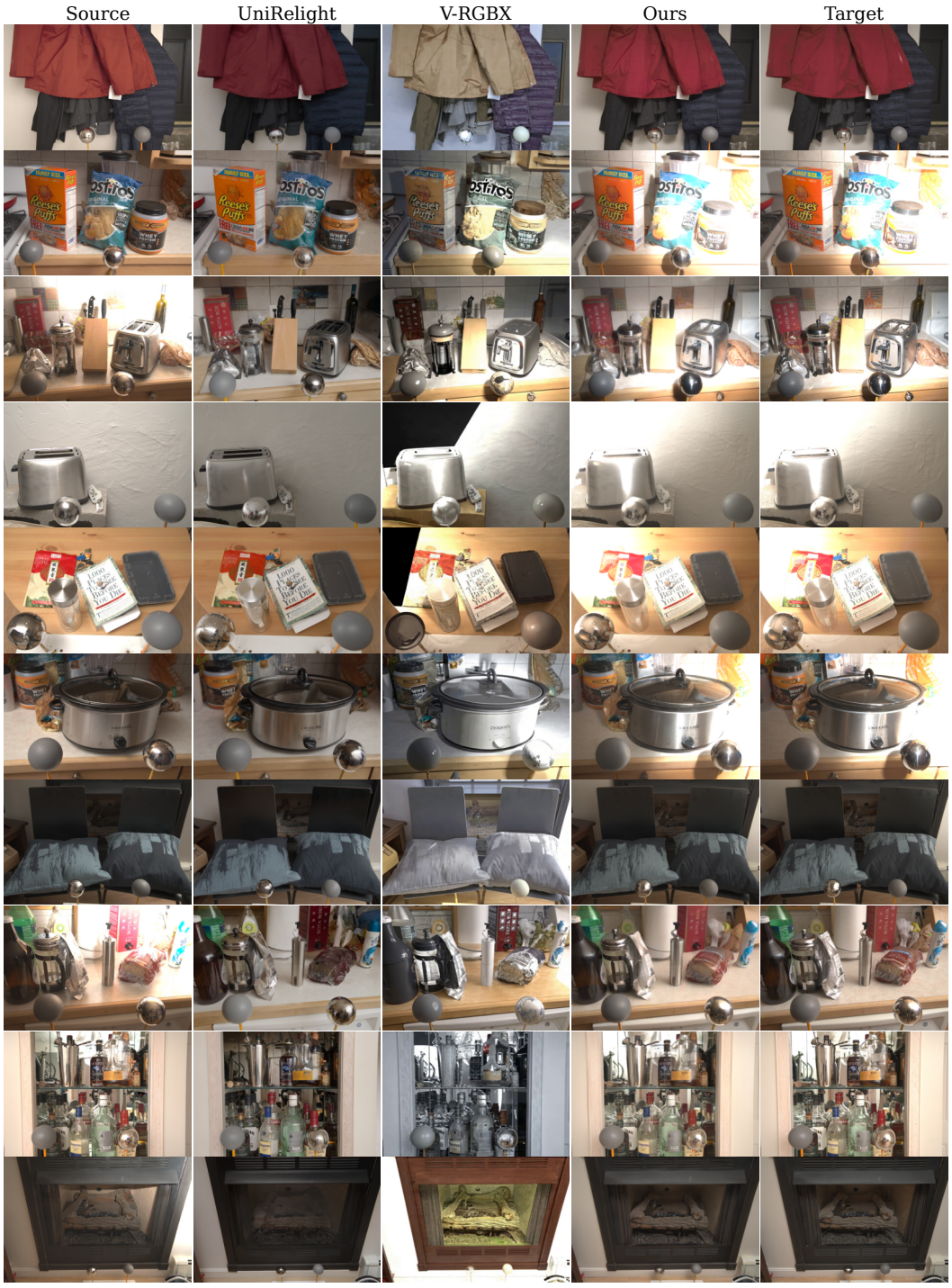

Figure 8: **Additional qualitative comparisons on the MIIW test split [34].** Each row shows a single source–target pair, with predictions from the most recent baseline of each conditioning group (UniRelight for environment-map methods; V-RGBX for shading-conditioned methods, see Tab. 1) and PIXLRELIGHT. Across all twelve scenes, PIXLRELIGHT retains source detail by construction and transfers only the lighting change implied by the conditioning intrinsics, including specular highlights on chrome spheres and shading gradients on diffuse surfaces.

