# NATIVE-HDR RELIGHTING WITH PIXEL-SPACE FLOW MODELS

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

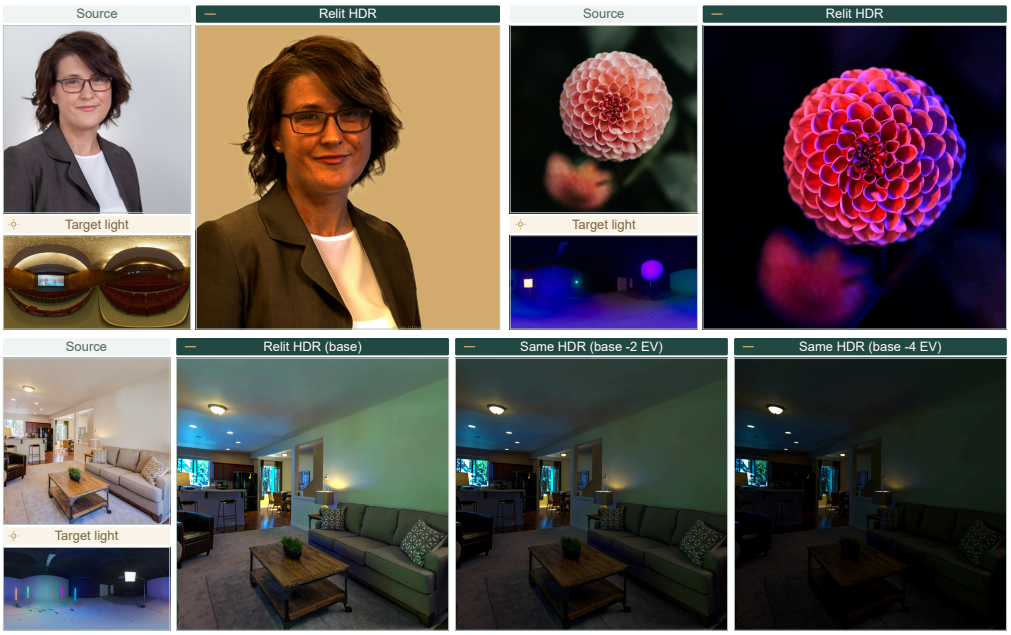



Figure 1: **Native HDR relighting.** Given an LDR source image and a target HDR environment, our model produces a relit HDR image.

## ABSTRACT

We present the first native high-dynamic-range (HDR) pixel-space model for single-image relighting. Fine-tuned only on paired synthetic single-object images, it generalizes far beyond this fine-tuning distribution and outperforms previous methods on even complex real-world scenes. This is achieved by three design choices. First, whereas previous methods modify the pretrained model's conditioning pathways to fit the relighting conditions, we instead conform the conditions to the model: express the conditions in a form the model fully accepts, which preserves the model's pretrained knowledge. Second, whereas most previous methods produce only low-dynamic-range (LDR) outputs through LDR-trained variational autoencoders (VAEs), our model predicts native HDR directly in pixel space. HDR more closely represents the physical nature of light transport, and predicting in pixel space avoids the lossy reconstruction and representational constraints of such VAEs, preserving fine source details and simplifying HDR fine-tuning. Third, native HDR output lets us formulate a linearity constraint grounded in the additivity of light transport, which maintains consistency better than previous formulations that operate in the latent space of LDR-trained VAEs.

## 1 INTRODUCTION

Lighting shapes the appearance of a scene through its interaction with geometry and materials. Single-image relighting reproduces a photographed scene under new illumination, enabling applications in image editing, compositing, and content creation. Recent methods, including Neural Gaffer (Jin et al., 2024), IC-Light (Zhang et al., 2025), DiffusionRenderer (Liang et al., 2025a), and UniRelight (He et al., 2025), fine-tune pretrained diffusion models for this task. Unlike traditional inverse-rendering-based approaches (Barron & Malik, 2015; Sengupta et al., 2019; Li et al., 2020),



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.

which explicitly recover geometry, materials, and illumination before re-rendering, these methods render with pretrained generative priors, often producing more realistic lighting effects and better capturing complex material appearance.

However, these models often generalize poorly, even those built on powerful pretrained generative models with rich visual knowledge, such as UniRelight and IC-Light: they produce compelling results when the source image and target lighting resemble their training data, but fail to synthesize plausible lighting effects consistent with scene geometry and materials when either departs substantially from the training distribution (Figure 8).

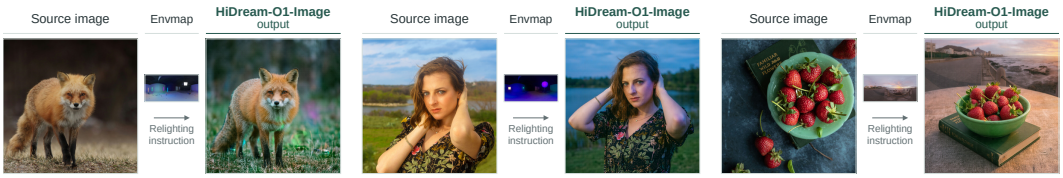



Figure 2: **Using text and images for relighting with HiDream-O1-Image** (Cai et al., 2026), a general-purpose image generation model, without further fine-tuning.

Yet without fine-tuning, a pretrained base model yields reasonably fine image details and promising generalization across diverse inputs (Figure 2). However, it remains very weak at accurately rendering object appearance under the target illumination, in particular plausible shadows and colors, and it leads to substantial identity drift. A key question is therefore: is there a better way to preserve and exploit the model's pretrained visual knowledge while teaching it to accurately capture how illumination interacts with object geometry and materials, enabling relighting that generalizes far beyond the fine-tuning distribution?

We investigate this question by fine-tuning the base model from Figure 2, HiDream-O1-Image (Cai et al., 2026). Our three design choices let the fine-tuned model keep the base model's fine image details and generalization while learning to relight accurately in native high dynamic range (HDR).

First, every condition enters the model through a pathway that the model was pretrained to use. Unlike previous methods, which fit the model to the relighting conditions by adding input channels (Zhang et al., 2025; He et al., 2025) or a dedicated lighting encoder (Liang et al., 2025a), we conform the conditions to the model. Because the model was pretrained to generate images from text and image conditions, we express both relighting conditions as images, optionally accompanied by text captions. This preserves the model's visual knowledge while it learns how illumination changes appearance. To demonstrate the role of this knowledge in generalization, we deliberately restrict fine-tuning to a minimal synthetic setting of single objects rendered under varying illumination: this setting provides basic relighting supervision but leaves a substantial gap to the complex content. Despite this restricted supervision, the model generalizes to multi-object compositions and complex real-world scenes, substantially outperforming previous relighting baselines.

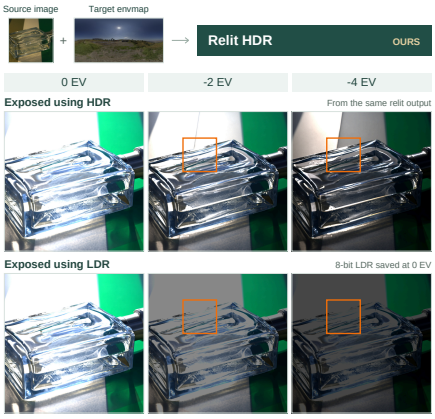



Figure 3: **Highlight retention after relighting.** Our HDR results can be re-exposed if too bright or dark. At $-2$ and $-4$ EV, the boxed reflections reveal highlight variations retained in HDR but clipped in the LDR export.

Second, our model predicts native HDR directly in pixel space. HDR is the natural mode of operation for relighting: it more closely represents the physical nature of light transport and preserves physically meaningful information that low dynamic range (LDR) images discard, which also facilitates post-processing (Figure 3). Yet most existing relighting methods produce only LDR outputs, as they build on latent diffusion models with LDR-trained variational autoencoders (VAEs) (Jin et al., 2024; Zhang et al., 2025; He et al., 2025). Like recent models such as PixelDiT (Yu et al., 2026a) and DLSS 5 (NVIDIA ADLR, 2026), our base model instead generates images directly in pixel space, without a VAE. This avoids the lossy reconstruction and representational constraints of LDR-trained VAEs: the model preserves fine source details, and we can fine-tune it directly in native HDR. To

provide HDR supervision for this fine-tuning, we create, to our knowledge, the first HDR relighting dataset of Objaverse objects (Deitke et al., 2023) to be openly released.

Third, native HDR output lets us formulate light-transport consistency analytically, since linear HDR responses are additive in illumination. Previous formulations such as IC-Light (Zhang et al., 2025) instead learn to combine predictions in the latent space of an LDR-trained VAE and assume that all components share the same noise. We derive how components with independent noises compose, including the corresponding variance and time rescaling. Our constraint needs no additional trainable module and maintains consistency better (Section 4), improving the linearity of our model's illumination response.

In summary, our contributions are:

- **Preserving pretrained knowledge for broad generalization.** We show that conforming the relighting conditions to the pretrained model lets it retain and exploit its visual knowledge while learning relighting from only synthetic single-object images.
- **Native HDR relighting.** We present the first native HDR pixel-space single-image relighting model, which operates in HDR because HDR is closer to how light behaves in the real world. For fine-tuning, we render HDR relighting data of Objaverse objects (Deitke et al., 2023), which we will openly release as, to our knowledge, the first publicly available dataset of its kind.
- **Analytical linearity constraint.** Building on native HDR output, we impose light-transport consistency as an analytical, architecture-agnostic constraint for flow-based generative models, maintaining it better than previous formulations.

## 2 Related Work

**Generative Single-Image Relighting.** Recent generative methods have explored relighting and illumination control across objects, portraits, scenes, and videos (Jin et al., 2024; Zeng et al., 2024; Kocsis et al., 2024; Ren et al., 2024; Zhang et al., 2025; Liang et al., 2025a; He et al., 2025; Bharadwaj et al., 2025; Magar et al., 2025; Wang et al., 2025b; Zhou et al., 2025; Li et al., 2026; Choi et al., 2026; Zhang et al., 2026). Among them, IC-Light (Zhang et al., 2025) fine-tunes Stable Diffusion 1.5 (Rombach et al., 2022), SDXL (Podell et al., 2024), and FLUX (Black Forest Labs, 2024), and UniRelight (He et al., 2025) fine-tunes Cosmos-Predict1 (NVIDIA et al., 2025). These methods keep their pretrained backbones but add conditioning inputs or modules outside the original model pathways, such as extra input channels (Zhang et al., 2025; He et al., 2025) or a dedicated lighting encoder (Liang et al., 2025a). In contrast, we reuse the pretrained model's existing pathways, thereby retaining its visual knowledge for generalization.

**HDR Image Generation and Reconstruction.** Learning-based HDR reconstruction has explored direct radiance prediction, multiscale processing, and intrinsic decomposition to recover information lost in LDR images (Eilertsen et al., 2017; Marnerides et al., 2018; Dille et al., 2024). Generative approaches have broadened HDR synthesis and reconstruction through text-conditioned panoramas, exposure bracketing, gain-map decomposition, and HDR-compatible representations (Chen et al., 2022; Phongthawee et al., 2024; Wang et al., 2025a; Guan et al., 2025; Wu et al., 2026; Yu et al., 2026b). Methods built on LDR-pretrained latent diffusion need additional processing to accommodate the original VAE: LEDiff generates and fuses latent exposures and fine-tunes the decoder (Wang et al., 2025a), whereas X2HDR and DiffHDR keep their VAEs frozen and map HDR radiance into PU21 and Log-Gamma representations, respectively (Wu et al., 2026; Yu et al., 2026b). We avoid such VAE-related processing by fine-tuning a pixel-space model to predict encoded HDR images directly, obtaining native HDR output.

**Pixel-Space Generative Models.** Diffusion models have generated images directly in pixel space using single-stage denoising or cascaded super-resolution (Ho et al., 2020; Dhariwal & Nichol, 2021; Saharia et al., 2022; Jabri et al., 2023). Recent work revisits this setting, avoiding the reconstruction bottleneck of image autoencoders, and makes high-resolution generation practical through improved prediction targets, perceptual supervision, architectures, and training recipes (Hoogeboom et al., 2023; 2025; Li & He, 2026; Ma et al., 2026; Yu et al., 2026a; Jiang et al., 2026; Cai et al., 2026). HiDream-O1-Image extends pixel-space generation to a unified multimodal model that processes text, visual conditions, and image patches within a shared Transformer (Cai et al., 2026). We use such a pixel-space model for relighting to preserve the source image's fine details.

# 3 METHOD

Given an LDR source image $S$ and a target HDR environment map $E$, our goal is to generate a native HDR image $H$ of the same scene from the same viewpoint under $E$. Using a pretrained image generator for this task raises two questions: how to represent and integrate the two conditions, and how to obtain native HDR output from a model pretrained on LDR images.

In practice, we fine-tune HiDream-O1-Image (Cai et al., 2026), a pretrained pixel-space generative model initialized from Qwen3-VL-8B-Instruct, for HDR relighting. HiDream follows the common design of unified multimodal models (Deng et al., 2025; Xie et al., 2025), combining semantic visual features with complementary low-level image representations—but in the form of raw pixel patches rather than VAE latents, for both visual understanding and image generation.

We conform both conditions to the pretrained model rather than modifying its conditioning pathways to fit them, which preserves its visual knowledge (Section 3.1). We then extend the method to predict native HDR directly in pixel space (Section 3.2). Native HDR output further enables us to formulate a linearity constraint grounded in the additivity of light transport that encourages additive responses to illumination (Section 3.3). Section 3.4 describes the training data, objective, and inference.

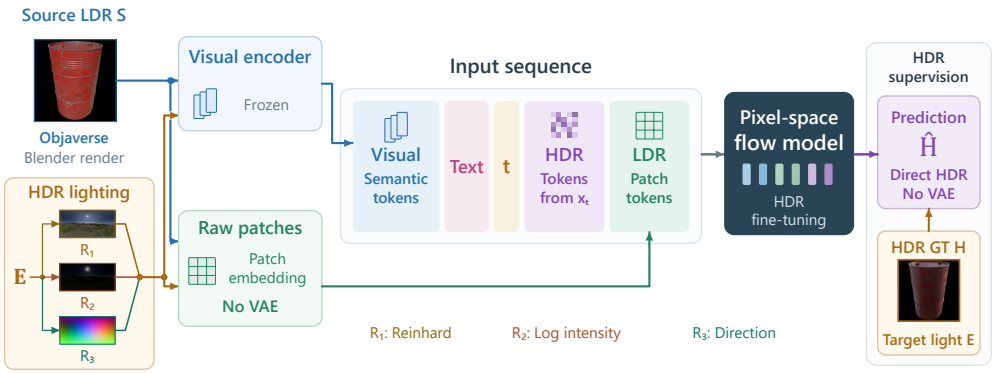



Figure 4: **Fine-tuning for native-HDR relighting.** We decompose the target HDR environment $E$ into three distinct LDR images $R_1, R_2, R_3$. Together with the LDR source $S$, these images enter both the visual encoder and raw-patch embedding. Paired HDR targets supervise the prediction.

## 3.1 CONFORMING RELIGHTING CONDITIONS TO THE PRETRAINED MODEL

The model's conditioning pathways, attention, and output resolution stay as pretrained. What we design is how the source and the target lighting are represented, routed through these pathways, and ordered, so that the model fully accepts them as conditions.

**Lighting representation.** To express incident light directions in the source view, we rotate $E$ into the source camera frame, following Neural Gaffer (Jin et al., 2024), and encode it with DiffusionRenderer's three lighting maps (Liang et al., 2025a): extended-Reinhard $R_1$ for illumination color, log-intensity $R_2$ for the strength of bright sources, and direction $R_3$ for incident directions. Represented as LDR images, the source and the three lighting maps each pass through the model's two existing image-conditioning pathways (Figure 4): features from the frozen visual encoder join the text as semantic context, while raw patches from the same images supply direct pixel information. The shared Transformer uses both to denoise the HDR target within one token sequence (Figure 5), so no conditioning module is added outside these pathways. Section 5 shows that the lighting inputs to both pathways contribute to the rendered lighting.

**Semantic context.** We retain the pretrained language–image interface to form the semantic context $\mathbf{z}_{\text{sem}}$ from *Visual* and *Text* tokens. The frozen Qwen3-VL visual encoder $\phi_V$ (Bai et al., 2025) independently extracts *Visual* tokens from $S, R_1, R_2, R_3$, while the pretrained text embedding maps the relighting instruction and optional source and lighting captions to *Text* tokens. Qwen3-VL's DeepStack variant (Meng et al., 2024; Bai et al., 2025) additionally supplies intermediate encoder features at the *Visual* positions. Within the shared Transformer, text hidden states integrate source content and target-light information into semantic context for HDR generation. Removing this text pathway substantially degrades generalization (Section 5).

**Patch embedding.** To help preserve source details and reproduce local color and shadow variations, we give the model direct access to raw pixel patches from the source and lighting images. A shared trainable embedding $\phi_L$ projects patches from $S, R_1, R_2, R_3$ into *LDR* tokens. The time token, *HDR* target tokens (Section 3.2), and *LDR* conditioning tokens form the pixel-space sequence $\mathbf{z}_{\text{pix}}$.

**Attention mechanism.** HiDream-O1-Image retains Qwen3-VL's causal attention over semantic visual and text tokens (Cai et al., 2026). As shown in Figure 5, its shared Transformer uses causal attention within $\mathbf{z}_{\text{sem}}$, with *Text* tokens following *Visual* tokens, and full attention within $\mathbf{z}_{\text{pix}}$. Tokens in $\mathbf{z}_{\text{pix}}$ can additionally attend to all of $\mathbf{z}_{\text{sem}}$, while semantic tokens cannot access $\mathbf{z}_{\text{pix}}$.

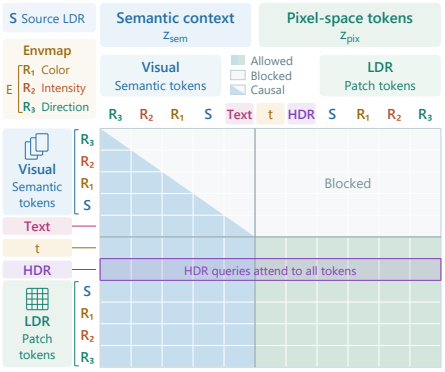



Figure 5: **Hybrid attention.** Triangles denote causal attention within a token span: each token reads itself and earlier tokens, but not later tokens. Pixel-space tokens attend to the full sequence.

Causal attention also determines how the model combines multiple images: later images can draw on earlier ones but not vice versa; multi-image studies link this asymmetry to a preference for later images and find that bidirectional visual interaction reduces it (Tian et al., 2025), and Qwen2-VL, for example, favors the final image in small image sets (Wu et al., 2025). We therefore keep causal attention and order the *Visual* tokens as $R_3, R_2, R_1, S$, expecting the source, placed last, to benefit from this preference while its representations incorporate all three lighting representations. Appendix F.2 shows that this choice preserves source backgrounds more reliably than source-first causal or full visual attention in challenging cases.

### 3.2 Native HDR Prediction in Pixel Space

We fine-tune the model to predict encoded HDR images directly in pixel space, thereby removing the representational constraints of LDR-trained VAEs.

**Flow matching with clean-target prediction.** Following JiT (Li & He, 2026), we use a clean-target ($x$-prediction) parameterization of flow matching (Lipman et al., 2023). For a clean target $x$, we construct the noisy state $x_t = tx + (1-t)\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ and $t \in [0, 1]$ progresses from noise to data. The model directly predicts $\hat{x} = f_\theta(x_t, t; S, R_1, R_2, R_3, \text{text})$, giving the velocity $v_\theta(x_t, t) = (\hat{x} - x_t)/(1-t)$, with the same conditioning left implicit. During training, we sample $\text{logit}(t) \sim \mathcal{N}(-0.8, 0.8^2)$ and minimize the time-weighted reconstruction loss $\mathcal{L}_{\text{FM}}$ below, which is equivalent to velocity matching under this parameterization.

**Log-Gamma encoding for HDR.** HDR radiance has a long-tailed distribution, so extreme intensities can dominate direct reconstruction losses. Following DiffHDR (Yu et al., 2026b), we compress its dynamic range with Log-Gamma encoding and perform reconstruction in the encoded coordinates; Section 5 shows that training on clipped linear HDR instead produces severe artifacts. The reconstruction objective and target encoding are

$$\mathcal{L}_{\text{FM}} = \mathbb{E}\left[\frac{\text{MSE}(\hat{x}, x)}{(1-t)^2}\right], \qquad x = T(H_c) = 2\left[\frac{\log(1+\gamma_{\text{LG}} H_c)}{\log(1+\gamma_{\text{LG}} M)}\right]^{1/\gamma_{\text{LG}}} - 1. \tag{1}$$

Here $H_c = \text{clip}(H, 0, M)$, $M = 50$, $\gamma_{\text{LG}} = 2.2$, and the HDR prediction is $\hat{H} = T^{-1}(\hat{x})$.

**HDR target embedding.** The target patches represent the evolving noisy HDR state $x_t$ at flow time $t$, whereas the source and lighting patches provide fixed LDR conditions. Their different roles and input distributions motivate an independent HDR embedding $\phi_H$. We initialize it from the pretrained patch projection and optimize it independently of $\phi_L$ to produce *HDR* tokens.

### 3.3 Light-Transport Linearity Constraint

**Physical additivity.** Linear HDR responses are additive in illumination for a fixed scene, viewpoint, and exposure. Following IC-Light (Zhang et al., 2025), our light-transport linearity constraint (LC) uses this additivity to constrain denoising. However, IC-Light's derivation queries all three lightings at one shared noisy input (Zhang et al., 2025, App. A, Eqs. 8–11), averaging rather than

adding the component predictions, and assumes that all components share the same noise rather than independent noises; we refine this treatment of component coupling and noise scaling. Specifically, we add the component paths under independent noises and rescale their sum, with a matching change of time, so that it starts from standard Gaussian noise and ends at the sum of the two HDR responses. The rescaled sum $H_t^{1+2}$ and its velocity $\widehat{v}_t^{H,1+2}$ are

$$H_t^{1+2} = \frac{\gamma_t}{\sqrt{2}}\left(x_{s_t}^1 + x_{s_t}^2\right) = (1-t)\epsilon + t(H_1 + H_2), \quad \widehat{v}_t^{H,1+2} = \frac{(\sqrt{2}-1)H_t^{1+2} + v_1 + v_2}{\gamma_t}, \tag{2}$$

where $x^1$ and $x^2$ are the flow paths from independent noises to the two responses $H_1$ and $H_2$, $v_1$ and $v_2$ are their velocities, $s_t$ is the changed time, $\gamma_t$ is the rescaling factor, and $\epsilon$ is standard Gaussian noise. Appendices H.1 and H.2 give the full derivation. In a controlled comparison, LC yields more accurate and more additive mixed-light predictions than an IC-Light-style constraint (Section 4).

**LC training.** Each LC example uses one source image $S$ and two environments $E_1, E_2$ from the training data, with HDR targets $H_1, H_2$. Adding $H_1+H_2$ as an extra flow-matching target for $E_{1+2} = E_1 + E_2$ would supervise only the endpoint of the flow. LC instead constrains the model's velocity field along the flow. Constraining the whole flow regularizes generation and, in our controlled studies, keeps predictions from drifting to extreme values (Appendix E.1.5). We use equation (2), taking $v_1, v_2$ from the model's own velocity predictions under $E_1$ and $E_2$, and train the model under $E_{1+2}$ to predict $\widehat{v}_t^{H,1+2}$ at $H_t^{1+2}$. Since light adds only in linear HDR, we form $H_t^{1+2}$ and $\widehat{v}_t^{H,1+2}$ after decoding with $T^{-1}$, then map them back with $T$ and its Jacobian $J_T$:

$$\widehat{v}_t^{1+2} = J_T(H_t^{1+2})\,\widehat{v}_t^{H,1+2}, \qquad \mathcal{L}_{\text{LC}} = \mathbb{E}\left[\left\|v_\theta(x_t^{1+2}, t) - \widehat{v}_t^{1+2}\right\|_2^2\right], \tag{3}$$

where $x_t^{1+2} = T(H_t^{1+2})$. LC is exact in linear HDR, but our Log-Gamma encoding distorts the Gaussian noise, so in practice LC is a soft constraint. Appendix H.3 details the construction in these coordinates and our practical choices, and Appendix G.4 gives the random weighting of $E_1, E_2$ and the numerical safeguards. In the full model, LC also makes mixed-light predictions more consistent with linear superposition (Section 5).

### 3.4 Training Data and Procedure

**Relighting HDR data generation.** Large-scale image datasets are predominantly LDR, with limited HDR supervision (Wang et al., 2025a; Yu et al., 2026b). Given the scarcity of HDR relighting data, we construct a synthetic HDR relighting dataset from Objaverse assets (Deitke et al., 2023) to fine-tune the pretrained model for native HDR relighting. By design, the dataset contains only synthetic single-object images (examples in Appendix G.1), to better highlight that the generalization comes from the pretrained model itself rather than from the data. Section 4 demonstrates this on complex real photographs. The dataset comprises two collections at different resolutions. For broad object coverage, the $512 \times 512$ collection retains approximately 125K objects with 16 viewpoints each, after we remove assets whose geometry, materials, or renders fail validity checks. For fine detail, the $2048 \times 2048$ collection rerenders about 200K views of 41K objects, selected with a vision-language model, directly at $2048 \times 2048$ under seven environments each; Appendix F.1 shows that these native targets preserve finer detail than targets resized from $512 \times 512$. Both collections use a lighting bank of 40K HDR environments generated by LuxDiT (Liang et al., 2025b) and 1.4K captured real-world environments. For each view, geometry, materials, camera, and framing remain fixed across illuminations. Appendix G.1 details data construction, filtering, pairing, and the optional captions.

**Perceptual supervision and objective.** Reconstruction in compressed coordinates does not fully capture visible structure and texture. Following PixelGen (Ma et al., 2026), we add LPIPS (Zhang et al., 2018) and DINOv2 (Oquab et al., 2024) supervision; Appendix F.1 shows that without it, results become smoother. These losses compare $\hat{H}$ and the decoded target $T^{-1}(x)$ after the same fixed tone mapping to LDR, with gradients through the HDR decoding step. With $n$ counting training examples and $\chi, \chi_{\text{LC}}$ controlling when the auxiliary losses are active, the full objective combines reconstruction (Section 3.2), perceptual supervision, and LC (Section 3.3):

$$\mathcal{L}_{\text{total}} = \mathbb{E}\left[\frac{\text{MSE}(\hat{x}, x)}{(1-t)^2} + \chi(n,t)\left(\lambda_p \ell_{\text{LPIPS}} + \lambda_d \ell_{\text{DINO}}\right)\right] + \chi_{\text{LC}}(n)\lambda_{\text{LC}}\mathcal{L}_{\text{LC}}. \tag{4}$$

**Training and inference.** Training progresses from the low-resolution collection, for broad object coverage, through mixed sampling to native high-resolution fine-tuning, for detailed HDR supervision. Throughout, we train and predict at the pretrained model's original output resolution of $2048 \times 2048$, which Section 5 shows is important for generalization. At inference, the fine-tuned model directly generates the relit HDR image from $S$ and the lighting maps of $E$. We use classifier-free guidance (CFG) (Ho & Salimans, 2022) with separate scales for the source and lighting conditions, which control the emergence of finer image details and more distinct shadows, respectively (Appendix F.3). Additional implementation details are provided in Appendix G.4, and we will release our training code and data upon acceptance.

