DynOLAT: OLAT Video Prediction for Controllable Dynamic Full-Body Relighting

TAOMING LIU*, State Key Lab of CAD&CG, Zhejiang University, China
YINJI SHENTU*, State Key Lab of CAD&CG, Zhejiang University, China
GUANGZHAO HE, Cornell University, United States of America
YUDONG JIN, State Key Lab of CAD&CG, Zhejiang University, China
YUKANG YE, 4DV.ai, China
SIYU ZHANG, 4DV.ai, China
FAN JIA, Ant Group, China
WEI HUA, Zhejiang University, China and Zhejiang Lab, China
HUJUN BAO, State Key Lab of CAD&CG, Zhejiang University, China
XIAOWEI ZHOU†, State Key Lab of CAD&CG, Zhejiang University, China

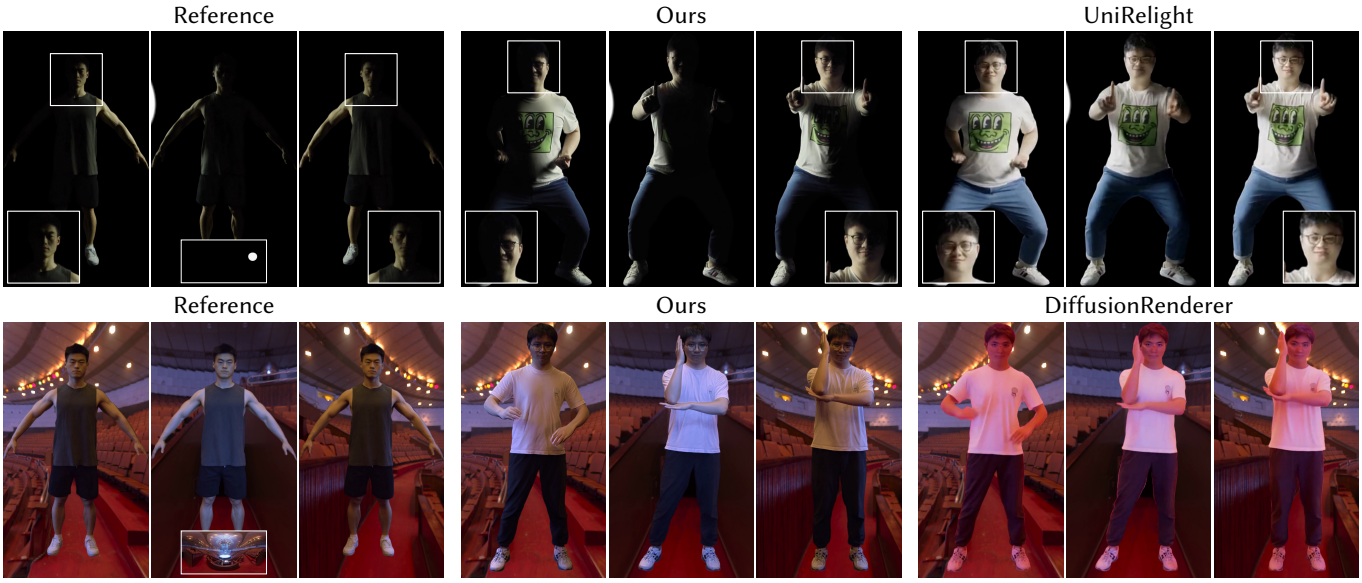


Fig. 1. **Controllable dynamic full-body relighting from a single video.** Each row compares a lighting reference (left), ours (middle), and a baseline (right). Top: ours better matches the direction and shadow orientation of a rotating point light than UniRelight [He et al. 2025]. Bottom: ours follows relative environment-light intensity without the spurious red cast produced by DiffusionRenderer [Liang et al. 2025] in the low-intensity red-chair region.

Accurate lighting response is essential for realistic dynamic human relighting, because shadows, highlights, and color changes should follow the target illumination in both direction and relative strength. Current learning-based relighting methods can generate visually plausible videos, but they often exhibit imprecise lighting responses: shadows and highlights may drift away from the target direction, and the output brightness may not faithfully follow the relative intensities in the target environment, making the result closer to appearance harmonization than controllable relighting. To address this issue, we propose DynOLAT, a generalizable OLAT video prediction framework for dynamic full-body humans. Instead of directly synthesizing a single relit appearance, our method predicts temporally consistent One-Light-At-a-Time (OLAT) basis videos from flat-lit inputs, and then linearly

*Equal contribution.
†Corresponding author.

Authors' Contact Information: Taoming Liu, taomingliu@zju.edu.cn, State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China; Yinji ShenTu, styj@zju.edu.cn, State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China; Guangzhao He, gh466@cornell.edu, Cornell University, New York, United States of America; Yudong Jin, krahetx@gmail.com, State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China; Yukang Ye, yeyukang@4dv.ai, 4DV.ai, Hangzhou, China; Siyu Zhang, derizhangzsy@gmail.com, 4DV.ai, Hangzhou, China; Fan Jia, jiafan.jf@antgroup.com, Ant Group, Hangzhou, China; Wei Hua, huawei@zhejianglab.org, Zhejiang University, Hangzhou, China and Zhejiang Lab, Hangzhou, China; Hujun Bao, bao@cad.zju.edu.cn, State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China; Xiaowei Zhou, xwzhou@zju.edu.cn, State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China.

This work is licensed under a Creative Commons Attribution 4.0 International License.

© 2026 Copyright held by the owner/author(s).
ACM 1557-7368/2026/12-ART209
https://doi.org/10.1145/3842500

combines the predicted bases according to the target environment map. This OLAT-based formulation makes the directional response explicit in the predicted bases, while the relative lighting strength is determined by the linear weights derived from the environment map. We further introduce a dynamic capture pipeline and the first large-scale dynamic full-body OLAT dataset, comprising over 24 million frames from 224 identity-outfit combinations, where different outfits of the same person define separate combinations. Experiments show that DynOLAT responds more faithfully to variations in target lighting direction and relative intensity than the tested baselines, enabling controllable video relighting while preserving fine appearance details and temporal coherence. The dataset and processing tools will be publicly released to foster future research.

CCS Concepts: • **Computing methodologies** → **Image-based rendering**; **Biometrics**; **Scene understanding**.

Additional Key Words and Phrases: Controllable relighting, full-body OLATs, latent bridge matching, video-to-video translation, HD full-body capture

ACM Reference Format:
Taoming Liu, Yinji ShenTu, Guangzhao He, Yudong Jin, Yukang Ye, Siyu Zhang, Fan Jia, Wei Hua, Hujun Bao, and Xiaowei Zhou. 2026. DynOLAT: OLAT Video Prediction for Controllable Dynamic Full-Body Relighting. *ACM Trans. Graph.* 45, 6, Article 209 (December 2026), 28 pages. https://doi.org/10.1145/3842500

1 Introduction

Realistic dynamic human relighting requires the synthesized subject to respond accurately to the specified illumination in addition to appearing visually plausible. In practical VR/AR, film production, immersive telepresence, and game-content workflows, artists expect a relighting method to place shadows, highlights, and color changes according to the target light direction and relative intensity. This paper studies this controllability-oriented setting: given a monocular flat-lit video of a full-body human captured from an arbitrary viewpoint, we aim to synthesize a temporally coherent video of the same performance under a new lighting environment.

Existing general relighting methods still struggle with two forms of lighting-response accuracy that are critical for production use. First, their directional response is often imprecise: even when the target light direction is specified, the synthesized shading may exhibit a noticeable angular offset. Second, their intensity response is often unreliable: the output brightness may not faithfully follow the relative intensities in the target environment, making the result behave more like appearance harmonization than physically controlled relighting. These failure modes, highlighted in Fig. 1 and Fig. 6, are not only local visual artifacts; they directly weaken control over the target light direction and relative strength.

The difficulty arises because monocular human relighting requires both appearance realism and disentangled light control. Traditional inverse-rendering methods [Barron and Malik 2015] attempt to estimate geometry, material, and illumination explicitly, but the problem is severely under-constrained for dynamic humans with spatially varying skin, hair, and clothing reflectance. Strong priors are therefore needed to resolve the ambiguity, often suppressing high-frequency details and producing overly smooth or artificial human appearance. Recent diffusion-based relighting methods [Bian et al. 2025; Fang et al. 2025; Liang et al. 2025] reduce the need for explicit decomposition by learning direct image or video mappings from large-scale data. While these models can produce visually plausible results, their lighting conditions are often encoded as holistic environment images or prompts, making it difficult to guarantee that output shadows and intensities follow the target illumination in a directionally and radiometrically faithful way.

OLAT representations provide a natural route to controllable relighting because they separate light transport into directional basis responses. A one-light-at-a-time (OLAT) image records the subject response to a single lighting direction; arbitrary environment illumination can then be formed by linearly combining the corresponding bases. DifFRelight [He et al. 2024] demonstrates that predicting OLAT representations can yield highly realistic and controllable facial relighting. Extending this idea from single facial images to dynamic full-body videos is non-trivial, because full-body motion introduces loose clothing deformation, self-occlusion, larger pose variation, and a lack of large-scale dynamic full-body OLAT supervision.

To address this gap, we propose DynOLAT, an OLAT video prediction framework for controllable dynamic full-body relighting. Given a flat-lit human video and a target lighting direction, our model predicts the corresponding OLAT basis video; the predicted bases are then combined in linear color space according to any target environment map. This formulation makes the target lighting explicit at two levels: the network learns directional basis responses, and the final relit video is assembled with environment-dependent intensity weights. As a result, the method is designed to preserve directional alignment and support explicit relative-intensity control, rather than relying on an unconstrained end-to-end generation of a single relit appearance.

Our technical design adapts a pretrained video model [Wan et al. 2025] to this OLAT prediction task while preserving source identity, details, and temporal coherence. We adopt Latent Bridge Matching (LBM) [Chadebec et al. 2025] to transport flat-lit video latents directly toward OLAT latents, avoiding the source-detail loss introduced by sampling from independent Gaussian noise. We further inject the target lighting direction through block-specific AdaLN modulation, enabling fine-grained directional control across the video transformer. For long sequences, a recurrent bridge context reuses previously generated OLAT frames as analytic overlap constraints, maintaining coherence beyond the training window.

To train this data-driven model, we build a dedicated dynamic OLAT acquisition pipeline. The system rapidly alternates flat and directional illumination and then aligns the resulting frames to form paired flat-lit/OLAT videos. Using this pipeline, we collect a large-scale, multi-view, dynamic full-body OLAT dataset containing diverse subjects, clothing, motions, and viewpoints. To the best of our knowledge, this is the first dataset of comparable scale that provides dynamic full-body OLAT supervision for learning temporally consistent lighting responses. We will publicly release the dataset and processing tools to support future research.

Extensive experiments demonstrate that DynOLAT improves controllable full-body relighting in dynamic sequences. Qualitative comparisons show that our method better follows the target light direction and relative illumination strength, while preserving fine body details and temporal consistency. Quantitative comparisons and ablation studies further validate the effectiveness of our design.

In summary, our main contributions are as follows:

(1) An OLAT video prediction framework that formulates dynamic full-body relighting as directionally decomposed light-response synthesis, enabling more faithful responses to target light direction and environment-map intensity.
(2) Tailored model designs that adapt Latent Bridge Matching for source-detail preservation, use block-specific lighting modulation for directional control, and employ recurrent bridge context for long-sequence temporal coherence.
(3) The first large-scale dynamic full-body OLAT dataset with sufficient subject, clothing, motion, and viewpoint diversity for learning temporally consistent human light transport.

2 Related Work

2.1 Human-centric Relighting

Relighting human performances from monocular video remains a fundamental challenge in computer graphics and vision. Existing methodologies can be broadly categorized into three paradigms: optimization-based inverse rendering, learning-based image relighting and learning-based video relighting.

Optimization-Based Inverse Rendering. Leveraging recent advances in 3D neural representations like NeRF [Mildenhall et al. 2021] and 3DGS [Kerbl et al. 2023], numerous works attempt to reconstruct human avatars and decompose intrinsic properties from multi-view images [Chen et al. 2024b; Chen and Liu 2022; Gao et al. 2024; Hu et al. 2024; Li et al. 2025; Ma et al. 2024; Singh et al. 2025; Xu et al. 2024a; Zhan et al. 2025]. By explicitly estimating geometry, material, and illumination, these methods provide a physically interpretable route to relighting control. However, standard physically-based rendering (PBR) approximations often fail to capture complex surface effects, such as subsurface scattering and intricate self-shadowing, which affect both appearance realism and the accuracy of the predicted lighting response. To circumvent these limitations, recent approaches utilize ground-truth OLAT scans to learn neural reflectance fields directly [Guo et al. 2019; Li et al. 2024; Singh et al. 2025; Wang et al. 2025b; Zhou et al. 2023]. While these approaches have achieved photorealistic results for both upper-body [Li et al. 2024; Saito et al. 2024] and full-body avatars [Guo et al. 2019; Singh et al. 2025; Teufel et al. 2026; Wang et al. 2025b; Zhou et al. 2023], they rely heavily on subject-specific optimization. This requirement for dense capture setups and per-identity training limits direct application to in-the-wild images or videos and makes controllable direction and intensity responses difficult for novel dynamic subjects.

