BodyReLux: Temporally Consistent Full-Body Video Relighting

LI MA, Eyeline Labs, United States of America
MINGMING HE, Eyeline Labs, United States of America
XUEMING YU, Eyeline Labs, United States of America
DAVID M. GEORGE, Eyeline Labs, United States of America
AHMET LEVENT TAŞEL, Eyeline Labs, Canada
PAUL DEBEVEC, Eyeline Labs and Netflix, United States of America
JULIEN PHILIP, Eyeline Labs, United Kingdom

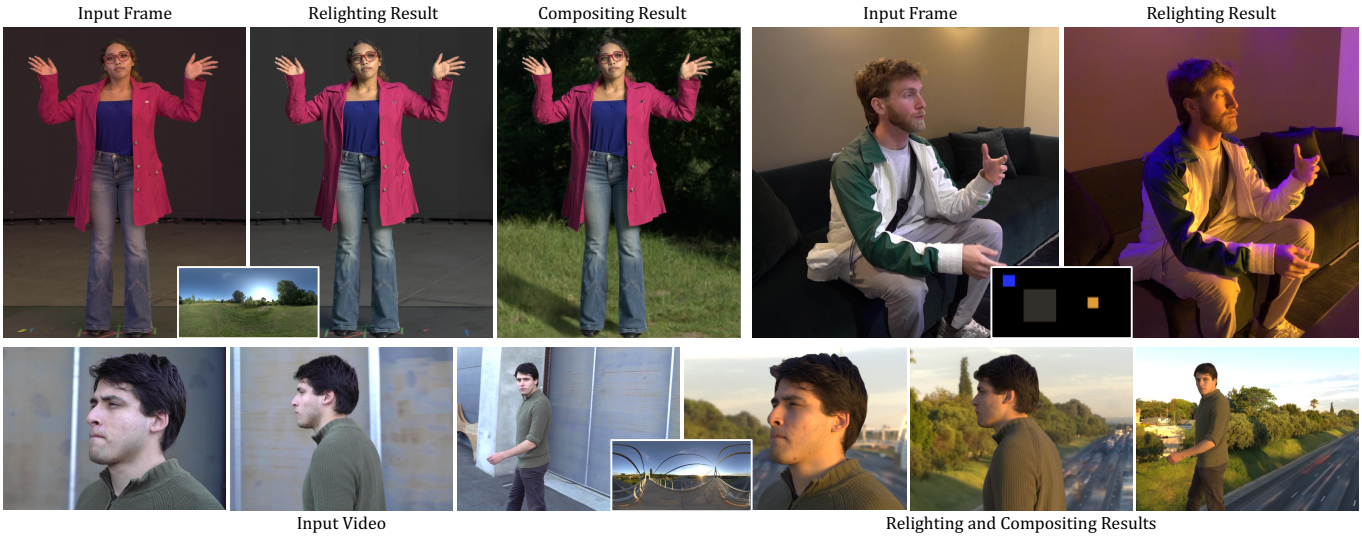

Fig. 1. **Input and relighting results of BodyReLux.** Given an input video of a subject with arbitrary lighting, and a target lighting condition, BodyReLux allows relighting the input video to the target lighting with a high level of photorealism and temporal consistency. The technique works for any framing (full body, upper body, and closeup), any resolution, and any frame length, including casually captured videos. Compositing the relit results onto appropriate backgrounds ("compositing results") produces plausible complete shots.

Being able to relight human performance is a fundamental task for post production and content creation. We present BodyReLux, a subject-specific video diffusion-based framework for relighting full-body human performances in a temporally consistent way. Our model is trained on a hybrid dataset of pixel-aligned video relighting pairs, covering a diverse combination of lighting conditions, performances and viewpoints. To acquire such dataset, we combine traditional static One-Light-at-a-Time (OLAT) capture and a novel dynamic performance capture in which two smoothly varying lighting sequences are rapidly interleaved. Because the lighting operates above the human flicker-fusion threshold, the interleaving does not appears to strobe. We train our video relighting model from a pretrained text-to-video model to fully leverage the generative priors for producing high quality videos. To achieve accurate lighting control, we introduce a new lighting conditioning method that represents each light source as a token. We further condition on sequences of lighting using masked attention to support dynamic lighting control. Together with a carefully designed data augmentation pipeline, we achieve photorealistic, robust, and temporally consistent video relighting of subject-specific human performances.

Additional Key Words and Phrases: Relighting, Video Diffusion Model

1 Introduction

Lighting is fundamental to cinema, essential for underscoring atmosphere, emotion, and composition. But effective lighting is difficult to achieve: on set, a team of trained professionals must move stands, ladders, and lights, adjust dimmers, barn doors, and lenses, and must place filters and scrims and flags and bounce cards. On location, the cinematography often must wait for the sun to be in the right positions or for a cloud to pass. In every case, many other talented craftpeople must wait for the lighting to be ready so the next scene can be shot. Being able to shoot scenes of a movie in whatever lighting is easily available, and, in post-production, changing the lighting to best serve the artistic storytelling needs of the scene, is highly appealing. It could save a great deal of time and resources, but, more importantly, it would allow turning the task of lighting

Authors' Contact Information: Li Ma, Eyeline Labs, Los Angeles, United States of America, li.ma@scanlinevfx.com; Mingming He, Eyeline Labs, Los Angeles, United States of America, mingming.he@scanlinevfx.com; Xueming Yu, Eyeline Labs, Los Angeles, United States of America, xueming.yu@scanlinevfx.com; David M. George, Eyeline Labs, Los Angeles, United States of America, david.george@scanlinevfx.com; Ahmet Levent Taşel, Eyeline Labs, Vancouver, Canada, ahmet.tasel@scanlinevfx.com; Paul Debevec, Eyeline Labs and Netflix, Los Angeles, United States of America, debevec@gmail; Julien Philip, Eyeline Labs, London, United Kingdom, julien.philip@scanlinevfx.com.

a shot as a malleable, decoupled and iterative task independent of budget, time and place constraints.

While appealing, production grade shot relighting is challenging. Many complex aspects of light transport such as shadows, specularities, interreflections, translucency, and scattering must be disentangled from the original scene's shape, material and performance, and simulated as under an entirely new lighting. Additionally, to satisfy the vision of the lighting designers, our method must be highly controllable.

While recent models have shown significant promise for relighting faces [Chaturvedi et al. 2025; He et al. 2024; Mei et al. 2025; Ren et al. 2024], relighting full-body performances in a controllable, realistic, and temporally stable way has remained elusive. Firstly, human bodies are harder to relight than faces: they exhibit many more degrees of freedom in motion, leading to complex pose variations as well as significantly stronger self-shadowing and interreflections. Moreover, full-body performances involve clothing, which introduces a vast range of material reflectance properties – from glossy black leather to translucent white mesh – whose appearance further changes as garments fold, stretch, and drape in complex ways during motion. As a result, high-quality video data with sufficient diversity in both lighting and body poses is crucial for full-body performance relighting. However, capturing such video data remains difficult, as it requires recording multiple instances of the same video under different lighting conditions. Rapid switching between lightings often introduces strong strobing, causing subject discomfort.

To address these challenges, we introduce a novel framework for capturing and relighting full-body performances. We propose to acquire high-quality video relighting training data by combining static One-Light-at-a-Time (OLAT) images [Debevec et al. 2000] with dynamic sequences captured under slowly evolving paired lighting conditions. This enables both highly controllable illumination through OLATs and a wide range of body poses and facial expressions through temporally coherent video data, providing supervision for training a video relighting model. Moreover, for the dynamic sequences, we exploit digitally bi-packed lighting patterns [Yu et al. 2025b] to produce paired video data without noticeable flicker to the subjects, enabling a comfortable capture environment.

We then finetune a video diffusion model using the high-quality relighting data. The strong generative priors of the diffusion model help produce photorealistic and temporally consistent relighting results. To accurately condition the model on illumination, we introduce OLAToken, a new lighting-conditioning mechanism that learns a permutation-invariant aggregation of per-light contributions, closely reflecting the compositional nature of physical illumination. In addition, we propose a dynamic lighting conditioning module that enables relighting under time-varying illumination, allowing lighting to change continuously throughout a performance.

To summarize, we first capture a subject under both static and dynamic poses with dynamic lighting patterns to produce paired video data with known illumination conditions. We then train a subject-specific video relighting model that can relight arbitrary poses of the same subject under arbitrary lighting in a long sequence, producing high-resolution, temporally coherent, photorealistic relighting effects across the whole body as shown in Fig. 1.

Our key contributions can be summarized as follows:

- A diffusion-based video relighting model that achieves state-of-the-art photorealistic, controllable, and temporally consistent subject-specific performance relighting.
- A novel lighting conditioning approach that supports both static and dynamic lighting control.
- A comfortable capture process that avoids flickering lighting.
- The use of a hybrid dataset that captures both lighting and pose diversity and generates pixel-aligned video relighting pairs, with wide, medium, and close-up views of the subjects.

2 Related Work

2.1 LED Spheres and Appearance Acquisition

Sophisticated LED Spheres [Debevec et al. 2002; LeGendre et al. 2016] have been proposed to give filmmakers controllable environmental lighting during shooting, but they do not offer the flexibility of relighting in postproduction. Relighting *after* shooting was addressed for static subjects with the first light stage Debevec et al. [2000], which acquired the reflectance field of a human face using one-light-at-a-time (OLAT) sequences for image-based relighting.