## 4 Experiments and Results

We evaluate native HDR benefits, physical correctness, real-world generalization, and light-transport consistency. Section 5 ablates the design choices in Section 3.

Table 1: **OLATverse.** Metrics at $512 \times 512$ over foreground masks, averaged over objects.

| Method | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|
| Neural Gaffer | 16.16 | 0.551 | 0.340 |
| DiLightNet | 15.82 | 0.493 | 0.340 |
| DiffusionRenderer-SVD | 15.57 | 0.590 | 0.285 |
| Light-X | 13.74 | 0.550 | 0.423 |
| IC-Light | 12.73 | 0.536 | 0.331 |
| UniRelight | 17.23 | 0.733 | 0.215 |
| Ours | **23.55** | **0.802** | **0.145** |

**Evaluation setup.** We test physical correctness on captured OLATverse objects (Zhou et al., 2026), captured indoor scenes from MIT Multi-Illumination (MIIW) (Murmann et al., 2019), and synthetic objects. Paired comparisons include Neural Gaffer (Jin et al., 2024), DiLightNet (Zeng et al., 2024), DiffusionRenderer-SVD (Liang et al., 2025a), Light-X (Liu et al., 2026), IC-Light (Zhang et al., 2025), and UniRelight (He et al., 2025).

**Physical correctness.** We evaluate 340 relighting pairs from 34 OLATverse objects. These real captures provide an out-of-distribution evaluation: no OLATverse data are used in training. Our model achieves better quantitative results than the baselines (Table 1), with a qualitative comparison in Figure 6. We also evaluate our model on our internal synthetic benchmark and MIIW. Detailed protocols and additional results for these benchmarks appear in Appendix C.

Improved results across all three dataset domains provide partial evidence for the physical correctness of the generated relighting. The evidence is partial because of ground-truth ambiguity: different material–light combinations can explain the same source $S$ yet yield different valid relightings under the same target environment $E$. Appendix A discusses this ambiguity in detail.

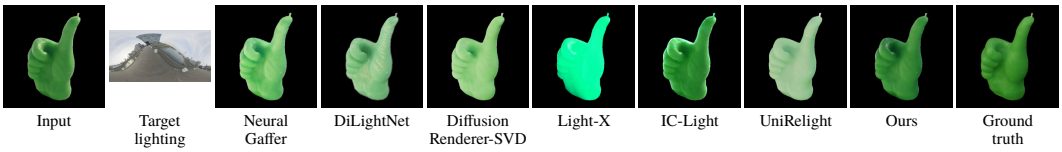



Figure 6: **Relighting a real OLATverse object.** Benchmark display and foreground masks.

**Real-world generalization.** We test generalization from synthetic single-object fine-tuning data (Section 3.4) to complex real photographs. These photographs lie outside our fine-tuning distribution, not the base model's pretraining distribution. RealPhoto24 comprises 24 photographs from Unsplash and Flickr, spanning portraits, animals, plants, food, everyday objects, vehicles, interiors, landscapes, and architecture (Figure 7). Relighting each under the same eight environment maps yields 192 source–lighting combinations without paired ground truth, for which we examine both target-light response and source-content preservation.

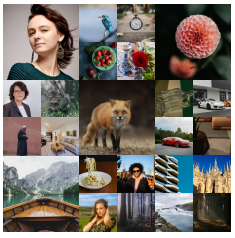


Figure 7: **RealPhoto24 source photographs.**

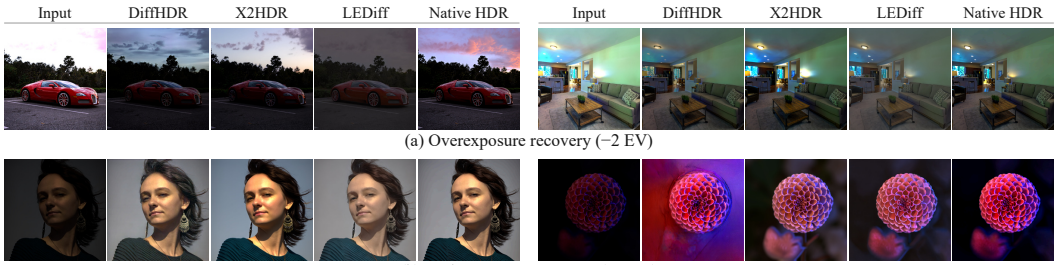
The results in Figure 8 show that our model changes the lighting plausibly while preserving source content, and does much better than the baselines. These results demonstrate generalization from the synthetic single-object supervision to complex scenes and diverse natural content. They provide evidence that the model retains and exploits pretrained visual knowledge while learning physically plausible lighting responses that preserve source structure and fine details. A masked-source probe (Appendix B.2) is consistent with the pretrained semantic and generative priors remaining accessible after HDR fine-tuning. Section 5 ablates the conditioning choices behind this generalization. Together, the appendix (Appendix B.1) and our

anonymous supplementary website provide comparisons for all 192 source–lighting combinations. We also provide relighting videos for 14 scenes, each under four lighting conditions.




Figure 8: **Relighting real-world photographs.** Additional examples appear in Figure 13.

**Benefits of native HDR relighting.** Native HDR output retains radiance that an LDR result cannot represent (Appendix D.2). To test whether a separate off-the-shelf HDR reconstruction approach can achieve results similar to our native HDR output, we export our predictions to LDR, reconstruct them with DiffHDR (Yu et al., 2026b), X2HDR (Wu et al., 2026), and LEDiff (Wang et al., 2025a), and compare the results with the originals (Appendix D.1). The reconstructions change the cloud structure and color of the car skies and the dahlia's colors and background (Figure 9), so they do not faithfully reproduce the native HDR results.




(a) Overexposure recovery (−2 EV)

(b) Underexposure recovery (+4 EV)

Figure 9: **Native HDR versus HDR reconstructed from LDR exports.** EV offsets are relative to the input exposure; the Input columns keep their original exposure.

**Light-transport consistency.** We test LC (Section 3.3) in a controlled study with a minimal conditional PixelDiT trained from scratch, comparing it with ordinary flow matching and an IC-Light-style constraint that learns to combine the component predictions (Zhang et al., 2025). LC predicts unobserved sums of observed lights more accurately in both coordinate systems (Table 2), and its predictions are also closer to the sum of its own single-light outputs (Appendix E.1). Section 5 evaluates LC in the full model.

Table 2: **Controlled study.** Relative HDR error (%).

| Method | Linear | Log-Gamma |
|---|---|---|
| Flow matching | 46.2 | 41.4 |
| IC-Light-style | 21.5 | 52.2 |
| LC (ours) | **3.7** | **23.5** |

## 5 Ablation Studies

We examine design choices for HDR fine-tuning and lighting conditioning; further ablations are in Appendix F. For fair comparison, these ablations use a reference budget of about 40 hours on 8 H200 GPUs, so their visual quality is lower than with full training. We then evaluate the linearity constraint in the full relighting model. Preserving detail during HDR fine-tuning depends on both the pretrained model's visual priors and the HDR supervision, while the lighting inputs to the semantic and pixel-space pathways (Section 3.1) shape the intensity and local shading of the rendered lighting (Figure 10; Table 3).

**HDR fine-tuning** (Figure 10a). The Log-Gamma encoding keeps extreme intensities from dominating the reconstruction loss (Section 3.2); training on clipped linear HDR instead (Raw Clip-10) produces severe artifacts. Keeping the pretrained configuration preserves generalization: removing the text pathway, which integrates source content and target light (Section 3.1), fails to render the

target lighting, and training and predicting at $512 \times 512$ instead of the pretrained $2048 \times 2048$ (Section 3.4) loses source detail.

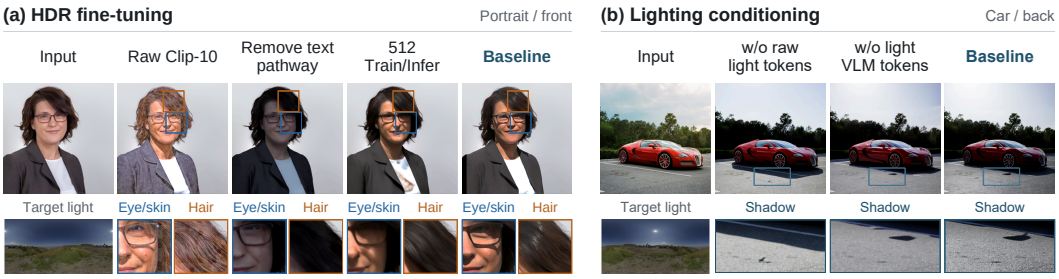



Figure 10: **HDR fine-tuning and lighting conditioning.** Each variant is compared with the baseline on the same input and target lighting. (a) Portrait under front lighting, with crops of the eye and skin and of the hair. (b) Car under back lighting, with crops of the side mirror's cast shadow. Settings are listed in Table 3.

**Lighting conditioning** (Figure 10b). The two pathways of Section 3.1 play complementary roles in rendering the cast shadow. The raw lighting patches, which supply direct pixel information, are needed for its strength: without them, the mirror's shadow is small and faint despite the strong sunlight. The lighting *Visual* tokens, which the semantic context integrates with the source content, are needed for its shape: without them, the shadow no longer follows the mirror. Only with both inputs is the shadow correct in strength and shape. Appendix F adds the room comparison, shows that altering these pathways can also disrupt source-background preservation, and further studies detail preservation, causal versus full visual attention, classifier-free guidance, and shared input projections.

Table 3: **Ablation configurations.**

| Configuration | Setting |
|---|---|
| Raw Clip-10 | Clip linear HDR to $[0, 10]$. |
| Remove text pathway | No *Text* pathway. |
| 512 Train/Infer | Train/output at $512^2$. |
| w/o raw light tokens | No LDR lighting patches. |
| w/o light VLM | No Qwen3-VL lighting *Visual* tokens. |
| Baseline | Train/output $2048^2$; full lighting inputs (Section 3.1). |

Raw Clip-10 and Remove text pathway use targets resized from $512^2$ to $2048^2$; the portrait baseline uses native $2048^2$ targets; full variants are in Table 13.

**Linearity constraint.** Complementing the controlled study in Section 4, we test whether LC (Section 3.3) makes the full model's mixed-light predictions more consistent with linear superposition. Models trained with and without LC relight each source under weighted mixtures of two environment maps. Output RMSE (HDR) and the LDR metrics compare each model's mixed-light output with the same generated, not measured, reference: the weighted sum of the two single-light outputs without LC. Additivity RMSE (HDR) compares it with the sum of the model's own single-light outputs. Across 24 mixtures over four scenes, LC lowers both HDR errors and improves the LDR metrics (Table 4 and Figure 11; settings in Appendix E.2.1).

Table 4: **Light composition.**

| Metric | w/o LC | With LC |
|---|---|---|
| Output RMSE (HDR) ↓ | 0.347 | **0.213** |
| Additivity RMSE (HDR) ↓ | 0.527 | **0.225** |
| PSNR (LDR, dB) ↑ | 28.05 | **32.59** |
| SSIM (LDR) ↑ | 0.961 | **0.982** |
| LPIPS (LDR) ↓ | 0.079 | **0.039** |

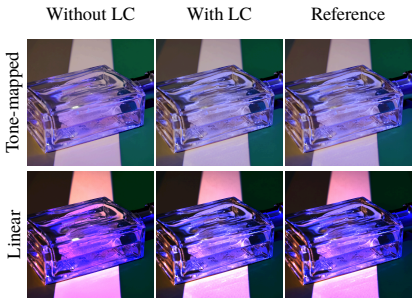



Figure 11: **Mixed-light predictions** (right).

## 6 Conclusion

We presented a native HDR pixel-space relighting model that generalizes from synthetic single-object training to complex real-world scenes. Conforming the relighting conditions to the pretrained model preserves the visual knowledge behind this generalization, pixel-space prediction keeps fine source details, and native HDR output enables an analytical light-transport consistency constraint.

### AI USE STATEMENT

Generative AI tools assisted with literature review, mathematical derivations and proofs, and interpretation of experimental results. They also assisted with model implementation and debugging, running experiments, manuscript drafting, translation, editing, and preparation of illustrative figures. For data preparation, ChatGPT and Codex assisted with developing and debugging data generation and cleaning code; neither was used to directly generate training images or captions. The core ideas of the method are the authors' own. The authors manually verified the mathematical derivations and validated the derived relationships through numerical toy experiments. The authors take responsibility for the correctness and originality of the final manuscript, code, and reported results.

## REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Jonathan T. Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1670–1687, 2015. doi: 10.1109/TPAMI.2014.2377712.

Shrisha Bharadwaj, Haiwen Feng, Giorgio Becherini, Victoria Fernandez Abrevaya, and Michael J. Black. GenLit: Reformulating single-image relighting as video generation. In *SIGGRAPH Asia 2025 Conference Papers*, 2025. doi: 10.1145/3757377.3763970. URL https://doi.org/10.1145/3757377.3763970.

Black Forest Labs. FLUX, 2024. URL https://github.com/black-forest-labs/flux.

Qi Cai, Jingwen Chen, Chengmin Gao, Zijian Gong, Yehao Li, Yingwei Pan, Yi Peng, Zhaofan Qiu, Kai Yu, Yiheng Zhang, Hao Ai, Siying Bai, Yang Chen, Zhihui Chen, Fengbin Gao, Ying Guo, Dong Li, Zhen Shen, Leilei Shi, Jing Wang, Siyu Wang, Yimeng Wang, Rui Zheng, Ting Yao, and Tao Mei. HiDream-O1-Image: A natively unified image generative foundation model with pixel-level unified transformer. *arXiv preprint arXiv:2605.11061*, 2026. URL https://arxiv.org/abs/2605.11061.

Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Text2Light: Zero-shot text-driven HDR panorama generation. *ACM Transactions on Graphics*, 41(6), 2022. doi: 10.1145/3550454.3555447. URL https://arxiv.org/abs/2209.09898.

Jun Myeong Choi, Jae Shin Yoon, Luchao Qi, Roni Sengupta, and Joon-Young Lee. Relightful video portrait harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23356–23366, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Choi_Relightful_Video_Portrait_Harmonization_CVPR_2026_paper.html.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13142–13153, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Deitke_Objaverse_A_Universe_of_Annotated_3D_Objects_CVPR_2023_paper.html.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. URL https://arxiv.org/abs/2505.14683.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021. URL https://arxiv.org/abs/2105.05233.

Sebastian Dille, Chris Careaga, and Yağız Aksoy. Intrinsic single-image HDR reconstruction. In *European Conference on Computer Vision*, 2024. URL https://yaksoy.github.io/intrinsicHDR/.

Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K. Mantiuk, and Jonas Unger. HDR image reconstruction from a single exposure using deep CNNs. *ACM Transactions on Graphics*, 36(6), 2017. URL https://computergraphics.on.liu.se/hdrcnn/.

Yuanshen Guan, Ruikang Xu, Yinuo Liao, Mingde Yao, Lizhi Wang, and Zhiwei Xiong. HDR image generation via gain map decomposed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17536–17545, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Guan_HDR_Image_Generation_via_Gain_Map_Decomposed_Diffusion_ICCV_2025_paper.html.

Kai He, Ruofan Liang, Jacob Munkberg, Jon Hasselgren, Nandita Vijaykumar, Alexander Keller, Sanja Fidler, Igor Gilitschenski, Zan Gojcic, and Zian Wang. UniRelight: Learning joint decomposition and synthesis for video relighting. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/9bb159d886a92a8c3375dc6cbb383333-Abstract-Conference.html.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. URL https://arxiv.org/abs/2006.11239.

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13213–13232. PMLR, 2023. URL https://proceedings.mlr.press/v202/hoogeboom23a.html.

Emiel Hoogeboom, Thomas Mensink, Jonathan Heek, Kay Lamerigts, Ruiqi Gao, and Tim Salimans. Simpler diffusion: 1.5 FID on ImageNet512 with pixel-space diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18062–18071, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Hoogeboom_Simpler_Diffusion_1.5_FID_on_ImageNet512_with_Pixel-space_Diffusion_CVPR_2025_paper.html.

Allan Jabri, David J. Fleet, and Ting Chen. Scalable adaptive computation for iterative generation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 14569–14589. PMLR, 2023. URL https://proceedings.mlr.press/v202/jabri23a.html.

Dengyang Jiang, Ruoyi Du, Zhennan Chen, Dongyang Liu, Zanyi Wang, Mingzhe Zheng, Xiangpeng Yang, Huanqia Cai, Aiming Hao, Yuming Jiang, Peng Gao, Harry Yang, and Steven Hoi. An empirical study of training pixel-space text-to-image diffusion models. *arXiv preprint arXiv:2608.16887*, 2026. URL https://arxiv.org/abs/2608.16887.

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, pp. 141129–141152, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ff7373914a96956f2a7cacbdf3b0b8d8-Abstract-Conference.html.

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, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Kocsis_LightIt_Illumination_Modeling_and_Control_for_Diffusion_Models_CVPR_2024_paper.html.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36115–36125, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_Back_to_Basics_Let_Denoising_Generative_Models_Denoise_CVPR_2026_paper.html.

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*, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Li_Inverse_Rendering_for_Complex_Indoor_Scenes_Shape_Spatially-Varying_Lighting_and_CVPR_2020_paper.html.

Zongming Li, Lianghui Zhu, Haocheng Shen, Longjin Ran, Wenyu Liu, and Xinggang Wang. TransLight: Image-guided customized lighting control with generative decoupling. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2508.14814.

Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Chih-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, and Zian Wang. Diffusion Renderer: Neural inverse and forward rendering with video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26069–26080, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liang_Diffusion_Renderer_Neural_Inverse_and_Forward_Rendering_with_Video_Diffusion_CVPR_2025_paper.html.

Ruofan Liang, Kai He, Zan Gojcic, Igor Gilitschenski, Sanja Fidler, Nandita Vijaykumar, and Zian Wang. LuxDiT: Lighting estimation with video diffusion transformer. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/07b02efbb28aeb2d339d2b7c80f0adb3-Abstract-Conference.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Tianqi Liu, Zhaoxi Chen, Zihao Huang, Shaocong Xu, Saining Zhang, Chongjie Ye, Bohan Li, Zhiguo Cao, Wei Li, Hao Zhao, and Ziwei Liu. Light-X: Generative 4D video rendering with camera and illumination control. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=VBew6vESGL.

Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3D captioning with pretrained models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 75307–75337, 2023. URL https://cap3d-um.github.io/.

Tiange Luo, Justin Johnson, and Honglak Lee. View selection for 3D captioning via diffusion ranking. In *European Conference on Computer Vision*, pp. 180–197, 2024. URL https://arxiv.org/abs/2404.07984.

Zehong Ma, Ruihan Xu, and Shiliang Zhang. PixelGen: Improving pixel diffusion with perceptual supervision. *arXiv preprint arXiv:2602.02493*, 2026. URL https://arxiv.org/abs/2602.02493.

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*, pp. 1–11, 2025. doi: 10.1145/3721238.3730696.

Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista. ExpandNet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. *Computer Graphics Forum*, 37(2):37–49, 2018. doi: 10.1111/cgf.13340. URL https://arxiv.org/abs/1803.02266.

Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. DeepStack: Deeply stacking visual tokens is surprisingly simple and effective for LMMs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 23464–23487, 2024. URL https://arxiv.org/abs/2406.04334.

Lukas Murmann, Michaël Gharbi, Miika Aittala, and Frédo 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. URL https://projects.csail.mit.edu/illumination/.

NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025. URL https://arxiv.org/abs/2501.03575.

NVIDIA ADLR. DLSS 5: Generative neural rendering. https://research.nvidia.com/labs/adlr/DLSS5/, September 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Pakkapon Phongthawee, Worameth Chinchuthakun, Nontaphat Sinsunthithet, Varun Jampani, Amit Raj, Pramook Khungurn, and Supasorn Suwajanakorn. DiffusionLight: Light probes for free by painting a chrome ball. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 98–108, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Phongthawee_DiffusionLight_Light_Probes_for_Free_by_Painting_a_Chrome_Ball_CVPR_2024_paper.html.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, pp. 1862–1874, 2024. URL https://arxiv.org/abs/2307.01952.

Mengwei Ren, Wei Xiong, Jae Shin Yoon, Zhixin Shu, Jianming Zhang, HyunJoon Jung, Guido Gerig, and He Zhang. Relightful harmonization: Lighting-aware portrait background replacement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6452–6462, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Ren_Relightful_Harmonization_Lighting-aware_Portrait_Background_Replacement_CVPR_2024_paper.html.

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*, 2022. URL https://arxiv.org/abs/2112.10752.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36479–36494, 2022. doi: 10.52202/068431-2643. URL https://arxiv.org/abs/2205.11487.

Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David W. Jacobs, and Jan Kautz. Neural inverse rendering of an indoor scene from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8598–8607, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Sengupta_Neural_Inverse_Rendering_of_an_Indoor_Scene_From_a_Single_ICCV_2019_paper.html.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Identifying and mitigating position bias of multi-image vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2503.13792.

Chao Wang, Zhihao Xia, Thomas Leimkühler, Karol Myszkowski, and Xuaner Zhang. LEDiff: Latent exposure diffusion for HDR generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_LEDiff_Latent_Exposure_Diffusion_for_HDR_Generation_CVPR_2025_paper.html.

Junying Wang, Jingyuan Liu, Xin Sun, Krishna Kumar Singh, Zhixin Shu, He Zhang, Jimei Yang, Nanxuan Zhao, Tuanfeng Y. Wang, Simon S. Chen, Ulrich Neumann, and Jae Shin Yoon. Comprehensive relighting: Generalizable and consistent monocular human relighting and harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 380–390, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_Comprehensive_Relighting_Generalizable_and_Consistent_Monocular_Human_Relighting_and_Harmonization_CVPR_2025_paper.html.

Ronghuan Wu, Wanchao Su, Kede Ma, Jing Liao, and Rafał K. Mantiuk. X2HDR: HDR image generation in a perceptually uniform space. *arXiv preprint arXiv:2602.04814*, 2026. URL https://arxiv.org/abs/2602.04814.

Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph E. Gonzalez, Trevor Darrell, and David M. Chan. Visual haystacks: A vision-centric needle-in-a-haystack benchmark. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2407.13766.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21469–21480, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Xiang_Structured_3D_Latents_for_Scalable_and_Versatile_3D_Generation_CVPR_2025_paper.html.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2506.15564.

Yongsheng Yu, Wei Xiong, Weili Nie, Yichen Sheng, Shiqiu Liu, and Jiebo Luo. PixelDiT: Pixel diffusion transformers for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14273–14282, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yu_PixelDiT_Pixel_Diffusion_Transformers_for_Image_Generation_CVPR_2026_paper.html.

Zhengming Yu, Li Ma, Mingming He, Leo Isikdogan, Yuancheng Xu, Dmitriy Smirnov, Pablo Salamanca, Dao Mi, Pablo Delgado, Ning Yu, Julien Philip, Xin Li, Wenping Wang, and Paul Debevec. DiffHDR: Re-exposing LDR videos with video diffusion models. In *European Conference on Computer Vision*, 2026b. URL https://arxiv.org/abs/2604.06161.

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*, pp. 1–12, 2024. doi: 10.1145/3641519.3657396. URL https://doi.org/10.1145/3641519.3657396.

Bohan Zhang, Huanwei Liang, Yuhan He, Hongteng Xu, Quxiao Chao, Luoqi Liu, Dixin Luo, and Ting Liu. Consistent feature transport for image relighting. *arXiv preprint arXiv:2607.17833*, 2026. URL https://arxiv.org/abs/2607.17833.

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. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1cce374fec7829c24a97ae0e91cd0999-Abstract-Conference.html.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Xilong Zhou, Jianchun Chen, Pramod Rao, Timo Teufel, Linjie Lyu, Tigran Minasian, Oleksandr Sotnychenko, Xiao-Xiao Long, Marc Habermann, and Christian Theobalt. OLATverse: A large-scale real-world object dataset with precise lighting control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28848–28859, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhou_OLATverse_A_Large-scale_Real-world_Object_Dataset_with_Precise_Lighting_Control_CVPR_2026_paper.html.

Yujie Zhou, Jiazi Bu, Pengyang Ling, Pan Zhang, Tong Wu, Qidong Huang, Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Anyi Rao, Jiaqi Wang, and Li Niu. Light-A-Video: Training-free video relighting via progressive light fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13315–13325, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhou_Light-A-Video_Training-free_Video_Relighting_via_Progressive_Light_Fusion_ICCV_2025_paper.html.

## Appendix Contents

We discuss the limitations and future work in Appendix A. The appendix also shows a selection of the qualitative results. Complete RealPhoto24 comparisons, covering all 24 photographs under the eight target lightings ($24 \times 8 = 192$ source–lighting combinations), are provided in the supplementary materials and on our anonymous website, together with videos and full-resolution PDFs:

`https://anonymous.4open.science/w/pixelhdr-supplement-B8C8/index.html`

For anonymous review, the website uses an anonymized address and neutral asset names, and omits author-identifying information from the pages and file metadata. Access requires no login, and we include no visitor analytics or tracking scripts.

A Limitations and Future Work . . . . . 17

**Additional results**

B Real-World Generalization (§4) . . . . . 17

B.1 Qualitative Comparisons on Real-World Photographs . . . . . 17

B.2 Generative Priors under Masked Source Inputs . . . . . 28

C Additional Paired Relighting Evaluations (§4) . . . . . 30

D Benefits of Native HDR Relighting (§4) . . . . . 34

D.1 HDR Reconstruction from Clipped Relighting Outputs . . . . . 34

D.2 Exposure Adjustment after Relighting . . . . . 42

E Light-Transport Consistency (§4, §5) . . . . . 45

E.1 Controlled Studies of Light Composition . . . . . 45

E.2 Full-Resolution Linearity Evaluation . . . . . 53

F Additional Ablation Studies (§5) . . . . . 57

F.1 Additional Ablation Examples . . . . . 57

F.2 Visual Attention Configuration . . . . . 58

F.3 Classifier-Free Guidance . . . . . 59

F.4 Background Preservation . . . . . 59

F.5 Shared Projection Ablation . . . . . 62

**Method details and derivations**

G Supplementary Method Details (§3) . . . . . 63

H Derivation of the Light-Transport Linearity Constraint (§3.3) . . . . . 65

I Consistency Gradients and Component-Prediction Bias . . . . . 68

# A Limitations and Future Work

**Training data.** Because we deliberately trained on simple synthetic data, we expect richer and more diverse datasets to further improve relighting quality and generalization.