Learning-Based Image Relighting. To achieve generalization across subjects, a classical paradigm [Daichi Tajima 2025; Ji et al. 2022; Kim et al. 2024; Mei et al. 2025; Nestmeyer et al. 2020; Pandey et al. 2021; Yeh et al. 2022; Zhang et al. 2024] adapts the physics-based inverse rendering workflow into 2D feature space. These methods typically employ a cascaded pipeline that first decomposes the input image into intrinsic components (e.g., albedo and normals) and subsequently synthesizes the relit output via a neural renderer. While this strategy explicitly models light transport, the multi-stage process is prone to error accumulation: inaccurate geometry or albedo can blur source details and also shift the resulting shadows, highlights, and brightness away from the target illumination. Conversely, end-to-end approaches [Liu et al. 2025a; Sun et al. 2019] map inputs directly to relit outputs. Due to the scarcity of paired real-world data under varying illumination, many methods rely on synthetic data [Ren et al. 2024; Zhang et al. 2025a] or auxiliary generation tasks [Chaturvedi et al. 2025], which can yield unrealistic shading or artifacts due to the domain gap. More importantly for controllable relighting, direct output synthesis often uses a target environment or background as a holistic condition, providing limited constraints that ensure the generated shadow direction and relative brightness follow the specified illumination. DifFRelight [He et al. 2024] and POLAR [Chen et al. 2026] fine-tune generative models on real OLAT data to predict OLATs for explicit lighting superposition, avoiding physical decomposition and synthetic-data domain gaps. In contrast, 3DPR [Rao et al. 2025] predicts portrait OLATs with a 3D head prior and linearly combines them for relighting. Nevertheless, applying these image-based solutions frame by frame to dynamic full-body sequences yields temporally unstable results, lacking the robustness required for video applications.

Learning-Based Video Relighting. To extend relighting to videos, earlier methods attempted to enforce temporal consistency via loss constraints [Zhang et al. 2021] or feature blending [Wang et al. 2025a]. However, given the geometric and textural complexity of dynamic full-body motion, independent frame processing can struggle to maintain coherence at high resolution. With the advent of video diffusion models [Blattmann et al. 2023; Brooks et al. 2024; Wan et al. 2025], the mainstream paradigm has shifted toward fine-tuning video diffusion models [Bian et al. 2025; Fang et al. 2025; Guo et al. 2025; He et al. 2025; Liang et al. 2025; Lin et al. 2025; Liu et al. 2025b; Mei et al. 2025; Zeng et al. 2025] or controlnets [Zhang et al. 2025b]. These methods typically inject condition frames by concatenating them with noisy latents, and represent lighting with environment maps, background images, or text-like conditions. However, this standard conditioning strategy poses a fundamental conflict: the stochastic denoising process, while excellent for diversity, introduces randomness that undermines the deterministic nature of relighting. This often leads to hallucinated details or identity drift, where the generated texture deviates from the source, and can also weaken the alignment between the output illumination and the target direction or intensity distribution. Instead of predicting a single relit video directly, our framework predicts OLAT basis videos and assembles the final illumination explicitly. To make this basis prediction source-aligned, we employ bridge matching [Chadebec et al. 2025] from the input video latents to the target OLAT latents, preserving fine source textures while learning a more deterministic light-response mapping. However, the success of such a data-driven video paradigm is heavily contingent on the availability of high-quality, temporally consistent training data.

2.2 LightStage and OLAT Acquisition

The LightStage architecture, pioneered by Debevec et al. [2000], revolutionized reflectance acquisition through precise, programmable lighting control. This setup inherently enables the capture of multi-view OLAT imagery. Fundamentally, relying on the linearity of light

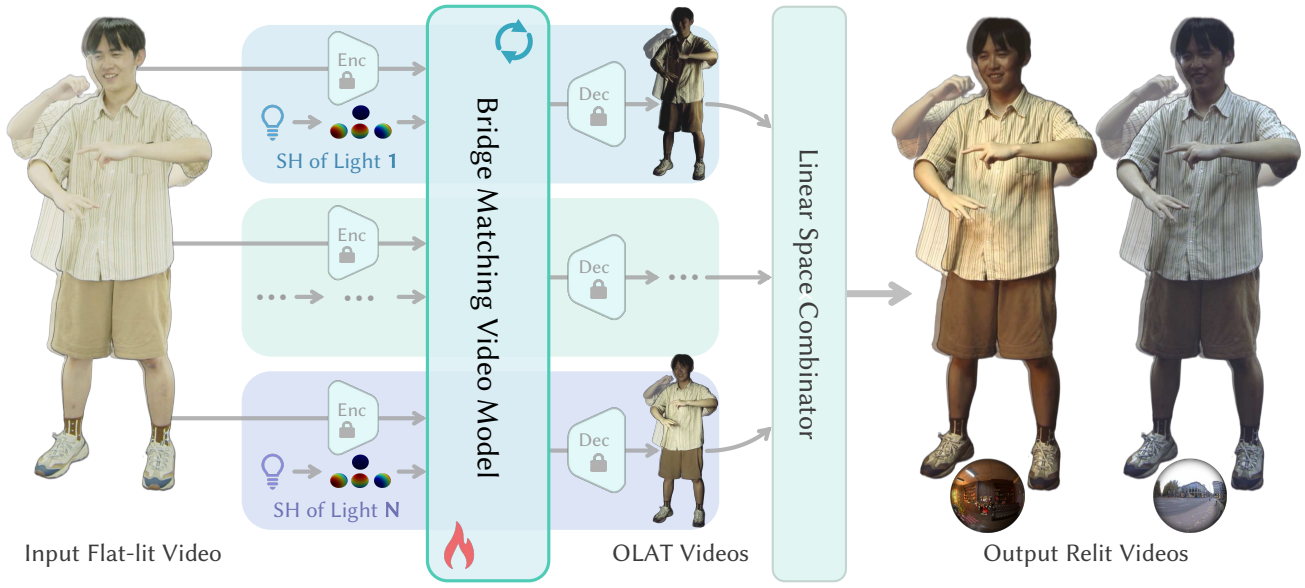


Fig. 2. **The overview of the relighting pipeline.** Our model encodes a flat-lit video into latent features and uses Spherical Harmonics (SH) embeddings of the target light direction to modulate the generation. Through an iterative process, it transforms the flat-lit latents into target OLAT representations. These generated bases are then linearly combined to achieve high-fidelity relighting under any environment map.

transport, the appearance of a subject under arbitrary environmental illumination can be synthesized via the linear superposition of these OLAT basis images. Consequently, LightStage data has become a "gold standard" benchmark for realistic relighting and has driven significant evolution in hardware design [Debevec 2008; Guo et al. 2019; He et al. 2024; Jones et al. 2006].

Despite its importance, the acquisition of such high-quality data remains heavily restricted by domain. While numerous works have explored OLAT-based frameworks, they predominantly focus on restricted domains such as facial appearance [Chen et al. 2026; He et al. 2024; Li et al. 2024], hands [Chen et al. 2024a], or static objects [Liu et al. 2023; Toschi et al. 2023; Zhou et al. 2025]. Research addressing full-body capture remains scarce [Martinez et al. 2024; Teufel et al. 2025; Wang et al. 2025b], and publicly available datasets are extremely limited. More critically, prior work on time-multiplexed dynamic OLAT capture has focused on faces and upper bodies [Debevec 2008; Zhang et al. 2021].

To the best of our knowledge, no existing public dataset provides dynamic full-body OLAT sequences at a comparable scale. This data scarcity becomes a bottleneck for training generalizable full-body relighting models, which require a substantial diversity of human materials and motions to avoid overfitting.

To bridge this gap, we present our OLAT dataset, which includes a dedicated motion-rich subset. This component provides extensive dynamic OLAT data that covers diverse human materials and motions while supplying the directional light-response supervision needed to train controllable video relighting models.

3 Method

We decompose dynamic human relighting into explicit light-response prediction and linear environment assembly. In the first stage, we employ our OLAT video prediction model to synthesize a set of basis OLAT videos from a flat-lit input sequence, where each basis corresponds to a specific lighting direction. In the second stage, we reconstruct the final relit video under any complex environment map by linearly accumulating these predicted OLAT components in the linear color space. This decomposition is central to our controllability goal: the lighting direction is handled by the predicted basis index, while the lighting intensity is handled by the environment-map weights. An overview of our proposed framework is illustrated in Fig. 2.

We first describe the architecture of our OLAT video prediction model in Sec. 3.1. We then explain how the flat-lit input frames (Sec. 3.2) and the target lighting direction (Sec. 3.3) are processed and injected into the network. Subsequently, we elaborate on our strategies for recovering high-frequency details (Sec. 3.4), which are necessary for producing reliable OLAT bases that retain the subject's material and texture response. We then discuss our solution for inference on long video sequences (Sec. 3.5). Finally, we present the image-based relighting (IBR) formulation that renders the final output via a linear combination of the predicted OLAT bases (Sec. 3.6).

3.1 Overview of OLAT Video Prediction Model

Given a flat-lit video sequence and a target lighting direction, our goal is to synthesize the corresponding OLAT video. Concretely, we operate within the compressed latent space of a pre-trained video diffusion model [Wan et al. 2025]. Adopting the formulation

of Chadebec et al. [2025], we train the model to learn a flow trajectory that directly transports the source flat-lit latents to the target OLAT latents (see Sec. 3.2 for details). During this process, the specific target lighting direction steers the generation by modulating the network features via Adaptive Layer Normalization (AdaLN) [Perez et al. 2018] (see Sec. 3.3 for details). Ultimately, the network predicts a velocity field to drive the transport of the input latents, which are subsequently decoded to yield the final high-fidelity OLAT video.

We explicitly model this process over video sequences rather than individual frames. While static image relighting is well-established, the straightforward frame-by-frame application of such methods to dynamic sequences inherently suffers from temporal flickering and inconsistent shadowing. By operating directly on video data, our model not only ensures temporal coherence and geometric consistency but also benefits from the diverse pose coverage naturally present in dynamic captures.

3.2 Bridge Matching for Video Latents

To predict a reliable OLAT basis video, it is crucial to feed the original video frames into the model in a way that preserves source appearance while supporting deterministic lighting edits. Prior diffusion-based relighting methods typically concatenate input frames with Gaussian noise along a specific axis, such as the channel dimension [Chaturvedi et al. 2025; Mei et al. 2025] or the frame dimension [Bian et al. 2025; He et al. 2025], to maintain the standard denoising process. While this strategy supports the generative workflow, the reliance on randomly sampled noise often leads to outputs that fail to preserve fine textures and details from the input. Such inconsistency is unacceptable for human-centric relighting: texture drift changes the subject appearance, and the same stochasticity can also perturb the predicted shadows and highlights that should correspond to a specific lighting direction. See Fig. 9 for comparative results.

Since relighting is inherently a pixel-aligned editing task, we draw inspiration from LBM [Chadebec et al. 2025] and replace the input noise with latent tokens encoded from the flat-lit frames, allowing direct token-wise alignment between model inputs and outputs. This source-to-target bridge helps preserve fine source textures while constraining the model to learn the lighting-dependent residual response needed for OLAT prediction.

Specifically, we implement this paradigm by constructing a stochastic bridge between the source and target distributions. We first encode the flat-lit video and the corresponding OLAT video into latent space via the encoder of a pretrained video variational autoencoder (VAE) [Kingma and Welling 2013], producing latents z_{flat} and z_{olat}. We then define the bridge interpolant z_t as:

$$z_t = (1-t)z_{\text{olat}} + t z_{\text{flat}} + \sigma\sqrt{t(1-t)}\epsilon\,, \tag{1}$$

where $t \in [0,1]$ is randomly sampled from a set of timesteps, $\epsilon \sim \mathcal{N}(0, I)$, and σ denotes the noise strength, which we set to 0.001 to enhance model robustness while keeping the input distribution within a reasonable range. To learn the trajectory of this bridge, we regress a vector field using a video DiT model v_Θ by minimizing the bridge matching objective:

$$\mathcal{L}_{\text{LBM}} = \mathbb{E}\left[\left\| \frac{(z_t - z_{\text{olat}})}{t} - v_\Theta(z_t, t)\right\|^2\right]. \tag{2}$$

At inference time, we feed the encoded flat-lit latents into the model and iteratively update them from $t = 1$ toward the distribution of the corresponding OLAT latents via:

$$\widehat{z}_{t+\Delta t} = z_t + v_\Theta(z_t, t)\Delta t\,, \tag{3}$$

where $\Delta t < 0$, and the noise term is removed to maximally preserve information from the original input. When the iteration reaches $t = 0$, we map the predicted $\widehat{z}_0$ back to the pixel space using the decoder of the video VAE.