Time-multiplexed illumination techniques [Chabert et al. 2006; Wenger et al. 2005] extend image-based relighting to dynamic subjects by looping though very rapidly repeating OLAT lighting patterns, but they require expensive ultra-high speed cameras (e.g. 1000+ fps) with limited spatial resolution, and tend to produce uncomfortable strobing illumination. Optical flow [Chabert et al. 2006; Peers et al. 2007] and tracked meshes [Hawkins et al. 2004] can be used to propagate lighting across multiple relit reference images. While showing promise, this approach relies heavily on the quality of the alignment and struggles with highly dynamic content. Recently, neural networks have been trained on OLAT reflectance data to implicitly learn a subject's appearance under different lighting [Bi et al. 2021; He et al. 2024; Meka et al. 2019, 2020] but have struggled with temporal stability.

2.2 Physics-Based Relighting

Instead of a purely data-driven approach, physics-based relighting recovers intrinsic or material attribute of images, and then use physics-based rendering (PBR) to relight them. Some early works infer a Cosine-based BRDF from color-gradient patterns [Fyffe 2009; Guo et al. 2019]. Retinex theory [Land and McCann 1971] is used as a hand-crafted prior to separate reflectance and illumination. High quality intrinsic decomposition from a single image or videos has been achieve with deep learning [Li et al. 2020; Liang et al. 2025; Zeng et al. 2024a; Zhu et al. 2022b,a] by training on large scale synthetic data. With either hand-crafted or data-driven priors, many works use differentiable rendering to jointly optimize materials in 3D to get a relightable 3D representation [CHEN et al. 2025; Jin et al. 2023; Liang et al. 2024; Srinivasan et al. 2021; Sun et al. 2025; Wu et al. 2025; Zhang et al. 2021a, 2022].

PBR based relighting often results in a synthetic looking appearance, especially for humans, whose skin and hair scatter light in complex ways. A neural renderer can be cascaded to PBR to fix these artifacts [Griffiths et al. 2022; Kim et al. 2024; Philip et al. 2019, 2021] but human appearance remains challenging. Some recent works have shown that Diffusion Models have priors strong enough to

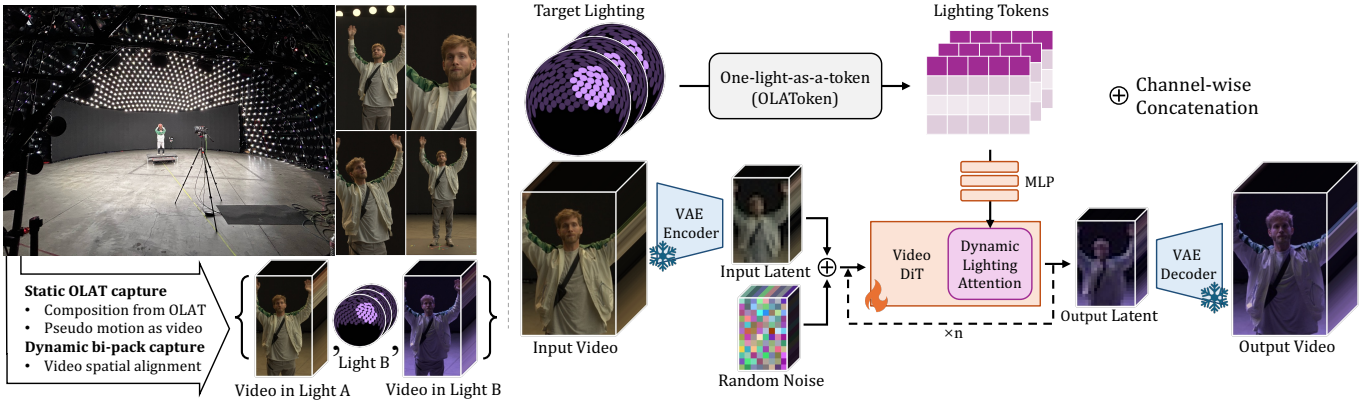

Fig. 2. **Overview of the method.** We capture static OLAT data and bi-packed video data of a subject moving inside a large LED sphere, resulting in a dataset of video relighting training tuples that consists of two pixel-aligned videos under different lighting conditions and the corresponding lighting sequences. We train a video diffusion model with a novel lighting conditioning module that supports dynamic lighting control.

bypass PBR and directly generate the final images from estimated materials and target lighting [Liang et al. 2025; Zeng et al. 2024a].

2.3 Deep Relighting

Training a relighting model using a deep neural network has been a promising approach to achieving robust and generalizable relighting [Zhou et al. 2019]. A common paradigm is to condition the model with intermediate buffers, such as gradient illumination [Meka et al. 2019], normal and albedo [Daichi Tajima 2025; Mei et al. 2024, 2023; Pandey et al. 2021; Wang et al. 2020], roughness [Daichi Tajima 2025; Kim et al. 2024], precomputed shading maps [Daichi Tajima 2025; Kim et al. 2024; Kocsis et al. 2024; Pandey et al. 2021; Philip et al. 2019; Zeng et al. 2024b] or shadow maps [Daichi Tajima 2025; Hou et al. 2021; Kocsis et al. 2024]. End-to-end prediction of the relighting results is also feasible if the model has a large capacity [Kim et al. 2024; Sun et al. 2019] or strong generative prior such as with a diffusion model [Bharadwaj et al. 2025; Chadebec et al. 2025; He et al. 2024; Jin et al. 2024; Magar et al. 2025; Mei et al. 2025; Zhang et al. 2025]. There is growing interest in using a large scale synthetic datasets to train relighting models [Chaturvedi et al. 2025; Poirier-Ginter et al. 2024; Yeh et al. 2022; Zhang et al. 2025]. Most of these methods either work on objects or single images and do not provide the quality required for professional video production.

To train a video relighting model, NVPR [Zhang et al. 2021b] acquired video relighting data at 1000fps Wenger et al. [2005]. To avoid using a high fps camera, Relumix [Wang et al. 2025a] and LightAVideo [Zhou et al. 2025] extend single image relighting models to video. Wang et al. [2025b] enforces temporal consistency by using flow-based warping. RelightVid [Fang et al. 2025] generates synthetic relighting pairs from in-the-wild videos through color augmentation. DiffusionRenderer [Liang et al. 2025] and Unirelight [He et al. 2025] use synthetic 3D assets to train and tend to generate synthetic-looking results. We believe real-world data is necessary for production grade relighting and propose a video capture process based on standard cinema cameras.

3 Method

An overview of our pipeline is shown in Fig 2. We first capture relighting data of the subject using a large LED Sphere. In addition to a set of traditional OLAT captures of static poses, we record paired video relighting training data using the digital bi-pack [Yu et al. 2025b] technique of interleaving two different lighting condition sequences. After data preprocessing, we get a dataset of video training triples $\{V_A, L_B, V_B\}$, where V_A and V_B are videos with pixel-aligned content but with two different lightings, L_A and L_B. From this dataset we train a video diffusion model for video-to-video translation that relights input videos to any target lighting condition.

3.1 Apparatus

We employ a large-scale LED Sphere with approximately 1600 custom LED light sources distributed over a spherical structure. Each light source has 216 high-power LEDs, spread across red, amber, green, blue, royal blue, and white. We use a real-time multi-spectral lighting reproduction algorithm [Yu et al. 2025a] to obtain high quality color rendition. The lights switch to the next lighting condition with sub-microsecond level accuracy through a wired TTL pulse, allowing a lighting sequence pre-loaded into onboard flash memory to be played back in precise synchronization with the camera shutter as in Wenger et al. [2005]. The stage also includes a black LED panel wall. However, due to its different intensity and color rendition, as well as the lack of synchronization with the strobe system, it is not used in our capture process.

We use five RED Komodo X cinema cameras placed near the periphery of the stage as in Fig. 3(c). The cameras are outfitted with a combination of medium and long focal length lenses to produce wide, medium, and closeup shots of the subject. The camera framed most tightly on the head uses a motorized pan-tilt system and optical tracking system to keep the subject's closeup in frame as they move. The cameras are configured to capture at 120fps at UHD resolution (3840×2160), with a 360° shutter angle and f/4 aperture.

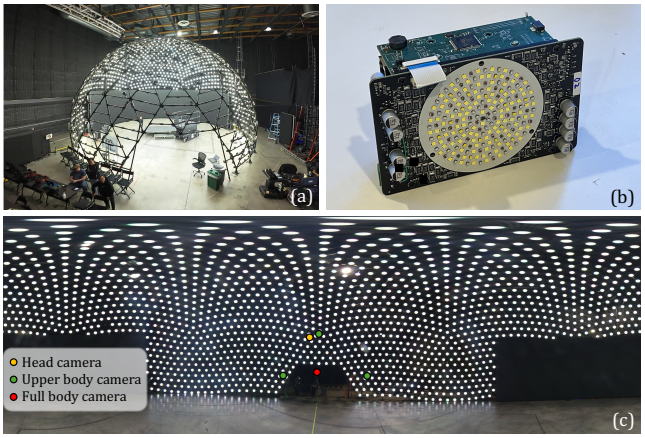

Fig. 3. The LED Sphere (a) consists of 1600 customized multi-spectrum lights (b). An equirectangular projection of all the lights can be seen in (c), where the distribution of cameras are visualized in colored dots.

3.2 Capture process

Our capture process is designed to comfortably record rich lighting data to train a subject-specific relighting model with the following criteria:

(1) Cover a varied set of static and dynamic lighting conditions
(2) Cover diverse view points and performances
(3) Contain video training pairs under two well-synchronized lighting conditions
(4) Avoid flickering light perception so the subjects remain comfortable

The biggest challenge is to design lighting patterns from which we can extract video training pairs (3), while at the same time guaranteeing the diversity of lighting patterns (1). Most existing works [Chabert et al. 2006; Wenger et al. 2005; Zhang et al. 2021b] capture time-multiplexed videos using high-fps (>1000fps) cameras so that multiple interleaved lighting conditions can be recorded closely together in time.