**Temporal consistency.** Although our model renders convincing shadows and lighting effects in single images, it is an image model. When it relights a video frame by frame, consecutive frames occasionally show visible lighting discontinuities, most noticeably in sharp shadows under sunlight. Extending our approach to video models, conditioning on the previous frame, or adding spatio-temporal attention may help.

**Ground-truth ambiguity.** Different material–light combinations can explain the same source $S$ yet yield different valid relightings under the same target environment $E$ (Figure 12). Agreement with a single measured target therefore provides partial evidence for physical correctness. Developing a better benchmark for verifying whether generated images obey physical laws remains future work.

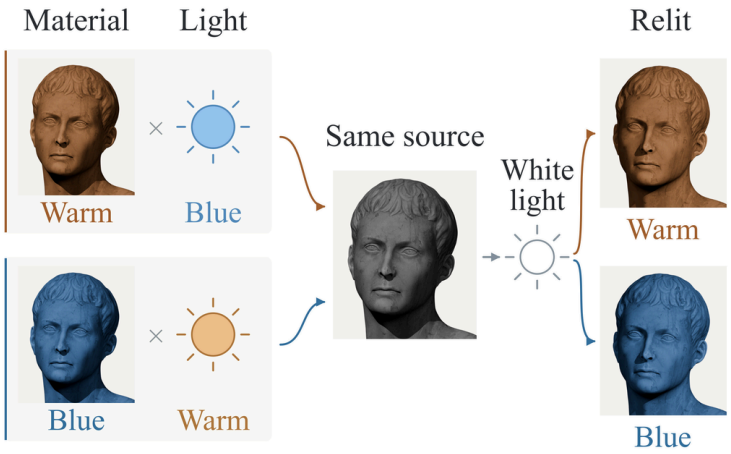



Figure 12: **Material–lighting ambiguity.**

**Linearity regularization.** Our Log-Gamma regularizer does not guarantee additive generated outputs. Its constructed probability path differs from ordinary flow matching, and a source image may leave scene properties undetermined. Together with gradients through the component targets, these can shift the supervised optimum (Appendices H and I). The improvements we report are empirical and specific to the evaluated settings.

**Beyond relighting.** Our light-transport consistency formulation is independent of the model architecture and may support linearity constraints in other flow-based generative tasks with known linear superposition between conditions and outputs.

# B Real-World Generalization

## B.1 Qualitative Comparisons on Real-World Photographs

RealPhoto24 comprises 24 high-quality Internet photographs collected from Unsplash and Flickr, without paired relighting ground truth. Each of the 24 photographs is paired with the same eight target environment maps, giving $24 \times 8 = 192$ source–lighting combinations, each compared across Ours and five baselines. This appendix shows a subset of these comparisons: four additional examples that complement Figure 8 (Figure 13), a close-up of fine-detail preservation (Figure 14), and four photographs under all eight target lightings (Figures 15 to 22). Comparisons for all 24 photographs are provided in the supplementary ZIP archive and additional-results PDF. Accompanying relighting

videos show selected source images under changing illumination. The fox in Figure 8 is photographed by Gary Bendig on Unsplash.[1]

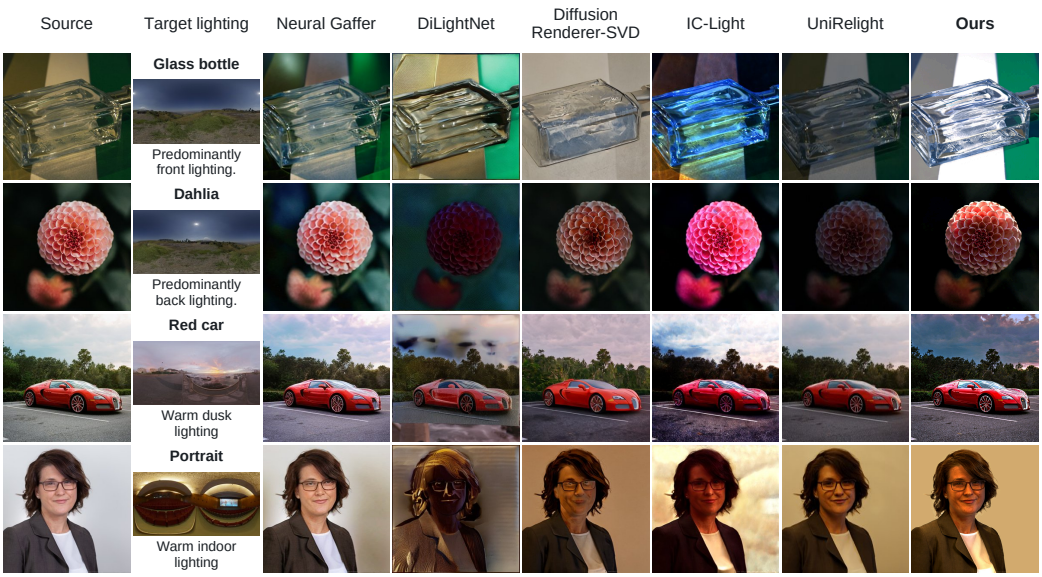



Figure 13: **Additional relighting comparisons on real-world photographs.** Rows show a glass bottle under front lighting, a dahlia under backlighting, a red car under warm dusk lighting, and a portrait under warm indoor lighting. Each row compares Ours with all five baselines. Front and back lighting are obtained by rotating the same HDR environment map containing a bright sun. These photographs have no paired relighting ground truth.

Under backlighting, our model preserves the dahlia's layered petals, with bright upper edges and a shaded center. DiffusionRenderer-SVD alters the dahlia's petal configuration, while IC-Light and UniRelight preserve the flower's broad structure but separate its backlit upper petals less clearly from the shaded front.

[1] https://unsplash.com/photos/red-fox-Yu-qayee34c.

The following pages compare four of the 24 RealPhoto24 photographs: a portrait, a cat, a red sports car, and a furnished living room. Each source is shown under eight target environment maps on two consecutive pages, with four lighting conditions per page. The rows show Ours, UniRelight, Neural Gaffer, DiLightNet, DiffusionRenderer-SVD, and IC-Light. The input photograph and target environment maps are displayed above the predictions. Figures 15, 17, 19 and 21 show the first lighting group for each source, with the remaining environments on the following figures. The other 20 photographs appear only in the supplementary ZIP archive and additional-results PDF.

**Display.** Our portrait HDR outputs use per-image histogram exposure and ACES 2.0/sRGB. Our cat, car, and living-room outputs use 0 EV, clipping to the display range followed by the sRGB transfer, without tone compression. Baseline images retain their method-specific output displays.

**Fine-detail preservation.** Figure 14 compares local source-detail preservation with UniRelight. Around the eye, our result retains a more distinct eyelid boundary and surrounding skin detail, while UniRelight smooths these features. The shoulder close-ups directly compare each output with the source: UniRelight displaces the outer shoulder contour and suppresses fine fabric texture, whereas our result follows the source boundary more closely and retains finer cloth structure while changing its shading.

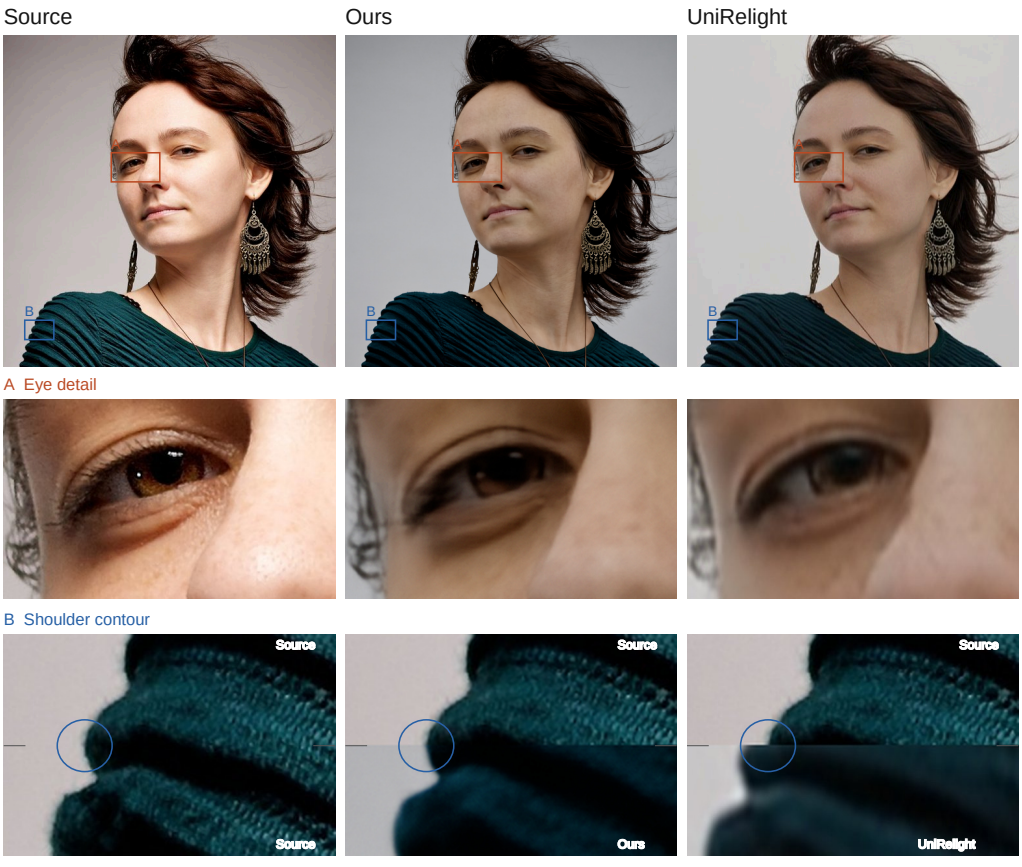



Figure 14: **Fine source-detail preservation on a RealPhoto24 portrait.** Columns show the source, Ours, and UniRelight. Identical crop coordinates are used throughout. Row A shows the eye close-ups. In row B, each panel places the source above the corresponding output; the first column repeats the source as a reference. The circles mark the same source-contour location.




Figure 15: **Portrait: directional and colored lighting.** Columns show front, side, back, and blue/magenta tube lights; rows compare the six methods.

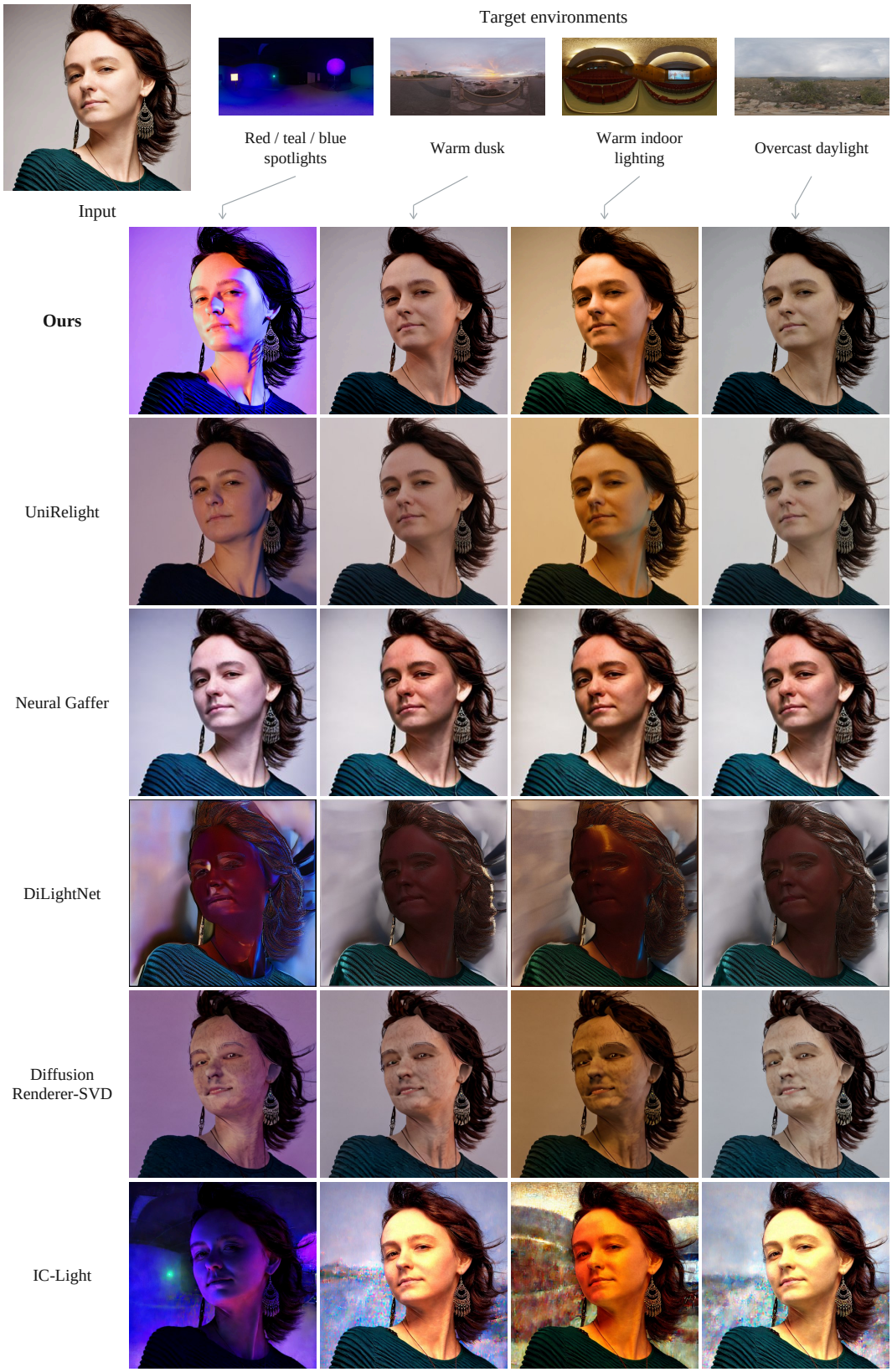



Figure 16: **Portrait: additional environments.** Columns show red/teal/blue spotlights, warm dusk, warm indoor lighting, and overcast daylight.

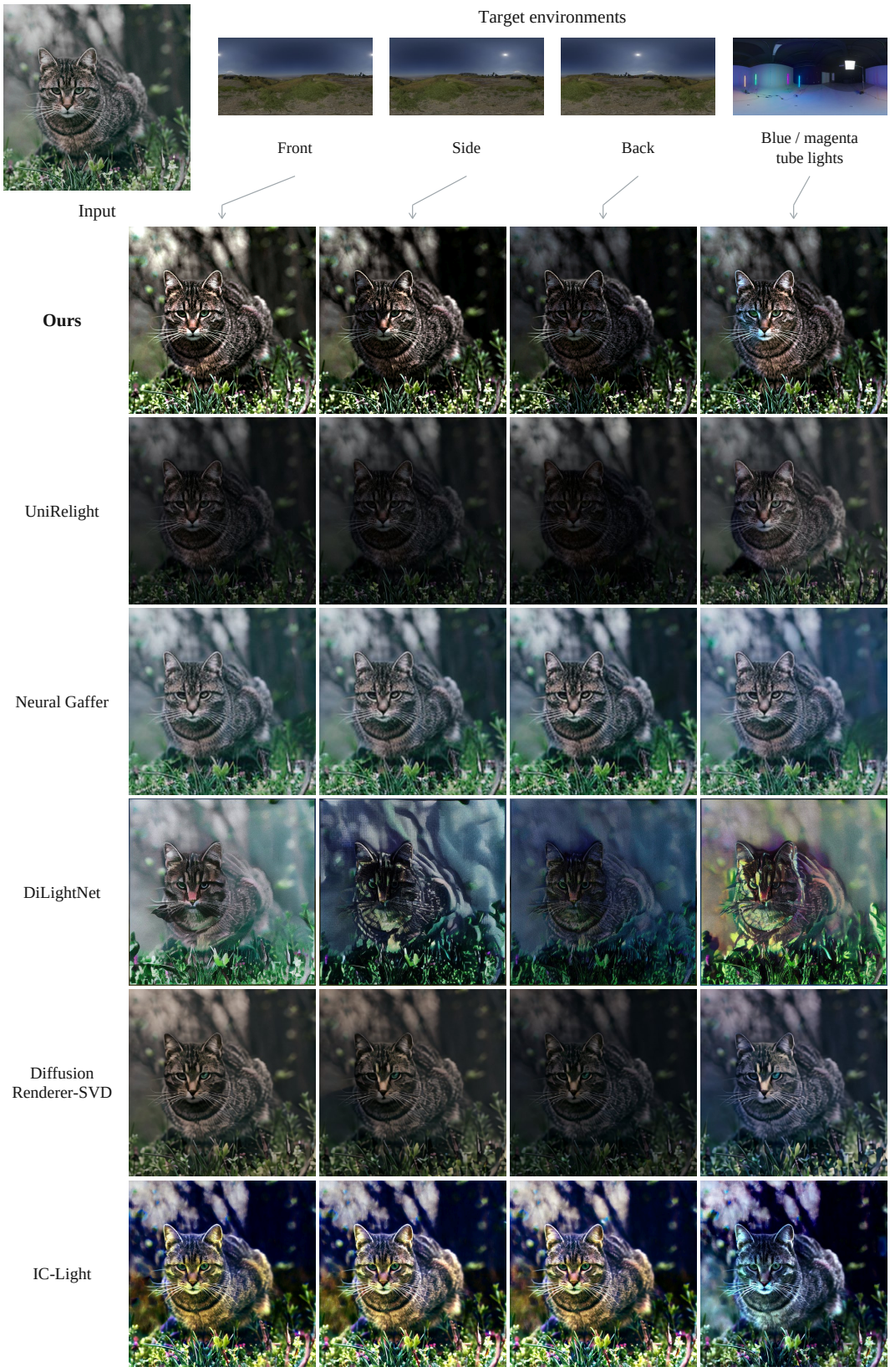



Figure 17: **Cat: directional and colored lighting.** Columns show front, side, back, and blue/magenta tube lights; rows compare the six methods.

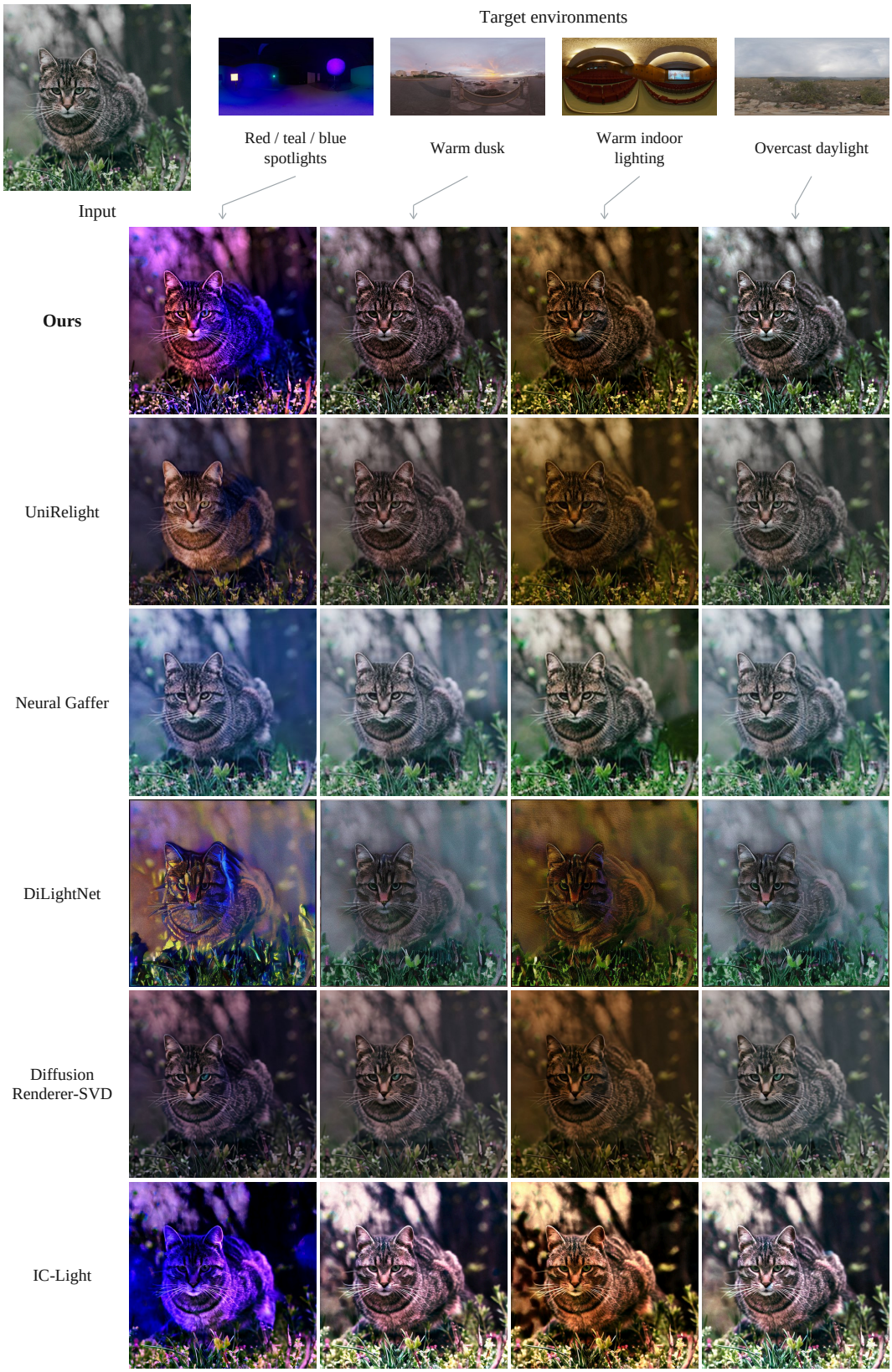



Figure 18: **Cat: additional environments.** Columns show red/teal/blue spotlights, warm dusk, warm indoor lighting, and overcast daylight.




Figure 19: **Red car: directional and colored lighting.** Columns show front, side, back, and blue/magenta tube lights.

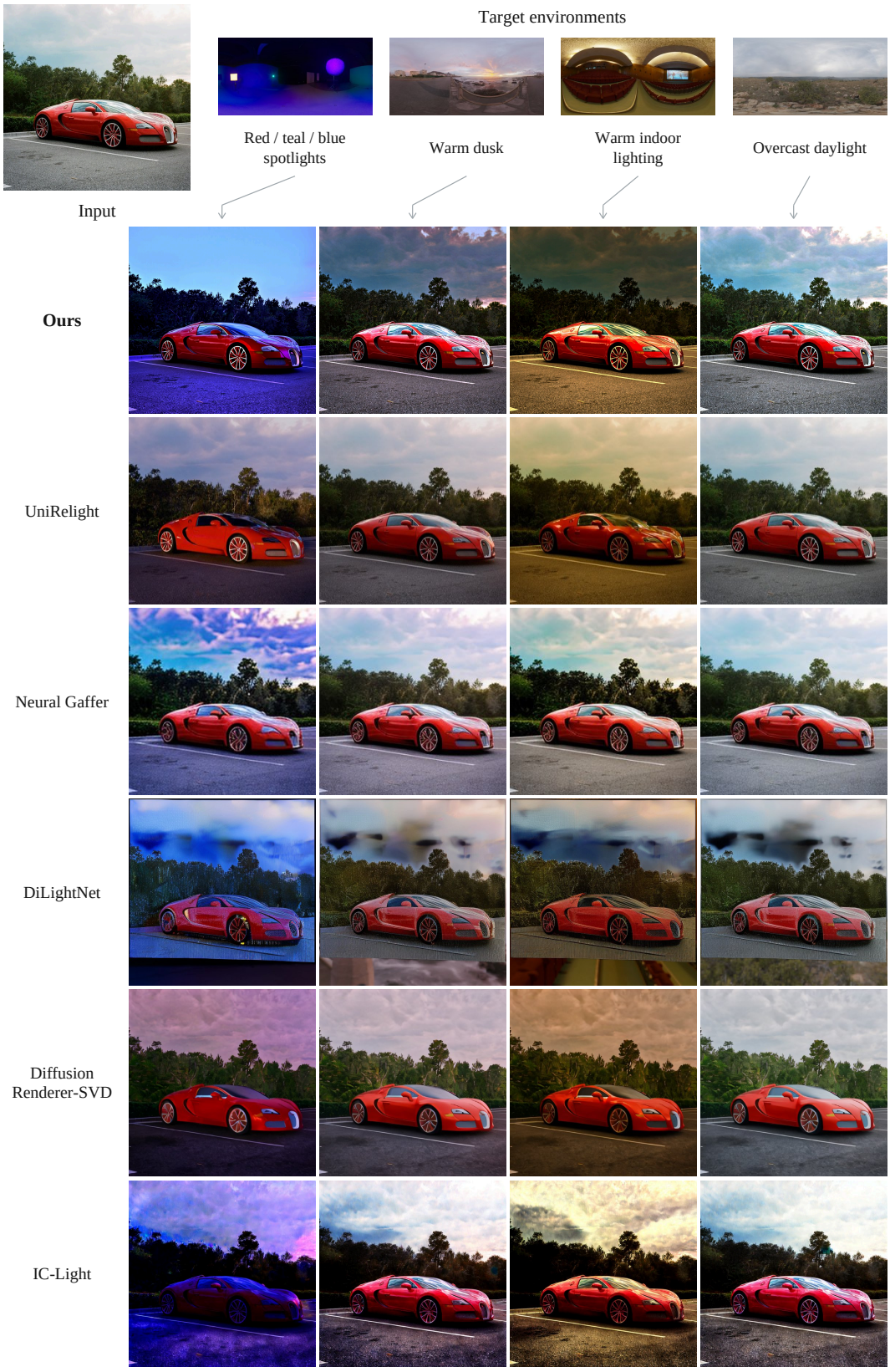



Figure 20: **Red car: additional environments.** Columns show red/teal/blue spotlights, warm dusk, warm indoor lighting, and overcast daylight.