3.3 Lighting Direction Conditioning

To achieve precise directional control while maintaining a compact representation, we condition the model directly on the 3D lighting direction vector. This approach serves as a robust alternative to explicit spatial encodings (e.g., shading maps [Pandey et al. 2021; Zeng et al. 2024] or light source maps [Choi et al. 2024]), avoiding their associated computational overhead and potential geometric ambiguity.

Encoding. To ensure geometric consistency with the visual content, we first rotate the world-space light vector into the camera coordinate system of the input video. Furthermore, to facilitate the learning of high-frequency illumination variations, we augment the raw direction vector d using Spherical Harmonics (SH) encoding $\mathcal{Y}(\cdot)$ [Tancik et al. 2023]. The final conditioning vector c_{d} is constructed by concatenating the raw vector with its positional encoding:

$$c_{\text{d}} = \text{d} \oplus \mathcal{Y}(\text{d})\,, \tag{4}$$

where $\oplus$ indicates concatenation along the channel dimension, and the SH degree is empirically set to 3.

Injection. We inject this lighting condition via an adaptive modulation mechanism. Within each transformer block, an individual linear projection layer maps the frequency-augmented vector c_{d} to the model's hidden dimension. This projected vector is subsequently added to the timestep embedding to modulate the input tokens by driving the scale, shift, and gate parameters of AdaLN [Perez et al. 2018]. To ensure the input distribution remains unperturbed at the onset of training, the parameters of these newly introduced linear layers are initialized to zero.

3.4 High-Frequency Details Enhancement

Accurate OLAT prediction requires preserving high-frequency human appearance in addition to estimating the correct lighting response. Fine wrinkles, hair strands, fabric texture, and sharp shadow boundaries are part of the basis response that will later be linearly combined under environment illumination. On one hand, the iterative denoising process of diffusion models excels at producing visually realistic results [Nichol et al. 2021; Rombach et al. 2021], but the randomness introduced by noise can degrade high-frequency details present in the input. On the other hand, while a bridge matching approach in latent space naturally preserves much of the input

pattern, it tends to excessively attenuate high-frequency details that should otherwise be preserved. To make the predicted OLAT bases both visually faithful and reliable for controllable linear assembly, we strengthen high-frequency details across four complementary dimensions: image-video joint training, dynamic resolution, pixel-space supervision, and sampling iterations.

Image-Video Joint Training. Although video data is essential for learning temporal dynamics, raw OLAT video captures frequently suffer from inherent quality degradations, such as motion blur, compression noise, and warping artifacts, which inevitably smooth out high-frequency details. In contrast, static OLAT imagery typically preserves superior textural fidelity and sharpness. To bridge this gap, we implement a joint training scheme that interleaves high-quality static images with video sequences. In this setup, static images guide the model to recover fine-grained structural and lighting details, while video data ensures temporal coherence.

Dynamic Resolution. We also observe that full-body human figures in videos tend to be much taller than typical 16:9 frames, meaning that a fixed training resolution often compresses the subject's effective pixels while leaving excessive background. We therefore use no fixed training resolution: for each sequence, we crop all flat-lit and OLAT frames to the union of the flat-lit foreground masks predicted by Zheng et al. [2024]. We then pad the crop, rather than resize the foreground, until both dimensions are divisible by 16 as required by the VAE.

Pixel-Space Supervision. Supervision using only a latent-space regression loss is insufficient to achieve high visual quality, as it does not emphasize the preservation of high-frequency structures that are perceptually important, whereas a perceptual loss can better retain these visually salient patterns. This supervision also benefits controllable relighting because blurred basis predictions can smear shadow boundaries and weaken the directional cues used in the final composition. Following Chadebec et al. [2025], we therefore incorporate a pixel-space LPIPS loss [Zhang et al. 2018]: the model's predicted OLAT latents $\widehat{z}_{\text{olat}}$ are decoded via the VAE into $\widehat{x}_{\text{olat}}$ and an LPIPS loss is computed between $\widehat{x}_{\text{olat}}$ and the ground truth x_{olat} in pixel space. In practice, to avoid excessive memory usage on video data, we apply a random cropping strategy, computing the perceptual loss only on patches when the video resolution exceeds a predefined threshold. The overall training objective is then defined as:

$$\mathcal{L} = \mathcal{L}_{\text{LBM}}(\mathcal{E}(x_{\text{flat}}), \mathcal{E}(x_{\text{olat}}), t) + \lambda \cdot \mathcal{L}_{\text{perceptual}}(\widehat{x}_{\text{olat}}, x_{\text{olat}}), \quad (5)$$

where x_{flat} and x_{olat} denote videos in pixel space, and $\mathcal{L}_{\text{perceptual}}$ refers to the frame-wise LPIPS loss.

Sampling Iterations. We find that simply reusing the single-step inference schedule from Chadebec et al. [2025] leads to overly smooth outputs, where high-frequency details such as wrinkles on human skin and clothing are not well preserved. Performing iterative sampling with 4 steps, however, sufficiently mitigates this issue and yields sharper OLAT bases for subsequent intensity-weighted composition.

3.5 Long Sequence Inference

Video generative models are typically constrained to a fixed context length F during training due to GPU memory constraints. Direct extrapolation to longer sequences is non-trivial; a naive approach of partitioning the video into independent chunks inevitably results in temporal discontinuities and flickering at the boundaries. Existing sliding-window strategies [Jin et al. 2025; Wu et al. 2024] achieve consistency by fusing latents across varying Gaussian denoising levels. These are incompatible with our framework, where the generative trajectory is governed by a structured source distribution rather than stochastic noise.

To address this, we propose a recurrent bridge context, a strategy tailored for consistency in bridge-matching-based generation. We employ a sliding window with length $W = F$ and stride S, leveraging previously generated OLAT frames to guide the inference process. Specifically, we define a "history context" corresponding to the overlapping region of length $W - S$. This context is populated by $z_{\text{olat}}^{\text{prev}} \in \mathbb{R}^{(W-S)\times c\times h\times w}$, which denotes the finalized OLAT latents derived from the preceding window. During the iterative sampling process of the current window, we enforce temporal continuity by overriding the estimated latents $\widehat{z}_t$ in this overlap region with the ideal bridge state derived from the history context:

$$\widehat{z}_t[i] = \begin{cases} (1-t)z_{\text{olat}}^{\text{prev}}[i] + t z_{\text{flat}}[i], & \text{if } i < W - S, \\ \widehat{z}_t[i], & \text{otherwise,} \end{cases} \quad (6)$$

where $z_{\text{flat}} \in \mathbb{R}^{W\times c\times h\times w}$ represents the encoded flat-lit frames of the current window. At each timestep, we replace the overlapping latents with the noise-free mean interpolant of Eq. 1, thereby constraining the current window to match the previously generated frames in the overlap region and propagating temporal context forward. Finally, we concatenate the context latents $z_{\text{olat}}^{\text{prev}}$ with the non-overlapping part of the predicted $\widehat{z}_0[W - S : W] \in \mathbb{R}^{S\times c\times h\times w}$ and decode the combined sequence via the video VAE.

This recursive strategy leverages the causal nature of the video VAE to extend bridge-matching inference beyond the training context while preserving consistent appearance and illumination responses without additional training. Results on sequences of up to 917 frames are provided in the supplementary material.

3.6 Linear Space Combination

To relight the dynamic performance under a target high dynamic range (HDR) environment map, we associate each predicted OLAT basis with a light direction in the camera coordinate system. We partition the environment map according to the nearest basis direction and integrate the linear HDR radiance within each region to obtain its weight. The final video is synthesized by performing a frame-by-frame linear combination of the predicted OLAT videos. Specifically, for each frame, the corresponding outputs from all basis videos are converted to linear space, multiplied by these weights, and accumulated to produce the relit result. This linear-space assembly is the key step for intensity controllability: the relative strength of each environment region is explicitly determined by its combination weight, rather than being implicitly inferred by the model from an unconstrained conditioning image. A single global scaling factor, typically set to 3, is applied per environment map to match



Fig. 3. **Data processing pipeline for the *Dynamic-subset*.** (1) Light Switch Sequence: One sequence adheres to a rigorous interleaved pattern: 18 flat-lit frames are paired with 9 frames of OLAT-A and 9 frames of OLAT-B, where A and B represent two distinct, spatially distant lighting directions. (2) OLAT to Flat-lit Warping: Within each sequence, we compute optical flow with AllTracker [Harley et al. 2025] and foreground masks with BiRefNet [Zheng et al. 2024] from the reference flat-lit frames. The OLAT frames are then warped to align with the flat-lit frames using the predicted flow to eliminate motion misalignment. (3) Training Pairs: Each sequence yields two temporally consistent video pairs. As shown, our light switching pattern successfully provides distinct illumination conditions for the identical motion sequence, ensuring robust disentanglement of light and geometry. For a detailed description of the processing pipeline, please refer to Sec. 4.3.

the dynamic range of the target illumination to the captured lighting bases, while preserving the relative weights among basis directions.

4 The Dataset

4.1 Overview

Training our controllable video relighting model requires large-scale paired video data consisting of a subject under flat lighting and the corresponding response to known OLAT illumination.

However, acquiring such ground truth is physically non-trivial, as a subject cannot simultaneously experience two distinct lighting conditions at the exact same instant. To address this, we employ a time-multiplexed acquisition protocol, rapidly alternating between flat-lit frames and OLAT frames. Since inevitable subject motion occurs during the interval between light switches, raw frames are inherently misaligned. We mitigate this by estimating optical flow between consecutive flat-lit frames and warping the intermediate OLAT frames to align spatially with the flat-lit references, thereby yielding pixel-aligned training pairs.

Crucially, to ensure the model learns robust representations, we devised a novel Light Switching Pattern. Our empirical findings suggest that training on a single fixed light direction leads to overfitting, where the model fails to disentangle illumination from motion features. To address this, we alternate between two distinct OLAT directions throughout a capture session. Specifically, we select the two lights with the greatest spatial distance within the LightStage to maximize illumination variance, while ensuring that all 101 distinct light sources are traversed exactly once across the full recording session of each subject, with no repetitions. Each OLAT direction is repeated for 9 consecutive cycles to ensure sufficient temporal sampling. Consequently, these two sets of OLAT frames, along with their interleaved flat-lit references, constitute a block of 18 pairs, which we define as a single sequence. The raw structure of such a sequence is visualized in Fig. 3(1), while the processed, aligned pairs are shown in Fig. 3(3). Across a full recording session for one subject, this process yields 51 such sequences for each camera view. Through this protocol, we acquired over 23 million valid frames across 189 identity-outfit combinations in the *Dynamic-subset*. We

avoid alternating among too many directions (e.g., three or more) as it introduces excessive temporal gaps between frames of the same lighting condition, which degrades the temporal coherence of the video and hinders training.

Additionally, we acquired complete OLAT sequences from 35 identity-outfit combinations in static poses. These dense captures support image-based relighting synthesis of ground-truth images under target environment maps, providing a controlled benchmark for image-level lighting evaluation.

Consequently, we organize our dataset into two primary components: the *Dynamic-subset* (comprising the motion-rich paired video data) and the *Static-subset* (comprising the comprehensive full OLAT sequences). Tab. 1 compares the scale and capture properties of our dataset with existing benchmarks. The supplementary material reports aggregate statistics for the 224 identity-outfit combinations and provides detailed dataset overviews.

All participants provided informed consent that explicitly permits research use and public release of their captured data. The release will exclude unnecessary personal metadata and will document the intended research use, privacy considerations, and potential misuse in synthetic-media pipelines. We plan to release the processed dataset upon publication under terms consistent with the participants' consent.