G Additional held-out tripod captures

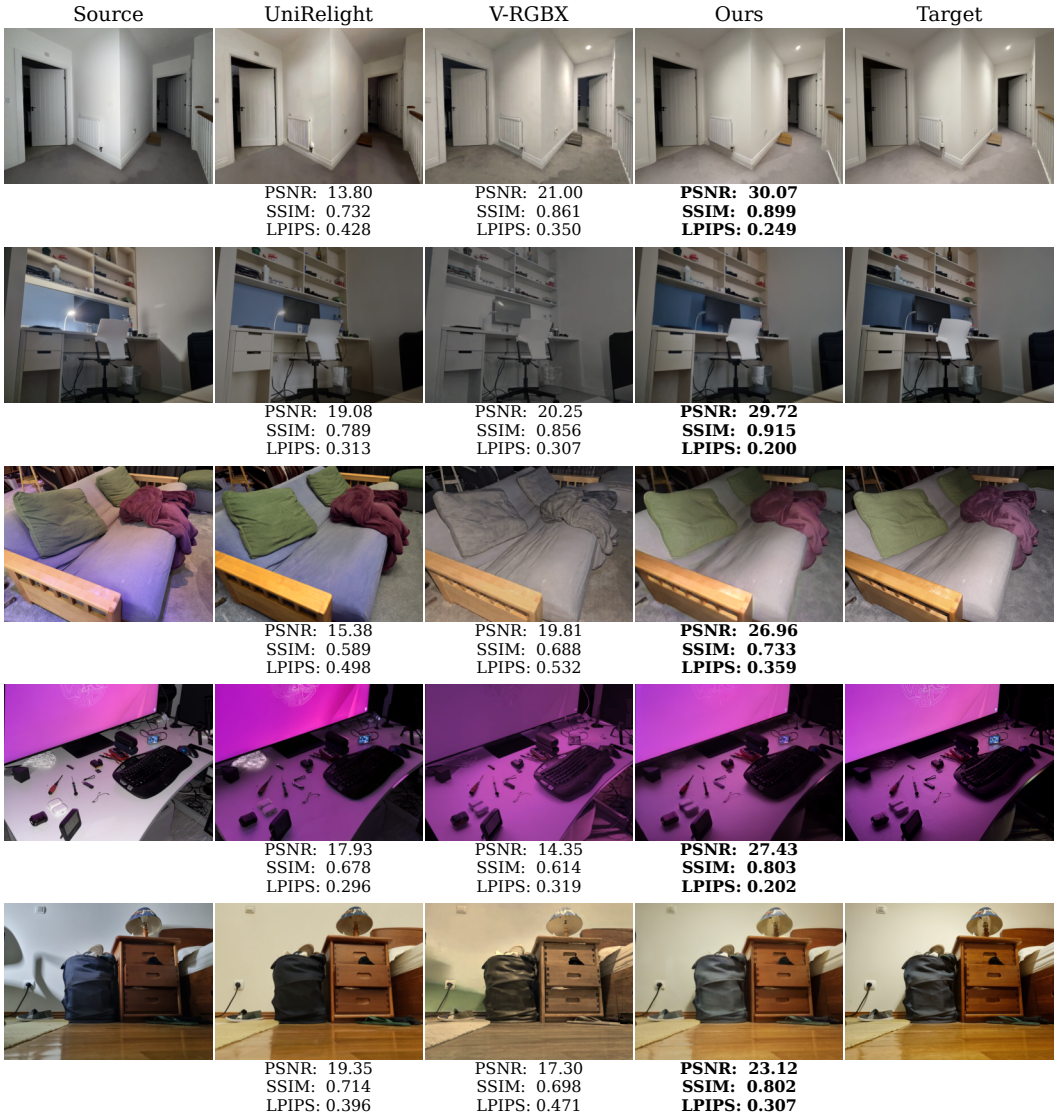

Figure 9: **Additional held-out tripod captures.** Further source–target pairs from the held-out set, beyond the three shown in Fig. 4. Columns: source, UniRelight (environment-map baseline), V-RGBX (shading baseline), PIXLRELIGHT, and target. PIXLRELIGHT is best on every metric in every scene.