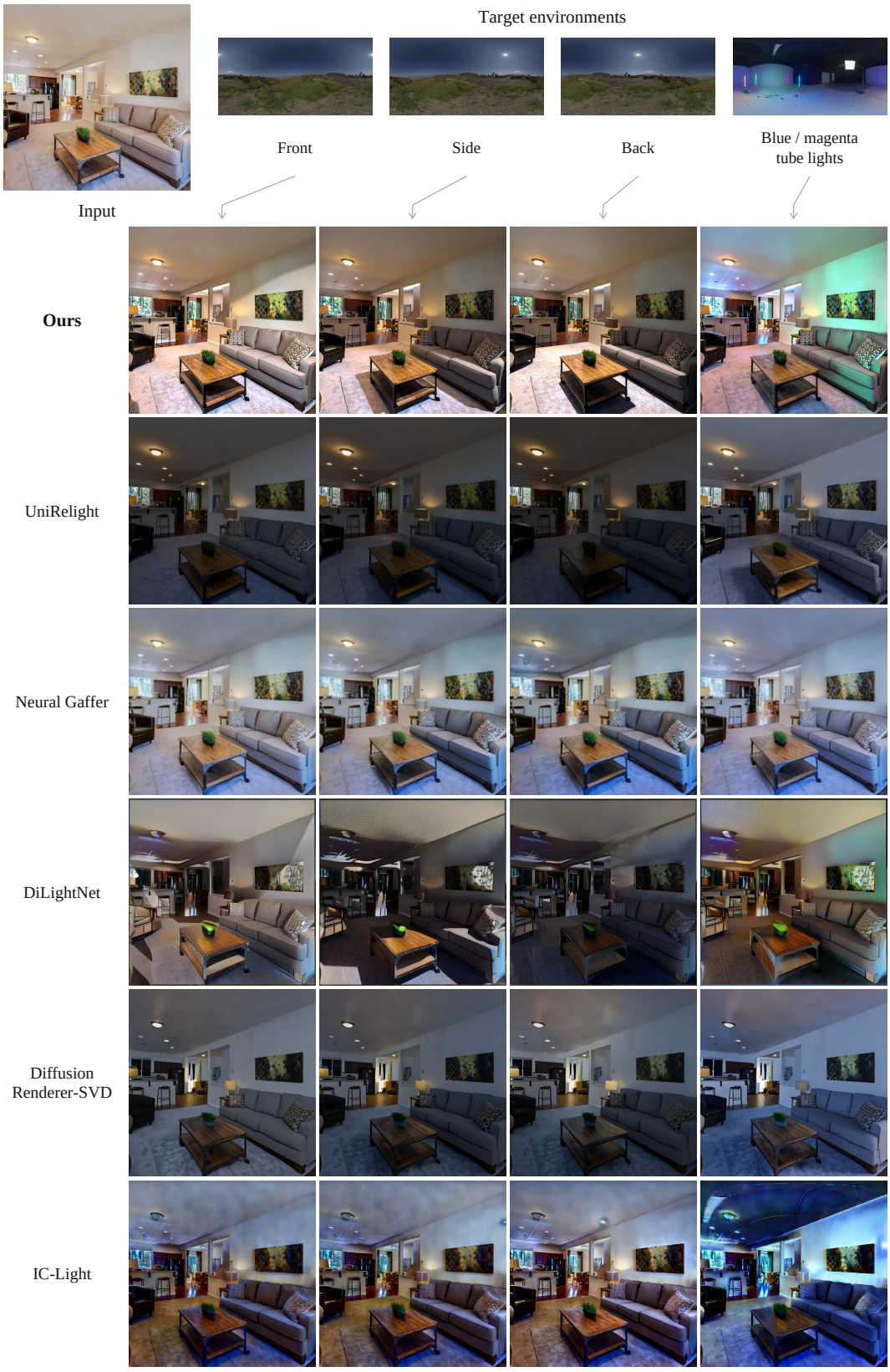



Figure 21: **Living room: directional and colored lighting.** Columns show front, side, back, and blue/magenta tube lights; all methods retain full-scene outputs.




Figure 22: **Living room: additional environments.** Columns show red/teal/blue spotlights, warm dusk, warm indoor lighting, and overcast daylight.

### B.2 Generative Priors under Masked Source Inputs

We probe whether the semantic and generative knowledge of the foundation model remains accessible after fine-tuning for HDR relighting. Starting from a photograph $I$ and its existing foreground mask $M$, we construct

$$I_{\text{masked}} = I \odot (1 - M), \tag{5}$$

where $\odot$ denotes pixelwise multiplication. In this example, $M$ covers the fox and nearby grass. Their appearance is removed by setting these regions to black, while the distant background and the foreground silhouette remain. As illustrated in Figure 23, the mask is used only to construct the RGB source image. The model then receives this image and a target environment map through its usual conditioning interface, without an additional mask input or task-specific training.

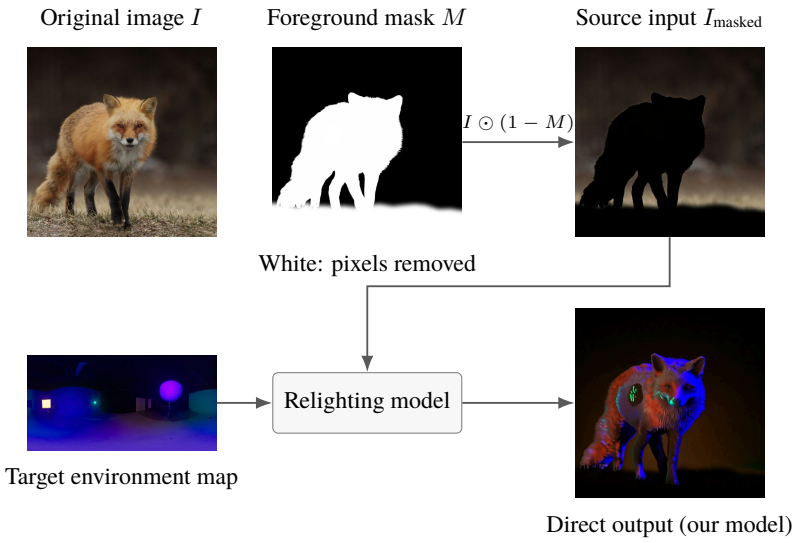



Figure 23: **Constructing a masked source input.** The existing mask removes the fox and nearby grass from the photograph. The resulting RGB image and a target environment map are passed to the relighting model. The illustrated prediction is a direct output of our model. Ours with caption additionally uses the original image and lighting captions. Source photograph: Gary Bendig, Unsplash (also used in Appendix B.1).

**Retained generative behavior.** Figure 24 compares our model and Ours with caption with four prior methods under four environment maps. Our model generates recognizable facial and fur details, particularly under red/teal/blue spotlights. Ours with caption produces more complete fox appearance and grass in several conditions. For Ours with caption, we retain the original complete-image caption, which explicitly describes a red fox, together with the corresponding lighting caption. The two configurations therefore illustrate generation with different available cues.

These observations are consistent with semantic and generative priors remaining accessible after HDR fine-tuning: the model can synthesize meaningful appearance even when the corresponding source pixels have been removed. This is a qualitative probe of one photograph. The remaining silhouette, background context, and the caption text given to Ours with caption still provide information about the missing content. Circular artifacts and incomplete appearance, especially under firelight, also show that generation is not a faithful reconstruction of the original photograph. The comparison does not isolate the contributions of pretraining, fine-tuning data, and text conditioning.




Figure 24: **Direct predictions from the same masked source under four environment maps.** All 24 results use the RGB input in Figure 23. Our model and Ours with caption synthesize object appearance in the removed region, with visible artifacts and failures. Ours with caption also receives the original complete-image caption, including "A red fox with orange fur and black legs stands on dry grass." Our HDR outputs are displayed at 0 EV using IEC sRGB; baselines retain their native LDR display. UniRelight uses the first frame of its native video.

# C Additional Paired Relighting Evaluations

We report internal synthetic and MIIW results as part of the paired evaluation introduced in Section 4, with protocols for all three benchmarks.

**OLATverse display protocol.** The 340 pairs comprise ten pairs for each of 34 objects. HDR sources are converted to LDR inputs using histogram-based auto-exposure and ACES 2.0/sRGB. For evaluation, HDR predictions and ground truth are resized to $512 \times 512$ with Blackman–Harris filtering and independently converted using the same display mapping. Baseline outputs are evaluated in their native LDR form. PSNR, SSIM, and spatial AlexNet LPIPS are computed in LDR display space using shared foreground masks, with ten-pair means averaged equally over objects. Our model uses 50-step FlowUniPC, shift 3, seed 32, initial noise standard deviation 1, and unit guidance. Public baselines retain their method-specific inference pipelines.

**Internal synthetic benchmark.** The final 20 Objaverse objects are selected by a Qwen-VL vision-language model from 80 randomly sampled held-out candidates outside the training set. Fifteen selected objects have ten lighting conditions, and five have eight; fixing one lighting condition as the source for each object gives nine or seven target conditions, respectively, yielding $15 \times 9 + 5 \times 7 = 170$ source–target pairs. We evaluate these pairs using fixed Reinhard tone mapping followed by the IEC sRGB transfer function and geometry masks on a $512 \times 512$ canvas. The retained image region is placed at its original coordinates, with zero-valued masked borders; it is not stretched to fill the canvas. Scores are averaged equally over all 170 pairs.

**MIIW.** We evaluate 375 pairs from 25 of the 30 scenes in the official MIIW test split (Murmann et al., 2019). Evaluation uses central $1000 \times 1000$ crops, with light probes masked as in UniRelight (He et al., 2025). HDR predictions use a fixed-exposure gamma-2.2 display mapping. PSNR, SSIM, and scalar AlexNet LPIPS are averaged equally over all pairs.

**A different lighting input.** MIIW specifies lighting differently from our training data: each condition is a flash aimed at a direction over the camera's upper hemisphere, recorded with chrome probes (Murmann et al., 2019). The model thus receives a probe-derived panorama rather than the full-sphere HDR environment maps it was trained on (Figure 25), so results without adaptation would be confounded by this input mismatch. For MIIW, we therefore adapt the model to this lighting input by fine-tuning on the official MIIW training split, keeping training and test scenes separate; UniRelight likewise uses MIIW for 14.3% of samples in its second training stage (He et al., 2025). With this adaptation, our model surpasses UniRelight on all three metrics under our protocol (Table 5). We keep this fine-tuning short, because longer fine-tuning degraded lighting control under full-sphere environment maps in qualitative comparisons.

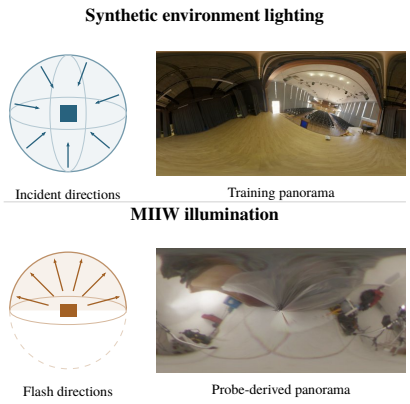



Figure 25: **Lighting acquisition.** Our training lighting is a full-sphere HDR environment map; MIIW lighting is a flash aimed over the upper hemisphere, recorded with chrome probes and converted to a panorama.

## C.1 Additional Qualitative Comparisons

The following examples show selected internal synthetic and MIIW comparisons. These selected qualitative examples are separate from the benchmark averages in Table 5. Internal synthetic examples use the same foreground mask for every method; MIIW examples retain the complete benchmark display.

Table 5: Additional paired evaluations. Each dataset uses its own display and metric protocol; scores are not pooled across datasets. MIIW scores use the model adapted to MIIW lighting by fine-tuning on its training split.

| | Internal (170 pairs) | | | MIIW (375 pairs) | | |
|---|---|---|---|---|---|---|
| Method | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
| Neural Gaffer | 17.29 | 0.568 | 0.291 | 16.73 | 0.668 | 0.465 |
| DiLightNet | 17.12 | 0.550 | 0.263 | 12.97 | 0.452 | 0.465 |
| DiffusionRenderer-SVD | 16.05 | 0.595 | 0.240 | 15.78 | 0.630 | 0.396 |
| Light-X | 13.27 | 0.520 | 0.345 | 14.69 | 0.630 | 0.518 |
| IC-Light | 12.19 | 0.489 | 0.312 | 14.26 | 0.573 | 0.368 |
| UniRelight | 16.22 | 0.670 | 0.196 | 19.49 | 0.764 | 0.311 |
| Ours | **24.33** | **0.812** | **0.127** | **21.37** | **0.804** | **0.220** |

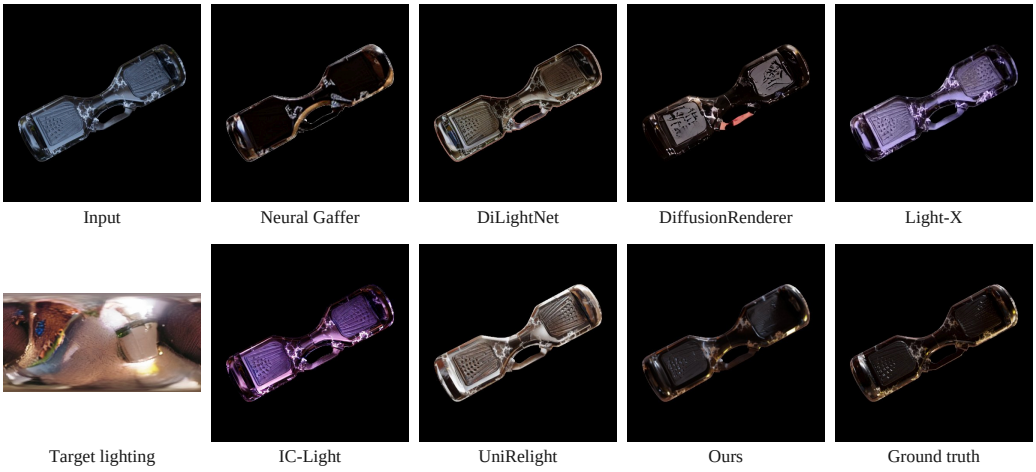



Figure 26: **Internal synthetic benchmark.** Comparison on one source–target pair.

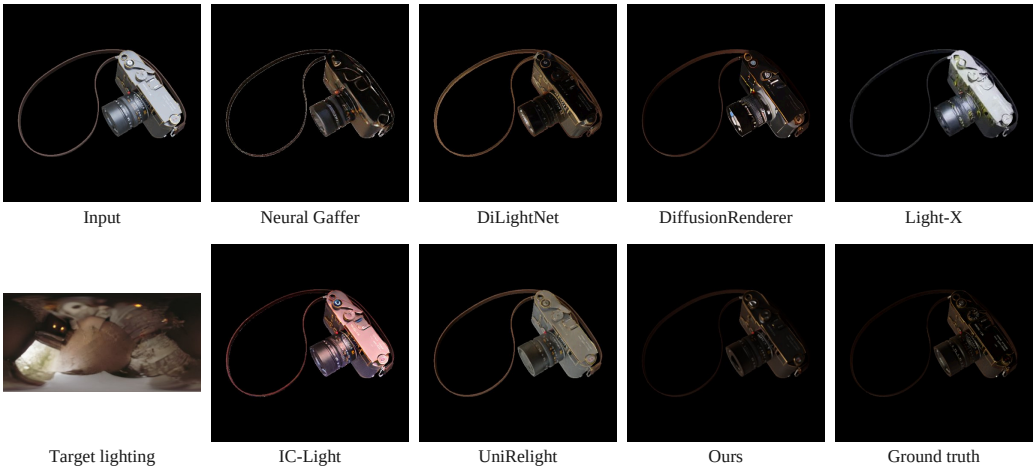



Figure 27: **Internal synthetic benchmark.** Comparison on one source–target pair.

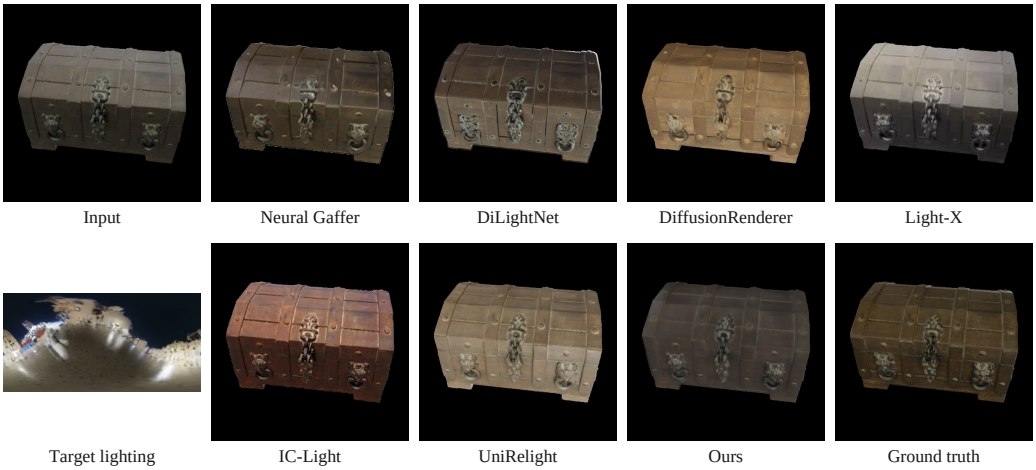



Figure 28: **Internal synthetic benchmark.** Comparison on one source–target pair.

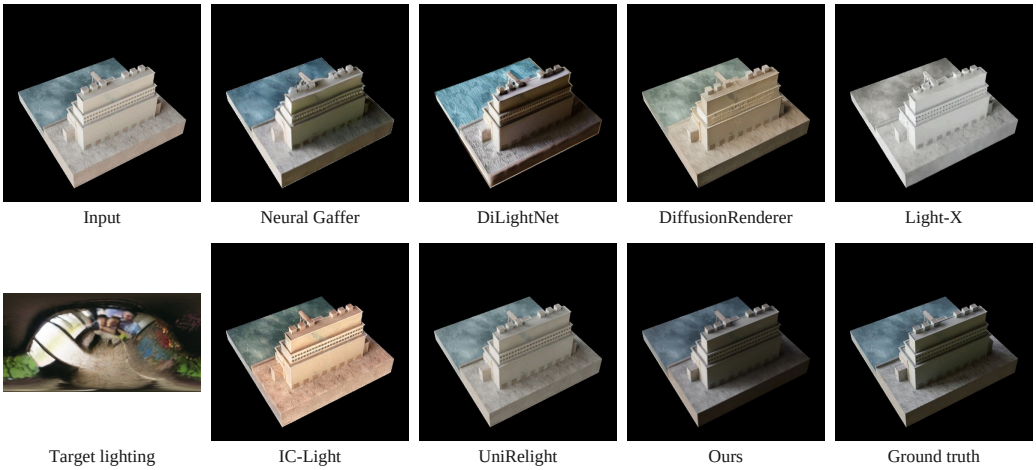



Figure 29: **Internal synthetic benchmark.** Comparison on one source–target pair.




Figure 30: **MIIW.** Comparison on one source–target pair.




Figure 31: **MIIW.** Comparison on one source–target pair.




Figure 32: **MIIW.** Comparison on one source–target pair.

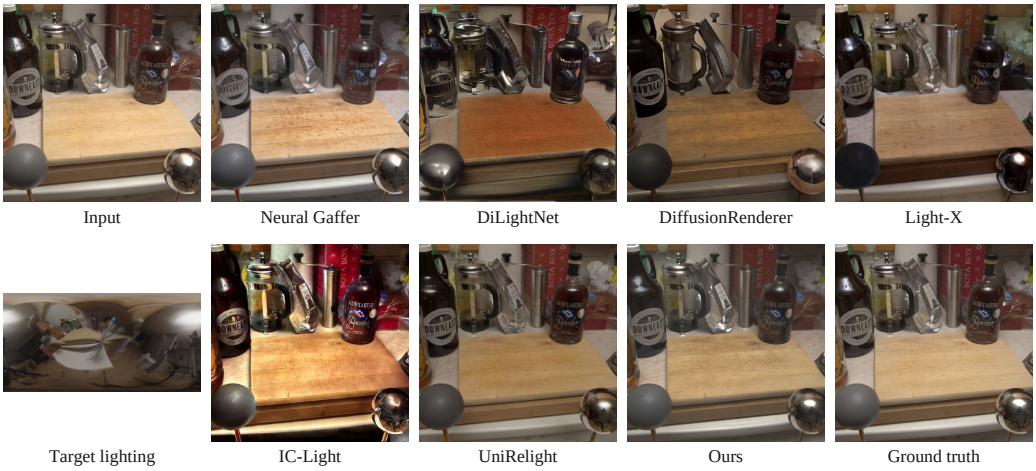



Figure 33: **MIIW.** Comparison on one source–target pair.




Figure 34: **MIIW.** Comparison on one source–target pair.

# D Benefits of Native HDR Relighting

## D.1 HDR Reconstruction from Clipped Relighting Outputs

**Motivation.** An alternative to native HDR relighting is to reconstruct HDR from an LDR relighting output. Comparing two independently trained relighting systems would mix differences in relighting quality with differences introduced by HDR reconstruction. We therefore hold the relighting result fixed and study the round trip

$$H \longrightarrow I \longrightarrow \widehat{H}_m, \tag{6}$$

where $H$ is a selected HDR relighting result, $I$ is its clipped, quantized LDR export, and $\widehat{H}_m$ is reconstructed by method $m$. This tests how well a reconstruction stage preserves the exposure-dependent appearance of the original output after the LDR bottleneck. The LDR image is a proxy obtained from our own result; no separate LDR relighting model is used in this experiment.

**Examples and input construction.** We use the living room from the teaser, the sunset sky, porcelain vases, and white car from our supplementary exposure examples, and a glass-bottle relighting output used in the main-paper qualitative results. These five examples cover localized light sources, large clipped sky regions, and reflective materials. All five references are native model predictions. Figure 9 in the main paper uses four further predictions: two overexposed ones exported at $0$ EV and two underexposed ones exported at $-4$ EV, reconstructed with DiffHDR, X2HDR, and LEDiff from shared $512 \times 512$ inputs, with one fitted brightness gain per reconstruction.

We clamp negative values in each $2048 \times 2048$ reference and area-resize it to $512 \times 512$ in linear RGB, obtaining $H$. Let $g$ be the IEC sRGB transfer function and $Q_8(z) = \text{round}(255z)/255$. The shared input is

$$I = Q_8(g(\text{clip}(H, 0, 1))), \qquad L = g^{-1}(I). \tag{7}$$

Here, $0$ EV means unit multiplication of the selected EXR, with no automatic exposure or preliminary darkening; it does not denote calibrated scene radiance. In particular, we use the unscaled room EXR, whereas its teaser display uses a $-1.5$ EV base. The reconstruction models receive the same quantized image information, with $L$ used for IntrinsicHDR's linear-input interface. No exposure value is supplied as a conditioning token, and no reference HDR is provided to the reconstruction models. Input clipping fractions are reported in Table 7.

**Reconstruction methods.** We use pretrained DiffHDR (Yu et al., 2026b), LEDiff (Wang et al., 2025a), IntrinsicHDR (Dille et al., 2024), and X2HDR (Wu et al., 2026), without fine-tuning on these examples. Table 6 lists the inference settings. DiffHDR uses its single-image path, which replicates the image into 125 frames; we retain the first output frame. LEDiff uses its highlight-reconstruction checkpoint and normal internal merge, without the optional additional

input-fusion helper. IntrinsicHDR uses the known inverse-sRGB transform in equation (7) in place of optional learned camera-response preprocessing; its reconstruction core is unchanged. X2HDR uses its LDR-to-HDR version of FLUX.1-dev and retains the official PU21 decoding and output normalization. Text-conditioned methods use empty positive prompts. We run one fixed seed per method and example.

Table 6: **HDR reconstruction settings.** All methods use the same $512 \times 512$ input information. Diffusion-based methods use empty positive prompts. Seed states are reset for each example.

| Method | Steps | Seed | Configuration |
|---|---|---|---|
| DiffHDR | 50 | 34 | Guidance 1; Wan2.1-VACE-14B with DiffHDR weights; single-image replication. |
| LEDiff | $50 + 50$ | 42 | Guidance 7.5; official highlight pipeline with two diffusion stages. |
| IntrinsicHDR | – | 0 | Official v1.0 reconstruction weights; known-sRGB input conversion. |
| X2HDR | 30 | 42 | Guidance 3.5; FLUX.1-dev with the official LDR-to-HDR model. |

**Input-only exposure alignment.** Reconstruction methods can produce different global output scales. We align each output to the unsaturated input using a single scalar, without fitting to $H$. Define linear luminance as $Y(X) = 0.2126X_R + 0.7152X_G + 0.0722X_B$. We first select pixels with $\max_c I_c < 0.95$ and $Y(L) > 0.01$, then retain those between the 10th and 90th input-luminance percentiles of this eligible set. This defines one mask $M$ shared by all methods for that example. For each reconstruction, we compute

$$a_m = \frac{\sum_{p \in M} Y(\widehat{H}_m(p))\, Y(L(p))}{\sum_{p \in M} Y(\widehat{H}_m(p))^2}, \qquad \widetilde{H}_m = a_m \widehat{H}_m. \tag{8}$$

All fitted gains are positive. Each gain is shared across RGB channels and frozen across the displayed exposures. We apply no gamma fit, per-channel correction, or local adjustment. The original reference is not rescaled by this procedure. Table 7 records all gains; these values are alignment parameters, not quality scores.

Table 7: **Input clipping and fixed exposure gains.** Clipped denotes the fraction of pixels with at least one linear reference channel at or above 1 before the 0-EV export. The remaining columns are the global RGB multipliers $a_m$ from equation (8); they are not reconstruction-quality metrics.

| Example | Clipped | DiffHDR | LEDiff | IntrinsicHDR | X2HDR |
|---|---|---|---|---|---|
| Living room | 6.78% | 1.0996 | 0.5746 | 0.9810 | 1.1978 |
| Sunset sky and red car | 36.76% | 0.3518 | 0.1740 | 0.8312 | 0.2213 |
| Porcelain vases | 23.74% | 0.7640 | 0.3239 | 0.7353 | 0.5450 |
| White car | 13.08% | 0.9199 | 0.1602 | 0.8490 | 0.2338 |
| Glass bottle | 40.94% | 0.7388 | 0.8090 | 0.8464 | 2.3029 |

**Exposure comparison.** We display $H$, the inverse-sRGB LDR input $L$, and each aligned reconstruction at $e \in \{0, -2, -4\}$ EV using the same operator:

$$D_e(X) = Q_8(\text{clip}(g(2^e \max(X, 0)), 0, 1)). \tag{9}$$

The three columns are exposure changes of one saved output, not separate inferences. Lowering exposure reveals variations retained above the LDR clipping threshold and exposes differences in reconstructed highlights. There is no additional tone mapping or exposure-specific normalization.

**Observations and interpretation.** Figures 35 to 39 show all five examples. In the room, the methods differ in the intensity and appearance of lamps and windows, with accompanying color changes. For the sunset, lowering exposure reveals orange-pink cloud structure in the reference; reconstructed skies differ in both structure and color. The porcelain, car, and glass examples similarly

show differences in highlight extent, gradients, and reflection color. The reconstruction methods can produce values beyond the LDR range, but this does not by itself imply recovery of the original highlight appearance.

These results illustrate the information lost when an existing HDR relighting output is reduced to LDR, and the resulting ambiguity of reconstructing its highlights. The selected HDR outputs are references for this round-trip test, not measured ground truth. Native HDR retains the pre-export information by construction, so this experiment alone does not establish more accurate scene radiance or a general ranking of relighting systems. Its conclusion is limited to preserving the exposure-dependent appearance of these five fixed outputs under the specified export and pretrained reconstruction settings.