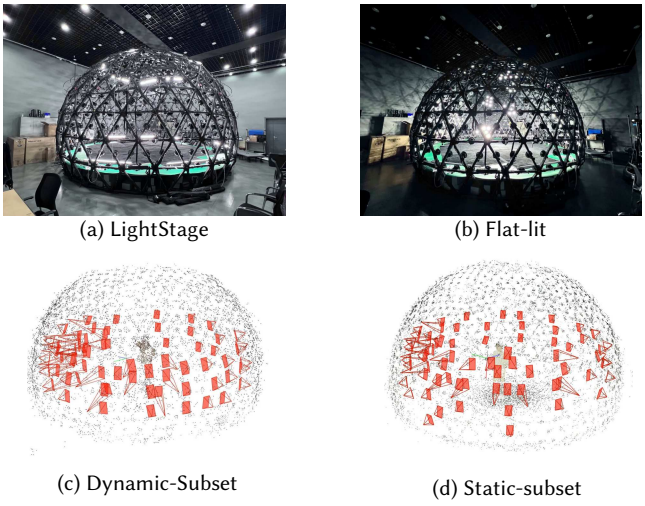
In the following subsections, we detail the technical implementation of our dataset construction. In Sec. 4.2, we describe the Capture Setup, detailing the unified LightStage hardware and the distinct camera array configurations tailored for each subset. In Sec. 4.3, we present the data-processing pipeline, focusing on motion-artifact mitigation and data-quality assessment.

4.2 Capture Setup

LightStage System. Our illumination system is shared across both subsets (Fig. 4(a)). It consists of over 800 LEDs grouped into 101 finite-area, individually addressable lighting units (referred to as One-Light units), driven by a custom microcontroller with microsecond precision. For dynamic acquisition, we configure the lighting clock cycle to 35 ms.

Camera Systems and Settings. We deployed two synchronized camera systems, one for each subset. We prioritized storage efficiency and bandwidth for the *Dynamic-subset* to enable large-scale recording, while prioritizing dynamic range and raw quality for the *Static-subset*.

- ***Dynamic-subset*:** We used 68 custom-built PTZ cameras (Sony IMX678) to record 4K H.264 video at 60 fps (60 Mbps bitrate, I-frame interval of 2). The array comprises 48 **short-lens** cameras (approximately 9 mm) distributed spherically and 20 **telephoto** cameras (approximately 34 mm) placed primarily in frontal positions to capture upper-body details. Exposure is fixed at 1/125 s (f/4.8) to reduce motion blur.
- ***Static-subset*:** We utilized 70 Blackmagic Micro Studio Camera 4K G2 units arranged in a similar spherical layout: 54 units (14/15 mm lenses, including six zenith/nadir cameras) and 16 telephoto units (47 mm lenses). The raw footage is stored as Blackmagic RAW. While this format provides superior fidelity, its prohibitive bitrate prevents further scaling. We use a faster shutter speed of 1/500 s (f/4.4, ISO 3200) to reduce motion blur and transcode the footage to 100 Mbps for processing.



(a) LightStage (b) Flat-lit (c) Dynamic-Subset (d) Static-subset

Fig. 4. **Capture system overview.** (a) Our LightStage. (b) The flat-lit illumination configuration. (c–d) Camera distributions for the *Dynamic-subset* (c) and the *Static-subset* (d), reconstructed using COLMAP [Schönberger and Frahm 2016].

The spatial distributions are visualized in Fig. 4(c) and (d). Representative images from these camera arrays are provided in the supplementary material.

Illumination Strategy. A major challenge in OLAT capture is the extreme dynamic range discrepancy between full-on lighting and single OLAT conditions. To address this discrepancy, we activate a uniformly sampled subset (1/6) of the lights to simulate ambient illumination (Fig. 4(b)). This preserves perceptual consistency within the camera's limited dynamic range. Note that all references to "flat-lit" in this paper refer to this flat-lit configuration.

Light Switching Patterns. The two subsets employ different switching patterns to serve their respective goals:

- ***Dynamic-subset* Pattern:** As introduced in Sec. 4.1, we repeat one OLAT direction for 9 cycles, interleaved with its spatially most distant counterpart for another 9 cycles (Fig. 3(3)). This variation encourages the model to decouple motion from lighting.
- ***Static-subset* Pattern:** We use 10 consecutive OLAT frames followed by 1 flat-lit frame. For each identity-outfit combination, we capture at least three static poses and several flat-lit dynamic sequences for relighting evaluation.

4.3 Data Processing

Given the complexity of massive-scale OLAT acquisition, especially for the *Dynamic-subset*, the raw data requires careful processing to achieve geometric alignment and photometric consistency. The overall pipeline is illustrated in Fig. 3. In the following sections, we

Table 1. **Comparison of our OLAT dataset with existing OLAT datasets.** To the best of our knowledge, this is the first dynamic full-body OLAT dataset. It contains 24M valid frames, compared with 0.85M frames in the previous largest full-body OLAT dataset [Teufel et al. 2025]. Column definitions: **Type**: subject category (upper-body or full-body); **Dynamic**: presence of dynamic OLAT motion sequences; **#Frames**: total frame count across all views; **#Subj.**: number of identity-outfit combinations; **#Views**: number of cameras, including short-lens and telephoto arrays; **Res.**: approximate image resolution, where 1K ≈ 1920 × 1080, 4K ≈ 3840 × 2160, and 8K ≈ 7680 × 4320.

Dataset	Type	Dynamic	#Frames	#Subj.	#Views	Res.
Dynamic OLAT [Zhang et al. 2021]	upper-body	✓	603k	4	1	4K
Goliath-4 [Martinez et al. 2024]	upper-body	✓	>1M	4	110	4K
NetFlix Data [Mei et al. 2025]	upper-body	✓	12M	67	36	4K
POLAR [Chen et al. 2026]	upper-body	✓	**28M**	220	32	4K
Ultrastage [Zhou et al. 2023]	full-body	✗	192k	100	32	8K
HumanOLAT [Teufel et al. 2025]	full-body	✗	850k	21	40	5K
Ours (Static-subset)	full-body	✗	1.06M	35	70	4K
Ours (Dynamic-subset)	full-body	✓	**23M**	189	68	4K
Ours (Total OLAT)	full-body	✓	**24M**	224	68/70	4K

detail the core algorithms for OLAT to flat-lit warping, followed by specific strategies for data quality assessment.

OLAT to Flat-lit Warping. We use optical flow to compensate for subject motion, including rapid actions in the *Dynamic-subset* and involuntary sway in the *Static-subset*, that causes misalignment between flat-lit and OLAT frames. We compute bidirectional optical flow between consecutive flat-lit frames [Harley et al. 2025]. Assuming linear motion within the short interval, we derive the flow field for the intermediate OLAT frame via temporal linear interpolation. Alongside flow estimation, we generate foreground masks for the flat-lit frames using BiRefNet [Zheng et al. 2024] and compute a unified bounding box.

Based on the computed flow, we apply different warping strategies tailored to the downstream application:

- **Training Pairs (Dynamic & Static):** To minimize warping artifacts, we warp each OLAT frame to its temporal nearest-neighbor flat-lit frame. This yields diverse, well-aligned pairs for learning-based models, preserving the original motion fidelity.
- **Quality Assessment Stack (Static only):** For quality evaluation, we warp all OLAT frames within a pose to the first flat-lit frame of the sequence, producing a geometrically aligned OLAT stack for the photometric analyses described below.

Quality Assessment. We utilize the geometrically aligned OLAT stack from the *Static-subset* to quantitatively assess the quality of our acquisition system. As illustrated in Fig. 5, our raw captures from both short-lens (a) and telephoto (b) cameras preserve high-frequency details. To assess geometric detail, we estimate surface normals via photometric stereo using gradient-descent optimization with outlier rejection. The resulting normal maps in Fig. 5(c) retain fine-scale skin texture and clothing wrinkles. To assess photometric linearity, we synthesize relit images by linearly combining the OLAT bases with environment-map weights after gamma correction. The


(a) Short-lens (b) Long-lens (c) Normal (d) Relight

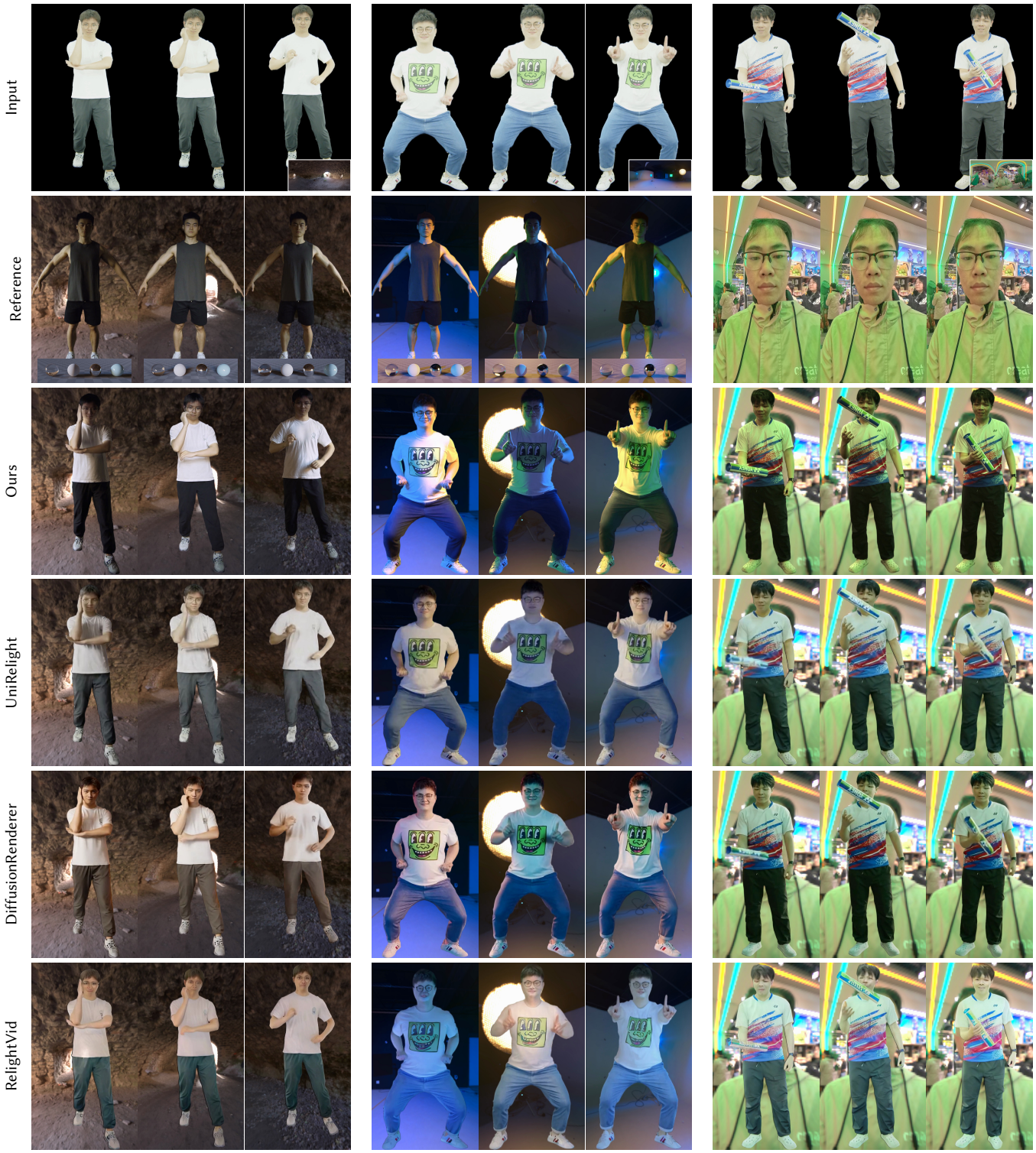
Fig. 5. **Quality assessment.** (a) Full-body view from the short-lens array. (b) High-resolution view from the telephoto array, capturing fine facial details. (c) Surface normals estimated via photometric stereo from the full OLAT stack, recovering pore-level geometry. (d) Image-based relighting result synthesized via linear superposition, illustrating the photometric consistency and dynamic range of our data.

synthesized results (d) preserve fine shading and self-shadowing details more faithfully than the evaluated neural relighting baselines.

5 Experiments

5.1 Implementation Details

Training Configuration. We perform full-parameter fine-tuning based on the Wan2.1 1.3B foundation model [Wan et al. 2025]. The model is trained for a total of 60,000 steps with a constant learning rate of 10^{-4} and a batch size of 8. During training, the target OLAT directions are randomly sampled from all available lighting directions. The probability of sampling static images for joint training follows a stepwise curriculum: it is set to 0.1, 0.2, and 0.3 for each consecutive 20,000-step interval. We use the same few-timestep sampling during training as Chadebec et al. [2025], where the discrete timesteps are sampled uniformly from the set $\{i/16\}_{i=1}^{16}$. For



(a) Results for Lighting Direction (b) Results on Lighting Intensity (c) Results on Real-world Lighting

Fig. 6. **Qualitative comparisons for video relighting.** We present results on our OLAT dataset under various environment maps.