To use standard cinema cameras, we record each subject performing a few-minute of movement sequences at 120 fps. As subjects perform, the LED Sphere digitally bi-packs [Yu et al. 2025b] two gradually transforming spherical environment lighting sequences, A and B, each running at 60 fps. This keeps high-frequency lighting changes above the human visual system's 60Hz flicker fusion frequency for the comfort of the subject as shown in Fig. 4. After temporally aligning sequences A and B, we have a plethora of paired video clips of the subject performing the same aligned motions but recorded under two different lighting conditions.

Since each lighting sequence evolves slowly, with smooth changes happening around once per second, the number of unique lighting conditions is limited. Therefore, we also capture OLATs of a small set of static poses [Debevec et al. 2000; He et al. 2024].

In summary, our capture process includes both static OLAT captures and dynamic bi-packed captures. We predefined 13 static poses and 9 dynamic performance sequences and ask the subject to perform accordingly. To get a diverse set of view points (2), we asked the subject to turn to the 4 cardinal directions in each take. Each take lasts around 40 seconds for a static OLAT capture and 70 seconds for a dynamic bi-pack capture, with a total capture time of 30 minutes per subject. The details of the predefined poses and lighting designs are presented in the supplementary material.

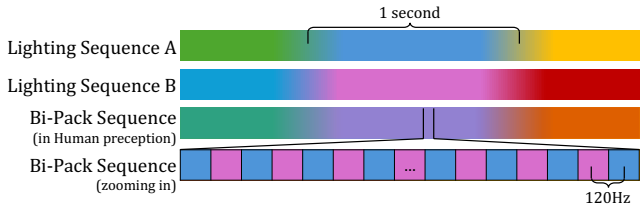

Fig. 4. **Visualization of bi-pack lighting sequence.** A bi-pack sequence consists of two lighting sequences that vary smoothly every 1 second, while rapidly alternating between the two at 120 Hz. Because the switching frequency exceeds the human flicker-fusion threshold, it appears as a mixture of two lightings evolving at 1 Hz.

3.3 Data Preprocessing

After capture, we get a set of video files from each camera. We extract linear EXRs and their corresponding lighting configurations. We preprocess this data to extract video relighting training tuples, $\{V_A, L_B, V_B\}$, where V_B represents a video sequence under lighting L_B, and V_A and V_B are pixel aligned and differ only in lighting.

For static OLAT capture, we align all the OLATs to a reference tracking frame as in [He et al. 2024; Wenger et al. 2005] to obtain pixel-aligned OLAT sequences. We then generate HDRI-lit images by compositing OLAT sequences in linear color space [Debevec et al. 2000]. For each take, we randomly select 50 HDRIs from Polyhaven [Poly Haven 2023] with a random horizontal rotation to create a dataset with diverse lighting. To get video pairs, we repeat the image along the time axis while adding translation and zoom motion.

For dynamic Bi-Pack captures, we extract two video sequences at 60fps with different lightings, noted as V_A and V_B. Since the corresponding frames in the two videos are still captured at slightly different times, they suffer from small misalignments, which could cause ambiguity while training. Therefore, for every two consecutive frames of the same lighting condition, we interpolate a middle frame using FILM [Reda et al. 2022a,b]. This results in two 120fps interpolated videos, V'_A and V'_B where the frames of V'_A are temporally aligned with those from V_B, and those from V'_B with V_A. Since these processed frames may exhibit artifacts, we only use them as conditioning to prevent our relighting model from learning to synthesize artifacts. Formally, we obtain two training pairs (V'_A, B, V_B) and (V'_B, A, V_A).

3.4 Video Relighting Models

Given a dataset of video training tuples $\{(V, L, V_L)\}$, we train a subject-specific video relighting model to relight video from any input lighting to any desired lighting.

Preliminary. For a practical capture process, our data covers just a sparse subset of viewpoints, motions, and lighting conditions. We thus require strong video priors to interpolate this space and leverage a pretrained video diffusion model, WAN2.2 5B Wan et al. [2025], as the backbone of our video relighting model. This model takes a text prompt and a random latent $z^{(T)} \in \mathbb{R}^{h \times w \times t \times c}$ and flattens it

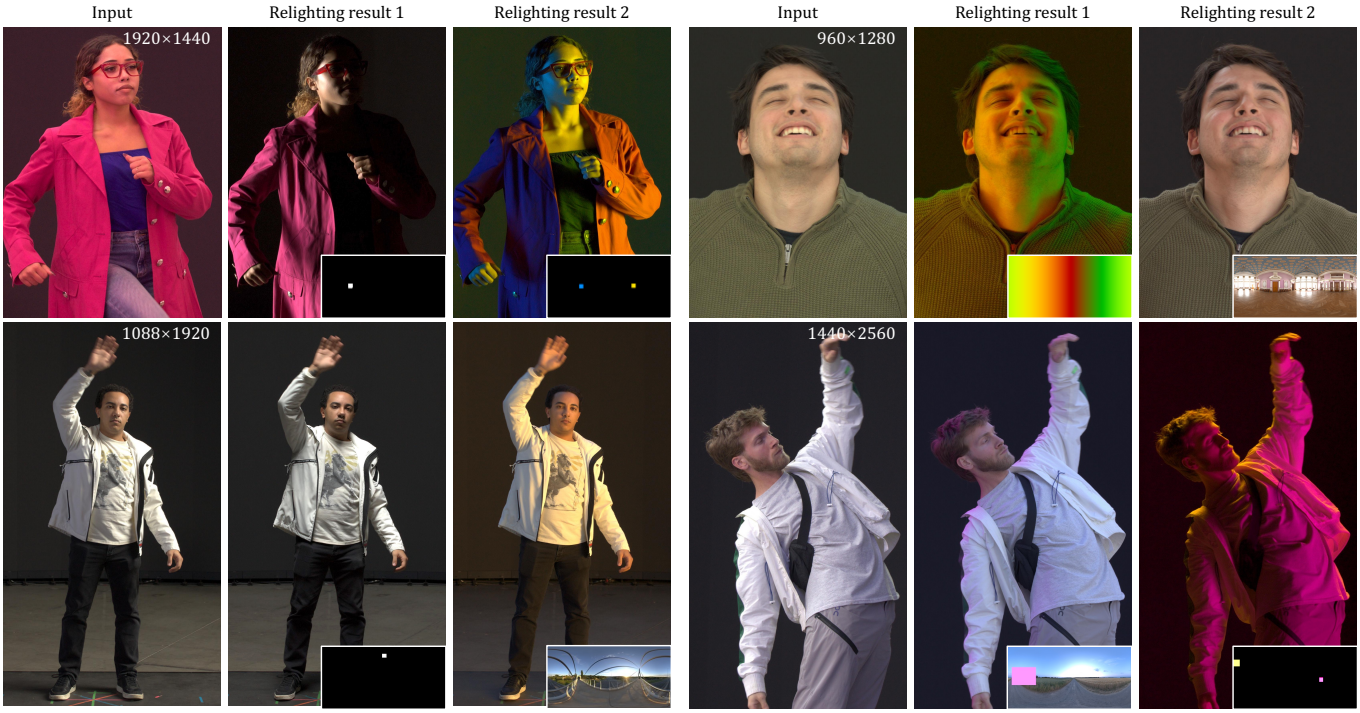

Fig. 5. **Relighting results.** We show input videos and relighting results under novel performance and lighting conditions. Each result is a single from from a relit video. Our method achieves photorealistic relighting under directional lighting, multiple point lights, gradient illumination, image-based lighting, and manually edited HDRI maps for various input aspect ratios and resolutions. The resolution is shown on the top right and the lighting condition is visualized on the bottom right as a lat-long map.

into a list of tokens. Then a Diffusion Transformer (DiT) iteratively de-noises the noisy latent to a clean one $z^{(0)}$, following the reversed flow matching process [Lipman et al. 2023]. The DiT is a cascade of transformer layers to which text tokens are cross-attentioned to achieve text conditioned generation. Finally the decoder of a variational autoencoder (VAE) is used to obtain an RGB video.

Spatial Conditioning. To achieve spatial conditioning of the input video V, inspired by several diffusion-based pixel-to-pixel translation frameworks [Brooks et al. 2023; He et al. 2024; Ke et al. 2023; Zhang et al. 2025], we concatenate the conditioning latent z_V with the noisy latent $z^{(t)}$ along the channel dimension, and duplicate the input channel number in the first convolution layer of the patchifier. We find this simple approach effectively preserves the details, while introducing minimal overhead to both training and inference.

One Light as a Token. Conditioning the model on the lighting L during diffusion introduces new challenges. To avoid generating hundreds of OLAT videos to be composed [He et al. 2024], we directly condition the model on a list of light sources to relight with a single inference. Existing works have explored the use of pixel-aligned shading maps rendered from an HDRI and estimated geometry [Pandey et al. 2021; Zeng et al. 2024b], but these are sensitive to errors and flickering, common with video data. Some encode HDRIs into lighting embeddings using either dedicated lighting encoder [Mei et al. 2025; Zhang et al. 2025] or pretrained image VAE [Chaturvedi et al. 2025; He et al. 2025; Liang et al. 2025]. These methods effectively compress the lighting along the channel dimensions, leading to information loss and inaccurate controls.