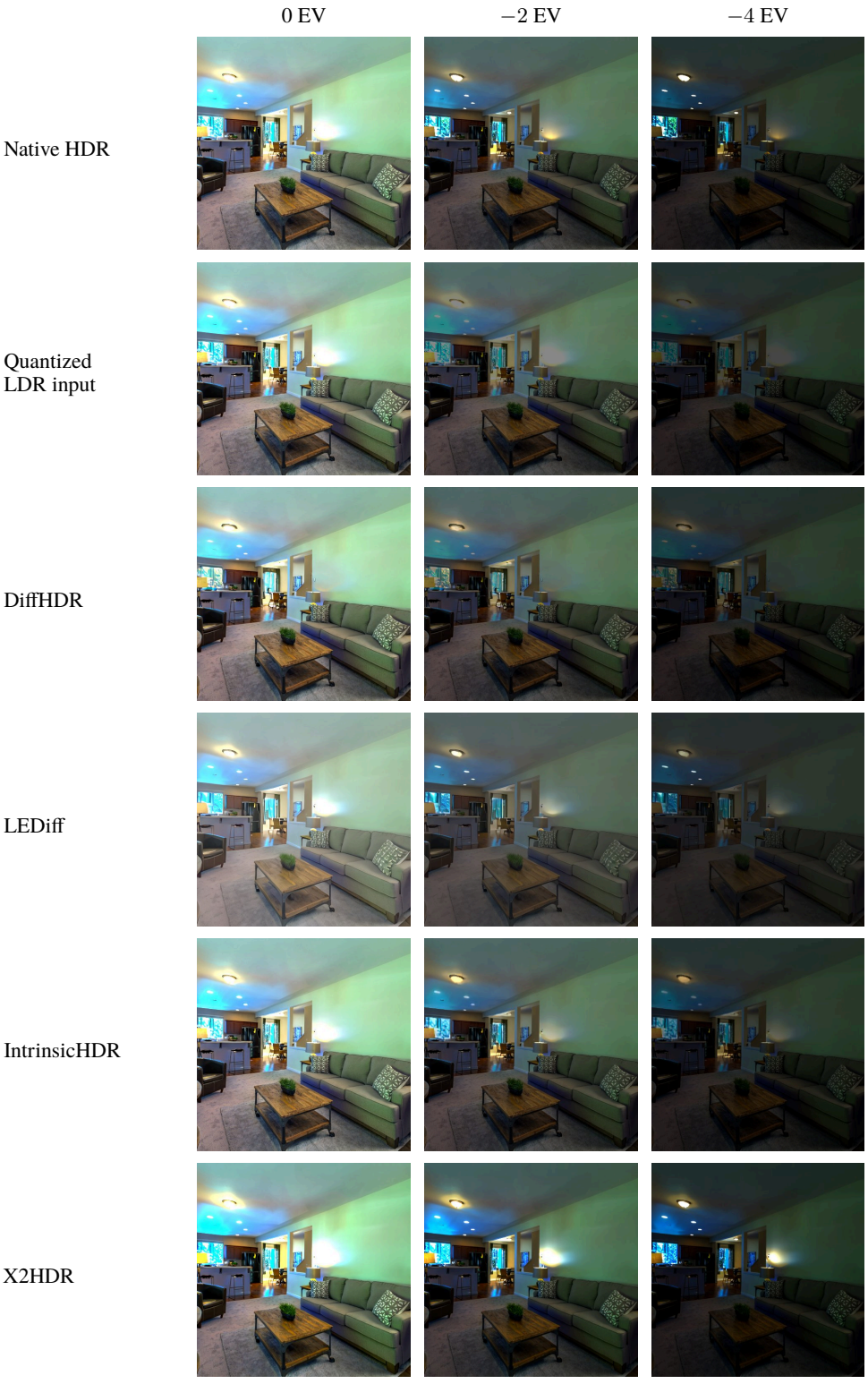



Figure 35: **Living room.** Lowering exposure reveals differences in lamp and window appearance. Reconstruction also changes some colors and the illumination of the right wall. Every row uses one fixed output displayed at three exposures. All reconstruction scales are fitted only to unsaturated input pixels and held fixed across columns.




Figure 36: **Sunset sky and red car.** The original HDR reveals orange-pink cloud variations at lower exposures. The reconstructions differ in cloud structure and color in the region clipped by the LDR export. Every row uses one fixed output displayed at three exposures. All reconstruction scales are fitted only to unsaturated input pixels and held fixed across columns.

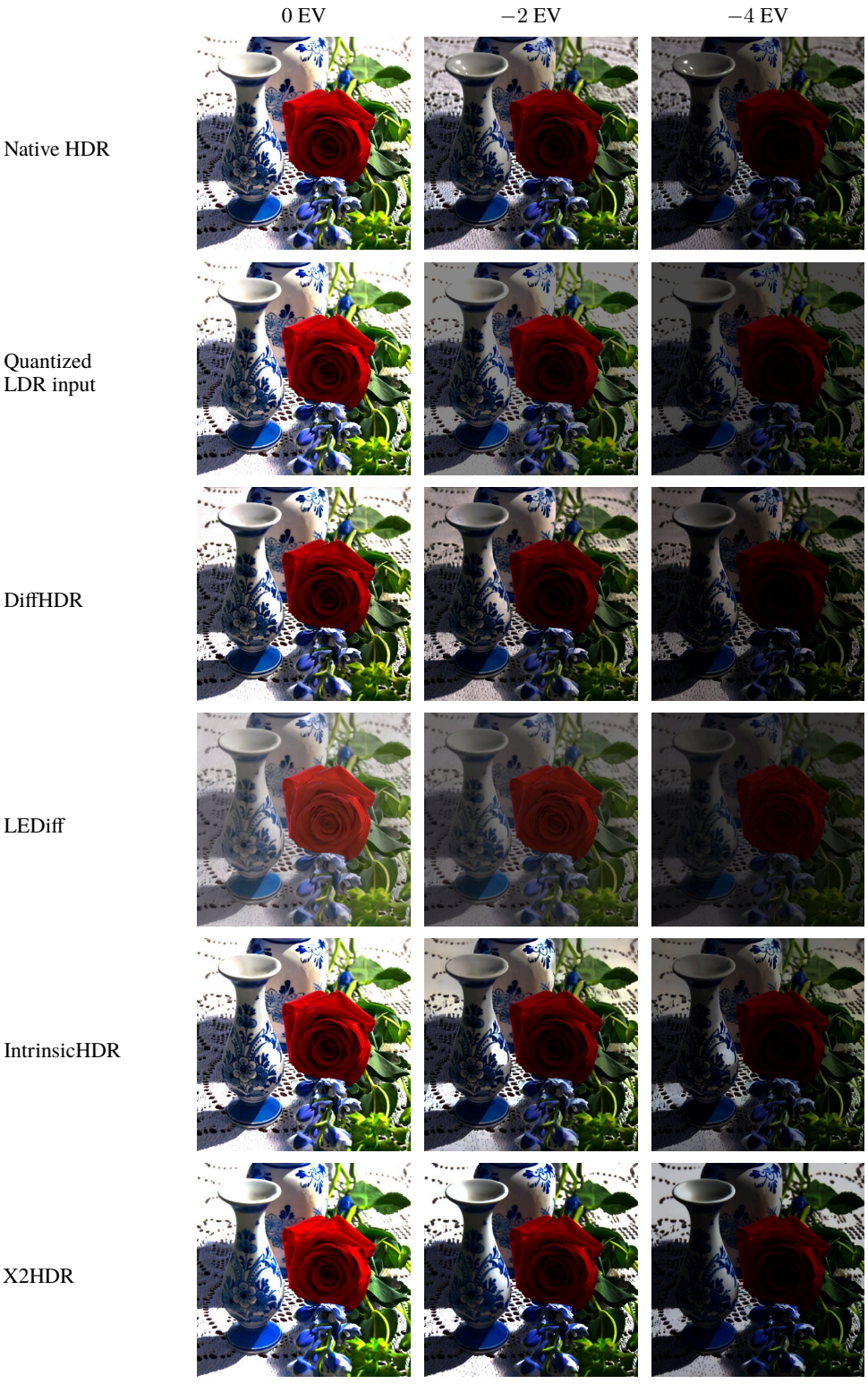



Figure 37: **Porcelain vases.** The outputs differ in the extent and intensity of glaze highlights and in local brightness gradients. These differences become clearer as exposure is lowered. Every row uses one fixed output displayed at three exposures. All reconstruction scales are fitted only to unsaturated input pixels and held fixed across columns.




Figure 38: **White car.** The clipped hood becomes flat gray when the LDR input is darkened. Reconstructions produce different highlight gradients and windshield colors. Every row uses one fixed output displayed at three exposures. All reconstruction scales are fitted only to unsaturated input pixels and held fixed across columns.

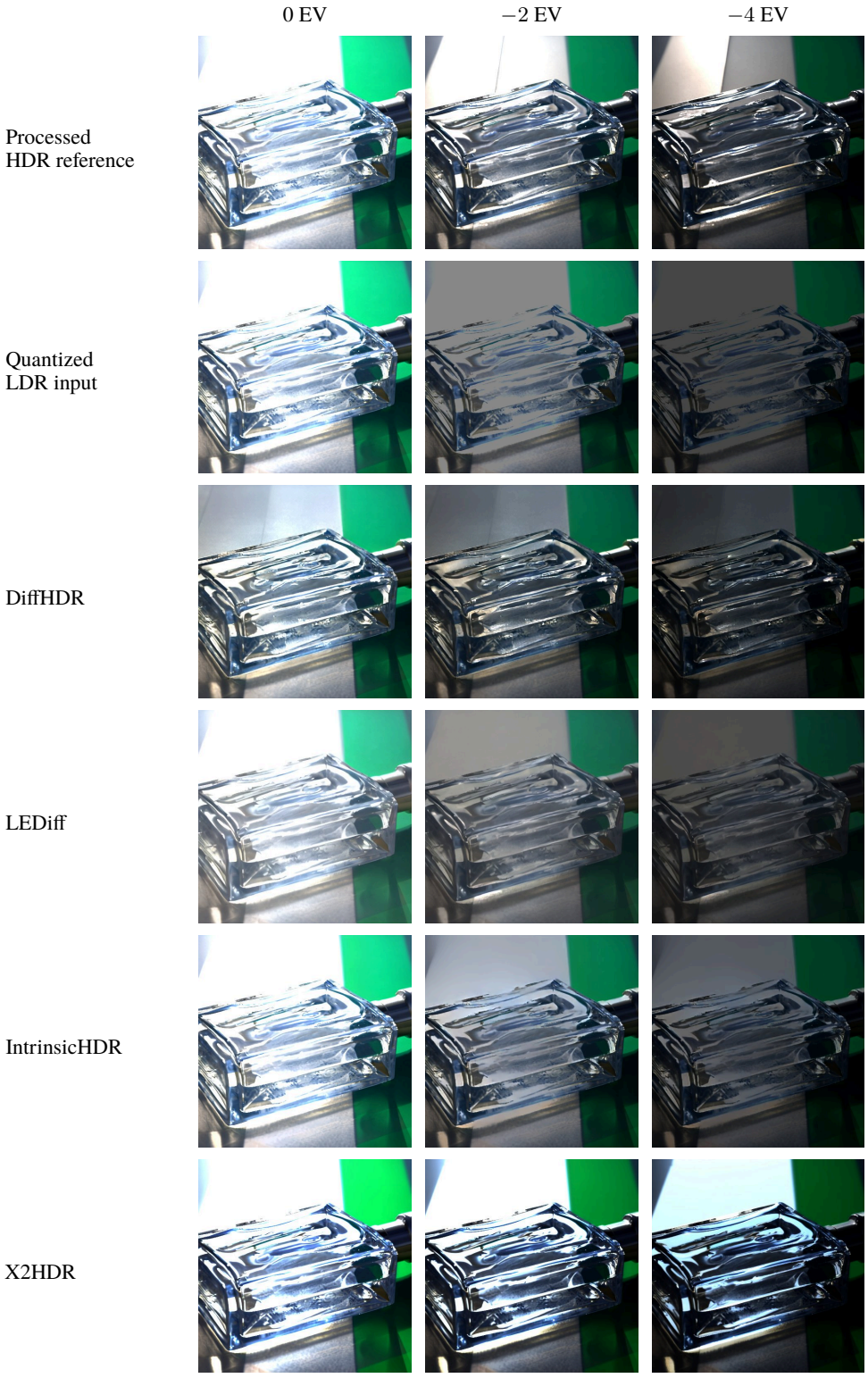



Figure 39: **Processed glass bottle.** This reference is the processed EXR used for the paper illustration, including existing highlight edits. Reconstructions differ in reflection structure, blue-cyan color, and background brightness; it is not measured ground truth. Every row uses one fixed output displayed at three exposures. All reconstruction scales are fitted only to unsaturated input pixels and held fixed across columns.

### D.2 Exposure Adjustment after Relighting

We show six HDR relighting results generated by our model, each displayed at four exposures. Source images and target environments are shown above. We compare native HDR with its clipped, 8-bit SDR export at 0 EV, undoing only the SDR's sRGB encoding before adjusting exposure. Native HDR retains highlight variations lost during export, enabling more flexible brightness adjustment.




Figure 40: **White car.** Four exposures of one HDR relighting result generated by our model (top), compared with its 0-EV SDR export (bottom).

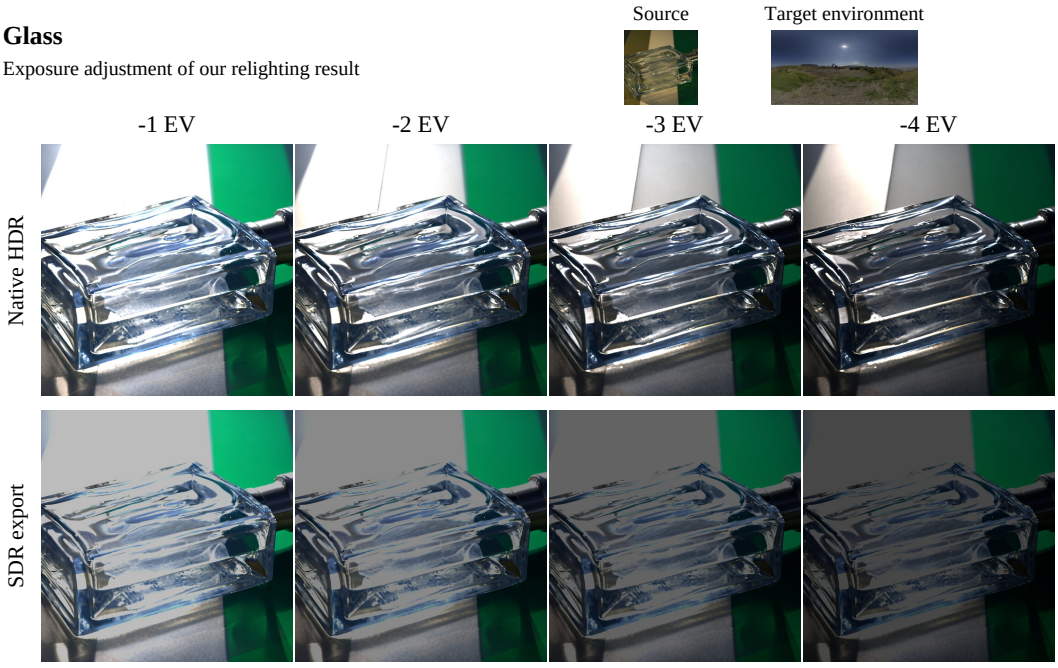



Figure 41: **Glass.** Four exposures of one HDR relighting result generated by our model (top), compared with its 0-EV SDR export (bottom).

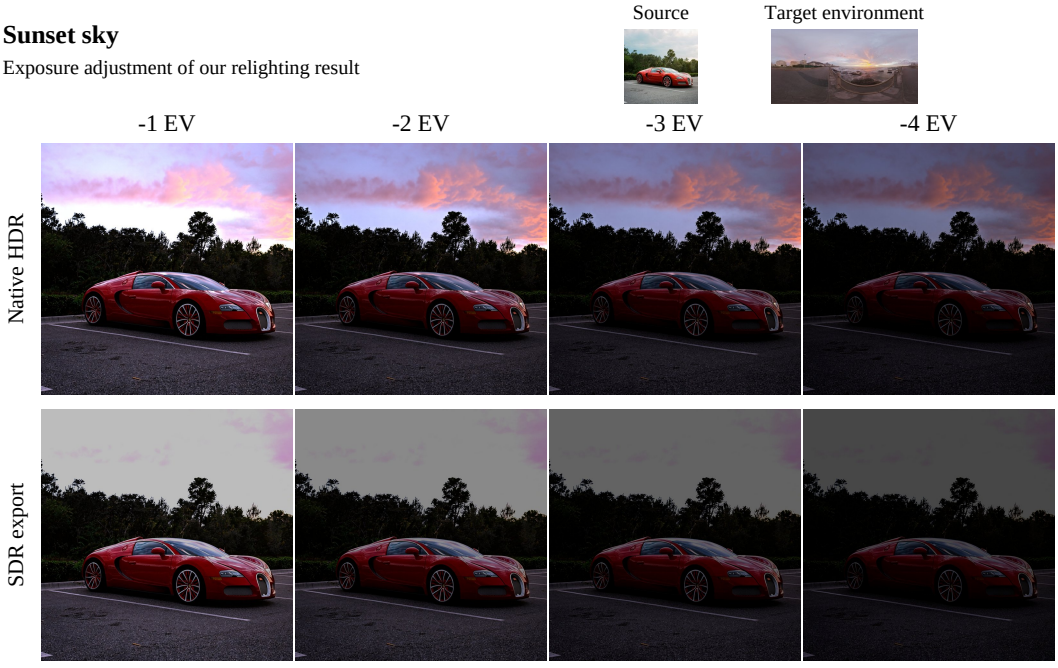



Figure 42: **Sunset sky.** Four exposures of one HDR relighting result generated by our model (top), compared with its 0-EV SDR export (bottom).

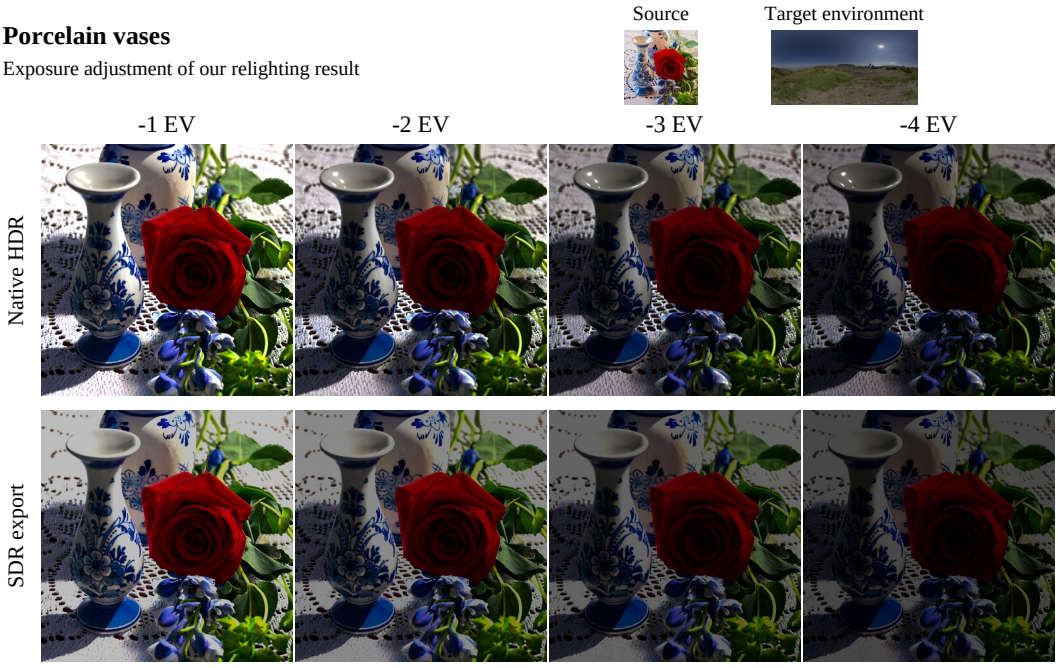



Figure 43: **Porcelain vases.** Four exposures of one HDR relighting result generated by our model (top), compared with its 0-EV SDR export (bottom).




Figure 44: **Colored interior.** Four exposures of one HDR relighting result generated by our model (top), compared with its 0-EV SDR export (bottom).




Figure 45: **Interior under overcast illumination.** Four exposures of one HDR relighting result generated by our model (top), compared with its 0-EV SDR export (bottom).

# E Light-Transport Consistency

## E.1 Controlled Studies of Light Composition

### E.1.1 Question, data, and prediction task

These studies compare LC with an IC-Light-style constraint for preserving physical additivity in a generative model in a controlled setting. We use rendered HDR responses of 2,361 scenes, each containing a randomly sampled Objaverse object (Deitke et al., 2023), with three lighting responses per scene, resized to $32 \times 32$ RGB. For each scene, ordinary flow matching observes only the first two lighting conditions. Their exact HDR sum is withheld from the ordinary supervised loss and provides a mixed-light evaluation target. The third lighting condition is also excluded from training; these experiments do not demonstrate recovery of an unseen lighting direction.

Every query uses a fixed reference image, the scene's first observed response, and a 1,536-dimensional environment-light condition. The two lights are added before encoding or normalization changes. The reference is held fixed across the two component queries and the summed-light query. This controlled, fixed-reference setting is distinct from arbitrary-source relighting in the full model.

The reduced PixelDiT receives six image channels: the three-channel noisy state and the three-channel reference. It predicts three velocity channels. Patch width is 256, pixel width 64, with eight patch-transformer blocks, four pixel-transformer blocks, patch size four, and eight groups. It is not the full-resolution production backbone.

### E.1.2 Coordinates and three training methods

We run separate studies in physical linear coordinates and Log-Gamma coordinates. Their encodings are

$$T_{\text{lin}}(H) = H, \tag{10}$$

$$T_{\text{LG}}(H) = 2\,\text{sgn}(H)\left[\frac{\log(1+\gamma_{\text{LG}}|H|)}{\log(1+\gamma_{\text{LG}}M)}\right]^{1/\gamma_{\text{LG}}} - 1, \tag{11}$$

$$T_{\text{LG}}^{-1}(x) = \frac{\text{sgn}(u)}{\gamma_{\text{LG}}}\left[(1+\gamma_{\text{LG}}M)^{|u|^{\gamma_{\text{LG}}}} - 1\right], \qquad u = \frac{x+1}{2}. \tag{12}$$

Here $M = 50$ and $\gamma_{\text{LG}} = 2.2$, with $M$ fixing the scale point $T_{\text{LG}}(M) = 1$. Values above $M$ map above 1 and remain recoverable through the inverse. The sign terms extend the coordinate map over the full real range, including negative intermediate states. The transform is applied elementwise, and evaluation uses its inverse to recover linear HDR before computing metrics. Below, $T$ denotes the encoding used in the respective study, either $T_{\text{lin}}$ or $T_{\text{LG}}$.

**Ordinary flow matching.** For an observed response $H$, Gaussian noise $n$, and $t \sim \mathcal{U}[0,1)$, the encoded state is $z_t = (1-t)n + tT(H)$ and the velocity target is $T(H) - n$. No ground-truth summed response is added to the ordinary supervised loss.

**Light-transport linearity constraint (LC).** The constraint queries two component states and a composed state. For component HDR responses $H_i$ and Gaussian noises $n_i$, define the path scale $\gamma_t = 1 + (\sqrt{2}-1)t$ and component query time $s_t = \sqrt{2}t/\gamma_t$. The states are

$$z_i = (1-s_t)n_i + s_t T(H_i), \qquad i \in \{1,2\}, \tag{13}$$

$$h_\Sigma = \frac{\gamma_t}{\sqrt{2}}\big(T^{-1}(z_1) + T^{-1}(z_2)\big), \qquad z_\Sigma = T(h_\Sigma). \tag{14}$$

Here $h_\Sigma$ is the composed linear-HDR state and $z_\Sigma$ its encoding. The two component queries use time $s_t$ and their respective lights; the composed query uses time $t$ and the sum of the lights. The LC target is the derivative of this physical composition, expressed in the model's coordinates through the encoding Jacobian. This is the trajectory-based construction discussed in Appendix H; the nonlinear-coordinate version is a soft consistency constraint, not a proof of matching intermediate flow marginals.

**IC-Light-style constraint.** We modify the learned combination constraint of IC-Light (Zhang et al., 2025) to this controlled toy using a two-input MLP. At the same three query locations, let $v_1, v_2$ be the component velocity predictions and $v_\Sigma$ the summed-light velocity prediction. A shared per-pixel network maps the concatenated component velocities to a composed velocity:

$$\phi : \mathbb{R}^6 \longrightarrow \mathbb{R}^3, \qquad \mathcal{L}_{\text{MLP}} = \mathbb{E}\left[\|v_\Sigma - \phi(v_1, v_2)\|_2^2\right]. \tag{15}$$

Its five linear layers have widths $6, 128, 128, 128, 128, 3$, with ReLU after the first four layers, biases, and default PyTorch linear initialization. These activation and initialization choices are implementation choices for this comparator, not a claim to reproduce unpublished details of IC-Light. The MLP receives neither state, time, light, reference, nor an analytic velocity correction. In Log-Gamma coordinates its inputs are the raw Log-Gamma velocities. All three backbone predictions and the MLP receive gradients; none is detached. The MLP is discarded at inference, which uses only the backbone's exponential moving average.

### E.1.3 Training and matched evaluation

Each method is trained from random initialization to 300,000 updates. The physical and Log-Gamma studies use seeds 20260741 and 20260803, respectively. The ordinary batch size is 64 and the consistency batch size is 32. Training uses AdamW with learning rate $2 \times 10^{-4}$, $(\beta_1, \beta_2) = (0.9, 0.999)$, zero weight decay, gradient-norm clipping at one, and backbone EMA decay 0.999. The consistency weight is $0.01\,\min(k/10000, 1)$ at update $k$. The MLP weights are initialized in an isolated random stream so they do not advance the backbone's initialization stream.

Table 8: Matched 300k reduced-PixelDiT evaluation. Pooled relative errors (%) on fixed cohorts, lower is better. The coordinate groups are independent studies.

| Coordinates | Training method | Single light | Mixed light | Additivity |
|---|---|---|---|---|
| Physical linear | Ordinary flow matching | 1.73 | 46.20 | 46.25 |
| | LC (ours) | 1.76 | 3.68 | 3.32 |
| | IC-Light-style constraint | 1.81 | 21.53 | 21.52 |
| Log-Gamma | Ordinary flow matching | 5.80 | 41.38 | 42.11 |
| | LC (ours) | 5.85 | 23.53 | 24.23 |
| | IC-Light-style constraint | 5.51 | 52.21 | 53.11 |

All ordinary-flow times are uniform on $[0, 1)$. Physical-coordinate LC and both IC-Light-style settings also use the full interval. The Log-Gamma LC setting uses $[0.3, 1)$ for the consistency time, whereas its IC-Light-style comparator uses $[0, 1)$. Consequently the Log-Gamma comparison is between these specified methods, not a strict one-factor module ablation.