Fig. 7. **Qualitative comparisons for in-the-wild video relighting.** We evaluate our method on (a) the Actor-HQ dataset [Işık et al. 2023] and (b) unstructured mobile phone captures.

the objective function, we utilize a VGG16-based LPIPS metric for the perceptual loss, setting the weight coefficient λ to 1.

Fixed Text Prompt. Throughout both training and inference, we utilize a fixed, descriptive text prompt to provide a consistent semantic context: *"Given: a reference video of a person and a lighting direction representing a specific point light source. Generate: a realistic relit video of the same person under the target lighting defined by one light. Requirements: preserve identity and appearance, maintain consistent clothing and proportions, correct perspective, realistic lighting, shadows, and no artifacts."*

Inference Configuration. During inference, we disable classifier-free guidance to rely primarily on our conditioning signals. To ensure temporal consistency, the recurrent bridge context is configured with a window length of $W = 9$ frames and a stride of $S = 4$, which is essential for stable video generation. We employ a uniform sampling strategy with 4 discrete timesteps, specifically $\{1, 0.75, 0.5, 0.25\}$. For each input sequence, we typically generate a set of 201 distinct OLAT videos to serve as the bases for relighting. These 201 bases preserve the zenith direction of the original 101-light configuration while doubling the azimuthal sampling density at each elevation. Additional results in the supplementary material show comparable quality under static lighting but smoother responses to rotating high-frequency illumination with 201 bases; 101 bases reduce generation and storage cost when finer sampling is unnecessary.

Hardware and Runtime. The entire training process is conducted on 8 NVIDIA A100 GPUs and takes approximately one week to complete. We additionally measure the inference cost of generating the OLAT basis set. For a 57-frame video at 1920×1088, generating one OLAT basis takes 119.45 seconds. When generation is distributed across eight A100 GPUs, producing 101 bases takes 26.89 minutes, while producing 201 bases takes 52.01 minutes. At 1280×720, generating one basis takes 61.08 seconds; with the same eight-GPU setup, producing 101 bases takes 14.53 minutes, while producing 201 bases takes 27.56 minutes. Because individual bases can be generated independently, they can be computed in parallel and cached for subsequent linear combinations without online network inference, although generation and storage costs scale approximately with their number.

5.2 Datasets and Metrics

We construct our training set by randomly selecting 100 identity-outfit combinations from the *Dynamic-subset* and 28 from the *Static-subset*. For evaluation, we hold out 54 and 7 identity-outfit combinations from the two subsets, respectively. For the *Dynamic-subset*, we use only views captured with short-lens cameras.

To ensure alignment with the pre-training distribution of the foundation model, all data are represented in the sRGB color space. Additionally, to isolate the subjects and eliminate background interference, we apply a uniform mask to all input frames, setting the background to black.

Following the evaluation protocol of He et al. [2024], we employ PSNR, SSIM, LPIPS [Zhang et al. 2018] (with an AlexNet backbone), and FLIP [Andersson et al. 2020] to quantitatively assess model performance.

5.3 Comparisons with Existing Methods

We compare our model with three video relighting approaches: the open-source RelightVid [Fang et al. 2025], the Cosmos-based DiffusionRenderer [Liang et al. 2025], and UniRelight [He et al. 2025]. To establish a comprehensive benchmark, we also include two recent single-image relighting methods: IC-Light [Zhang et al. 2025a] and LBM [Chadebec et al. 2025]. For the open-source baselines, we resize input videos or images to their recommended resolutions and adapt the lighting inputs to each method's requirements. For DiffusionRenderer and UniRelight, we use the full HDR environment map as the condition. For RelightVid, IC-Light, and LBM, which rely



Fig. 8. **Qualitative comparisons for image relighting.** Compared with the evaluated baselines, our result retains more of the input's fine-grained structure and more closely matches the captured OLAT-based ground truth. The top row highlights the specular response on the face and the shading contrast along the subject's rear arm.

on a background image, we sample the corresponding field of view from the HDR map and convert it to sRGB.

We qualitatively evaluate video relighting on our OLAT dataset and in-the-wild sequences in Figs. 6 and 7. Reference images are generated via image-based relighting from captured OLAT images, together with material balls rendered under the target environment maps.

Fig. 6 qualitatively compares responses to target direction (a), intensity distributions (b), and real-world lighting (c). Our results follow the target changes more closely than the evaluated baselines. The supplementary material further shows how our method responds to different intensity scalings of the same environment map.

To reduce the domain gap between the required flat-lit input and in-the-wild videos, we apply the delighting model described in the supplementary material before OLAT prediction. Consequently, beyond controlled setups, Fig. 7 highlights our strong generalization capabilities to novel, in-the-wild scenarios. Our model robustly preserves fine details and subject identity far better than the baselines. This superiority is evident in the fine-grained environmental interactions it synthesizes, such as the physically precise shadows in (a) and the accurate reflected bounce light from the background red curtain in the first frame of (b). Overall, these results confirm that our method synthesizes more photorealistic appearances and physically plausible lighting effects compared to existing approaches.

For quantitative lighting evaluation, we use the *Static-subset*, whose complete captured OLAT sequences allow us to synthesize ground-truth images under target environment maps via image-based relighting. In this setting, we focus our comparison on two representative image-based methods (IC-Light [Zhang et al. 2025a] and LBM [Chadebec et al. 2025]) and two video-based methods (DiffusionRenderer [Liang et al. 2025] and UniRelight [He et al. 2025]).

Image-relighting results are reported in Tab. 2 and Fig. 8, with an additional qualitative HumanOLAT comparison provided in the supplementary material. We resize all outputs and ground truth to the minimum evaluated resolution and mask their backgrounds. Following Jin et al. [2024], each predicted color channel is rescaled to the ground-truth mean to resolve global illumination–surface-color scale ambiguity. This calibration is used in quantitative evaluation to ensure a consistent comparison across methods. The results demonstrate that our method outperforms the baselines across most metrics and preserves significantly richer fine-grained details.

Following Lux Post Facto [Mei et al. 2025], we evaluate temporal consistency using LPIPS between adjacent frames and warping error (WE) computed with optical flow from RAFT [Teed and Deng 2020] and AllTracker [Harley et al. 2025]. Tab. 3 shows that our method achieves excellent temporal consistency.

5.4 Ablation Studies

We conduct three ablation studies to investigate the effectiveness of key components in our proposed framework. All quantitative evaluations are performed on a representative subset of 200 test sequences randomly sampled from our *Dynamic-subset*. Specifically, these sequences consist of paired flat-lit inputs and ground-truth

Table 2. **Quantitative comparison on image relighting.** To ensure a fair assessment, we employ a channel-aligned evaluation strategy to mitigate the inherent scale ambiguity between lighting intensity and surface color. Under this protocol, our method better matches the captured OLAT-based relighting ground truth across most metrics, indicating improved light-response fidelity.

Methods	PSNR↑	SSIM↑	LPIPS↓	FLIP↓
IC-Light [Zhang et al. 2025a]	19.51	.8760	.1322	.1226
LBM [Chadebec et al. 2025]	23.08	.9185	.1002	.0874
UniRelight [He et al. 2025]	22.91	.9070	.0928	.0889
DiffusionRenderer [Liang et al. 2025]	23.11	.9057	.0900	.0841
Ours	26.23	.9169	.0567	.0602

Best result Second-best result

Table 3. **Quantitative comparison of temporal consistency.**

Methods	LPIPS (temp.)↓	WE(RAFT)↓	WE(alltracker)↓
RelightVid [Fang et al. 2025]	.02763	.00107	.00112
UniRelight [He et al. 2025]	.02660	.00079	.00082
DiffusionRenderer [Liang et al. 2025]	.02629	.00078	.00080
Ours	.02532	.00065	.00060

Best result Second-best result

OLAT videos, and all reported results (both quantitative and qualitative) evaluate the quality of the synthesized OLAT videos. To ensure fair comparison while managing computational costs, all ablated versions are trained for 20,000 steps with a sampling ratio of 0.1 for images and 0.9 for videos. Consistent with the main experiments, since our model focuses exclusively on the foreground subject, all quantitative metrics are computed on background-masked images to prevent background noise from biasing the results.

Latent Bridge Matching for Input Frames. To validate the efficacy of the LBM paradigm [Chadebec et al. 2025], we compare it against standard diffusion conditioning strategies that rely on noise concatenation. Specifically, we implemented two baseline variants where the flat-lit input latents are concatenated with the noisy latents along either the channel dimension or the frame dimension. For the channel-wise concatenation baseline, we expanded the input channels of the first convolutional layer by a factor of two and initialized the weights of the newly added channels to zero for training stability. Quantitative results in Tab. 4 and qualitative comparisons in Fig. 9 reveal a significant performance gap. The noise concatenation methods tend to degrade source fidelity, often producing artifacts or losing fine-grained texture details due to the destructive nature of the initial Gaussian noise. In contrast, by establishing a direct transport trajectory from the source, bridge matching for video latents effectively preserves the structural integrity and high-frequency details of the input, yielding significantly higher fidelity in the final relit video.

Lighting Conditioning. We primarily evaluate our design against two alternative mechanisms: cross-attention and global AdaLN. The cross-attention method, following He et al. [2024], injects SH encodings by replacing the original text prompt embeddings, which risks disrupting the inherent input distribution of the backbone. The global AdaLN scheme employs a single shared projection layer for all blocks, offering only coarse, uniform control. In contrast, our block-specific AdaLN strategy preserves the integrity of the input features while enabling finer-grained, block-wise lighting modulation. Quantitative results in Tab. 4 and qualitative comparisons in Fig. 10 demonstrate that our proposed conditioning scheme achieves more precise directional control and yields more photorealistic predictions.

Table 4. **Quantitative comparison of ablated variants.** We evaluate OLAT prediction under alternative input-frame and lighting-conditioning designs. The full model achieves the best overall performance across the reported perceptual and structural metrics.

Methods	PSNR↑	SSIM↑	LPIPS↓	FLIP↓
Cross Attn	24.94	.9364	.0981	.0696
Global AdaLN	26.80	.9473	.0780	.0596
Channel Concat	27.13	.9434	.0815	.0589
Frame Concat	27.67	.9481	.0750	.0556
Ours	27.61	.9514	.0727	.0554

Best result Second-best result

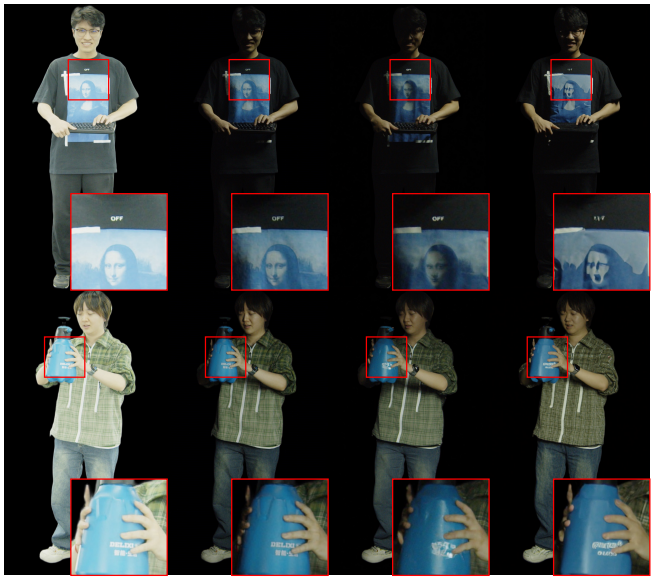


Fig. 9. **Ablation study on input frames processing strategies.** We compare our method against concatenation along the channel and frame dimensions. Results show that Latent Bridge Matching strategy excels at preserving input details.

Long Sequence Inference. We compare our recurrent bridge context with Position Interpolation (PI) [Chen et al. 2023] and sliding iterative sampling [Jin et al. 2025] in Fig. 11. PI introduces out-of-distribution positional embeddings, while heterogeneous latent timesteps cause significant lighting drift in sliding iterative sampling.

