
Supplement for UniRelight: Learning Joint Decomposition and Synthesis for Video Relighting

Kai He^{1,2,3} Ruofan Liang^{1,2,3} Jacob Munkberg¹ Jon Hasselgren¹

Nandita Vijaykumar^{2,3} Alexander Keller¹ Sanja Fidler^{1,2,3}

Igor Gilitschenski^{2,3†} Zan Gojcic^{1†} Zian Wang^{1,2,3†}

¹NVIDIA ²University of Toronto ³Vector Institute

In this Appendix, we first discuss the broader impact of our project ([Appendix A](#)). We then provide additional implementation details of our model and experiments ([Appendix B](#)), followed by further results and analysis ([Appendix C](#)). Please refer to the [accompanying video](#) for additional qualitative results and comparisons.

A Broader Impact

We present UniRelight, a generative framework that jointly estimates albedo and synthesizes relit videos from a single input, enabling diverse lighting manipulation across both synthetic and real-world scenes. This capability can support a range of applications, including creative content generation, visual effects, virtual production, and potentially data augmentation for training more robust computer vision models in domains such as robotics and autonomous driving.

As with all generative video models, UniRelight may reflect biases present in its training data. Such biases could lead to relighting results that fail to generalize to underrepresented scene types or lighting conditions. Furthermore, tools for lighting manipulation carry the risk of misuse, such as altering or misrepresenting visual content in sensitive contexts like surveillance or media.

We discourage the use of UniRelight in applications where relighting may contribute to misinformation, misattribution, or privacy violations. In human-centric use cases, we recommend careful dataset curation to ensure fair representation across skin tones, races, and gender identities. Practitioners are encouraged to critically assess and de-bias training data to mitigate unintended harms where appropriate.

B Experimental Details

B.1 Implementation Details

We fine-tune our models based on Cosmos-Predict1-7B-Video2World [2], a pre-trained DiT video diffusion model.

The encoded latents \mathbf{z}^I , \mathbf{z}^a , \mathbf{z}^E , $\mathcal{E}(\mathbf{E}_{\text{ldr}})$, $\mathcal{E}(\mathbf{E}_{\text{log}})$, and $\mathcal{E}(\mathbf{E}_{\text{dir}})$ are all in $\mathbb{R}^{l \times h \times w \times C}$, where $C = 16$. We use $C_{\text{emb}} = 3$ as the dimension of the type embedding. Thus, the concatenated tokens have a channel dimension of $16 + 1 + (16 + 1) \times 3 + 3 = 71$, where each latent has an associated binary condition mask (added as an extra channel), and the lighting features (\mathbf{E}_{ldr} , \mathbf{E}_{log} , \mathbf{E}_{dir}) each include a condition mask to indicate whether they are provided.

We adopt an image-video co-training strategy and train the model in two stages. Firstly, we train with only synthetic data by mixing the image data (sampling one frame from the video data) and the video data in a ratio of 1 : 1 for 15,000 iterations. Then we train with all data with random

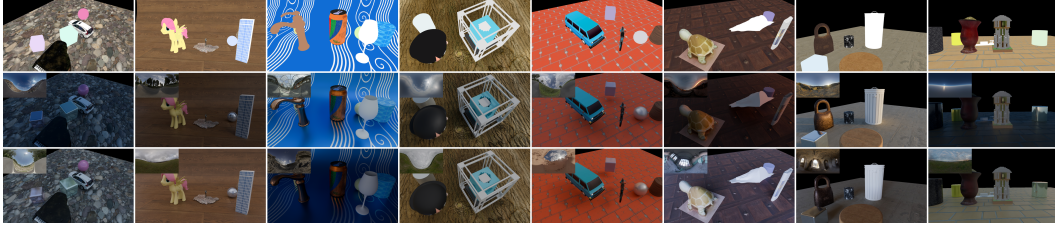


Figure 1: **Synthetic data visualization.** Randomly sampled example images are shown from our synthetic rendering data. The top-most images show albedo maps, while the bottom two rows display rendered scenes under two different illuminations with the corresponding environment maps.

Table 1: Quantitative comparison with IC-Light.

	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
IC-Light [4]	18.08	0.834	0.096
Ours	23.19	0.901	0.079

Table 2: Evaluation of inference runtime cost.

	Runtime cost (seconds) \downarrow
DiffusionRenderer [1]	566.6
DiffusionRenderer (Cosmos)	780.0
Ours	445.5

sampling, including synthetic video data, synthetic image data, real-world auto-labeled data, and MIT multi-illumination data, in a ratio of 8 : 1 : 3 : 2 for 12,000 iterations. We augment the real-world auto-labeled data with random flipping.