A good lighting condition should preserve the full dynamic range and directionality. Therefore, we introduce **OLAToken**, which stands for One-Light-as-a-Token. We treat each light source as one light embedding containing its intensity and direction in their own channels. Each light token is passed through a small MLP to match the channel dimension of the model, and is fed into the DiT through cross-attention. This simple design introduces two strong inductive biases that approximate the physical structure of light transport: (1) *Permutation invariance.* Similar to a list of lights, the order of the tokens does not affect the final results in cross-attention. And (2) *Compositionality.* Although cross-attention is not inherently linear as lighting should be, it behaves as a context-dependent weighted summation, which allows the model to approximate the physical superposition of illumination. This approach is more general than the lighting encoder used in Mei et al. [2025]. It allows flexible combinations of light sources from different directions, without the constraints of fixed resolution and orientation.

Formally,a target lighting condition is represented as a list of light sources $L = \{(\mathbf{l}_i, \mathbf{d}_i)\}$, where $\mathbf{l}_i \in \mathbb{R}^3$ is the linear intensity of the RGB channels, and $\mathbf{d}_i$ is the light direction in camera space. For efficiency, we downsample the lights by evenly sampling K directions over the hemisphere, and then aggregate every light source in L to the

closest points to form a new light source $(\mathbf{I}_j, \mathbf{D}_j)$:

$$\mathbf{I}_j = \sum_{i \in \mathbb{J}} \mathbf{l}_i, \text{ and } \mathbf{D}_j = normalize(\frac{\sum_{i \in \mathbb{J}} \mathbf{d}_i \|\mathbf{l}_i\|_2}{\sum_{i \in \mathbb{J}} \|\mathbf{l}_i\|_2}), \tag{1}$$

where $\mathbb{J}$ is a set of indices for all the closest light sources close to j-th point. Then an OLAToken $\mathbf{T}_j$ is computed as:

$$\mathbf{T}_j = \mathcal{D}(\mathbf{D}_j) \oplus \mathcal{I}(\mathbf{I}_j). \tag{2}$$

Here, $\mathcal{D}$ is a directional encoding represented by Fourier Features similar to NeRF [Mildenhall et al. 2020]. $\mathcal{I}$ is a color encoding function. We use a series of different gamma functions (log spacing from 1/3 to 3) to enhance the expressiveness of the color intensity. $\oplus$ indicates channel concatenation.

Dynamic lighting attention. Importantly, we want the ability to condition the video relighting model on dynamic lighting where the lighting condition is specified for each frame, such as an environment rotating around the subject, or the subject walking into sunlight. We show in the experiments that when adding time embeddings to the OLATokens, conditions of specific frames tend to leak into other frames. To address this, we use block diagonal attention masks during cross-attention such that the tokens from a specific frame only attend to the light tokens corresponding to that frame. In practice, to avoid computation where the mask is empty, we flatten the temporal dimension with the batch dimension to achieve frame-wise cross-attention.

Loss and training. With the input video condition and light condition, we convert the video generation backbone into a video relighting model. At training time, we freeze the VAE encoder and decoder, and finetune only the video diffusion DiT. We use a standard flow-matching training loss:

$$\mathcal{L} = \|\hat{\epsilon}(\mathbf{z}^{(t)}; \mathcal{D}(V), L, t) - \epsilon\|_2, \tag{3}$$

where $\mathbf{z}^{(t)}$ is the noisy latent at time step t and epsilon is the added noise. $\hat{\epsilon}$ indicates the noise predicted by the DiT model, conditioned on the input latent $\mathcal{D}(V)$, lighting L and time step t.

To support arbitrary aspect ratios and frame length during training, we randomly select an arbitrary aspect ratio between $1:2$ and $2:1$, and a sequence length from 9 to 37. We then determine the spatial resolution by limiting the number of tokens to 5k and randomly crop the input video. We achieve frame rate augmentation by retiming using the nearest neighbor. By applying these data augmentations, we obtain a video relighting model that is robust to different aspect ratios, framing, resolutions, and frame rates.

We implement our model using DiffSynth [modelscope Community 2026] based on the WAN2.2 5B model [Wan et al. 2025]. We train our relighting model on 8 NVIDIA A100 GPUs with 80GB memory each, for 100K iterations with a batch size of 8 taking around 3 days. We use the Adam Optimizer with a learning rate of 4e-5.

Long video inference. While our model is trained on videos that have at most 37 frames, we empirically find the model is robust to videos with up to 100 frames. To infer on videos of unconstrained length, we apply the relighting model on overlapping windows and utilize MultiDiffusion [Bar-Tal et al. 2023] to combine the predictions into temporally consistent video results.

Table 1. **Quantitative comparisons.** Ours achieve best relighting results while being the most temporally consistent. We also measure the inference time per frame on one A100 machine and ours achieve a reasonable speed.

name	PSNR↑	SSIM↑	LPIPS↓	T-PSNR↑	Inf. time↓
Ours	**23.43**	**0.9441**	**0.04208**	**27.90**	1.8s
DiffRelight+	21.75	0.9425	0.08095	26.34	4.2min
Switchlight3	16.88	0.8814	0.11945	26.84	1.7s
LuxPostFacto	12.84	0.8407	0.13341	27.18	8.7s
Allfreq	11.85	0.8139	0.20248	21.95	**0.72s**

4 Experiments

We conduct several experiments to evaluate the performance of BodyReLux. We capture static OLAT and bi-packed video relighting data from four performers and train a single subject-specific relighting model on all four subjects. We also hold out several bi-packed lighting sequences from the training data to produce input videos with the same performance in a ground-truth alternate lighting condition, which we use to evaluate the relighting results both qualitatively and quantitatively. During training, we hold out one camera view to ensure that the evaluation sequences include novel viewpoints.

For qualitative evaluation, we additionally capture several dynamic sequences for each subject in in-the-wild settings using a Sony Alpha α7S at 50 fps with auto exposure enabled. We record performances in a variety of natural environments and ask each subject to perform freely as in a real-world scenario.

4.1 Main Results

We demonstrate the effectiveness of our method on several test sequences. Example frames are shown in Fig. 1 and Fig. 5. Thanks to a unified light representation using OLAToken, BodyReLux achieves photorealistic relighting under a wide range of lighting conditions, including single and multiple point lights, gradient illumination, HDRI, and manually edited HDRI. The method is also robust to input videos with varying aspect ratios and resolutions.

Qualitative video frames are shown in Fig. 7; we refer the readers to the supplementary video for dynamic relighting examples. Using our long-video inference strategy, our method can process videos exceeding 200 frames. Because of the dynamic lighting conditioning, BodyReLux can relight sequences under smoothly varying lighting conditions that are not observed during training. Notably, when the input video itself contains changing illumination, our method still produces temporally consistent relighting results.

Moreover, although trained only on stage-captured data, our model generalizes well to in-the-wild captures, as shown in Fig. 1 and Fig. 6. We evaluate several challenging scenarios that are unseen during training, including diverse performances, unusual framings, and camera motions. Our method achieves photorealistic quality across almost all the examples. Although our model is trained only to relight the subjects, we find that as a by-product it often relights the backgrounds somewhat plausibly. We attribute this effect to the priors of the base model, which is trained to generate plausible videos.

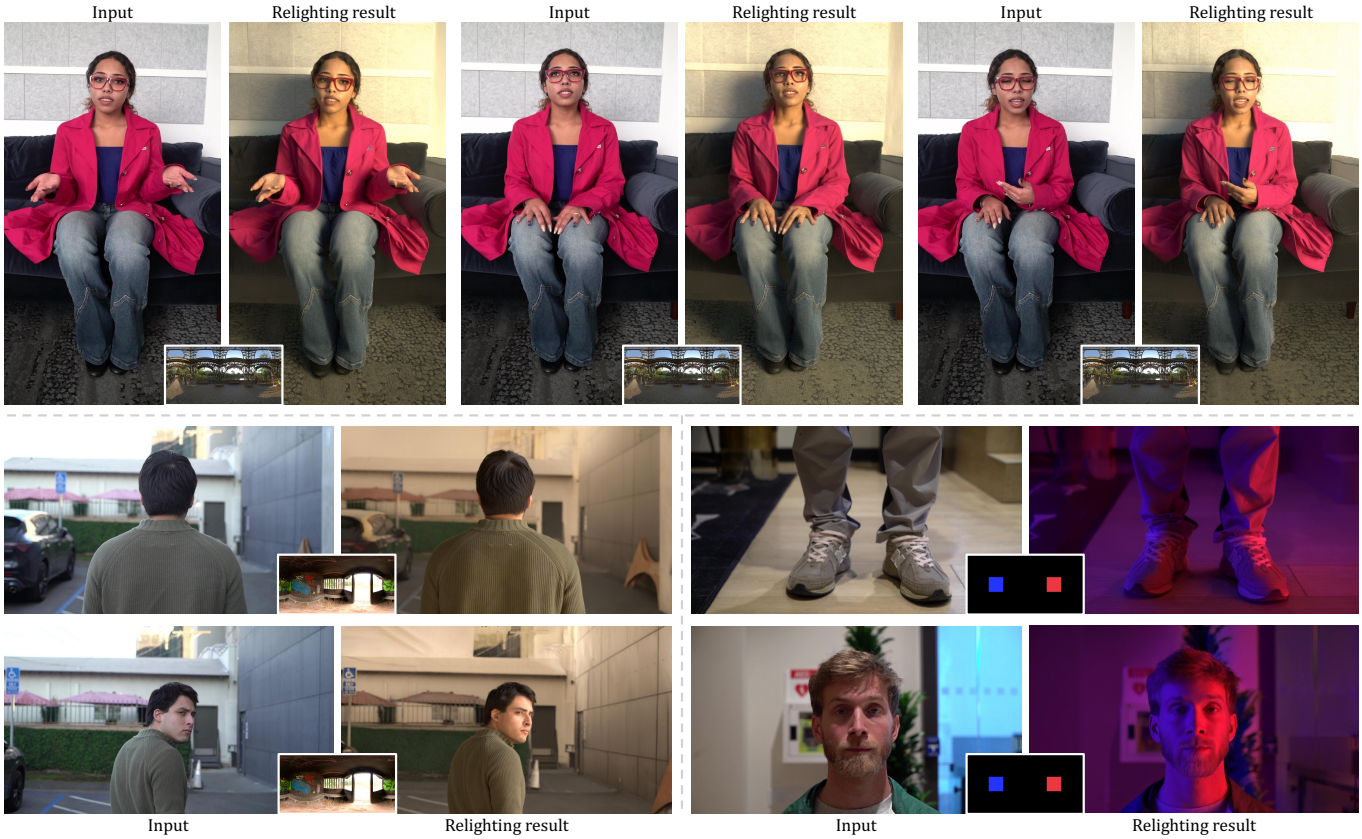

Fig. 6. **In-the-wild relighting results.** We show video relighting results for in-the-wild captures.