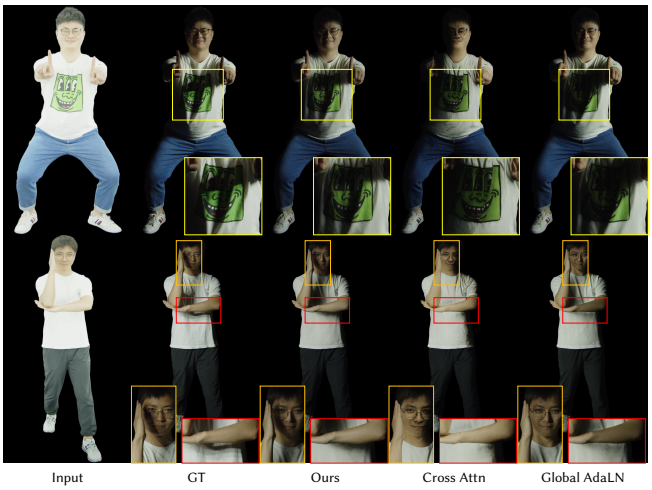


Fig. 10. **Ablation study on lighting conditioning mechanisms.** We compare cross-attention, global AdaLN, and the proposed block-specific AdaLN strategy. Our result better matches the target OLAT image in the shown shadow placement and local details.

5.5 Additional Qualitative Results

We further demonstrate the versatility and robustness of our model on unseen subjects. As illustrated in Fig. 12, our model synthesizes precise shading and self-shadowing effects under continuously rotating lighting directions, maintaining physical plausibility. Fig. 13 validates temporal stability, showcasing a dynamic performance coherently illuminated by a fixed directional light source across the sequence. Finally, Fig. 14 highlights the geometric consistency of our predictions, successfully generalizing to diverse camera viewpoints for a specified lighting direction. For more extensive visualization results and dynamic comparisons, please refer to our supplementary video.

6 Conclusion

In this work, we have presented a comprehensive framework for controllable dynamic full-body relighting from monocular flat-lit videos. By shifting the paradigm from direct relit-image generation to data-driven OLAT video prediction, our approach decomposes human light transport into directional basis responses and reconstructs the final appearance through linear-space intensity weighting. This formulation makes direction explicit in the OLAT bases and relative intensity explicit in the environment-map weights while preserving appearance details and temporal coherence. The accompanying large-scale dynamic full-body OLAT dataset provides the directional supervision needed to learn these responses across diverse identities, outfits, motions, and views.

Current limitations include the flat-lit input assumption, sensitivity of delighting to strong shadows or colored illumination, and the reconstruction ceiling of the pretrained video VAE. The explicit OLAT representation also incurs basis-dependent generation and storage costs, while cross-OLAT and cross-view consistency remains implicit rather than architecturally guaranteed. Finally, the finite-area capture lights limit directional resolution; missing dynamic relighting ground truth and reliable pixel-level lighting-angle annotations further constrain quantitative evaluation. Further discussion and failure analysis are provided in the supplementary material.

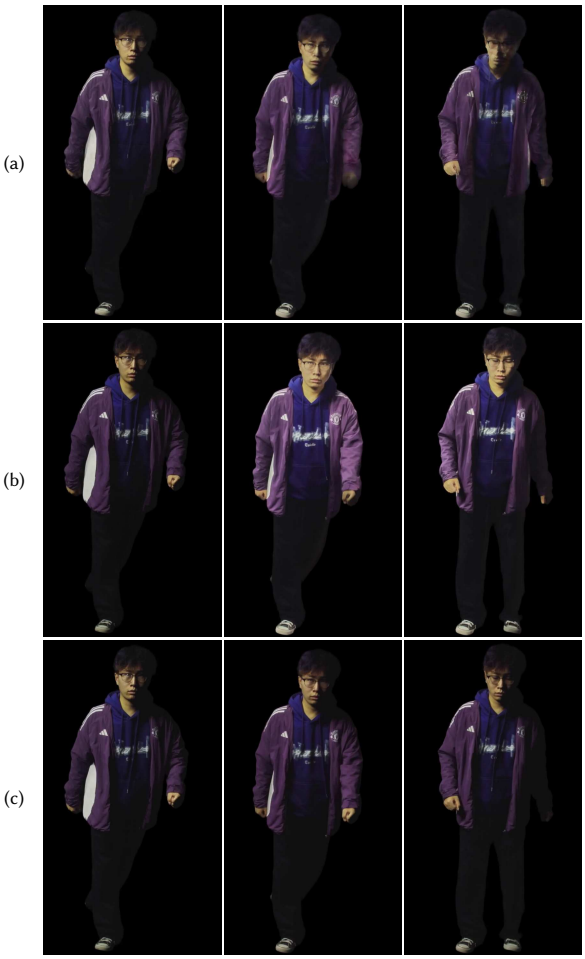


Fig. 11. **Ablation study on strategies for length extrapolation:** (a) position interpolation [Chen et al. 2023], (b) sliding iterative sampling [Jin et al. 2025], and (c) our recurrent bridge context. Results show that our method avoids the visible artifacts and lighting drift of the two alternatives.

These boundaries motivate more robust delighting, higher-fidelity video autoencoders, efficient basis prediction with explicit consistency, finer illumination capture, and dynamic lighting benchmarks. The dataset may further support neural rendering and controllable generative video editing.

Acknowledgments

We thank Dongli Tan for LightStage camera calibration and Linxian Wu for providing paired outdoor environment panoramas and reference images captured under the same lighting conditions.

This work was partially supported by National Key R&D Program of China (No. 2024YFB2809105), NSFC (No. U24B20154), Zhejiang Provincial Natural Science Foundation of China (No. LR25F020003), and Information Technology Center and State Key Lab of CAD&CG, Zhejiang University.

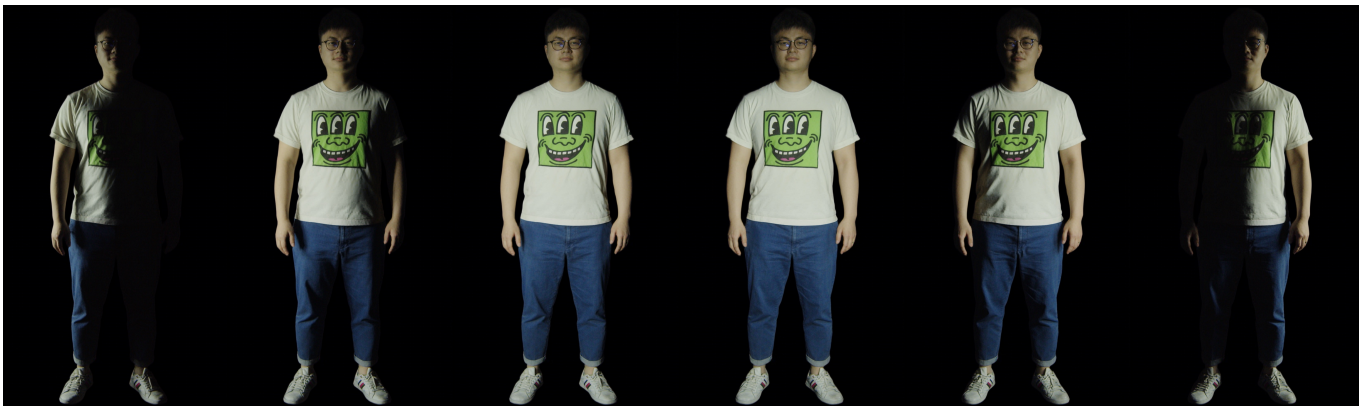


Fig. 12. **OLAT synthesis under rotating lighting directions.** We visualize the predicted OLAT results for a specific video frame of an unseen subject under varying lighting directions. Note the physically plausible movement of cast shadows and the preservation of high-frequency clothing details as the light source rotates.



Fig. 13. **Temporal consistency in dynamic OLAT generation.** Selected frames from a generated OLAT video sequence under a fixed directional light, sampled at 10-frame intervals. Our model maintains temporally coherent shading throughout the complex motion.



Fig. 14. **Generalization to arbitrary viewpoints.** We present OLAT predictions for the same subject captured from distinct camera views. For each view, the model successfully synthesizes the corresponding OLAT image under a specified lighting direction, demonstrating robust generalization across diverse observation viewpoints.

References

Pontus Andersson, Jim Nilsson, Tomas Akenine-Möller, Magnus Oskarsson, Kalle Åström, and Mark D Fairchild. 2020. FLIP: A Difference Evaluator for Alternating Images. *Proc. ACM Comput. Graph. Interact. Tech.* 3, 2 (2020), 15–1.

Jonathan T. Barron and Jitendra Malik. 2015. Shape, Illumination, and Reflectance from Shading. *TPAMI* (2015).

Weikang Bian, Xiaoyu Shi, Zhaoyang Huang, Jianhong Bai, Qinghe Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Hongsheng Li. 2025. RelightMaster: Precise Video Relighting with Multi-plane Light Images. *arXiv preprint arXiv:2511.06271* (2025).

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. 2023. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. arXiv:2311.15127 [cs.CV] https://arxiv.org/abs/2311.15127

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. 2024. Video generation models as world simulators. (2024). https://openai.com/research/video-generation-models-as-world-simulators

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).

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending Context Window of Large Language Models via Positional Interpolation. arXiv:2306.15595 [cs.CL] https://arxiv.org/abs/2306.15595

Yushuo Chen, Zerong Zheng, Zhe Li, Chao Xu, and Yebin Liu. 2024b. MeshAvatar: Learning High-quality Triangular Human Avatars from Multi-view Videos. arXiv:2407.08414 [cs.CV] https://arxiv.org/abs/2407.08414

Zhaoxi Chen and Ziwei Liu. 2022. Relighting4D: Neural Relightable Human from Videos. arXiv:2207.07104 [cs.CV] https://arxiv.org/abs/2207.07104

Zhaoxi Chen, Gyeongsik Moon, Kaiwen Guo, Chen Cao, Stanislav Pidhorskyi, Tomas Simon, Rohan Joshi, Yuan Dong, Yichen Xu, Bernardo Pires, et al. 2024a. URhand: Universal relightable hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 119–129.

Zhuo Chen, Chengqun Yang, Zhuo Su, Zheng Lv, Jingnan Gao, Xiaoyuan Zhang, Xiaokang Yang, and Yichao Yan. 2026. POLAR: A Portrait OLAT Dataset and Generative Framework for Illumination-Aware Face Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 28871–28881.

Jun Myeong Choi, Annie Wang, Pieter Peers, Anand Bhattad, and Roni Sengupta. 2024. ScribbleLight: Single Image Indoor Relighting with Scribbles. *arXiv preprint arXiv:2411.17696* (2024).

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, Westley Sarokin, and Mark Sagar. 2000. *Acquiring the Reflectance Field of a Human Face* (1 ed.). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3596711.3596762

Paul E Debevec. 2008. Performance relighting and reflectance transformation with time-multiplexed illumination. US Patent 7,436,403.

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 preprint arXiv:2501.16330* (2025).

Jian Gao, Chun Gu, Youtian Lin, Zhihao Li, Hao Zhu, Xun Cao, Li Zhang, and Yao Yao. 2024. Relightable 3d gaussians: Realistic point cloud relighting with brdf decomposition and ray tracing. In *European Conference on Computer Vision*. Springer, 73–89.

Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. 2019. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (ToG)* 38, 6 (2019), 1–19.

Mingtao Guo, Guanyu Xing, and Yanli Liu. 2025. High-Fidelity Relightable Monocular Portrait Animation with Lighting-Controllable Video Diffusion Model. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*. 228–238.

Adam W. Harley, Yang You, Xinglong Sun, Yang Zheng, Nikhil Raghuraman, Yunqi Gu, Sheldon Liang, Wen-Hsuan Chu, Achal Dave, Pavel Tokmakov, Suya You, Rares Ambrus, Katerina Fragkiadaki, and Leonidas J. Guibas. 2025. AllTracker: Efficient Dense Point Tracking at High Resolution. In *ICCV*.

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* (Tokyo, Japan) *(SA '24)*. Association for Computing Machinery, New York, NY, USA, Article 11, 12 pages. doi:10.1145/3680528.3687644

Shoukang Hu, Tao Hu, and Ziwei Liu. 2024. GauHuman: Articulated Gaussian Splatting from Monocular Human Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20418–20431.

Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. 2023. HumanRF: High-Fidelity Neural Radiance Fields for Humans in Motion. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–12. doi:10.1145/3592415

Chaonan Ji, Tao Yu, Kaiwen Guo, Jingxin Liu, and Yebin Liu. 2022. Geometry-aware Single-image Full-body Human Relighting. (October 2022).

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*.

Yudong Jin, Sida Peng, Xuan Wang, Tao Xie, Zhen Xu, Yifan Yang, Yujun Shen, Hujun Bao, and Xiaowei Zhou. 2025. Diffuman4d: 4d consistent human view synthesis from sparse-view videos with spatio-temporal diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11047–11057.