All training is done with a batch size of 64, using the AdamW optimizer with a learning rate of 2×10^{-5} , with mixed-precision (BF16) training at a resolution of 480×848 pixels. The AdamW optimizer was employed with a weight decay of 0.1. The exponential decay rates for the moment estimates β are set to 0.9 for the first moment and 0.99 for the second moment, with ϵ at 1×10^{-10} . The total training of two stages takes around 4 days on 32 A100 GPUs.

During inference, we use 35 denoising steps. We do not apply classifier-free guidance (CFG), as we empirically found that inference without CFG yields more accurate and visually consistent results.

Baseline configurations. Since DiLightNet [3] requires a text prompt per example, we use meta/llama-3.2-90b-vision-instruct¹ to generate a short prompt for each example in the datasets based on the first image in each clip with the instruction “What is in this image? Describe the materials. Be concise and produce an answer with a few sentences, no more than 50 words.”

As each of the baselines generates videos in different resolutions, for UNet-based baselines, we run inference on the model with the video first resized to 486×864 and then center-cropped to a resolution of 448×832 ; for our DiT-based model, we run inference on the model with the resized video with resolution of 486×864 and then center-cropped to 448×832 to align the results.

Quantitative comparison configurations. For quantitative evaluation, we apply background masks to the synthetic dataset to focus on foreground appearance. For the MIT multi-illumination dataset, we follow the dataset protocol and mask out light probes in all outputs before computing metrics.

B.2 User Study Details

We conducted two user studies on Amazon Mechanical Turk to evaluate the perceptual quality of relighting results.

MIT multi-illumination dataset is a public benchmark with ground truth relighting. Participants were shown three images: a ground truth relit image and two relighting results—one generated by our method and one by a baseline model. Their task was to choose the result that more closely resembled the ground truth, considering attributes such as transparency, shadows, and reflections.

The exact instructions shown to participants were as follows:

¹<https://www.llama.com/>



Figure 2: **Qualitative comparison with IC-Light [4].** We provide the environmental background used for IC-Light conditioning, with the reference environment ball on the left. Our method produces higher-quality and more accurate relighting results.

Carefully compare Image A, the Reference Image, and Image B. Your task is to determine which image (A or B) is more similar to the Reference Image.

To make an informed decision, you may zoom in to examine the details. Pay close attention to aspects such as lighting, reflections, and shadows, as these can affect how natural the image appears.

Once you have compared the images, select the one that best matches the Reference Image.

☐ Image A

☐ Image B

We evaluated 30 scenes from the test set, comparing our method against four baselines. The study was repeated three times with 11 unique participants in each run. In total, this resulted in $30 \times 4 \times 11 \times 3 = 3960$ individual comparisons.

StreetScenes Dataset. This dataset contains 19 urban street scenes without ground-truth relighting. Participants were shown a reference image along with two relit videos generated by different ablated versions of our method.

The instructions presented to participants were as follows:

In this study, you will be shown a Reference Image and two videos – Video A and Video B – that changes the lighting of the scene. Your task is to watch both videos and choose which one (A or B) you think has more realistic shadows and reflections. To make an informed decision, you may zoom in to examine the details. Pay close attention to aspects such as lighting, reflections, and shadows, as these can affect how natural the image appears.

Once you have compared the videos, select the one that has more realistic lighting effects.

☐ Video A

☐ Video B

We evaluated 19 scenes, comparing a base version against two ablated versions of our method. The study was repeated three times with 11 unique participants in each run. In total, this resulted in $19 \times 2 \times 11 \times 3 = 1254$ individual comparisons.