The primary evaluation uses 512 fixed samples per evaluation cohort for the physical study and 256 for Log-Gamma coordinates. Within each coordinate study, methods share saved scene indices, light pairs, and initial noise tensors. The coordinate studies use separate evaluation cohorts; all comparisons in Table 8 are within a coordinate study.

We use DPM-Solver++ with 50 model queries, second-order multistep updates, the time-uniform flow schedule, flow shift three, lower-order final update, and no terminal denoising query. The same saved component noises $n_1, n_2$ are used for all methods, and the summed-light query starts from $(n_1 + n_2)/\sqrt{2}$. We decode all generated endpoints to unclipped linear HDR before evaluation.

**Metrics.** For sample $j$, let $H_{1j}, H_{2j}$ denote the observed HDR component responses, $P_{1j}, P_{2j}, P_{\Sigma j}$ the generated component and summed-light responses, and $R_j = H_{1j} + H_{2j}$ the true summed response. We report pooled relative errors over the entire fixed cohort:

$$e_{\text{mix}} = \sqrt{\frac{\sum_j \|P_{\Sigma j} - R_j\|_2^2}{\sum_j \|R_j\|_2^2}}, \tag{16}$$

$$e_{\text{add}} = \sqrt{\frac{\sum_j \|P_{\Sigma j} - (P_{1j} + P_{2j})\|_2^2}{\sum_j \|P_{1j} + P_{2j}\|_2^2}}. \tag{17}$$

Observed single-light prediction error is defined analogously against the real observed response on its own fixed cohort. These are pooled errors, not averages of per-image relative errors. The first evaluates physical image prediction; the second evaluates generated self-consistency. Neither is a tone-mapped PNG difference or a physical-velocity closure diagnostic.

**Results and limitations.** In Table 8, LC reduces mixed-light prediction error in physical coordinates from $46.20\%$ to $3.68\%$, compared with $21.53\%$ for the IC-Light-style constraint. In Log-Gamma coordinates the corresponding values are $41.38\%$, $23.53\%$, and $52.21\%$. The IC-Light-style comparator attains low observed single-light error without establishing additive mixed-light prediction, particularly in Log-Gamma coordinates. This supports LC in these specific reduced-model studies.

The two coordinate studies have different seeds and evaluation cohorts and do not establish a coordinate winner. These single-seed results also do not establish robustness across seeds, unseen-light-direction recovery, or quality at full resolution.

### E.1.4 Qualitative comparison

The following figures compare ordinary flow matching, the IC-Light-style constraint, and LC using the same 300k models and DPM-Solver++50 sampler as in Table 8. Each row shows the true summed HDR response and the three predictions, with scene, reference, light pair, and initial noise shared across methods.

We retain a fixed selection of nine scenes per coordinate study to illustrate improvements in mixed-light prediction. Table 8 reports aggregate performance on the full evaluation cohorts.

For display, each row uses a shared scale derived from the true mixed response's positive 99th percentile, $p_{99}$, applying $\operatorname{asinh}(H/(0.1p_{99}))/\operatorname{asinh}(10)$ followed by display clipping. Images are enlarged by nearest-neighbor interpolation. The percentages below the predictions report mixed-light error in unclipped linear HDR; all methods in a row use the same display transformation.

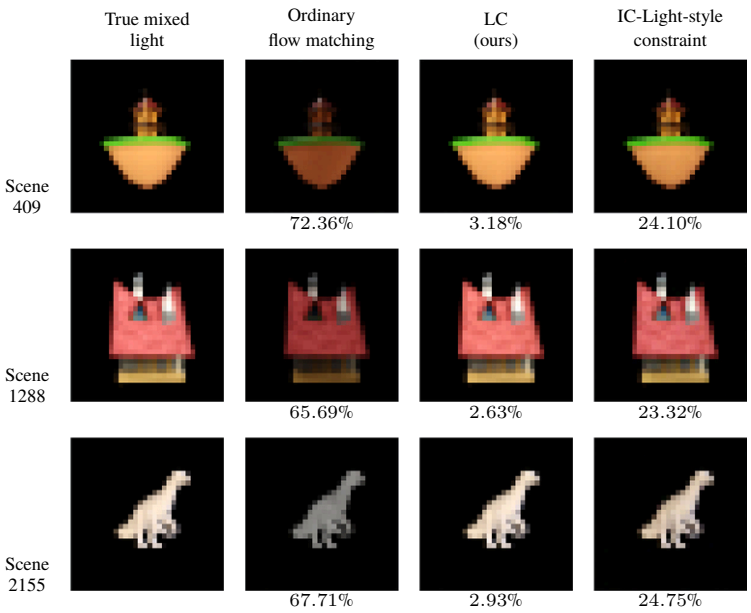



Figure 46: Physical linear coordinates: examples 1–3, compared at 300k updates with DPM-Solver++50. Percentages are mixed-light prediction errors in linear HDR. Within each row, all methods share the same scene, reference, noise, and display scale.

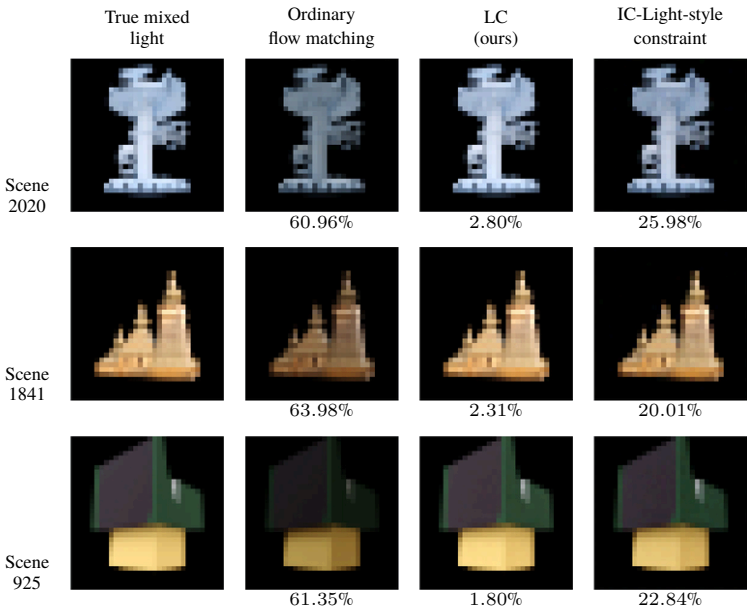



Figure 47: Physical linear coordinates: examples 4–6, compared at 300k updates with DPM-Solver++50. Percentages are mixed-light prediction errors in linear HDR. Within each row, all methods share the same scene, reference, noise, and display scale.

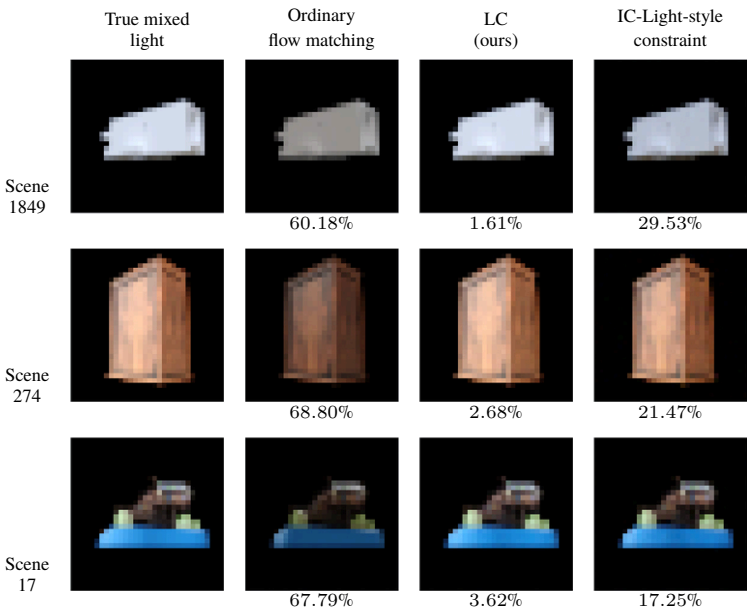



Figure 48: Physical linear coordinates: examples 7–9, compared at 300k updates with DPM-Solver++50. Percentages are mixed-light prediction errors in linear HDR. Within each row, all methods share the same scene, reference, noise, and display scale.

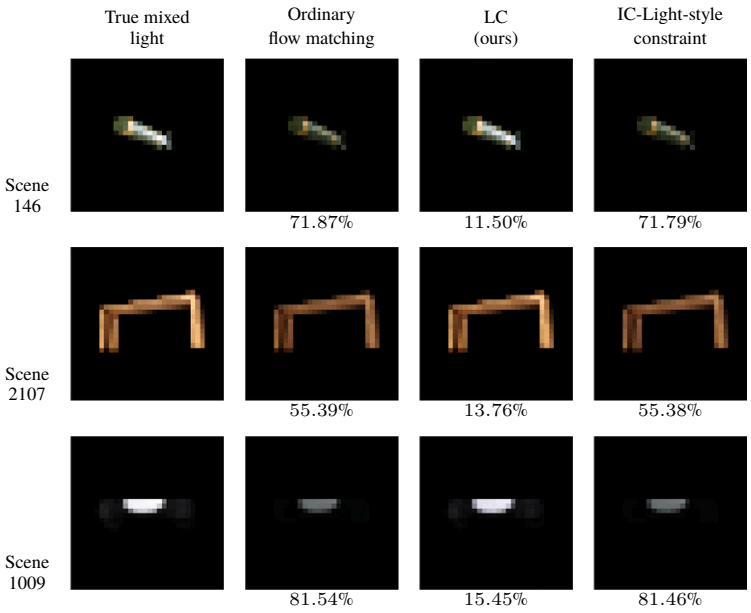



Figure 49: Log-Gamma coordinates: examples 1–3, compared at 300k updates with DPM-Solver++50. Percentages are mixed-light prediction errors in linear HDR. Within each row, all methods share the same scene, reference, noise, and display scale.

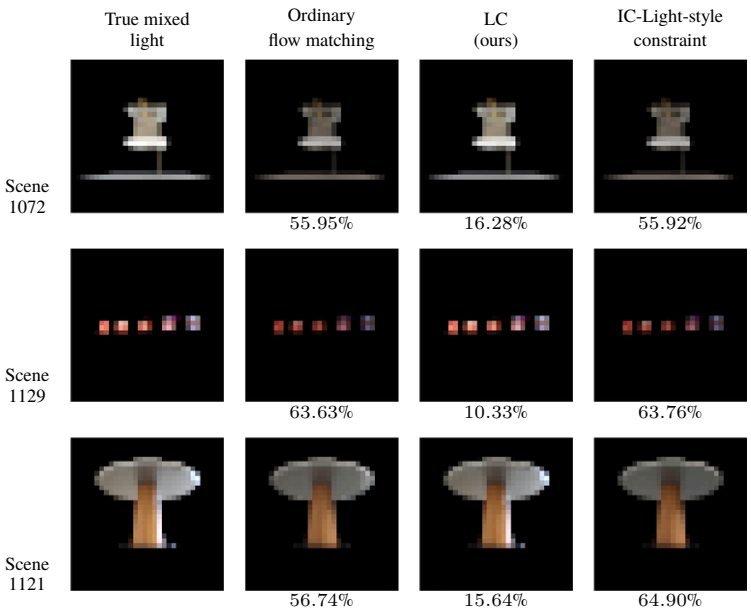



Figure 50: Log-Gamma coordinates: examples 4–6, compared at 300k updates with DPM-Solver++50. Percentages are mixed-light prediction errors in linear HDR. Within each row, all methods share the same scene, reference, noise, and display scale.

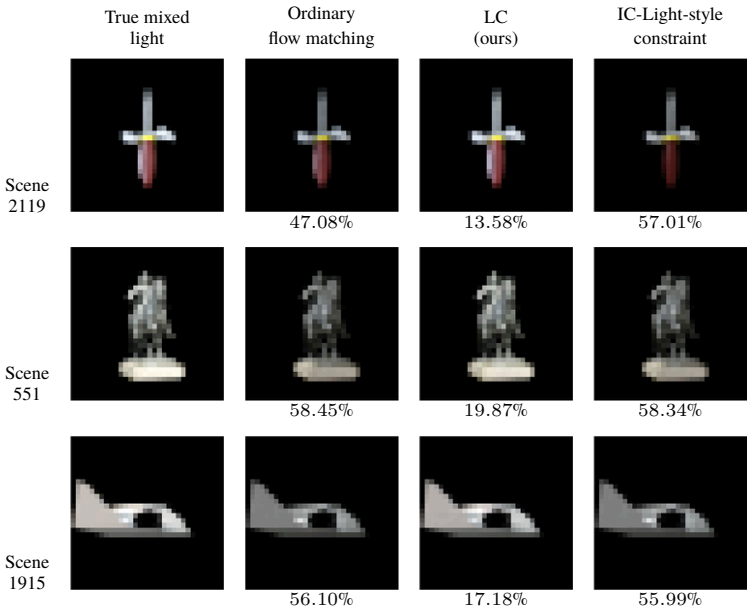



Figure 51: Log-Gamma coordinates: examples 7–9, compared at 300k updates with DPM-Solver++50. Percentages are mixed-light prediction errors in linear HDR. Within each row, all methods share the same scene, reference, noise, and display scale.

### E.1.5 Endpoint supervision and pathwise consistency

Data augmentation (DA) retains the original flow matching (FM) objective and augments its training data with exact summed HDR responses. It therefore supplies additional *endpoint supervision*. The light-transport linearity constraint (LC; Section 3.3) imposes a *pathwise constraint* on the composition of model predictions along the flow trajectories. DA and LC are orthogonal training choices and can be used together. We evaluate ordinary FM, DA, LC, and DA+LC under equal measured GPU training time to examine their individual and combined effects.

**Controlled setting.** This separate CNN experiment uses one object, a fixed camera, and eight observed light–response pairs formed from nonnegative mixtures of 64 rendered environment responses. Only the eight pairs are available for training. The CNN has width 64, four residual blocks, and no normalization; it predicts flow velocity at $32 \times 32$ in linear HDR coordinates. All configurations share the same initialization (seed 17).

For auxiliary training pairs, one component is an observed anchor with unit coefficient sum. The second is a nonnegative mixture of the eight anchors, scaled so that the total coefficient sum is uniform from one to four. We call this coefficient sum the lighting intensity below. DA directly supervises the exact summed response $H_1 + H_2$. LC instead constructs noisy component states from $H_1, H_2$ and applies the physical composition constraint from Appendix H, without using their sum as a supervised target. All three predictions receive gradients. The combined configuration uses both objectives. Ordinary supervision samples the eight observed anchors.

**Training budget and evaluation.** All configurations run in alternation on one H200 for approximately 701 seconds each. CUDA-synchronized timing includes forward, backward, optimizer, and EMA updates; sampling, logging, saving, and evaluation are excluded. Each update uses 160 model-example presentations. The baseline uses 160 observed examples. The other configurations use 64 observed examples and allocate the remaining 96 presentations to DA, LC, or equally to both; each LC tuple requires three queries. Training uses Adam at $10^{-3}$ and EMA decay 0.999. DA and LC weights rise linearly over 10,000 updates to 0.5 and 0.01; LC times are uniform on $[0, 1)$. Table 9 reports the resulting update counts.

Evaluation uses 512 fixed samples per intensity interval and 100 Heun steps, shared across configurations. Test mixtures use coefficient directions withheld from auxiliary training but remain inside the span of the eight observed lights. We report pooled relative HDR prediction error against the exact summed response, using the same definition as $e_{\text{mix}}$ above.

Table 9: **Endpoint supervision and pathwise consistency at equal training time.** Pooled relative HDR prediction error (%) for held-out mixtures inside the observed lighting span; lower is better. Intensity is the sum of mixture coefficients. The lower rows report the measured training budget.

| Lighting intensity | Baseline | DA | LC | DA+LC |
|---|---|---|---|---|
| 0–1 | 122.00 | 54.52 | 56.51 | **52.75** |
| 1–4 | 73.21 | 51.29 | **35.66** | 50.46 |
| 4–5 | 82.37 | 51.09 | **32.53** | 49.88 |
| 5–6 | 85.81 | 49.81 | **30.75** | 47.42 |
| 6–7 | 88.43 | 48.17 | **31.91** | 45.79 |
| 7–8 | 90.42 | 48.62 | **45.60** | 45.76 |
| Training time (s) | 700.977 | 700.978 | 700.980 | 700.988 |
| Optimizer updates | 66,348 | 70,032 | 54,692 | 33,778 |

**Results.** LC gives the lowest prediction error from intensity one through eight, including beyond the auxiliary training range: at intensities one to four, error is 35.66% with LC versus 51.29% with DA. DA+LC is best below unit intensity. This supports more effective intensity extrapolation by LC within the known lighting span at equal training time. However, all methods retain substantial errors on observed anchors (about 52%) and unseen basis directions (48–49%). This single-scene, single-seed result does not establish recovery of unseen lighting directions or a general advantage over DA.

**Extrapolation in linear and logarithmic coordinates.** To test whether LC improves extrapolation beyond ordinary FM supervision, we compare FM with FM+LC in both linear and signed logarithmic coordinates. This toy uses 96 scenes, four nonnegative lighting coefficients, and 64-dimensional HDR responses. Known lights span two dimensions per scene; ordinary FM supervision covers intensities from 0.25 to one. For each coordinate system and seed, both configurations start from the same model after 50,000 FM updates and continue for another 50,000 updates with the same ordinary supervision. LC uses two exact component responses synthesized from the known lighting bases, with one component inside the ordinary training range. Their summed response is not added as an FM target.

Table 10: **LC improves extrapolation in both coordinate systems.** Relative RMSE in physical HDR, pooled over three seeds after 50,000 continuation updates. Test intensities are above two through four within the known lighting span. Lower is better.

| Coordinates | Training | Endpoint error | Additivity error |
|---|---|---|---|
| Linear | FM | 0.191120 | 0.105086 |
| | FM+LC | **0.100305** | **0.055344** |
| Signed logarithmic | FM | 1.916865 | 1.308039 |
| | FM+LC | **0.289170** | **0.167742** |

At intensities two to four within the known lighting span, LC reduces endpoint prediction and additivity errors by 47.5% and 47.3% in linear coordinates, and by 84.9% and 87.2% in signed logarithmic coordinates (Table 10). Thus, the pathwise constraint improves both coordinate formulations, with larger relative gains in the logarithmic formulation in this experiment.

**Suppressing extreme extrapolation errors.** A higher-dimensional toy tests whether LC controls growing prediction errors during continued FM training. It uses 192 scenes, 16 lighting coefficients, four known lighting bases per scene, and $16 \times 16$ HDR images. Both CNN configurations start from the same model after 50,000 ordinary FM updates (seed 29) and receive 10,000 further updates. They share ordinary supervision; FM+LC adds a constraint within the known lighting span at intensities up to four. We evaluate intensities above four through eight outside that span, with normalized distance to the span at most 0.25.

Table 11: **LC suppresses extreme errors during continued FM training.** Both models use linear coordinates. Endpoint prediction error is relative RMSE in physical HDR on the same 384 stress-test samples; lower is better.

| Continuation updates | FM | FM+LC |
|---|---|---|
| 0 | 504.270 | 504.270 |
| 1,000 | 887.839 | 299.704 |
| 2,000 | 1,491.539 | 59.004 |
| 3,000 | 1,980.459 | 16.126 |
| 5,000 | 5,324.095 | 6.632 |
| 7,500 | 29,815.320 | 3.906 |
| 10,000 | 107,920.074 | **2.045726** |

Starting from the same error of 504.270, ordinary FM deteriorates to 107,920.074, while FM+LC improves to 2.045726 (Table 11). LC therefore suppresses the growth of extreme HDR prediction errors in this stress test, beyond the intensity range used for constraint training.

### E.2 Full-Resolution Linearity Evaluation

We compare our full-resolution 8B relighting model with and without LC using single-light intensity scaling and two-light composition. The reduced-model constraint studies are presented separately in Appendix E.1.

The six sources are a portrait, dahlia, strawberries, perfume bottle, car, and room. Intensity scaling uses front, back, and side unit-light conditions for these six sources; the four-source composition test is specified in Appendix E.2.1. Source, text, and initial noise are held fixed throughout each comparison. Following Section 5, let $\hat{H}_{\text{woLC}}(E)$ denote the linear-RGB output of the model trained without LC after sampling and HDR decoding under target environment map $E$. Let $\hat{H}_{\text{wiLC}}(E)$ denote the corresponding output of the model trained with LC. All references below are constructed from model outputs rather than measured ground-truth relighting. Metrics use all RGB channels of the native $2048 \times 2048$ image, including the background; no foreground mask or per-image exposure fitting is applied.

**Single-light intensity scaling.** For front, back, and side illumination, we multiply the HDR environment by $k \in \{0.5, 1.5, 2, 2.5\}$ before light encoding, without renormalizing individual conditions. Thus $k$ changes the physical light input, not the lighting CFG value.

For each single-light response, we use the corresponding unit-light HDR output without LC, $H_{\text{unit}} = \hat{H}_{\text{woLC}}(E)$, as the common reference. The baseline fitted gain is

$$g_{\text{woLC}}(k) = \frac{\langle \hat{H}_{\text{woLC}}(kE), H_{\text{unit}} \rangle}{\|H_{\text{unit}}\|_2^2}. \tag{18}$$

Linear homogeneity implies $g_{\text{woLC}}(k) = k$. The unit value $g_{\text{woLC}}(1) = 1$ follows by definition. This statistic weights RGB values by the unit-light image; it is not a mean-luminance ratio. For the with-LC model, let $\hat{H}_{\text{wiLC}}(kE)$ denote its rendered linear-HDR output under the same front, back, and side intensity-scaling tests. We compute $g_{\text{wiLC}}(k)$ with the same fitted-gain statistic against $H_{\text{unit}}$. This keeps both models on a common HDR reference rather than renormalizing each model at $k = 1$; with-LC unit-input values therefore need not equal one.

Table 12: Overall brightness response under single-light intensity scaling. Entries are fitted linear-HDR gains relative to the corresponding unit-light HDR output without LC, averaged over the evaluated sources for each lighting direction. Ideal scaling gives response $k$.

| | Without LC | | | With LC | | |
|---|---|---|---|---|---|---|
| Input scale $k$ | Front | Back | Side | Front | Back | Side |
| 0.5 | 0.866 | 1.091 | 0.887 | 0.617 | 0.698 | 0.622 |
| 1 | 1.000 | 1.000 | 1.000 | 1.043 | 1.021 | 1.032 |
| 1.5 | 1.081 | 0.958 | 1.069 | 1.369 | 1.276 | 1.338 |
| 2 | 1.136 | 0.958 | 1.115 | 1.783 | 1.592 | 1.716 |
| 2.5 | 1.240 | 0.992 | 1.167 | 2.104 | 1.893 | 2.039 |

Table 12 summarizes the overall response for both models. Figure 52 shows portrait examples for back and side lighting at $2.5E$. The with-LC panels are rendered outputs $\hat{H}_{\text{wiLC}}(2.5E)$, and the linear references are constructed as $H_{\text{linear}}(2.5) = 2.5H_{\text{unit}}$. All panels use the same display transform and exposure within each comparison.

**Interpretation.** Table 12 shows improved intensity response with LC, though high-scale responses remain sublinear. The single-light previews (Figure 53) use linear-HDR references with a shared $-2$ EV ACES display; the gains in Table 12 use native linear RGB. Results apply to this model, source set, and fixed noise, and do not isolate decoder support limits from other nonlinearities.

### E.2.1 Composition with a Shared Weighted-Output Reference

In a separate comparison, both models are evaluated on four sources: portrait, perfume bottle, car, and room. We use red/teal/blue spotlights and front lighting as $E_1, E_2$, source and lighting guidance 1.5, text guidance 1, and the same saved initial-noise tensor (seed 0). The six mixing-weight pairs $(c_1, c_2)$ are $(0.17, 0.62)$, $(0.43, 0.11)$, $(0.08, 0.89)$, $(0.76, 0.48)$, $(0.91, 0.22)$, and $(0.55, 0.73)$. For each source $S$, we form the mixed environment $E_\Sigma = c_1E_1 + c_2E_2$, giving 24 mixtures in total. The




Figure 52: **Back- and side-light response with and without LC.** The With LC panels are rendered outputs at $2.5E$, while the linear references scale the corresponding unit-light HDR output without LC, $H_{\text{unit}}$, by 2.5 before tone mapping. All images use the same ACES view, with shared exposure within each row: 0 EV for back lighting and $-2$ EV for side lighting.

shared reference weights the separately generated unit-light HDR outputs without LC:

$$\hat{H}_{\text{ref}} = c_1 \hat{H}_{\text{woLC}}(E_1) + c_2 \hat{H}_{\text{woLC}}(E_2), \qquad P_j = \hat{H}_j(E_\Sigma), \quad j \in \{\text{woLC}, \text{wiLC}\}. \tag{19}$$

The weights are applied to the generated HDR images before addition and display mapping. Both models are compared to this same generated reference, which is not measured ground truth.

Table 4 reports two errors on native $2048 \times 2048$ linear-RGB outputs. Output RMSE (HDR) is the absolute root mean squared error against the shared reference, computed per image and averaged over the 24 mixtures. Additivity RMSE (HDR) is each model's relative error $e_{\text{add}}$ (Appendix E.1), computed from its own mixed-light and single-light outputs and pooled over the 24 mixtures. PSNR, SSIM, and LPIPS (LDR) use the corresponding fixed-exposure PNGs after ACES 2.0 SDR display mapping, at 0, $+1$, and $+2.2$ EV, and are computed per image and exposure before averaging. All pixels and channels are retained, with no metric mask, resizing, or exposure alignment. The peak signal-to-noise ratio uses display values in $[0, 1]$, and the structural similarity uses a $7 \times 7$ window. The perceptual distance uses calibrated VGG LPIPS (Zhang et al., 2018). The complete metric table and a selected glass-bottle comparison are presented in Section 5 (Table 4 and Figure 11).