Andrew Jones, Andrew Gardner, Mark Bolas, Ian Mcdowall, and Paul Debevec. 2006. Simulating spatially varying lighting on a live performance. In *The 3rd European Conference on Visual Media Production (CVMP 2006)-Part of the 2nd Multimedia Conference 2006*. IET, 127–133.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* 42, 4 (2023), 139–1.

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*. 25096–25106.

Diederik P. Kingma and Max Welling. 2013. Auto-Encoding Variational Bayes. *CoRR* abs/1312.6114 (2013). https://api.semanticscholar.org/CorpusID:216078090

Junxuan Li, Chen Cao, Gabriel Schwartz, Rawal Khirodkar, Christian Richardt, Tomas Simon, Yaser Sheikh, and Shunsuke Saito. 2024. Uravatar: Universal relightable gaussian codec avatars. In *SIGGRAPH Asia 2024 Conference Papers*. 1–11.

Linzhou Li, Yumeng Li, Yanlin Weng, Youyi Zheng, and Kun Zhou. 2025. RGBAvatar: Reduced Gaussian Blendshapes for Online Modeling of Head Avatars. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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)*.

Yuanze Lin, Yi-Wen Chen, Yi-Hsuan Tsai, Ronald Clark, and Ming-Hsuan Yang. 2025. IllumiCraft: Unified Geometry and Illumination Diffusion for Controllable Video Generation. *arXiv preprint arXiv:2506.03150* (2025).

Isabella Liu, Linghao Chen, Ziyang Fu, Liwen Wu, Haian Jin, Zhong Li, Chin Ming Ryan Wong, Yi Xu, Ravi Ramamoorthi, Zexiang Xu, et al. 2023. Openillumination: A multi-illumination dataset for inverse rendering evaluation on real objects. *Advances in Neural Information Processing Systems* 36 (2023), 36951–36962.

Ropeway Liu, Hangjie Yuan, Bo Dong, Jiazheng Xing, Jinwang Wang, Rui Zhao, Yan Xing, Weihua Chen, and Fan Wang. 2025b. UniLumos: Fast and Unified Image and Video Relighting with Physics-Plausible Feedback. *arXiv preprint arXiv:2511.01678* (2025).

Yong Liu, Wenpeng Xiao, Qianqian Wang, Junlin Chen, Shiyin Wang, Yitong Wang, Xinglong Wu, and Yansong Tang. 2025a. DreamLight: Towards Harmonious and Consistent Image Relighting. *arXiv preprint arXiv:2506.14549* (2025).

Shengjie Ma, Yanlin Weng, Tianjia Shao, and Kun Zhou. 2024. 3D Gaussian Blendshapes for Head Avatar Animation. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers 24 (SIGGRAPH '24)*. ACM, 1–10. doi:10.1145/3641519.3657462

Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, Chenghui Li, Shih-En Wei, Rohan Joshi, Wyatt Borsos, Tomas Simon, Jason Saragih, Paul Theodosis, Alexander Greene, Anjani Josyula, Silvio Mano Maeta, Andrew I. Jewett, Simon Venshtain, Christopher Heilman, Yueh-Tung Chen, Sidi Fu, Mohamed Ezzeldin A. Elshaer, Tingfang Du, Longhua Wu, Shen-Chi Chen, Kai Kang, Michael Wu, Youssef Emad, Steven Longay, Ashley Brewer, Hitesh Shah, James Booth, Taylor Koska, Kayla Haidle, Matt Andromalos, Joanna Hsu, Thomas Dauer, Peter Selednik, Tim Godisart, Scott Ardisson, Matthew Cipperly, Ben Humberston, Lon Farr, Bob Hansen, Peihong Guo, Dave Braun, Steven Krenn, He Wen, Lucas Evans, Natalia Fadeeva, Matthew Stewart, Gabriel Schwartz, Divam Gupta, Gyeongsik Moon, Kaiwen Guo, Yuan Dong, Yichen Xu, Takaaki Shiratori, Fabian Prada, Bernardo R. Pires, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, and Yaser Sheikh. 2024. Codec Avatar Studio: Paired Human Captures for Complete, Driveable, and Generalizable Avatars. *NeurIPS Track on Datasets and Benchmarks* (2024).

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. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2025).

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas M. Lehrmann. 2020. Learning Physics-guided Face Relighting under Directional Light. arXiv:1906.03355 [cs.CV] https://arxiv.org/abs/1906.03355

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:245335086

Rohit Pandey, Sergio Orts Escolano, Chloe Legendre, Christian Häne, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. 2021. Total relighting: learning to relight portraits for background replacement. *ACM Trans. Graph.* 40, 4, Article 43 (July 2021), 21 pages. doi:10.1145/3450626.3459872

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. 2018. FiLM: Visual Reasoning with a General Conditioning Layer. In *AAAI*.

Pexels. 2026. Pexels Videos. https://www.pexels.com/videos/. Accessed August 10, 2026.

Pramod Rao, Abhimitra Meka, Xilong Zhou, Gereon Fox, Mallikarjun B R, Fangneng Zhan, Tim Weyrich, Bernd Bickel, Hanspeter Pfister, Wojciech Matusik, Thabo Beeler, Mohamed Elgharib, Marc Habermann, and Christian Theobalt. 2025. 3DPR: Single Image 3D Portrait Relighting with Generative Priors. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers (SA Conference Papers '25)*. Association for Computing Machinery, New York, NY, USA, Article 108, 12 pages. doi:10.1145/3757377.3763962

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*. 6452–6462.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]

Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. 2024. Relightable gaussian codec avatars. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 130–141.

Johannes Lutz Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kunwar Maheep Singh, Jianchun Chen, Vladislav Golyanik, Stephan J. Garbin, Thabo Beeler, Rishabh Dabral, Marc Habermann, and Christian Theobalt. 2025. Relightable Holoported Characters: Capturing and Relighting Dynamic Human Performance from Sparse Views. arXiv:2512.00255 [cs.CV] https://arxiv.org/abs/2512.00255

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

Mingxing Tan and Quoc V. Le. 2020. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv:1905.11946 [cs.LG] https://arxiv.org/abs/1905.11946

Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, et al. 2023. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–12.

Zachary Teed and Jia Deng. 2020. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. arXiv:2003.12039 [cs.CV] https://arxiv.org/abs/2003.12039

Timo Teufel, Pulkit Gera, Xilong Zhou, Umar Iqbal, Pramod Rao, Jan Kautz, Vladislav Golyanik, and Christian Theobalt. 2025. HumanOLAT: A Large-Scale Dataset for Full-Body Human Relighting and Novel-View Synthesis. In *International Conference on Computer Vision (ICCV)*.

Timo Teufel, Xilong Zhou, Umar Iqbal, Jan Kautz, Marc Habermann, Vladislav Golyanik, and Christian Theobalt. 2026. D-Rex : Diffusion Rendering for Relightable Expressive Avatars. *ArXiv* abs/2604.27871 (2026). https://api.semanticscholar.org/CorpusID:287914898

Marco Toschi, Riccardo De Matteo, Riccardo Spezialetti, Daniele De Gregorio, Luigi Di Stefano, and Samuele Salti. 2023. Relight my nerf: A dataset for novel view synthesis and relighting of real world objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 20762–20772.

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, et al. 2025a. Comprehensive Relighting: Generalizable and Consistent Monocular Human Relighting and Harmonization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 380–390.

Shaofei Wang, Tomas Simon, Igor Santesteban, Timur Bagautdinov, Junxuan Li, Vasu Agrawal, Fabian Prada, Shoou-I Yu, Pace Nalbone, Matt Gramlich, Roman Lubachersky, Chenglei Wu, Javier Romero, Jason Saragih, Michael Zollhoefer, Andreas Geiger, Siyu Tang, and Shunsuke Saito. 2025b. Relightable Full-Body Gaussian Codec Avatars. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '25)*. Association for Computing Machinery, New York, NY, USA, Article 146, 12 pages. doi:10.1145/3721238.3730739

Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo, Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, and Dahua Lin. 2024. HumanVid: Demystifying Training Data for Camera-controllable Human Image Animation. In *NeurIPS*.

Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T. Barron, and Aleksander Holynski. 2024. CAT4D: Create Anything in 4D with Multi-View Video Diffusion Models. *arXiv:2411.18613* (2024).

Zhen Xu, Sida Peng, Chen Geng, Linzhan Mou, Zihan Yan, Jiaming Sun, Hujun Bao, and Xiaowei Zhou. 2024a. Relightable and animatable neural avatar from sparse-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 990–1000.

Zhen Xu, Tao Xie, Sida Peng, Haotong Lin, Qing Shuai, Zhiyuan Yu, Guangzhao He, Jiaming Sun, Hujun Bao, and Xiaowei Zhou. 2023. EasyVolcap: Accelerating Neural Volumetric Video Research. (2023).

Zhen Xu, Yinghao Xu, Zhiyuan Yu, Sida Peng, Jiaming Sun, Hujun Bao, and Xiaowei Zhou. 2024b. Representing Long Volumetric Video with Temporal Gaussian Hierarchy. *ACM Transactions on Graphics* 43, 6 (November 2024). https://zju3dv.github.io/longvolcap

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).

Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. 2024. DiLightNet: Fine-grained Lighting Control for Diffusion-based Image Generation. In *ACM SIGGRAPH 2024 Conference Papers*.

Jianshu Zeng, Yuxuan Liu, Yutong Feng, Chenxuan Miao, Zixiang Gao, Jiwang Qu, Jianzhang Zhang, Bin Wang, and Kun Yuan. 2025. Lumen: Consistent Video Relighting and Harmonious Background Replacement with Video Generative Models. *arXiv preprint arXiv:2508.12945* (2025). https://arxiv.org/abs/2508.12945

Youyi Zhan, Tianjia Shao, Yin Yang, and Kun Zhou. 2025. Real-time High-fidelity Gaussian Human Avatars with Position-based Interpolation of Spatially Distributed MLPs. *arXiv preprint arXiv:2504.12909* (2025).

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2025a. 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. 2021. Neural video portrait relighting in real-time via consistency modeling. In *Proceedings of the IEEE/CVF international conference on computer vision*. 802–812.

Qian Zhang, Vikas Thamizharasan, and James Tompkin. 2024. Learning physically based material and lighting decompositions for face editing. *Computational Visual Media* 10, 2 (2024), 295–308. doi:10.1007/s41095-022-0309-1

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

Yuxin Zhang, Dandan Zheng, Biao Gong, Shiwen Wang, Jingdong Chen, Ming Yang, Weiming Dong, and Changsheng Xu. 2025b. LumiSculpt: Enabling Consistent Portrait Lighting in Video Generation. arXiv:2410.22979 [cs.CV] https://arxiv.org/abs/2410.22979

Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. 2024. Bilateral Reference for High-Resolution Dichotomous Image Segmentation. *CAAI Artificial Intelligence Research* 3 (2024), 9150038.

Taotao Zhou, Kai He, Di Wu, Teng Xu, Qixuan Zhang, Kuixiang Shao, Wenzheng Chen, Lan Xu, and Jingyi Yu. 2023. Relightable neural human assets from multi-view gradient illuminations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4315–4327.

Xilong Zhou, Jianchun Chen, Pramod Rao, Timo Teufel, Linjie Lyu, Tigran Minasian, Oleksandr Sotnychenko, Xiaoxiao Long, Marc Habermann, and Christian Theobalt. 2025. OLATverse: A Large-scale Real-world Object Dataset with Precise Lighting Control. *arXiv preprint arXiv:2511.02483* (2025).

Supplementary Material



(a) OLAT (b) Smooth RS (c) Flat (d) Sharp RS

Fig. 15. **Analysis of rolling shutter artifacts.** (a–c) Three consecutive frames contain an OLAT frame, a rolling-shutter (RS) frame, and a flat-lit frame. (d) An example of a "distinct" RS frame from the same sequence where the exposure boundary is clearly visible. Due to the vertical orientation of our sensors, the electronic rolling shutter scans horizontally (right-to-left). Ideally, this results in a sharp transition boundary as seen in (d). However, as shown in (b), high-frequency boundary lines are often smoothed out by video compression algorithms. Additionally, (a) illustrates a failure case for heuristic brightness-based detectors: direct opposing illumination causes pixel intensities to exceed those of the flat-lit frame, complicating simple thresholding methods.