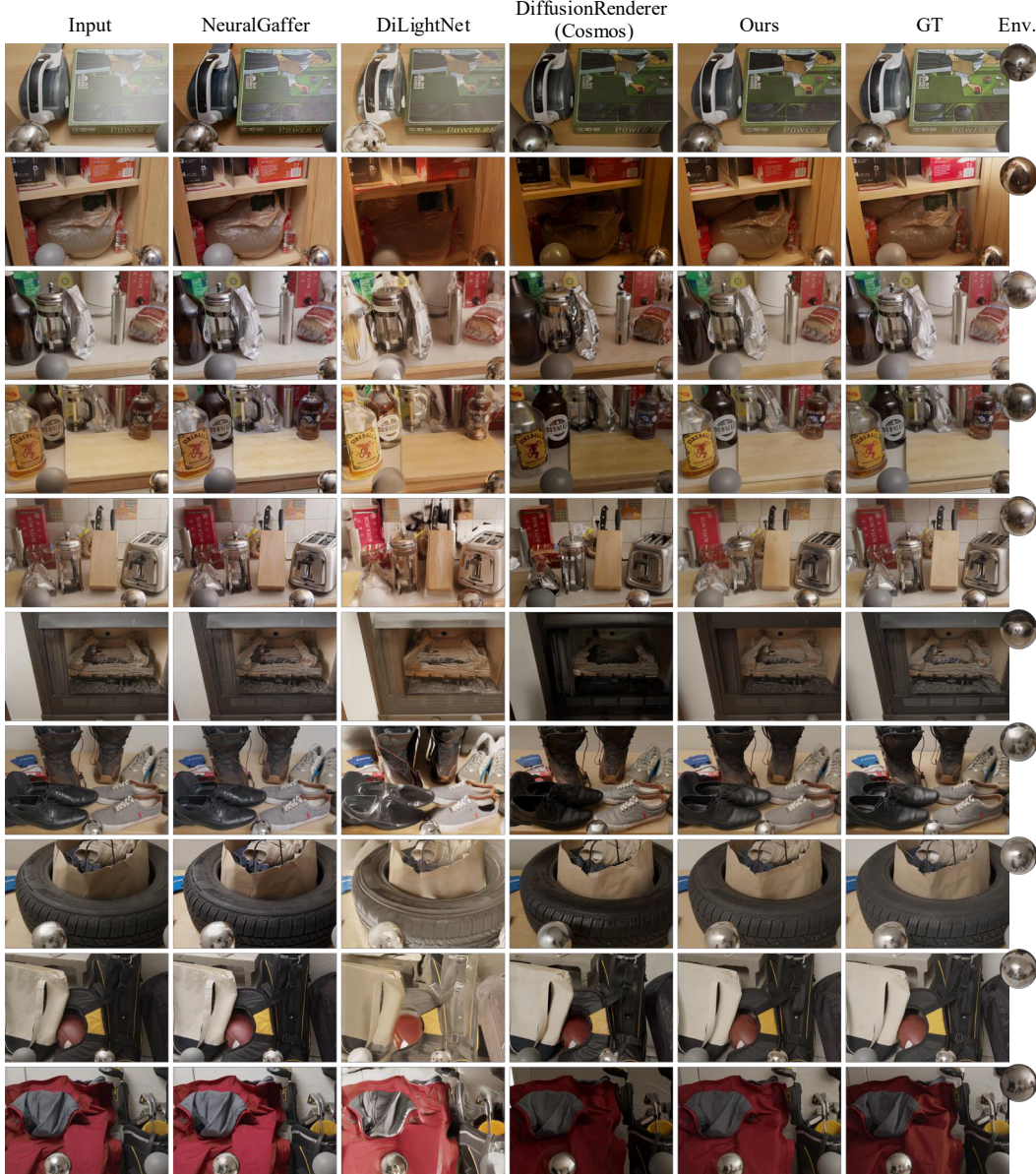


Figure 3: **Additional qualitative comparison on MIT multi-illumination dataset.** Our method consistently achieves more accurate relighting results than all baselines on the MIT multi-illumination dataset, demonstrating strong capability in relighting complex materials.

B.3 Synthetic data visualization

We show a synthetic data visualization in Figure 1. Each scene contains albedo videos, two environment maps, and pairwise videos rendered under the environment maps.

C Additional Results

Runtime cost. We evaluate the inference runtime of our model on a 57-frame video at a resolution of 480×848 . The overall inference time for performing 35 denoising steps, including VAE encoding and decoding, is 445.5 seconds, measured on a single A100 GPU.

To contextualize this cost, we compare our method against two baselines: DiffusionRenderer [1] and its Cosmos-based variant. All methods use 35 denoising steps for consistency. DiffusionRenderer is run at 448×832 resolution (slightly smaller than ours but divisible by 32 to fit its architecture) while the Cosmos variant is run at our native resolution.



Figure 4: **Additional qualitative results under point-light illumination.** The bottom right of each column indicates the target lighting conditions. Our results show strong robustness of our method under point-light illumination.

For baseline methods, the total runtime is the sum of the inverse rendering and forward rendering durations. Notably, DiffusionRenderer requires five separate inverse rendering passes and one forward rendering pass per video, resulting in significantly higher computational cost. In contrast, our approach performs joint relighting and albedo estimation in a single pass and is correspondingly faster. Full timing results are shown in Table 2.

Comparison with IC-Light [4]. We compare our method with the single-image relighting approach IC-Light [4] on object-centric synthetic data, as shown in Table 1 and Figure 2. Note that the two methods follow different relighting formulations: IC-Light is designed for object relighting using background context as the primary cue—without access to an explicit environment map, while our method is conditioned on full HDR illumination, but is not specifically tuned for object-centric data.