**Display settings for the selected comparisons.** Figure 11 shows the glass bottle in two rows: an ACES 2.0 triplet (top) and a direct-linear triplet (bottom). Each triplet orders the images as without LC, with LC, and shared reference. The ACES 2.0 displays use 0 EV. Direct linear displays map the native $2048 \times 2048$ linear-RGB values $H$ to 8-bit codes as $I = \text{round}[255\ \text{clip}(H/U, 0, 1)]$. $U$ is a selected percentile of per-pixel $\max(R, G, B)$ values pooled across the without-LC, with-LC, and reference images. This single scalar is shared by all three images and all RGB channels, with no mask, per-image gain fitting, gamma encoding, or tone curve. The glass bottle uses the 65th percentile, $(c_1, c_2) = (0.91, 0.22)$, and $U = 1.0074$. The mapping is linear before clipping and quantization: channels above $U$ saturate, so these regions do not reveal radiance differences above that bound. Linearity here refers to stored image codes, not the screen's luminance response. Linear-HDR root mean squared error is computed on the original HDR values before display mapping.

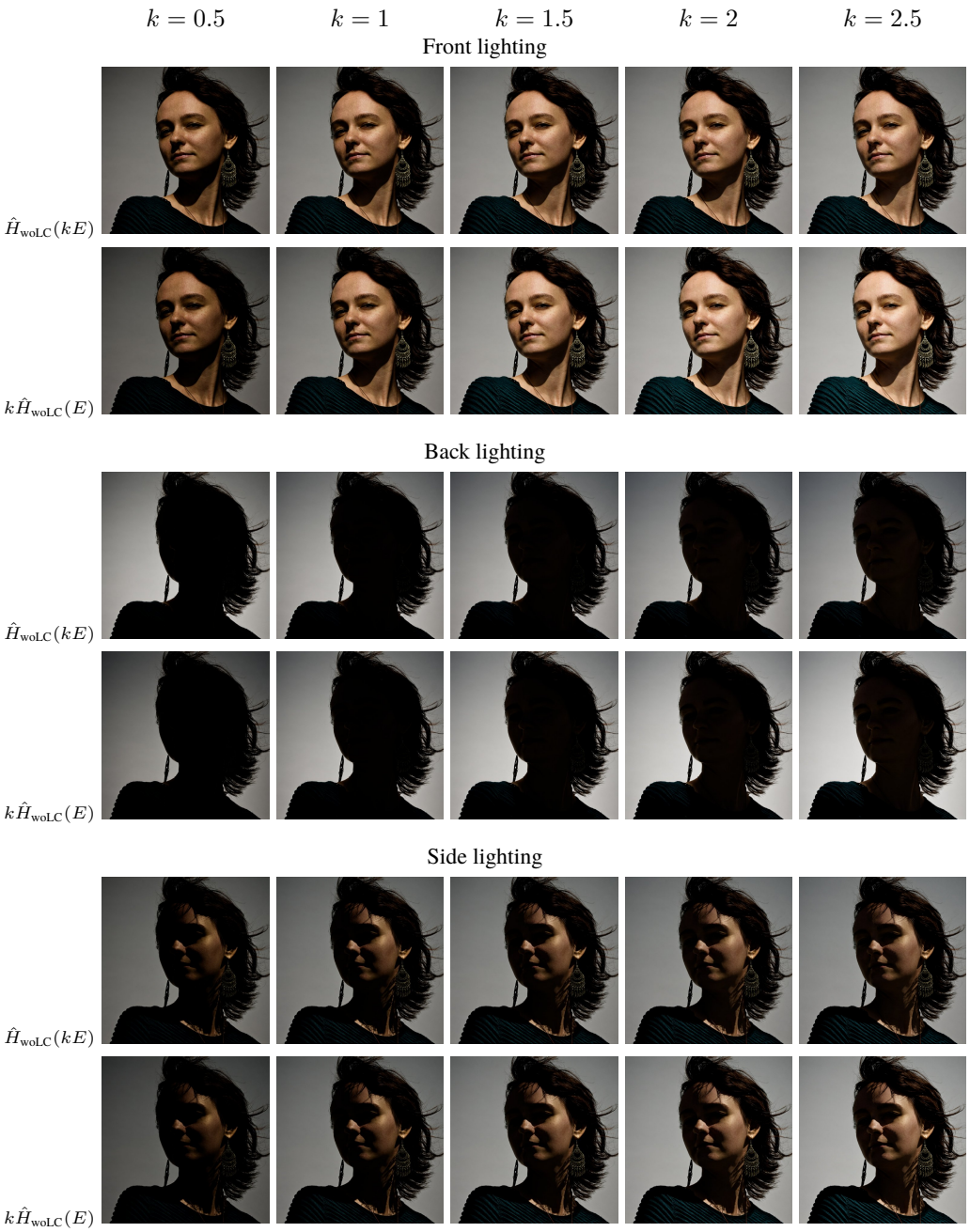



Figure 53: Portrait intensity sweep for our 8B relighting model without the linearity constraint. For each direction, the upper row is the model output $\hat{H}_{\text{woLC}}(kE)$ and the lower row is the linear reference $k\hat{H}_{\text{woLC}}(E)$ formed from the unit-light output. The unit-light panel is therefore identical in both rows. All panels use the same $-2$ EV exposure and ACES 2 display transform, with no per-panel exposure adjustment. Numerical responses in Table 12 are measured on native linear RGB before this display transform.

# F Additional Ablation Studies

## F.1 Additional Ablation Examples

**Detail preservation.** Figure 54 provides the complete portrait comparison, including the perceptual-supervision and training-target variants omitted from Figure 10a (settings in Table 13). Without LPIPS and DINOv2 supervision (w/o Perceptual), hair strands and skin texture become smoother, supporting their inclusion (Section 3.4). With the same training members and the same $2048 \times 2048$ training and output resolution, native $2048^2$ targets (2048 Data) preserve noticeably finer detail than targets resized from $512^2$ (512 Data). The resolution and detail of the HDR training data thus directly affect the detail of the generated results.

Table 13: **Complete detail ablation settings.**

| Configuration | Setting |
|---|---|
| Raw Clip-10 | Clip linear HDR to $[0, 10]$. |
| Remove text pathway | *Text* pathway removed. |
| 512 Train/Infer | Train/output: $512 \times 512$. |
| w/o Perceptual | No LPIPS or DINOv2 loss. |
| Data variants (train/output: $2048 \times 2048$) | |
| 512 Data | Same members; targets resized from $512^2$ to $2048^2$. |
| 2048 Data | Same members; targets kept at native $2048^2$. |

512 Data is the visual reference for the first four variants. The final pair changes only target resolution preparation.

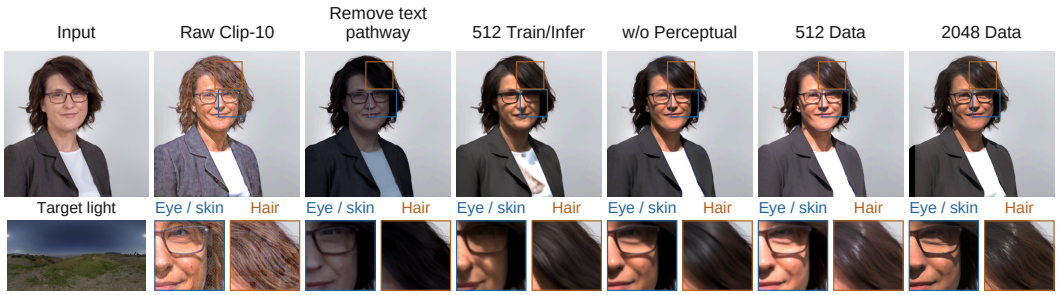



Figure 54: **HDR fine-tuning and detail preservation.** Portrait under front lighting. Settings in Table 13.

Figure 55 complements the portrait in Section 5 using the same configurations (Table 13).




Figure 55: **Additional detail preservation ablation.** Strawberries under side lighting, with crops of seeds and leaves.

**Lighting conditioning and attention.** Figure 56 gives the complete Room and Car comparisons; Section 5 discusses the car. In the room, without raw lighting patches (w/o raw light tokens), the ceiling fixture appears unlit at both 0 EV and $-2$ EV, and the room is much brighter overall. This suggests that the raw patches convey the intensity of the target illumination, including its balance against the scene's own light sources. Without lighting *Visual* tokens (w/o light VLM), the room remains plausible and the fixture stays lit, suggesting that raw lighting patches, which interact globally through full attention in the pixel-space sequence, can convey the overall lighting; local effects such as the car's cast shadow degrade, however. With both inputs, the baseline keeps the fixture lit, with its bright core preserved at $-2$ EV.

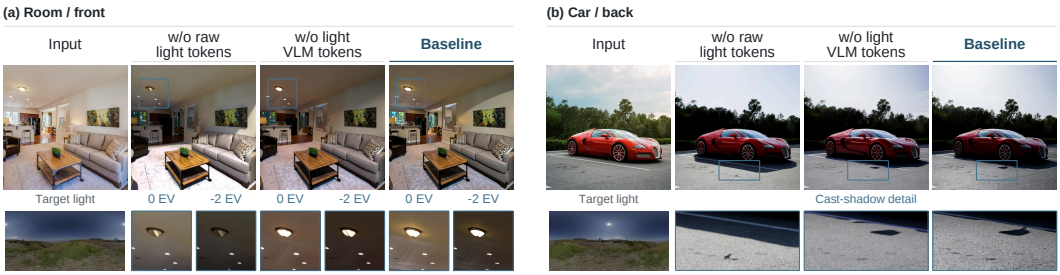



Figure 56: **Lighting conditioning.** Baseline is compared with removing lighting VLM tokens and removing raw light tokens. Ceiling crops show HDR predictions at 0 and $-2$ EV; car crops show cast shadows. Settings in Table 3.

Figure 57 provides additional qualitative ablations on a cat under backlighting and a watch under side lighting, complementing the examples in Section 5. *Envmap encoding* replaces the extended-Reinhard and log-intensity lighting maps with Log-Gamma ($M = 50$) and direct sRGB encodings, retaining the XYZ direction map. *VLM edge isolation* jointly blocks lighting-to-source *Visual* and *Text*-to-*Visual* attention, and *Edge isolation + no raw light* additionally removes lighting LDR pixel-space patch tokens. *Shared projection* ties the HDR and LDR input projections, as examined in Appendix F.5; the remaining settings follow Table 3.

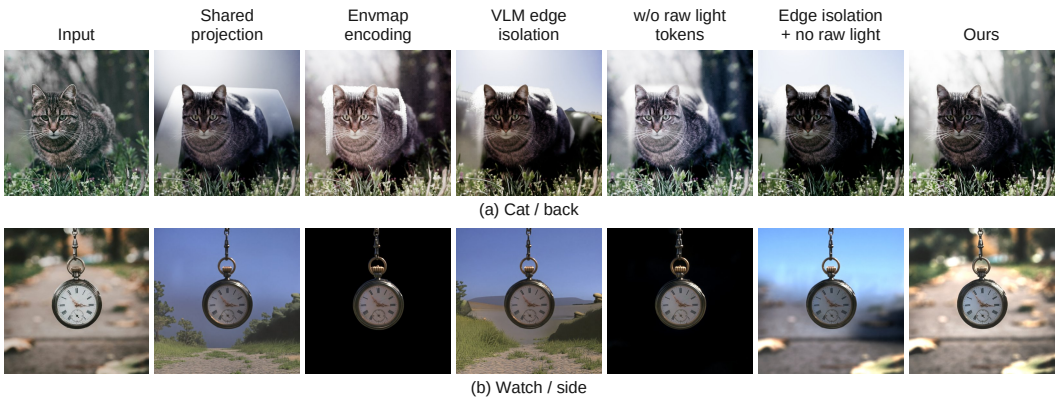



Figure 57: **Additional lighting ablations.** Cat under backlighting and watch under side lighting.

### F.2 Visual Attention Configuration

To test whether the visual attention configuration affects source preservation, we compare our source-last causal attention (Section 3.1) with source-first causal attention (Source-first), which places the source first among the semantic *Visual* tokens, and full attention among the semantic *Visual* tokens (Visual Full). Keeping causal attention follows the pretrained model, but visual conditioning does not inherently require it: unified models such as BAGEL and Show-o2 attend bidirectionally within visual representations while retaining causal sequence modeling (Deng et al., 2025; Xie

et al., 2025). Both causal and full attention relight most inputs plausibly (Figure 58), so preserving pretrained visual knowledge does not require every aspect of the configuration to remain unchanged. All three configurations differ, however, in challenging cases such as the backlit pocket watch and cat (Figure 59): Source-first loses parts of the original backgrounds, and Visual Full introduces environment-map scenery behind the pocket watch. Source-last causal attention keeps the original backgrounds while changing their illumination, supporting this choice.

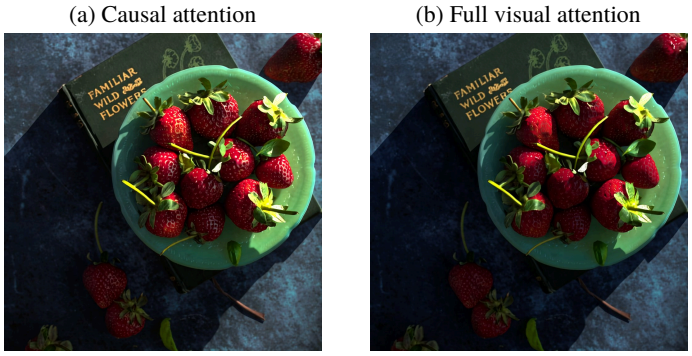



Figure 58: **Causal versus full visual attention results** under the same source image and target illumination.




Figure 59: **Effect of the visual attention configuration on source preservation in challenging backlit cases.**

### F.3 Classifier-Free Guidance

At inference, separate CFG scales control the influence of the source and lighting conditions (Section 3.4). Raising the source scale from 1 to 1.5 yields finer image details, and raising the lighting scale by the same amount yields more distinct shadows cast by hair (Figure 60). Stronger source guidance also mitigates background leakage (Figure 62).

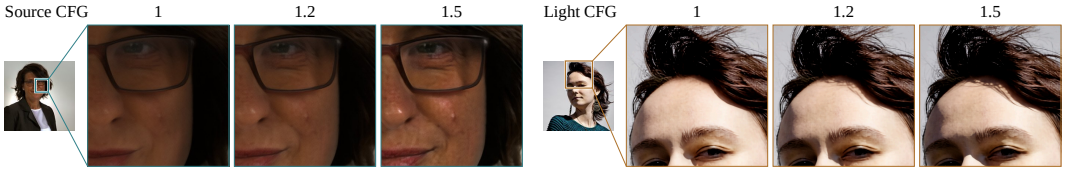



Figure 60: **Source and lighting guidance.** Boxes locate the crops in the full outputs (CFG 1.5).

### F.4 Background Preservation

Relighting sometimes darkens or replaces the background of a natural photograph. A plausible contributor is the prevalence of black-background object renders in training, without a dedicated background-preservation objective. We consider three complementary mitigations.

**Per-image captions.** In Figure 61, *Ours with caption* retains more of the garden path, vegetation, and foliage than *our model*. Our model uses synthetic training data; Ours with caption additionally uses per-image captions.




Figure 61: **Background preservation.** Rows show a pocket watch, a cat, and a dahlia. Within each pair, the source image, target lighting, random seed, and guidance scales are matched. Each prediction uses a global display exposure matched on the subject, followed by the sRGB transfer. Ours with caption retains more surrounding scene content in these selected examples.

**Stronger source guidance.** Increasing source classifier-free guidance strengthens the input image's influence at inference. In Figure 62, the original blurred foliage behind the parrots is replaced by window-like or chair-like structures at a source CFG of 1. Increasing it to 1.5 suppresses these structures across three target lighting conditions and better retains the foliage and defocused background parrot. This mitigates background leakage, although exact background preservation is not guaranteed.

**Background-aware captions.** Explicitly describing the existing background provides a further preservation cue. In Figure 63, captions that describe blurred foliage and background parrots and request preservation of the source scene reduce unwanted wall and doorway structures under warm indoor lighting. The augmented captions also specify that the environment maps should provide illumination cues. None of these three strategies guarantees exact background retention.

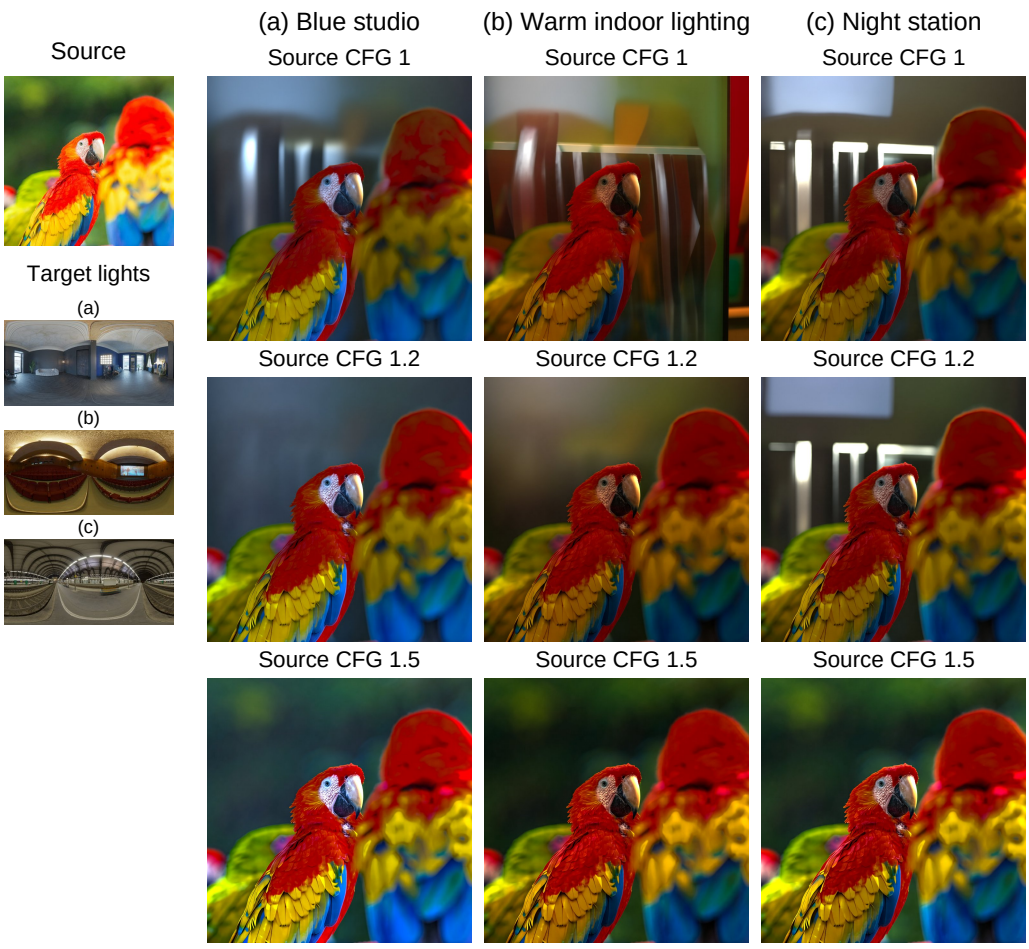



Figure 62: **Source guidance mitigates background leakage.** The same model, Ours with caption, is evaluated under three target lights. Each column compares source CFG 1, 1.2, and 1.5, fixing the source, lighting, text, random seed, other guidance scales, and display exposure.

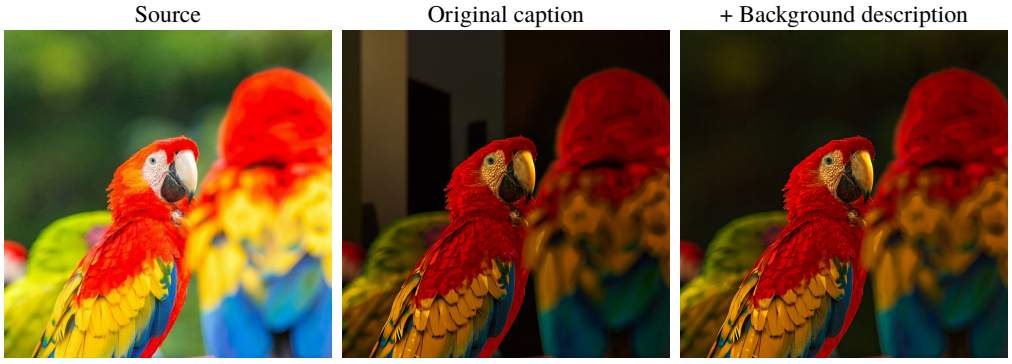



Figure 63: **Background descriptions improve scene preservation.** Describing the source foliage and defocused background parrots, together with scene-preservation instructions, suppresses the wall and doorway present with the original captions. The model, input images, saved initial noise, guidance scales, and display exposure are fixed; only the captions change.

### F.5 Shared Projection Ablation

We compare the baseline's separate HDR and reference-patch projections with *Shared projection*, which uses one projection for both inputs. Figure 64 shows the selected sofa and glass-bottle examples under directional illumination. Separate projections retain more distinct rug patterns and glass reflections, particularly in the backlit bottle.

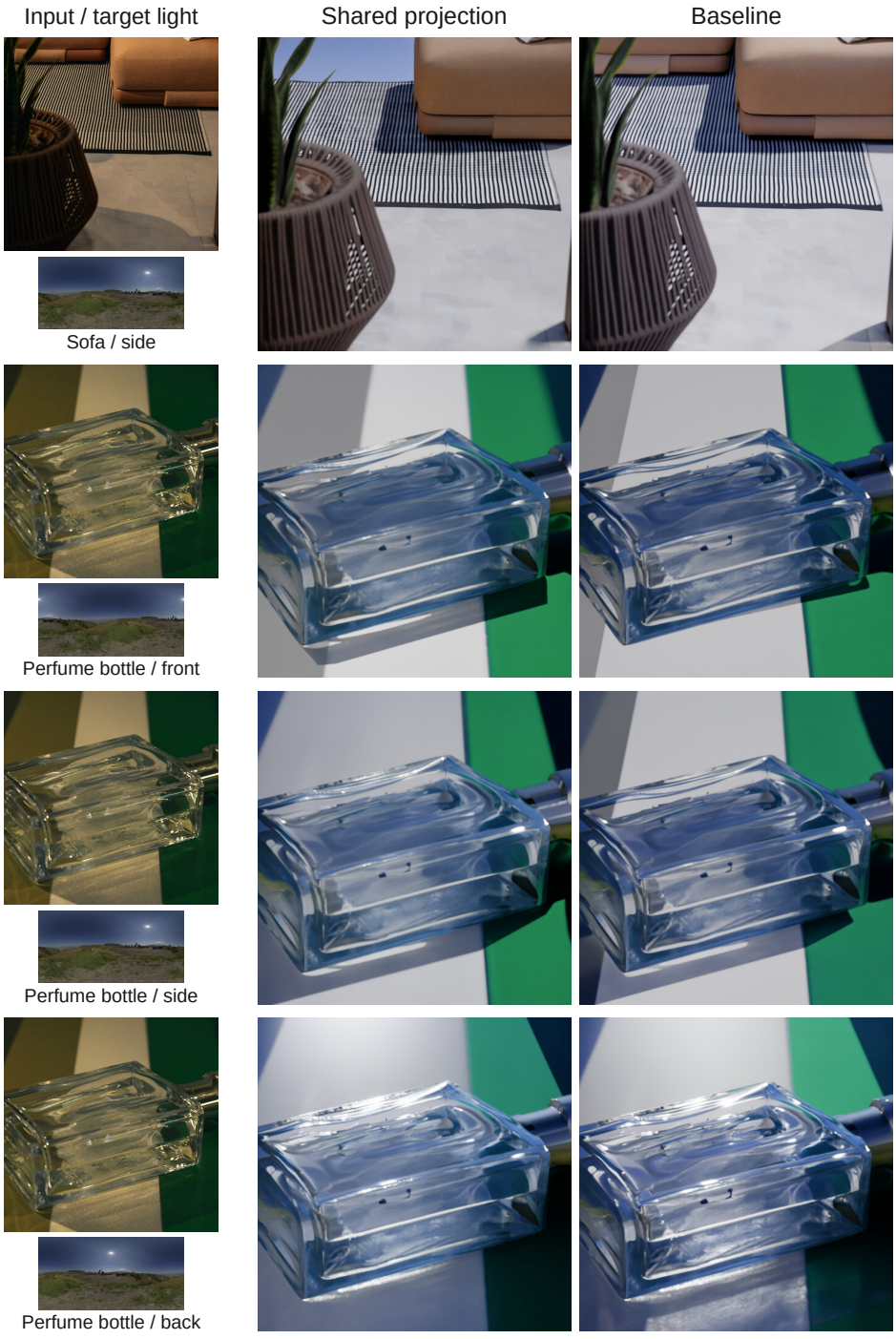



Figure 64: **Shared versus separate input projections.** Each row compares the two configurations for the same source and target light.

# G Supplementary Method Details

## G.1 Training Data and Pair Construction

**Synthetic HDR collections.** The $512^2$ collection contains $124{,}544$ objects with 16 viewpoints each, drawn from the union of the object lists of TRELLIS (Xiang et al., 2025), Neural Gaffer (Jin et al., 2024), and DiLightNet (Zeng et al., 2024). Asset screening removes meshes with empty or degenerate geometry, non-finite vertices, or missing materials or textures, and render validation checks that each object has all expected views and lightings, finite HDR values, valid masks, and consistent camera and lighting metadata. For native high-resolution data, a frozen vision-language model scores views from the broader low-resolution render pool on five equally weighted criteria: composition, structural completeness, detail and material, readability, and overall value as a training image. The 200,000 top-ranked views are rerendered directly at $2048^2$ under original, LuxDiT, and captured environments, yielding $41{,}003$ objects and $199{,}763$ views. This collection is neither an upsampled copy nor a strict object subset of the $512^2$ training collection. The collections share $33{,}412$ object identities and together cover $132{,}135$ distinct objects. Figure 65 shows example environment maps and renders.



Figure 65: **Training data examples.** HDR environment maps (top) and synthetic single-object renders (bottom), tone mapped for display.

**Relighting pairs.** Source and target depict the same object and viewpoint under different illuminations, with geometry, materials, and framing fixed. Training pairs are sampled from both collections using valid radiometric inputs, drawing each example from the $512^2$ collection with probability $0.3$ and from the native $2048^2$ collection with probability $0.7$. Low-resolution targets are resized to the shared prediction grid, while native high-resolution targets are used directly.

**Optional descriptions.** Inference requires a source image and a target HDR environment map; content and lighting descriptions are optional. Content descriptions reuse the Cap3D captions (Luo et al., 2023; 2024) where available, with additional content and environment descriptions generated offline by a vision-language model. The relighting instruction remains present when descriptions are omitted.

## G.2 Radiometric Preprocessing

**HDR coordinates.** Following DiffHDR (Yu et al., 2026b), we use the Log-Gamma encoding introduced in Section 3.2, and we set $M = 50$ and $\gamma_{\mathrm{LG}} = 2.2$. Training targets are clipped to $[0, M]$ before encoding. Perceptual supervision and final export decode predictions after clipping encoded values to $[-1, 1]$; the prediction head, reconstruction loss, and intermediate sampling states use unclipped predictions. For LC, the transform, inverse, and Jacobians instead use the full-real signed extension in Appendix H.3. Constraint states and their decoded radiance are not clipped by this mapping.