H Additional qualitative results on DL3DV scenes

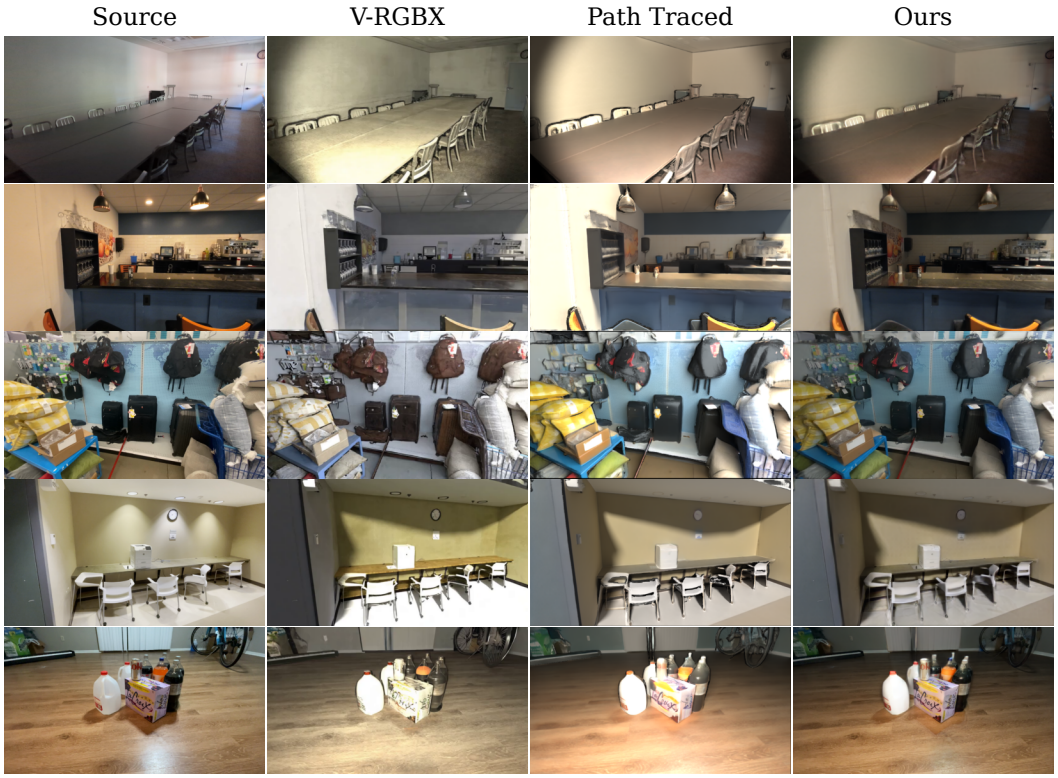

Figure 10: **Additional relighting comparisons on DL3DV scenes [30].** Each row: a single source image, the most recent shading-conditioned baseline (V-RGBX), Blender's full RGB render of the reconstructed scene under the authored lighting (Path Traced), and PIXLRELIGHT. PIXLRELIGHT transfers the authored lighting while preserving the source's photographic detail.