Our method shows improved quantitative and qualitative performance. Since IC-Light relies on background appearance as its primary cue and has less information about the surrounding lighting, it may retain input-specific effects in its outputs—such as specular highlights and shadows, which can limit accuracy under novel lighting conditions. In contrast, our method produces more faithful relighting results, with sharper specular highlights, more realistic shadows, and improved visual fidelity.

Additional qualitative comparison on the MIT multi-illumination dataset. We provide additional qualitative comparisons on the MIT Multi-Illumination dataset in Figure 3. To ensure a fair comparison, we include results from our re-implemented version of DiffusionRenderer using the Cosmos backbone, which achieves higher visual fidelity than the original implementation. Our method consistently produces more accurate transparency, specular highlights, and shadows across scenes, demonstrating strong capability in handling complex materials and outperforming all baselines in visual quality.

Additional qualitative results under point-light illumination. We further evaluate the robustness of our method in extreme cases, such as point-light illuminations, which do not exist in our training data. As shown in Figure 4, our method produces high-quality relighting results, demonstrating the strong robustness and generalization capability of our method.

Additional qualitative results on real scenes. We present additional results on real scenes in Figure 5. Our method produces high-quality albedo and relighting results with realistic specular highlights and shadows under target lighting conditions.

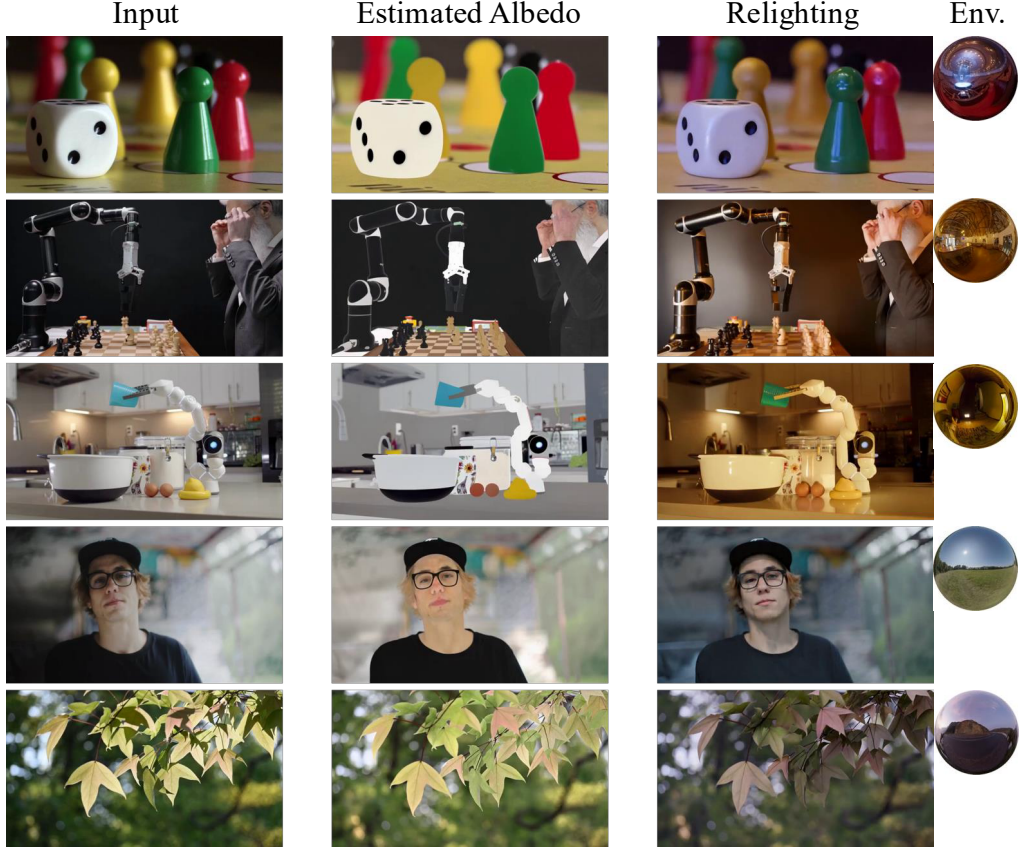


Figure 5: **Additional qualitative results on real scenes.** Our method provides high-quality albedo estimation and realistic relighting results.

References

- [1] Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Zhi-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, and Zian Wang. Diffusionrenderer: Neural inverse and forward rendering with video diffusion models. In *CVPR*, 2025. 2, 4
- [2] NVIDIA. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025. 1
- [3] Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. DiLightNet: fine-grained lighting control for diffusion-based image generation. In *SIGGRAPH*, 2024. 2
- [4] 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 *ICLR*, 2025. 2, 3, 5