4.2 Comparisons

We compare our method with several baselines. Closest to ours is DifFRelight [He et al. 2024], an image diffusion-based, subject-specific relighting model for facial performances. We re-implement DifFRelight using our network backbone and retrain it on our data to align the experiment settings; this baseline is denoted as DifFRelight+. We also compare against generalized human-centric relighting models: Switchlight3 [Kim et al. 2024], LuxPostFacto [Mei et al. 2025], and Allfreq [Daichi Tajima 2025]. Since not all methods support dynamic lighting, we evaluate all approaches using clips with constant lighting conditions. We report PSNR, SSIM, and LPIPS [Zhang et al. 2018] between relighting results and ground truth in Tab. 1. As our focus is human-centric relighting, we use the Sapiens model [Khirodkar et al. 2024] to extract human segmentation masks, and compute all metrics only within the masked regions. We measure temporal consistency using T-PSNR, which is defined as the PSNR between the current frame and its optical-flow–warped neighboring frame. We also report per-frame inference time at 1080p resolution. The quantitative results show that our method achieves the highest relighting quality among all baselines, along with the best temporal consistency, while remaining reasonably efficient. DifFRelight+ performs slightly worse due to the lack of video-based training and is significantly slower at inference, as it requires running the model hundreds of times to generate all OLATs followed by post-compositing. Overall, subject-specific methods substantially outperform generalized models, highlighting the importance of subject-specific training for high-quality relighting.

We further present qualitative comparisons in Fig. 8. Our method produces photorealistic relighting results that are closest to the ground truth. In comparison, DifFRelight+ achieves reasonable quality but often introduces artifacts in ambiguous regions such as occlusions and fast-moving parts. This limitation arises because DifFRelight+ operates on single frames. SwitchLight3 produces plausible relighting but generates overly glossy and unnatural appearances, particularly on skin regions. This is likely caused by the physically based rendering priors it is using, which do not capture the complex reflectance of human skin. Other generalized relighting methods, such as LuxPostFacto, achieve plausible temporal consistency but exhibit pronounced artifacts and significantly alter subject identity due to the lack of subject-specific training data.

4.3 Ablation Studies

We conduct ablation studies to evaluate key design choices in our method. Experiments are performed on all evaluation sequences from the four subjects, which include novel performances, camera viewpoints, and lighting conditions. The test set consists of 50% sequences with constant lighting and 50% with dynamic lighting. A summary of the quantitative results is provided in Tab. 2. Additional experiments are provided in the supplementary material.

Fig. 7. **Video Relighting results.** We show video relighting results of BodyReLux under static lighting and dynamic lighting conditions. Note that our method also works for input video with dynamic lighting conditions, while still being able to produce consistent relighting results.

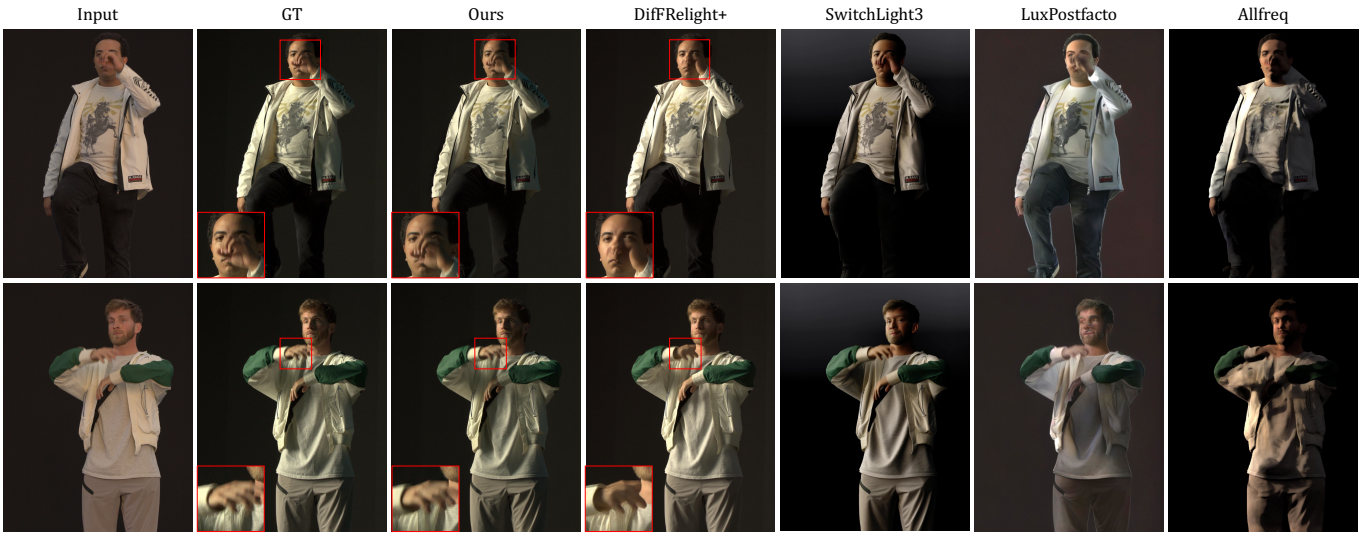

Fig. 8. **Qualitative comparisons.** We compare ground truth and predicted relighting results. Ours achieves the most photorealistic relighting results. DifFRelight+ achieves reasonable results, but tends to produce artifacts in fast-moving regions, as highlighted in red boxes. Other generalized relighting models produce larger errors compared to the ground truth.

Table 2. **Quantitative ablation results.** Showing the effectiveness of static OLAT, bi-pack video data, as well as the lighting conditioning method. The best and second-best results are highlighted in **bold** and underline, respectively.

	PSNR↑	SSIM↑	LPIPS↓
Full	<u>22.62</u>	<u>0.9369</u>	**0.04895**
w/o video data	21.50	0.9338	0.07463
w/o OLAT data	22.03	0.9275	0.05685
w/o OLAToken	20.39	0.9312	0.08011
w/o dyn. cond.	22.31	0.9368	0.04913
w/o alignment	21.45	0.9260	0.05692
w/o pretrained weight	17.69	0.9055	0.09099
w/ WAN2.1 1.3B	**22.88**	**0.9388**	<u>0.05011</u>

Fig. 9. **Ablation of different types of training data.** Without bi-packed video data, the model fails to produce plausible results when trained only on pseudo motion derived from static OLAT captures. Without OLAT data, the model cannot accurately relight scenes under extreme lighting conditions.

Ablation on training data. We evaluate the necessity of using bi-packed video captures and static OLAT captures (w/o video data and w/o OLAT data). Visual results are shown in Fig. 9. Both data types are indeed necessary to achieve photorealistic, robust, and accurate relighting. Without bi-packed video data (w/o video data), the model is trained purely on pseudo-videos generated from transformed static images and fails to generalize to realistic human dynamics. On the other hand, static OLATs provide a broader distribution of lighting conditions. Without them (w/o OLAT data), the model struggles to produce plausible relighting results, particularly under extreme lighting conditions such as highly directional illumination.

Ablation of lighting conditioning. We evaluate the effectiveness of OLAToken conditioning by replacing it with an HDRI-based conditioning scheme (w/o OLAToken). Following prior work [Mei et al. 2025; Zhang et al. 2025], we implement the HDRI-based conditioning by encoding a 2D HDRI into a 1D embedding using a shallow convolutional neural network (CNN), which is then fed to the DiT backbone via cross-attention. To support dynamic lighting, we encode HDRI sequences frame by frame and apply dynamic lighting attention. As qualitative results show in Fig. 10, without OLAToken, the relighting results are less accurate than the ground truth, which is also reflected in the quantitative results reported in Tab. 2.

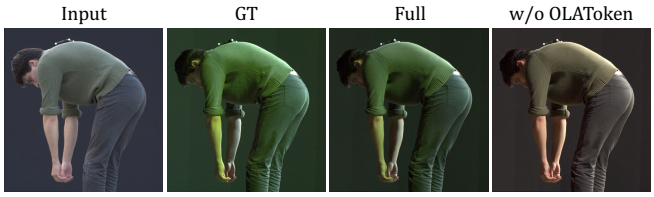

Fig. 10. **OLAToken conditioning.** Our OLAToken conditioning method helps improve the lighting accuracy.