**Illumination references.** We align each HDR environment to the camera frame using its recorded orientation and strength. The equirectangular representation follows OpenCV camera coordinates

(right, down, forward), without environment-wise intensity normalization. Following the three-map representation of DiffusionRenderer (Liang et al., 2025a), fixed extended-Reinhard and logarithmic views expose complementary intensity ranges, while a direction image records angular location. Together with the source image, these bounded references enter the shared conditioning interface.

**Source preparation.** Display-referred sources retain their values. HDR sources are exposed and tone-mapped to form display-referred references. Foreground support is used only for source preparation; it is neither a model input nor a loss weight. Source exposure does not rescale the HDR target or illumination, and display adjustments do not change the radiance of the exported HDR output.

### G.3 Model Architecture

**Shared references and independent HDR embedding.** The source and three lighting references share the pretrained Qwen3-VL visual encoder $\phi_V$ and the LDR patch projection $\phi_L$. The HDR projection $\phi_H$ is initialized by copying $\phi_L$ after loading the pretrained weights, then optimized independently. The shared decoder combines visual features and raw image patches to predict the clean encoded HDR target directly in pixel space, without a VAE.

**Attention.** Visual reference tokens and task text form semantic context with causal attention. Pixel-space tokens contain the time embedding, noisy target, and raw source and lighting patches. They attend to the full valid sequence, while semantic tokens cannot attend to pixel-space tokens. Each reference is encoded independently by the shared visual encoder.

**Optimization scope.** The shared decoder, LDR and HDR patch projections, time embedding, and clean-target readout are trained. The text embedding and language-model head remain frozen. The visual encoder and perceptual networks are also frozen and kept in evaluation mode.

### G.4 Optimization and Sampling

**Optimization overview.** Training progresses from the low-resolution collection through mixed sampling to native high-resolution fine-tuning, while the prediction grid remains $2048^2$. We optimize the trainable parameters with AdamW using the reconstruction and perceptual objectives in Section 3.4, with a constant learning rate of $2 \times 10^{-5}$ and no warm-up. Training runs on eight GPUs for about five days as consecutive 24-hour jobs; each optimizer update aggregates 128 examples per GPU, giving a global batch of 1,024 examples. The flow parameterization and time distribution are defined in Section 3.2. Additional experiments with captions are reported in Appendix F.4.

**Perceptual supervision.** LPIPS and DINOv2 compare the decoded prediction and retained target after the same fixed display mapping $D(H) = g(H/(1 + H))$, where $g$ is the sRGB transfer function. The feature extractors remain frozen: target features are evaluated without gradients, while predicted-image features propagate gradients through HDR decoding to the model. The factor $(1 - t)^{-2}$ applies only to the reconstruction term, not the perceptual terms. We use $\lambda_p = 0.1$ and $\lambda_d = 0.01$. The gate $\chi(n, t) = \mathbf{1}[n \geq 101{,}152]\,\mathbf{1}[t \geq 0.3]$ activates the perceptual terms after the first $101{,}152$ training examples (about 100 optimizer updates), and only for $t \geq 0.3$, since at noisier times the clean-image prediction is too uncertain for a meaningful perceptual comparison.

**LC configuration.** The effective global batch for LC contains $128$ light pairs. Each pair uses two environments and retained HDR responses of the same scene and view, with the source fixed. Independent coefficients $c_1, c_2 \sim \mathcal{U}[0, 1]$ scale each environment and its corresponding response. LC queries retain all source and lighting conditions, without condition dropout or classifier-free guidance. Two component states use independent standard Gaussian noises and are queried at time $s_t$. Their decoded states and Jacobian-transformed velocities determine the composed radiance state and velocity target. A third query at time $t$, conditioned on the summed environment, matches this target in model coordinates. All three queries share parameters and receive gradients; no complete sampling trajectory is required. This composition and its limitations are derived in Appendix H.

We use $\lambda_{\text{LC}} = 0.005$ and apply LC only for $t > 0.3$ after activation by $\chi_{\text{LC}}(n) = \mathbf{1}[n > 404{,}508]$, that is, after the first $404{,}508$ training examples (about 400 optimizer updates). Here $n$ is the number of

examples committed before the current update, counted from the start of fine-tuning. LC is averaged over pixels and color channels, then over light pairs at $t > 0.3$, using the validity mask below. It remains a soft constraint combined with paired supervision; the time threshold does not restore exact flow marginals under nonlinear encoding.

**Numerical stability of LC.** Inverse Log-Gamma decoding and Jacobian conversion can amplify extreme values. We apply the empirical check $-1 < x_t < 3$ pixelwise and channelwise to noisy encoded states used for LC. Out-of-range entries are excluded from the consistency loss, while valid entries in the same pair remain in use. Delayed activation lets the model first learn paired relighting, and the restriction $t > 0.3$ reduces the initial-noise contribution. Global gradient-norm clipping at $0.1$ further bounds the update size.

**Inference.** We use classifier-free guidance with separate source and illumination scales. For null, source-only, and fully conditioned predictions $f_\varnothing, f_S, f_{S,E}$, respectively,

$$\hat{x}_{\text{guided}} = f_\varnothing + s_{\text{src}}(f_S - f_\varnothing) + s_{\text{light}}(f_{S,E} - f_S). \tag{20}$$

During supervised training, each example keeps both conditions with probability $0.85$, keeps only the source or only the lighting with probability $0.05$ each, and drops both with probability $0.05$, which supports these branches. Inference integrates the induced flow from Gaussian noise, with optional captions provided through the same conditioning interface. Outputs are linear RGB HDR EXRs; exposure adjustments for display do not alter the exported radiance.

Upon acceptance, we will provide full implementation details, hyperparameter settings, and training code.

## H Derivation of the Light-Transport Linearity Constraint

Linear light transport makes the HDR image under the combined lighting $E_1 + E_2$ equal to the sum of the images under $E_1$ and $E_2$. LC turns this identity into a training target: from the model's predictions under $E_1$ and $E_2$, it builds a velocity target for the model's prediction under $E_1 + E_2$. Appendix H.1 derives this target in linear HDR coordinates, where it is exact under stated assumptions. Appendix H.2 explains why the two component noises are independent and what the construction requires when the source image does not determine the scene. Appendix H.3 gives the version used in training, in Log-Gamma coordinates, and explains why it is only approximate there. Throughout, the source image $S$, the viewpoint, and the exposure are fixed.

### H.1 Light Composition in Linear HDR Coordinates

**Setting and assumptions.** In this subsection, states and velocities are in linear HDR. For a fixed scene, let $H_1$ and $H_2$ be the HDR images under environments $E_1$ and $E_2$, and let $H_\Sigma$ be the image under $E_\Sigma = E_1 + E_2$. The main text writes $1+2$ in place of $\Sigma$. If every radiance contribution scales with the illumination (no fixed emission) and nothing is clipped, linear light transport gives

$$H_\Sigma = H_1 + H_2. \tag{21}$$

The lighting weights $c_i$ used in training are absorbed into $E_i$ and $H_i$. We first assume that the conditioning determines the responses; Appendix H.2 discusses what changes when it does not.

**Adding two flows.** Flow matching (Section 3.2) moves along straight paths from Gaussian noise to the clean image. For the two lightings, with independent noises $\epsilon_1, \epsilon_2 \sim \mathcal{N}(0, I)$ and component time $s \in [0, 1]$, these paths and their velocities are

$$x_s^i = (1-s)\epsilon_i + sH_i, \qquad \frac{\mathrm{d}x_s^i}{\mathrm{d}s} = H_i - \epsilon_i, \qquad i \in \{1, 2\}. \tag{22}$$

We write $v^i(x, s)$ for the flow-matching velocity field of lighting $i$, the velocity a perfect model predicts at state $x$ and time $s$; for a deterministic response, $v^i(x, s) = (H_i - x)/(1 - s)$, which equals $H_i - \epsilon_i$ on the path. Adding the two paths gives a path that ends at the physical sum and whose velocity is the sum of the two velocities:

$$x_s^1 + x_s^2 = (1-s)(\epsilon_1 + \epsilon_2) + s(H_1 + H_2), \qquad \frac{\mathrm{d}}{\mathrm{d}s}\big(x_s^1 + x_s^2\big) = v^1(x_s^1, s) + v^2(x_s^2, s). \tag{23}$$

This sum is not yet a flow-matching path, because it starts from $\epsilon_1 + \epsilon_2 \sim \mathcal{N}(0, 2I)$ instead of standard Gaussian noise.

**Restoring the Gaussian start.** Dividing the sum by $\sqrt{2}$ restores the noise, $\epsilon_\Sigma = (\epsilon_1 + \epsilon_2)/\sqrt{2} \sim \mathcal{N}(0, I)$, but also divides the endpoint by $\sqrt{2}$. We undo the endpoint scaling by stretching space and time. With the path scale $\gamma_t = 1 + (\sqrt{2} - 1)t$, which is unrelated to the Log-Gamma exponent $\gamma_{\mathrm{LG}}$, and the component time $s_t = \sqrt{2}t/\gamma_t$, the composed state is

$$H_t^\Sigma = \frac{\gamma_t}{\sqrt{2}}\big(x_{s_t}^1 + x_{s_t}^2\big) = (1-t)\epsilon_\Sigma + t(H_1 + H_2), \qquad t \in [0, 1]. \tag{24}$$

The second equality uses $\gamma_t(1 - s_t) = 1 - t$ and $\gamma_t s_t/\sqrt{2} = t$. The composed state lies on the ordinary flow-matching path of the summed image $H_1 + H_2$: it starts from standard Gaussian noise and ends at the physical sum. The components are queried at $s_t \geq t$, with equality at $t = 0$ and $t = 1$.

**Velocity target.** Differentiating equation (24), with $\dot{\gamma}_t = \sqrt{2} - 1$ and $\dot{s}_t = \sqrt{2}/\gamma_t^2$, expresses the velocity of the composed state through the two component velocities:

$$\widehat{v}_t^{H,\Sigma} = \frac{\sqrt{2} - 1}{\gamma_t} H_t^\Sigma + \frac{v^1(x_{s_t}^1, s_t) + v^2(x_{s_t}^2, s_t)}{\gamma_t}. \tag{25}$$

In training, $v^1$ and $v^2$ are the model's own predictions under $E_1$ and $E_2$. A model queried at the composed state under $E_\Sigma$ sees $H_t^\Sigma$, $t$, and its conditions, but not the pair $(x_{s_t}^1, x_{s_t}^2)$ that produced $H_t^\Sigma$. Squared regression onto $\widehat{v}_t^{H,\Sigma}$ therefore learns the average of the target over all pairs that give the same composed state:

$$v_\Sigma^\star(H, t) = \mathbb{E}\big[\widehat{v}_t^{H,\Sigma} \;\big|\; H_t^\Sigma = H\big] = \arg\min_v \; \mathbb{E}\big\|v(H_t^\Sigma, t) - \widehat{v}_t^{H,\Sigma}\big\|_2^2, \tag{26}$$

where the minimum is over unrestricted functions and the component fields are held fixed. When $v^1$ and $v^2$ are the exact fields of deterministic responses, $v_\Sigma^\star$ is exactly the flow-matching velocity field of the summed image along equation (24).

**The same target in terms of clean predictions.** The model predicts clean images, and its velocity at state $x$ and time $t$ is $(f - x)/(1 - t)$ for the predicted clean image $f$ (Section 3.2). Let $f_1$ and $f_2$ be the clean predictions at the two component states, so that $v^i(x_{s_t}^i, s_t) = (f_i - x_{s_t}^i)/(1 - s_t)$. Substituting this into equation (25) and using the two identities above simplifies the target to

$$\widehat{v}_t^{H,\Sigma} = \frac{f_1 + f_2 - H_t^\Sigma}{1 - t}. \tag{27}$$

The target is the flow-matching velocity that points from the composed state to the sum of the two clean component predictions. If $f_\Sigma$ is the model's clean prediction at the composed state, so that its velocity is $v_\Sigma = (f_\Sigma - H_t^\Sigma)/(1 - t)$, the linear-coordinate loss is

$$\mathcal{L}_{\text{linear}} = \mathbb{E}\big\|v_\Sigma - \widehat{v}_t^{H,\Sigma}\big\|_2^2 = \mathbb{E}\left[\frac{\|f_\Sigma - f_1 - f_2\|_2^2}{(1-t)^2}\right]. \tag{28}$$

In linear coordinates, LC asks the clean prediction under $E_\Sigma$ to equal the sum of the clean predictions under $E_1$ and $E_2$, each made at the coupled noisy states and times above. This holds for any predictions.

**Use in training.** One network produces all three predictions, and gradients also flow into the component predictions (Appendix I). We add the loss with a finite weight to the paired supervision. Because the regression argument above holds only for fixed component fields, it does not guarantee that the supervised optimum stays unchanged.

### H.2 Noise Coupling and Scene Uncertainty

The physical identity equation (21) relates clean images. To use it for denoising, we must also decide how the noisy states of the two components are drawn together, that is, choose a *coupling*. The coupling determines both the distribution of the composed state and the scale of its endpoint.

**Independent or shared noise.** Both components can start from standard Gaussian noise while their noises are independent or identical. The sum of the two noises then has different variance,

$$\epsilon_1 + \epsilon_2 \sim \begin{cases} \mathcal{N}(0, 2I), & \epsilon_1, \epsilon_2 \text{ independent}, \\ \mathcal{N}(0, 4I), & \epsilon_1 = \epsilon_2, \end{cases} \tag{29}$$

so restoring standard noise requires dividing the summed path by $\sqrt{2}$ in the first case and by 2 in the second. Dividing by 2 turns the endpoint into the average of the two images:

$$\frac{x_s^1 + x_s^2}{2} = (1-s)\epsilon + s\,\frac{H_1 + H_2}{2}, \qquad \epsilon_1 = \epsilon_2 = \epsilon. \tag{30}$$

Even with identical noise, the two component states are different points, since $x_s^1 - x_s^2 = s(H_1 - H_2)$. We use independent noises together with the space and time rescaling of equation (24), which keeps standard Gaussian noise at the start and the unscaled sum $H_1 + H_2$ at the end.

**IC-Light's shared-input derivation.** The derivation in IC-Light's Appendix A (Eqs. 8–11) (Zhang et al., 2025) evaluates all three lightings at one shared noisy input $x$. In our notation, each clean prediction is $\hat{H}_i = (x - b_s\hat{\epsilon}_i)/a_s$ for $i \in \{1, 2, \Sigma\}$, with common signal and noise coefficients $a_s, b_s > 0$, and the derivation's factor of two gives

$$2\hat{\epsilon}_\Sigma = \hat{\epsilon}_1 + \hat{\epsilon}_2 \quad \Longrightarrow \quad \hat{H}_\Sigma = \frac{\hat{H}_1 + \hat{H}_2}{2}, \tag{31}$$

which is the average, not the sum, of the component predictions. Evaluating all three lightings at one input also does not match the shared-noise path equation (30), whose component states differ.

**Scene uncertainty.** So far the conditioning determined $H_1$ and $H_2$. A source photograph, however, can admit several material–light explanations. The two responses then share unknown scene properties even though their noises are independent. Each component field $v^i$ averages the velocity over these possibilities using only its own state $x_{s_t}^i$, whereas the correct field at the composed state would average using both states. Equation (25) uses the component fields in place of the sample velocities $H_i - \epsilon_i$, and this is exact only if seeing the other component does not change the expected velocity of each component, $\mathbb{E}[H_i - \epsilon_i \mid x_{s_t}^1, x_{s_t}^2] = \mathbb{E}[H_i - \epsilon_i \mid x_{s_t}^i]$. Deterministic responses satisfy this condition, and so does a suitable conditional independence of the two component paths; independent noises alone do not. Otherwise, the composed target can be biased even in linear coordinates.

### H.3 The Construction in Log-Gamma Coordinates

The model works in Log-Gamma coordinates $x = T(H)$ (Section 3.2), but light adds only in linear HDR. LC therefore decodes the component states, composes them in linear HDR as in Appendix H.1, and maps the result back.

**Encoding and Jacobians.** For LC we use the signed extension of the Log-Gamma encoding to all real values, without the clipping applied to training targets and exported images (Appendix G.2):

$$T(H) = 2\,\mathrm{sgn}(H)\left[\frac{\log(1 + \gamma_{\mathrm{LG}}|H|)}{\log(1 + \gamma_{\mathrm{LG}} M)}\right]^{1/\gamma_{\mathrm{LG}}} - 1, \tag{32}$$

$$T^{-1}(x) = \frac{\mathrm{sgn}(u)}{\gamma_{\mathrm{LG}}}\left[(1 + \gamma_{\mathrm{LG}} M)^{|u|^{\gamma_{\mathrm{LG}}}} - 1\right], \qquad u = \frac{x+1}{2}, \tag{33}$$

with $M = 50$ and $\gamma_{\mathrm{LG}} = 2.2$, so that $T(M) = 1$. Values above $M$ map above 1 and can still be inverted, and the sign terms allow the negative values that occur in noisy states. Velocities convert between the two coordinates with the Jacobians of $T^{-1}$ and $T$,

$$\dot{H} = J_{T^{-1}}(x)\, v, \qquad v = J_T(H)\, \dot{H}. \tag{34}$$

Because $T$ acts on each channel of each pixel separately, both Jacobians are diagonal. With $\ell = \log(1 + \gamma_{\mathrm{LG}} M)$, $r(H) = \log(1 + \gamma_{\mathrm{LG}}|H|)/\ell$, and $u = (x+1)/2$, their diagonal entries are

$$T'(H) = \frac{2}{\ell}\,\frac{r(H)^{1/\gamma_{\mathrm{LG}} - 1}}{1 + \gamma_{\mathrm{LG}}|H|}, \qquad (T^{-1})'(x) = \frac{\ell}{2}\,|u|^{\gamma_{\mathrm{LG}} - 1}(1 + \gamma_{\mathrm{LG}} M)^{|u|^{\gamma_{\mathrm{LG}}}}, \tag{35}$$

so converting a velocity is an elementwise multiplication. $T'$ diverges at $H = 0$, that is, at $x = -1$, and the conversion is used away from this point.

**Composed state and target.** In training, $H_1$ and $H_2$ are the HDR targets, clipped to $[0, M]$ (Appendix G.2); where clipping is active, equation (21) need not hold. The component states follow the model's own paths in encoded coordinates, $x^i_{s_t} = (1 - s_t)\epsilon_i + s_t T(H_i)$ with independent $\epsilon_i \sim \mathcal{N}(0, I)$, and the model predicts their velocities $v_i = v_\theta(x^i_{s_t}, s_t; \mathcal{C}_{E_i})$, where $\mathcal{C}_{E_i}$ contains the source $S$, the lighting maps of $E_i$, and the text. We decode the states, $H^i_{s_t} = T^{-1}(x^i_{s_t})$, and compose state and velocity in linear HDR with the rescaling of equations (24) and (25):

$$H^\Sigma_t = \frac{\gamma_t}{\sqrt{2}}\big(H^1_{s_t} + H^2_{s_t}\big), \qquad \widehat{v}^{H,\Sigma}_t = \frac{(\sqrt{2}-1)H^\Sigma_t + J_{T^{-1}}(x^1_{s_t})\, v_1 + J_{T^{-1}}(x^2_{s_t})\, v_2}{\gamma_t}. \tag{36}$$

Both are then mapped back to model coordinates:

$$x^\Sigma_t = T(H^\Sigma_t), \qquad \widehat{v}^\Sigma_t = J_T(H^\Sigma_t)\, \widehat{v}^{H,\Sigma}_t. \tag{37}$$

The LC loss equation (3) compares the model's velocity $v_\Sigma = v_\theta(x^\Sigma_t, t; \mathcal{C}_{E_1+E_2})$ at this state with $\widehat{v}^\Sigma_t$. With $T$ the identity, the construction reduces to that of Appendix H.1.

**Why the construction is not exact.** The model is trained on straight paths in encoded coordinates. For the summed image, such a path would be

$$\bar{x}^\Sigma_t = (1 - t)\epsilon_\Sigma + tT(H_1 + H_2), \qquad \epsilon_\Sigma \sim \mathcal{N}(0, I). \tag{38}$$

The composed state $x^\Sigma_t$ ends at the same point, $T(H_1 + H_2)$, but it is formed by decoding, adding in linear HDR, and re-encoding, so it follows a different path. At $t = 0$, for example, it equals $T\big((T^{-1}(\epsilon_1) + T^{-1}(\epsilon_2))/\sqrt{2}\big)$, which is not standard Gaussian. The Jacobians differentiate this constructed path correctly, but its velocity is not the velocity along the path on which the model is trained. Even with exact component predictions and deterministic, perfectly additive responses, the average LC target at a state can therefore differ from the correct flow-matching velocity there, and LC can pull an otherwise correct model away from it. In Log-Gamma coordinates, we use LC as a regularizer, not as an exact identity, and judge it by its measured effect on relighting and additivity (Appendices E.1 and E.2).

**Practical choices.** We apply LC only for $t > 0.3$ to reduce the influence of the initial noise; this does not remove the path difference above. For numerical stability, entries of the noisy states outside $-1 < x < 3$ are left out of the loss (Appendix G.4), which also changes which states the loss averages over. All three predictions share parameters and receive gradients, which can also shift the component predictions (Appendix I); in the controlled studies, single-light errors change little with LC (Table 8). The construction adds no trainable module and is not used at inference.

# I Consistency Gradients and Component-Prediction Bias

The LC target is built from the model's own component predictions, and training sends gradients into those predictions as well. This section shows that such gradients can lower the consistency loss by changing the component predictions, at the cost of their fit to the data. The effect is separate from the two limits discussed in Appendix H (scene uncertainty and the Log-Gamma path), which remain even when the component predictions are held fixed.

## I.1 Conditional-Variance Decomposition

Let $Z$ collect everything the summed-light prediction sees: its state, time, source, and combined lighting. Let $Y$ additionally contain the sampled component states. Write $P_\theta(Z)$ for the predicted velocity $v_\Sigma$ and $U_\theta(Y)$ for its target $\widehat{v}^\Sigma_t$ (Appendix H.3), which depends on $\theta$ through the component predictions. Assume that the sampling distribution of $(Z, Y)$ does not depend on $\theta$ and that second moments are finite, and let

$$m_\theta(Z) = \mathbb{E}[U_\theta(Y) \mid Z]. \tag{39}$$

Adding and subtracting $m_\theta$ splits the loss into two terms:

$$\begin{aligned} \mathcal{L}_{\text{cons}} &= \mathbb{E}\|P_\theta(Z) - U_\theta(Y)\|_2^2 \\ &= \mathbb{E}\|P_\theta(Z) - m_\theta(Z)\|_2^2 + \mathbb{E}\,\mathrm{Var}(U_\theta \mid Z). \end{aligned} \tag{40}$$

Here the variance is the trace of the conditional covariance, and the cross term vanishes because $\mathbb{E}[U_\theta - m_\theta(Z) \mid Z] = 0$.

If the target were frozen, the second term would be a constant, and squared regression would move $P_\theta$ toward the conditional mean of the constructed target. As Appendix H.3 shows, this mean need not be the flow-matching velocity. With live component predictions, both terms depend on $\theta$. Under a change $\delta U$ of the target,

$$\delta\, \mathbb{E}\, \mathrm{Var}(U_\theta \mid Z) = 2\mathbb{E}\langle U_\theta - m_\theta(Z), \delta U\rangle, \tag{41}$$

because $\delta m_\theta(Z) = \mathbb{E}[\delta U \mid Z]$ and its cross term also vanishes. The loss can therefore also decrease by making the targets from different component pairs agree with each other, not only by improving the summed-light prediction. With $J_P = \partial P_\theta/\partial\theta$ and $J_U = \partial U_\theta/\partial\theta$, the full gradient is

$$\nabla_\theta \mathcal{L}_{\text{cons}} = 2\mathbb{E}[(J_P - J_U)^\top (P_\theta - U_\theta)]. \tag{42}$$

Stopping gradients through the target removes the $J_U$ term. Because the three predictions share parameters, the component predictions can still change through later updates.

### I.2 A Closed-Form Example with Data Supervision

A one-dimensional example shows that the effect remains even with data supervision. Let $X_1, X_2$ be independent standard normal variables and $Z = X_1 + X_2$. A component prediction $U_a = aX_1$ has supervised optimum $a = 1$, and a summed-light prediction $P_b = bZ$ matches the conditional mean of $U_a$ when $b = a/2$. With data weight $\lambda_d > 0$ and consistency weight $\lambda_c > 0$, and using $\mathbb{E}[Z^2] = 2$ and $\mathbb{E}[ZX_1] = 1$,

$$\begin{aligned}\mathcal{L}(a,b) &= \lambda_d(a-1)^2 + \lambda_c\mathbb{E}(bZ - aX_1)^2 \\ &= \lambda_d(a-1)^2 + \lambda_c(2b^2 + a^2 - 2ab).\end{aligned} \tag{43}$$

Minimizing over $b$ gives $b = a/2$, and then

$$\mathcal{L}(a) = \lambda_d(a-1)^2 + \frac{\lambda_c}{2}a^2, \qquad a^* = \frac{\lambda_d}{\lambda_d + \lambda_c/2} < 1, \qquad b^* = \frac{a^*}{2}. \tag{44}$$

The consistency term moves the component away from its data optimum, $a^* < 1$, although data supervision is present. Here $\mathrm{Var}(aX_1 \mid Z) = a^2/2$, so shrinking $a$ directly reduces the variance term. Detaching $U_a$ leaves $a$ to the supervised loss alone and recovers $a = 1$ and $b = 1/2$.

### I.3 When the Mechanism Applies

The effect can differ across the range of conditions. Suppose, for example, that the lighting condition is a scalar intensity, that component intensities $u_1, u_2$ lie in $[0.05, 1]$, and that the summed condition is $u_1 + u_2 \in [0.1, 2]$. Intensities in $[0.05, 0.1)$ then occur only as components, so their consistency gradients come only through the component side. If the target has irreducible conditional variance, this asymmetry puts extra pressure on the component predictions in that range.

Conversely, if the conditioning determines the physical response and the coupled paths satisfy $U^*(Y) = P^*(Z)$ for every sample, then

$$\mathrm{Var}(U^* \mid Z) = 0, \qquad P^* - U^* = 0, \tag{45}$$

so both gradient terms vanish at the correct solution, and no detachment is needed to avoid this particular bias. This holds in linear coordinates with deterministic responses and exact component predictions (Appendix H.1). Additivity of the clean images alone does not give it: it also depends on the coupling (Appendix H.2) and on the coordinates. In particular, deterministic responses do not remove the Log-Gamma path difference (Appendix H.3), and detaching the target corrects neither that difference nor an incompatible joint distribution of the components. These are separate reasons why LC can move the optimum of the supervised objective.