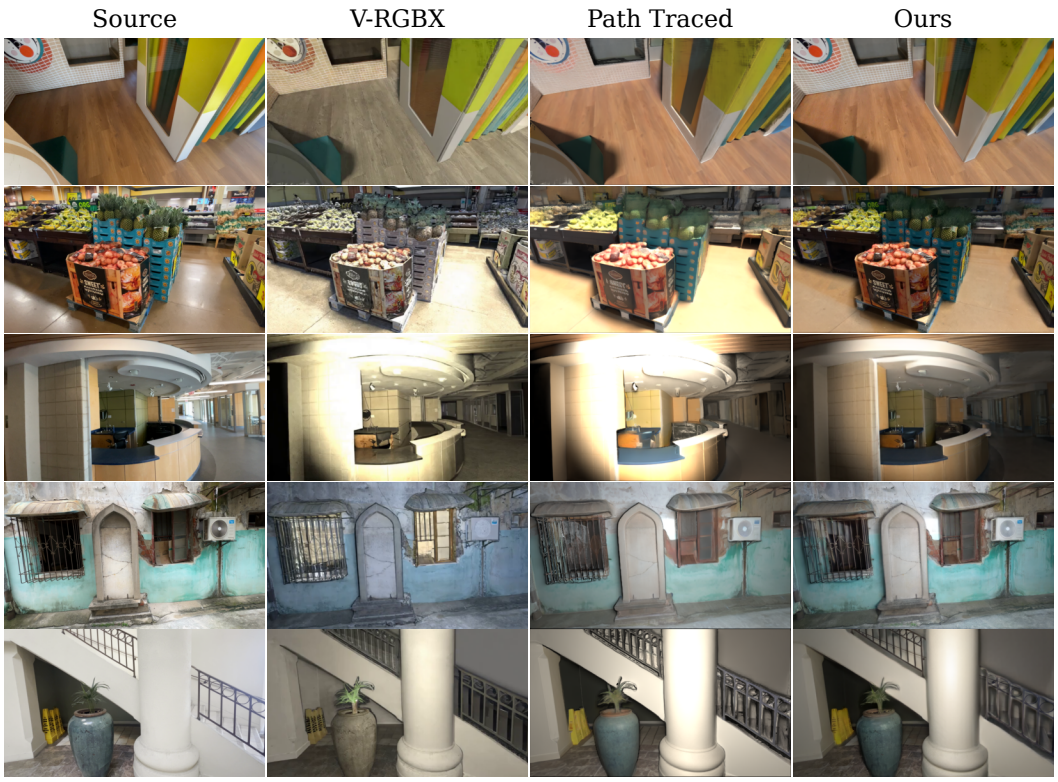

Figure 11: **Additional relighting comparisons on DL3DV scenes [30].** Each row: a single source image, the most recent shading-conditioned baseline (V-RGBX), Blender's full RGB render of the reconstructed scene under the authored lighting (Path Traced), and PIXLRELIGHT. PIXLRELIGHT transfers the authored lighting while preserving the source's photographic detail.

I Additional qualitative ablation results

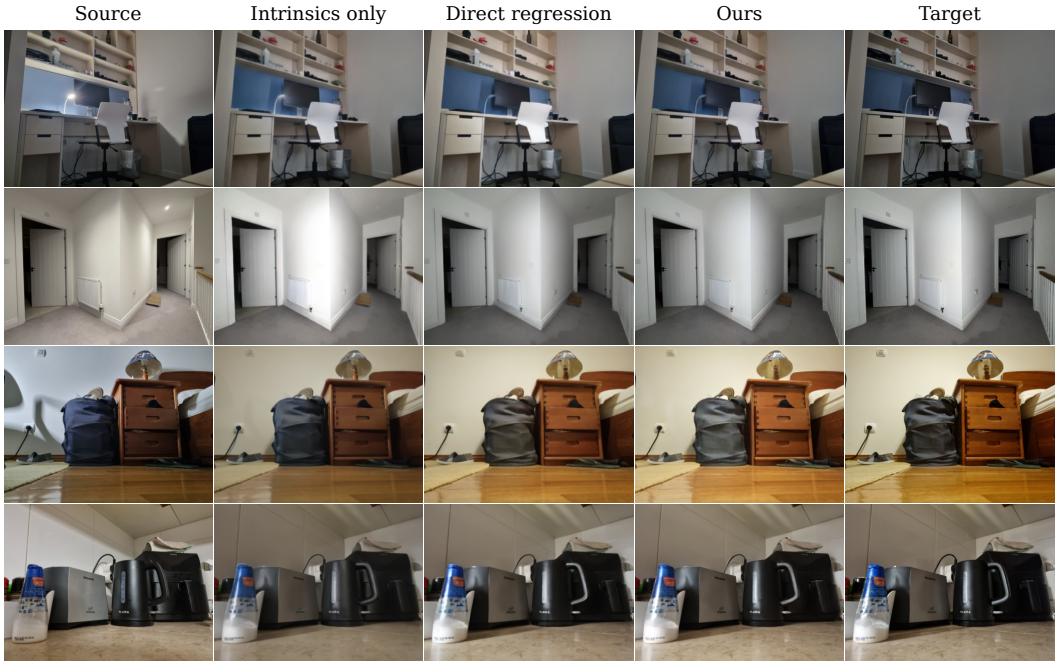

Figure 12: **Additional ablation comparisons on the held-out tripod captures.** Both variants are trained from scratch and differ from the full model in exactly one component. *Intrinsics-only*: the source ViT branch is removed; the source enters the network only through the modulation head. *Direct regression*: the modulation of eq. (4) is replaced by a sigmoid-activated RGB regression. Ours is the full model.