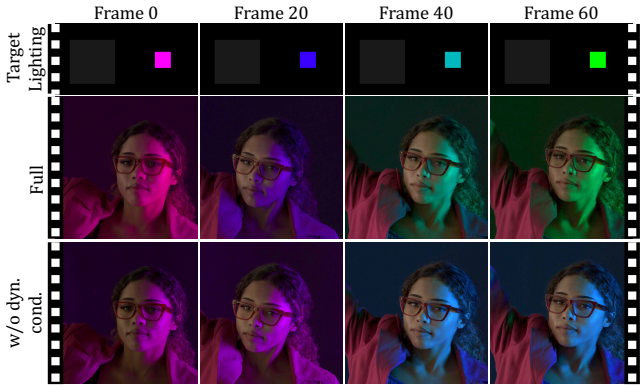

Fig. 11. **Dynamic lighting conditioning.** Without using dynamic lighting conditioning, lighting conditions tend to leak across frames.

We further evaluate our dynamic lighting conditioning by replacing it with a straightforward temporal conditioning scheme (w/o dyn. cond.). Specifically, we remove the temporal attention mask, and add RoPE [Su et al. 2024] time embedding to each lighting token. We show an example of dynamic lighting with smoothly changing light colors in Fig. 11. The results show that without dynamic lighting attention, lighting conditions at different frames leak into each others. This indicates that the model fails to correctly reason about the correspondence between the video frames and the lighting sequence relying solely on time embedding.

Ablation of frame alignment. In bi-pack video capture, frames under different lighting conditions are recorded with a slight temporal offset. We evaluate the importance of aligning input and target frames when generating the training data for video relighting by removing this step (w/o alignment). The results are shown in Tab. 2 and Fig. 12. Surprisingly, even without alignment, the model produces visually plausible relighting results. However, the relit outputs exhibit temporal misalignment with respect to the input video. In practical production settings, such misalignment can complicate

Fig. 12. **Effectiveness of alignment.** Without alignment of video relighting pairs, the model achieves similar quality but suffers from misalignment from the input.

downstream compositing, making frame alignment a necessary component of our pipeline.

Ablation of pretraining prior. We evaluate the effectiveness of using pretrained DiT weights by comparing our approach against a model trained from random initialization. The numerical results are shown in Tab. 2, and one example is shown in Fig. 13. Without prior knowledge embedded in the pretrained weights, the model produces less accurate relighting results characterized by blurry details and diffused specular highlights.

Ablation of other WAN variants. We further evaluate our method using a different diffusion backbone, WAN2.1 1.3B, as shown in Tab. 2. This variant achieves comparable performance, yielding slightly higher PSNR and SSIM, and lower LPIPS. This suggests that our method is largely backbone-agnostic. However, WAN2.1 uses a VAE with a smaller compression ratio, making it approximately 5× slower inference at the same resolution despite having fewer parameters. We do not evaluate the 14B variant due to its prohibitively expensive train cost and substantial memory requirements. Overall, WAN2.2 5B provides a favorable trade-off between quality, training cost and inference efficiency.

4.4 Additional Analysis

Linearity of relighting. Since the lighting transport is linear, we can evaluate the accuracy of lighting control by testing the linearity of the output image $\mathcal{R}(L)$ with respect to the input lighting condition L. For brevity, we denote the relighting model as $\mathcal{R}(L)$ and omit the input image. We assess linearity through two tasks, light combination and exposure scaling. If the relighting function $\mathcal{R}$ is linear with respect to L, it should satisfy $\mathcal{R}(L_A + L_B) = \mathcal{R}(L_A) + \mathcal{R}(L_B)$, and $\mathcal{R}(\alpha L) = \alpha\mathcal{R}(L)$, where α is a scalar. Accordingly, we evaluate the linearity by comparing $\mathcal{R}(L_A + L_B)$ with $\mathcal{R}(L_A) + \mathcal{R}(L_B)$ and $\mathcal{R}(\alpha L)$ with $\alpha\mathcal{R}(L)$. The results are shown in Fig. 14. As the model is trained in the sRGB color space, all outputs are converted to linear space for evaluation and then back to sRGB for visualization. The

Fig. 13. **Effectiveness of pretrained weight.** Without loading the pretrained weight, the relighting result tends to be blurry with fewer specular reflections and lower relighting accuracy.

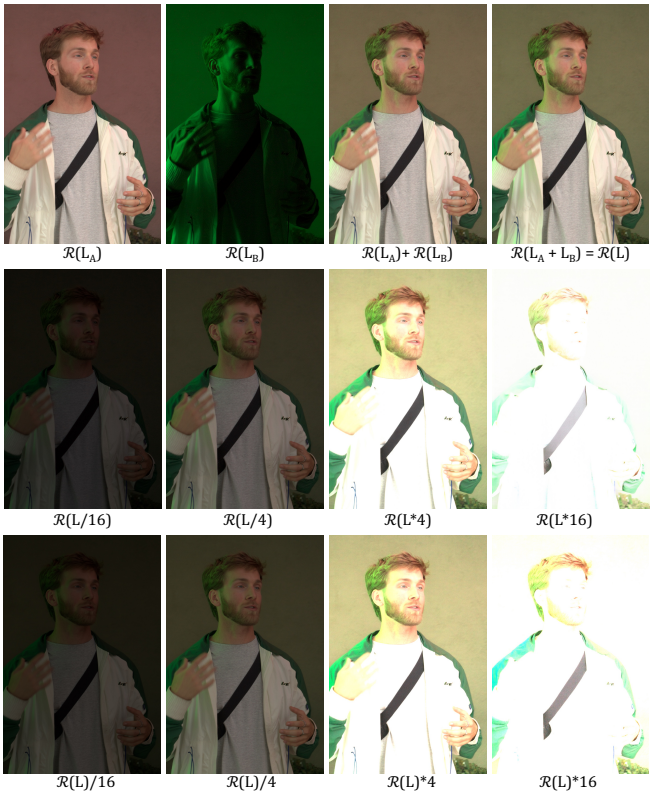

Fig. 14. **Linearity test of our relighting model.** Our relighting model produces visually linear results with respect to the input lighting conditions. Here $\mathcal{R}$ indicates our video relighting model.

results indicate that our method exhibits near-linear behavior under both lighting addition and exposure scaling.

Repeatability of results. We assess the repeatability of our method by conducting three independent training experiments. Specifically, we randomly pick two mutually exclusive subsets from all of the captures of a subject, A and B, each containing 6 static OLAT poses and 4 dynamic bi-pack sequences. We train separate models on each subset, as well as an additional model using the full set of captures.

Table 3. **Repeatability evaluation.** Performance remains consistent across models trained on different subsets, demonstrating strong repeatability. Training on fewer captures results in only a minor degradation in quality.

Training data	PSNR↑	SSIM↑	LPIPS↓
All captures	**21.91**	**0.9462**	**0.05070**
Subset A	21.47	0.9436	0.05418
Subset B	21.52	0.9431	0.05499

Fig. 15. **Repeatability across capture subsets.** Models trained on all captures or disjoint subsets produce consistent relighting results. Error maps (insets) show strong pixel alignment of the outputs.

Quantitative results are reported in Tab. 3, and qualitative comparisons are shown in Fig. 15. The results demonstrate consistent performance across different capture subsets, indicating strong repeatability. Using fewer captures leads to a slight degradation in quality, yet the relighting capability remains stable.

In Fig. 15, we additionally visualize error maps. The outputs are almost perfectly pixel-aligned with the input. Minor misalignment (typically on the order of 1-3 pixels around some edges) can occasionally occur, likely due to the spatial compression of the VAE.

Multiple subjects inference. Although our training data consists exclusively of single-subject sequences, we test our relighting model on a two-subject scene, as shown in Fig. 16. We compare against two individual inference results on crops of each subject, and find the results to be visually consistent. This suggests that our model generalizes naturally to multi-subject scenarios. However, more complex interactions—such as severe occlusions and inter-subject shadow casting—remain challenging. We leave these cases for future work.

5 Limitations, Future Work and Conclusion

We presented a new framework for photo-real, temporally consistent human performance relighting. We propose a new video capture process to train a video diffusion-based relighting model with a novel lighting conditioning mechanism achieving accurate relighting. We achieve the best relighting quality compared to baselines, and our ablation studies show the effectiveness of our design choices.

Our method still has a few limitations. First, our LED Sphere only lights from the upper hemisphere and lights are a few meters from the subjects. Therefore, our model is not able to relight with light from the lower hemisphere, or with near-field lighting [Jones et al. 2006]. Second, our method is subject-specific and not designed to relight unseen identities. We show one example in Fig. 17, where

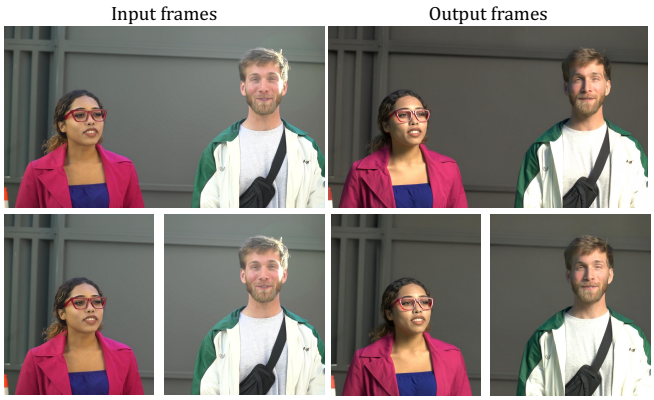

Fig. 16. **Relighting results on multi-subject shot.** We compare relighting results on a two-subject video (first row), with relighting them individually by cropping (second row). The model is able to generalize to multi-subject shot even if it's only trained on single-subject data.

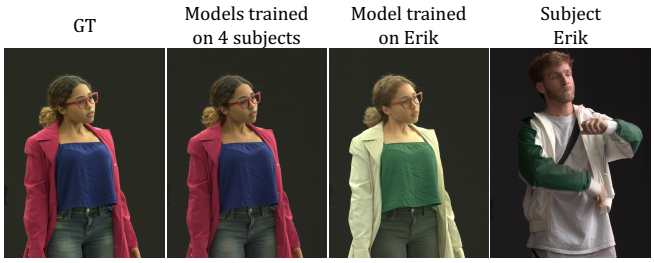

Fig. 17. **Relighting results from out-of-distribution identities.** When inferred on unseen identities, the model tends to bake in subject-specific features such as cloth colors, hair color, and facial features.

we train our model for one subject, and infer on a different subject. Interestingly, it still produces reasonable relighting, but it transfers features from the training subject to the new subject, including the color of the jacket and hair, and facial features like the beard. This indicates that the model may learn to relight based on semantic features from the input videos, rather than purely overfitting to relight one subject.

For future work, we hope to scale the data capture to more identities, performances, props, and scenes to train a generalized subject and scene relighting model.

Acknowledgments

We would like to thank Jeffrey Shapiro for his ongoing support; Alek Nieberlein and Samuel Price for stage operation; Jennifer Lao for performer coordination; Lauren Wilson for stage scheduling; Emmett Steven for infrastructure support; Scott McDonald for Red camera control; the Software Department for stage software development, and our performers: Pablo Salamanca, David George, Erik Patten and Anica Rose.

References

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. 2023. MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. *arXiv preprint arXiv:2302.08113* (2023).

Shrisha Bharadwaj, Haiwen Feng, Giorgio Becherini, Victoria Fernandez Abrevaya, and Michael J. Black. 2025. GenLit: Reformulating Single Image Relighting as Video Generation. In *SIGGRAPH Asia Conference Papers '25 (SIGGRAPH Asia Conference Papers '25)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3757377.3763970

Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn Mcphail, Ravi Ramamoorthi, Yaser Sheikh, and Jason Saragih. 2021. Deep relightable appearance models for animatable faces. *ACM Trans. Graph.* 40, 4, Article 89 (July 2021), 15 pages. doi:10.1145/3450626.3459829

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18392–18402.

Charles-Félix Chabert, Per Einarsson, Andrew Jones, Bruce Lamond, Wan-Chun Ma, Sebastian Sylwan, Tim Hawkins, and Paul Debevec. 2006. Relighting human locomotion with flowed reflectance fields. In *ACM SIGGRAPH 2006 Sketches* (Boston, Massachusetts) *(SIGGRAPH '06)*. Association for Computing Machinery, New York, NY, USA, 76–es. doi:10.1145/1179849.1179944

Clément Chadebec, Onur Tasar, Sanjeev Sreetharan, and Benjamin Aubin. 2025. LBM: Latent Bridge Matching for Fast Image-to-Image Translation. *arXiv preprint arXiv:2503.07535* (2025).

Sumit Chaturvedi, Mengwei Ren, Yannick Hold-Geoffroy, Jingyuan Liu, Julie Dorsey, and Zhixin Shu. 2025. SynthLight: Portrait Relighting with Diffusion Model by Learning to Re-render Synthetic Faces. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2025).

HONGZE CHEN, Zehong Lin, and Jun Zhang. 2025. GI-GS: Global Illumination Decomposition on Gaussian Splatting for Inverse Rendering. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=hJIEtJlvhL

Yuki Endo Daichi Tajima, Yoshihiro Kanamori. 2025. All-frequency Full-body Human Image Relighting. *Computer Graphics Forum (Proc. of Eurographics 2025)* 44, 2 (2025), e70007.

Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. 2000. Acquiring the reflectance field of a human face. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '00)*. ACM Press/Addison-Wesley Publishing Co., USA, 145–156. doi:10.1145/344779.344855

Paul Debevec, Andreas Wenger, Chris Tchou, Andrew Gardner, Jamie Waese, and Tim Hawkins. 2002. A lighting reproduction approach to live-action compositing. *ACM Trans. Graph.* 21, 3 (July 2002), 547–556. doi:10.1145/566654.566614

Ye Fang, Zeyi Sun, Shangzhan Zhang, Tong Wu, Yinghao Xu, Pan Zhang, Jiaqi Wang, Gordon Wetzstein, and Dahua Lin. 2025. RelightVid: Temporal-Consistent Diffusion Model for Video Relighting. arXiv:2501.16330 [cs.CV] https://arxiv.org/abs/2501.16330

Graham Fyffe. 2009. Cosine lobe based relighting from gradient illumination photographs. In *SIGGRAPH '09: Posters* (New Orleans, Louisiana) *(SIGGRAPH '09)*. Association for Computing Machinery, New York, NY, USA, Article 80, 1 pages. doi:10.1145/1599301.1599381

David Griffiths, Tobias Ritschel, and Julien Philip. 2022. OutCast: Single Image Relighting with Cast Shadows. *Computer Graphics Forum* 43 (2022).

Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, Danhang Tang, Anastasia Tkach, Adarsh Kowdle, Emily Cooper, Mingsong Dou, Sean Fanello, Graham Fyffe, Christoph Rhemann, Jonathan Taylor, Paul Debevec, and Shahram Izadi. 2019. The relightables: volumetric performance capture of humans with realistic relighting. *ACM Trans. Graph.* 38, 6, Article 217 (Nov. 2019), 19 pages. doi:10.1145/3355089.3356571

Tim Hawkins, Andreas Wenger, Chris Tchou, Andrew Gardner, Fredrik Göransson, and Paul Debevec. 2004. Animatable facial reflectance fields. In *Proceedings of the Fifteenth Eurographics Conference on Rendering Techniques* (Norrköping, Sweden) *(EGSR'04)*. Eurographics Association, Goslar, DEU, 309–319.

Kai He, Ruofan Liang, Jacob Munkberg, Jon Hasselgren, Nandita Vijaykumar, Alexander Keller, Sanja Fidler, Igor Gilitschenski, Zan Gojcic, and Zian Wang. 2025. UniRelight: Learning Joint Decomposition and Synthesis for Video Relighting. *arXiv preprint arXiv:2506.15673* (2025).

Mingming He, Pascal Clausen, Ahmet Levent Taşel, Li Ma, Oliver Pilarski, Wenqi Xian, Laszlo Rikker, Xueming Yu, Ryan Burgert, Ning Yu, and Paul Debevec. 2024. DifFRelight: Diffusion-Based Facial Performance Relighting. In *SIGGRAPH Asia 2024 Conference Papers (SA '24)*. Association for Computing Machinery, New York, NY, USA, Article 11, 12 pages. doi:10.1145/3680528.3687644

Andrew Hou, Ze Zhang, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. 2021. Towards High Fidelity Face Relighting with Realistic Shadows. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Haian Jin, Yuan Li, Fujun Luan, Yuanbo Xiangli, Sai Bi, Kai Zhang, Zexiang Xu, Jin Sun, and Noah Snavely. 2024. Neural Gaffer: Relighting Any Object via Diffusion. In *Advances in Neural Information Processing Systems*.

Haian Jin, Isabella Liu, Peijia Xu, Xiaoshuai Zhang, Songfang Han, Sai Bi, Xiaowei Zhou, Zexiang Xu, and Hao Su. 2023. TensoIR: Tensorial Inverse Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Andrew Jones, Andrew Gardner, Mark Bolas, Ian Mcdowall, and Paul Debevec. 2006. Simulating Spatially Varying Lighting on a Live Performance. In *CVMP*. 127 – 133. doi:10.1049/cp:20061934

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. 2023. Repurposing diffusion-based image generators for monocular depth estimation. *arXiv preprint arXiv:2312.02145* (2023).

Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. 2024. Sapiens: Foundation for Human Vision Models. *arXiv preprint arXiv:2408.12569* (2024).

Hoon Kim, Minje Jang, Wonjun Yoon, Jisoo Lee, Donghyun Na, and Sanghyun Woo. 2024. SwitchLight: Co-design of Physics-driven Architecture and Pre-training Framework for Human Portrait Relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 25096–25106.

Peter Kocsis, Julien Philip, Kalyan Sunkavalli, Matthias Nießner, and Yannick Hold-Geoffroy. 2024. LightIt: Illumination Modeling and Control for Diffusion Models. In *CVPR*.

Edwin Herbert Land and John J. McCann. 1971. Lightness and retinex theory. *Journal of the Optical Society of America* 61 1 (1971), 1–11. https://api.semanticscholar.org/CorpusID:14430259

Chloe LeGendre, Xueming Yu, Dai Liu, Jay Busch, Andrew Jones, Sumanta Pattanaik, and Paul Debevec. 2016. Practical multispectral lighting reproduction. *ACM Trans. Graph.* 35, 4, Article 32 (July 2016), 11 pages. doi:10.1145/2897824.2925934

Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. 2020. 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 (CVPR)*.

Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Zhi-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, and Zian Wang. 2025. DiffusionRenderer: Neural Inverse and Forward Rendering with Video Diffusion Models. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhihao Liang, Qi Zhang, Ying Feng, Ying Shan, and Kui Jia. 2024. GS-IR: 3D Gaussian Splatting for Inverse Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 21644–21653.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. 2023. Flow Matching for Generative Modeling. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=PqvMRDCJT9t

Nadav Magar, Amir Hertz, Eric Tabellion, Yael Pritch, Alex Rav-Acha, Ariel Shamir, and Yedid Hoshen. 2025. LightLab: Controlling Light Sources in Images with Diffusion Models. (2025). arXiv:arXiv:2505.09608 doi:10.1145/3721238.3730696

Yiqun Mei, Mingming He, Li Ma, Julien Philip, Wenqi Xian, David M George, Xueming Yu, Gabriel Dedic, Ahmet Levent Taşel, Ning Yu, Vishal M. Patel, and Paul Debevec. 2025. Lux Post Facto: Learning Portrait Performance Relighting with Conditional Video Diffusion and a Hybrid Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5510–5522.

Yiqun Mei, Yu Zeng, He Zhang, Zhixin Shu, Xuaner Zhang, Sai Bi, Jianming Zhang, HyunJoon Jung, and Vishal M. Patel. 2024. Holo-Relighting: Controllable Volumetric Portrait Relighting from a Single Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4263–4273.

Yiqun Mei, He Zhang, Xuaner Zhang, Jianming Zhang, Zhixin Shu, Yilin Wang, Zijun Wei, Shi Yan, HyunJoon Jung, and Vishal M. Patel. 2023. LightPainter: Interactive Portrait Relighting With Freehand Scribble. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 195–205.

Abhimitra Meka, Christian Häne, Rohit Pandey, Michael Zollhöfer, Sean Fanello, Graham Fyffe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, Peter Denny, Sofien Bouaziz, Peter Lincoln, Matt Whalen, Geoff Harvey, Jonathan Taylor, Shahram Izadi, Andrea Tagliasacchi, Paul Debevec, Christian Theobalt, Julien Valentin, and Christoph Rhemann. 2019. Deep reflectance fields: high-quality facial reflectance field inference from color gradient illumination. *ACM Trans. Graph.* 38, 4, Article 77 (July 2019), 12 pages. doi:10.1145/3306346.3323027

Abhimitra Meka, Rohit Pandey, Christian Häne, Sergio Orts-Escolano, Peter Barnum, Philip David-Son, Daniel Erickson, Yinda Zhang, Jonathan Taylor, Sofien Bouaziz, Chloe Legendre, Wan-Chun Ma, Ryan Overbeck, Thabo Beeler, Paul Debevec, Shahram Izadi, Christian Theobalt, Christoph Rhemann, and Sean Fanello. 2020. Deep relightable textures: volumetric performance capture with neural rendering. *ACM Trans. Graph.* 39, 6, Article 259 (Nov. 2020), 21 pages. doi:10.1145/3414685.3417814

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.

modelscope Community. 2026. DiffSynth-Studio: Enjoy the magic of Diffusion models! https://github.com/modelscope/DiffSynth-Studio. GitHub repository, accessed January 18, 2026.

Rohit Pandey, Sergio Orts-Escolano, Chloe LeGendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. 2021. Total Relighting: Learning to Relight Portraits for Background Replacement. *ACM Transactions on Graphics (Proceedings SIGGRAPH)* 40, 4. doi:10.1145/3450626.3459872

Pieter Peers, Naoki Tamura, Wojciech Matusik, and Paul Debevec. 2007. Post-production facial performance relighting using reflectance transfer. *ACM Trans. Graph.* 26, 3 (July 2007), 52–es. doi:10.1145/1276377.1276442

Julien Philip, Michaël Gharbi, Tinghui Zhou, Alexei A. Efros, and George Drettakis. 2019. Multi-view relighting using a geometry-aware network. *ACM Trans. Graph.* 38, 4, Article 78 (July 2019), 14 pages. doi:10.1145/3306346.3323013

Julien Philip, Sébastien Morgenthaler, Michaël Gharbi, and George Drettakis. 2021. Free-viewpoint Indoor Neural Relighting from Multi-view Stereo. *ACM Transactions on Graphics* (2021). http://www-sop.inria.fr/reves/Basilic/2021/PMGD21

Yohan Poirier-Ginter, Alban Gauthier, Julien Philip, Jean-François Lalonde, and George Drettakis. 2024. A Diffusion Approach to Radiance Field Relighting using Multi-Illumination Synthesis. *Computer Graphics Forum* (2024). doi:10.1111/cgf.15147

Poly Haven. 2023. Poly Haven HDRI. https://polyhaven.com/hdris Accessed: 2026-01-04.

Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. 2022a. FILM: Frame Interpolation for Large Motion. In *European Conference on Computer Vision (ECCV)*.

Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. 2022b. Tensorflow 2 Implementation of "FILM: Frame Interpolation for Large Motion". https://github.com/google-research/frame-interpolation

Mengwei Ren, Wei Xiong, Jae Shin Yoon, Zhixin Shu, Jianming Zhang, HyunJoon Jung, Guido Gerig, and He Zhang. 2024. Relightful Harmonization: Lighting-aware Portrait Background Replacement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6452–6462.

Pratul P. Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T. Barron. 2021. NeRV: Neural Reflectance and Visibility Fields for Relighting and View Synthesis. In *CVPR*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomput.* 568, C (Feb. 2024), 12 pages. doi:10.1016/j.neucom.2023.127063

Hanxiao Sun, Yupeng Gao, Jin Xie, Jian Yang, and Beibei Wang. 2025. SVG-IR: Spatially-Varying Gaussian Splatting for Inverse Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16143–16152.

Tiancheng Sun, Jonathan T. Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. 2019. Single image portrait relighting. *ACM Trans. Graph.* 38, 4, Article 79 (July 2019), 12 pages. doi:10.1145/3306346.3323008

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314* (2025).

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. 2025b. Comprehensive Relighting: Generalizable and Consistent Monocular Human Relighting and Harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 380–390.

Lezhong Wang, Shutong Jin, Ruiqi Cui, Anders Bjorholm Dahl, Jeppe Revall Frisvad, and Siavash Bigdeli. 2025a. ReLumix: Extending Image Relighting to Video via Video Diffusion Models. arXiv:2509.23769 [cs.GR] https://arxiv.org/abs/2509.23769

Zhibo Wang, Xin Yu, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. 2020. Single image portrait relighting via explicit multiple reflectance channel modeling. *ACM Trans. Graph.* 39, 6, Article 220 (Nov. 2020), 13 pages. doi:10.1145/3414685.3417824

Andreas Wenger, Andrew Gardner, Chris Tchou, Jonas Unger, Tim Hawkins, and Paul Debevec. 2005. Performance relighting and reflectance transformation with time-multiplexed illumination. *ACM Trans. Graph.* 24, 3 (July 2005), 756–764. doi:10.1145/1073204.1073258

Zirui Wu, Jianteng Chen, Laijian Li, Shaoteng Wu, Zhikai Zhu, Kang Xu, Martin R. Oswald, and Jie Song. 2025. 3D Gaussian Inverse Rendering with Approximated Global Illumination. https://arxiv.org/abs/2504.01358

Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. 2022. Learning to Relight Portrait Images via a Virtual Light Stage and Synthetic-to-Real Adaptation. *ACM Transactions on Graphics (TOG)* (2022).

Xueming Yu, David George, John Millward, and Paul Debevec. 2025a. Real-Time Multispectral Lighting Reproduction. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Posters (SIGGRAPH Posters '25)*. Association for Computing Machinery, New York, NY, USA, Article 35, 3 pages. doi:10.1145/3721250.3743035

Xueming Yu, Mingming He, David George, Avani Joshi, and Paul Debevec. 2025b. Digital Bi-Pack: Recording Live-Action under Two Near-Simultaneous Lighting Conditions. In *Proceedings of the SIGGRAPH Asia 2025 Technical Communications (SA Technical Communications '25)*. Association for Computing Machinery, New York, NY, USA, Article 19, 4 pages. doi:10.1145/3757376.3771405

Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. 2024b. DiLightNet: Fine-grained Lighting Control for Diffusion-based Image Generation. In *ACM SIGGRAPH 2024 Conference Papers* (Denver, CO, USA) *(SIGGRAPH '24)*. Association for Computing Machinery, New York, NY, USA, Article 73, 12 pages. doi:10.1145/3641519.3657396

Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. 2024a. RGB↔X: Image decomposition and synthesis using material- and lighting-aware diffusion models. In *ACM SIGGRAPH 2024 Conference Papers* (Denver, CO, USA) *(SIGGRAPH '24)*. Association for Computing Machinery, New York, NY, USA, Article 75, 11 pages. doi:10.1145/3641519.3657445

Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. 2021a. PhySG: Inverse Rendering with Spherical Gaussians for Physics-based Material Editing and Relighting. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2025. Scaling In-the-Wild Training for Diffusion-based Illumination Harmonization and Editing by Imposing Consistent Light Transport. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=u1cQYxRI1H

Longwen Zhang, Qixuan Zhang, Minye Wu, Jingyi Yu, and Lan Xu. 2021b. Neural Video Portrait Relighting in Real-Time via Consistency Modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 802–812.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. 2022. Modeling Indirect Illumination for Inverse Rendering. In *CVPR*.

Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W. Jacobs. 2019. Deep Single-Image Portrait Relighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Yujie Zhou, Jiazi Bu, Pengyang Ling, Pan Zhang, Tong Wu, Qidong Huang, Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, et al. 2025. Light-A-Video: Training-free Video Relighting via Progressive Light Fusion. *arXiv preprint arXiv:2502.08590* (2025).

Jingsen Zhu, Fujun Luan, Yuchi Huo, Zihao Lin, Zhihua Zhong, Dianbing Xi, Rui Wang, Hujun Bao, Jiaxiang Zheng, and Rui Tang. 2022b. Learning-Based Inverse Rendering of Complex Indoor Scenes with Differentiable Monte Carlo Raytracing. In *SIGGRAPH Asia 2022 Conference Papers*. ACM, Article 6, 8 pages. https://doi.org/10.1145/3550469.3555407

Rui Zhu, Zhengqin Li, Janarbek Matai, Fatih Porikli, and Manmohan Chandraker. 2022a. IRISformer: Dense Vision Transformers for Single-Image Inverse Rendering in Indoor Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2822–2831.