(a) w/o Correction (b) w/ Correction (c) w/o Correction (d) w/ Correction

Fig. 16. **Evaluation of Color Correction.** (a) and (b) show the raw and color-corrected frames of the calibration sequence, respectively. (c) and (d) demonstrate the generalization of the computed CCM to a human subject within the same view. A significant artifact is observed in the white sleeve region in (c) and (d). The CCM causes originally neutral white pixels (near saturation) to shift significantly towards a darker, tinted hue, violating the scene's white balance. This severe deviation on neutral tones motivates our decision to keep CCMs optional.

A Dataset Statistics

Participants were 18–30 years old. Tab. 5 summarizes the combined dataset composition.

B Supplementary of Data Processing

B.1 Rolling Shutter Detection and Sequence Segmentation

The light-switch cycle is intentionally set slightly longer than two camera frame periods (approximately 33.3 ms at 60 fps), which facilitates rolling-shutter detection. In addition to the target flat-lit and OLAT frames, our raw footage inevitably contains frames affected by rolling-shutter (RS) artifacts. Both subsets rely on the precise identification of valid frames to correctly extract lighting sequences. However, identifying RS frames is non-trivial for the *Dynamic-subset*: due to H.264 compression, the distinct RS boundary is blurred by spatial artifacts (Fig. 15(b)), rendering traditional spatial stitching or gradient-based detection infeasible.

To distinguish valid frames from artifacts, we developed a deep learning-based classifier. We annotated 1,000 frames using three classes: flat-lit, OLAT, and rolling shutter. An EfficientNet-based [Tan and Le 2020] classifier was trained with label smoothing to mitigate the small sample size. Crucially, we modified the network to accept high-resolution inputs (512×512) to preserve the high-frequency features of the RS bands, avoiding aliasing during downsampling. This robust classifier achieves near-perfect accuracy, serving as the foundation for our sequence segmentation: For the *Dynamic-subset*, leveraging the frame-wise classifications, we segment the video stream into 51 discrete sequences, strictly adhering to the fixed structure of 18 interleaved flat-lit–OLAT pairs per sequence. For the *Static-subset*, the classifier identifies the extended intervals of flat-lit frames (approx. 120 frames) that act as delimiters between distinct static poses.

B.2 Optional Color Calibration.

Although we computed Color Correction Matrices (CCMs) using X-Rite charts, applying the full CCMs can introduce severe color casts, especially by shifting white regions toward darker hues (see Fig. 16). We therefore provide the CCMs as optional metadata rather than applying them to the primary training data.

C Inference-time Processing

OLAT post-processing. The dynamic range of generated videos differs from that of common Internet videos and may leave weak responses in OLAT bases that should be nearly dark. Therefore, after OLAT prediction and before linear environment-map assembly, we apply a deterministic foreground-aware post-processing step to the predicted OLAT videos. Within the foreground, we define brightness as the maximum RGB value and quadratically attenuate responses below 0.1. For directions prone to back-light leakage, we apply stronger sixth-power attenuation below 0.3 together with a 5×5 Gaussian blur. The directional risk score is centered at an azimuth of $-18°$, is maximal over $[-36°, 0°]$, and decreases with a cubic smoothstep to zero at an angular distance of $36°$; the zenith basis is excluded. We blend the default quadratic and stronger corrections according to this risk score while leaving pixels outside the foreground mask unchanged. This procedure reduces dark noise and spurious back-light leakage while preserving bright, valid light responses used for relighting.

Delighting for in-the-wild inputs. Since our main OLAT predictor is trained on controlled flat-lit videos, in-the-wild videos with cast shadows or colored illumination can introduce a domain gap. For in-the-wild testing, we first pass the input video through a lightweight

Table 5. **Combined dataset composition.** All counts refer to the 224 identity-outfit combinations, not to unique participants. For accessories and props, *None* and *Any accessory/prop* are mutually exclusive; indented rows are multi-label breakdowns of *Any accessory/prop* and may overlap.

Category	Count	Distribution
Gender		
Male	**135**	
Female	**89**	
Accessory / prop		
None	**118**	
Any accessory/prop	**106**	
Glasses	**87**	
Handheld prop	**20**	
Hat/cap	**8**	
Other accessory/prop	**8**	

Category	Count	Distribution
Top garment		
Outerwear	**93**	
Long-sleeve	**84**	
T-shirt	**38**	
Dress/one-piece	**5**	
Sleeveless	**3**	
Other	**1**	
Bottom garment		
Full-length pants	**208**	
Shorts	**11**	
Skirt/dress	**5**	

Category	Count	Distribution
Dominant top color		
White/light	**45**	
Gray	**40**	
Multicolor	**39**	
Black/dark	**33**	
Blue	**23**	
Yellow/brown	**22**	
Green	**11**	
Red/pink	**7**	
Purple	**4**	

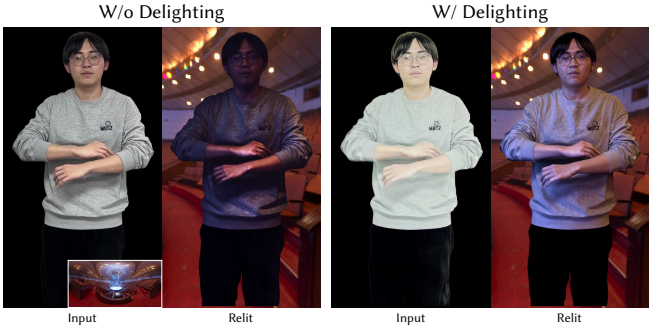


Fig. 17. **Effect of the delighting module.** The comparison illustrates that the module adapts in-the-wild inputs toward the flat-lit domain used by our OLAT predictor, making downstream relighting more consistent with the training setting.

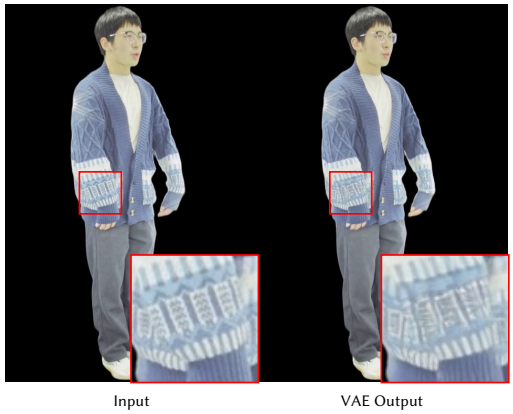


Fig. 18. **Reconstruction limitations of the video VAE.** We compare the ground-truth input frame (left) with the VAE reconstruction (right). The reconstruction exhibits a noticeable loss of high-frequency details, particularly in regions with rapid motion.

delighting module that maps uncontrolled illumination toward our flat-lit training domain. The module is implemented as a LoRA adaptation of Wan2.1 1.3B [Wan et al. 2025], trained as a video-to-video translation model using our generated relit videos as inputs and the corresponding flat-lit videos as targets. We use a rank-32 LoRA and a 16-step LBM schedule with bridge-noise standard deviation 0.001. Using AdamW with a constant learning-rate schedule, we train on more than 1,000 paired clips for 2,000 steps with a learning rate of 10^{-4} and a global batch size of 8. We train separate 57-frame video and single-frame variants under the same settings, and use four LBM steps at inference. As shown in Fig. 17, this preprocessing suppresses pre-existing illumination patterns before OLAT prediction, improving robustness on real-world videos without changing the core relighting model.

Foreground compositing. For all visualizations that composite relit subjects onto target backgrounds, we refine the foreground masks using BiRefNet's `refine_foreground` option [Zheng et al. 2024] before alpha compositing. This step improves high-resolution foreground boundaries and reduces dark contour artifacts around the subjects.

D Limitations and Failure Analysis

A key component of our framework is the latent video diffusion model, which relies on a pretrained video VAE for encoding and decoding frames. Consequently, output fidelity is bounded by the reconstruction quality of this VAE. We conducted an analysis to isolate this bottleneck. As illustrated in Fig. 18, while the VAE performs adequately for static or slow-moving content, it exhibits a noticeable loss of high-frequency details in scenarios involving rapid or large-scale motions (e.g., blurring of facial features or fabric textures). This degradation is inherent to the compression artifacts of the VAE and is independent of our relighting prediction. Future improvements in video autoencoder architectures will directly benefit our framework's output fidelity.

The flat-lit input assumption is practical for controlled production but limits direct use on arbitrary consumer footage. Although the delighting module reduces this domain gap, it remains preliminary: cast shadows or strong colored illumination may not be fully removed and can be interpreted by the OLAT predictor as material-

or geometry-dependent responses, propagating residual artifacts to the relit outputs. The in-the-wild results therefore demonstrate feasibility under varied inputs rather than uniform robustness to uncontrolled illumination.

Controllability through explicit OLAT bases also introduces computational and consistency tradeoffs. Generation time and storage grow approximately linearly with the number of lighting directions; parallel prediction reduces wall-clock time and cached bases can be reused, but neither removes the total prediction and storage burden. Because bases and views are generated independently, their consistency arises implicitly from the common flat-lit input and training data rather than from an explicit architectural constraint.

The capture system imposes complementary representation and evaluation limits. Each One-Light group spans a finite area, which smooths the incident lighting direction and may under-resolve sharp specular responses. Moreover, the system cannot record dynamic sequences under arbitrary environment illumination, leaving no dynamic ground-truth relit videos for target-light evaluation. Accordingly, static image metrics measure image-level agreement under target lighting, whereas the reported video metrics measure temporal smoothness rather than target-light correctness. Precise angular evaluation of directional control is also difficult because a single ground-truth lighting angle cannot be assigned reliably to the spatially varying shadows and highlights on a deforming human; we therefore do not report a shadow-direction offset metric.

E Additional Experimental Results

We provide additional qualitative results for the quality–cost tradeoff (Fig. 19), in-the-wild inputs (Fig. 20), long-sequence inference (Fig. 21), cross-dataset evaluation (Fig. 22), and relative intensity control (Fig. 23).

F Extended 4D Relighting

To probe the implicit consistency discussed above, we optimize a LongVolcap 4D Gaussian Splatting (4DGS) representation from independently relit multi-view videos [Xu et al. 2023, 2024b]. The plausible novel-view reconstruction in Fig. 24 suggests that the generated videos retain sufficient implicit multi-view coherence for this downstream example.

G Dataset overview

We provide several visualizations to illustrate the composition and quality of our dataset. Specifically, Fig. 25 and Fig. 26 show the basic training units: multi-view video pairs for the *Dynamic-subset* and OLAT images for the *Static-subset*. The multi-view snapshots for individual subjects are displayed in Fig. 27 (*Dynamic-subset*) and Fig. 28 (*Static-subset*). Furthermore, we present a gallery of all captured subjects in the *Dynamic-subset* and the *Static-subset* in Fig. 29 and Fig. 30.



Fig. 19. **Quality–cost tradeoff between 101 and 201 OLAT bases.** (a) Under a static environment map, 101 and 201 bases provide comparable visual quality. (b) For a rotating environment map, 201 bases respond more smoothly to lighting changes; this is visible in the evolution of the clothing shadow between Frames 44 and 45.

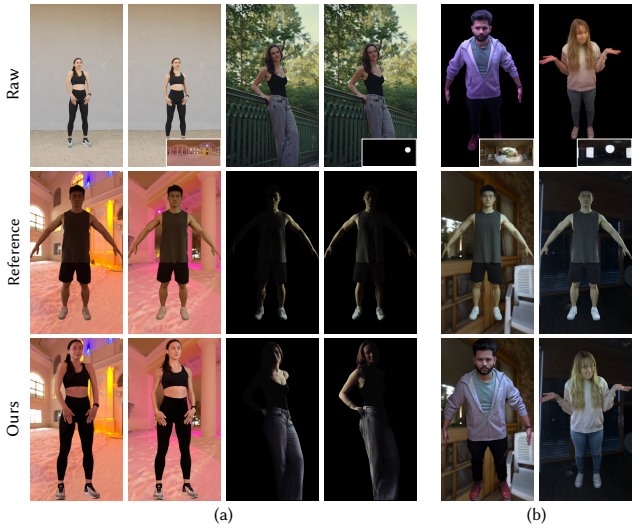


Fig. 20. **Additional in-the-wild relighting results.** (a) Video inputs sourced from Pexels [Pexels 2026] and HumanVid [Wang et al. 2024]. (b) Images captured under environment illumination from HumanOLAT [Teufel et al. 2025].



Fig. 21. **Long-sequence relighting.** We show two relit videos, each containing 917 frames, under two target environment maps. Both examples maintain stable subject appearance and lighting evolution over the complete sequences.



Fig. 22. **Cross-dataset qualitative comparison on HumanOLAT [Teufel et al. 2025].** All displayed method outputs are brightness-aligned to the corresponding ground truth for visualization.

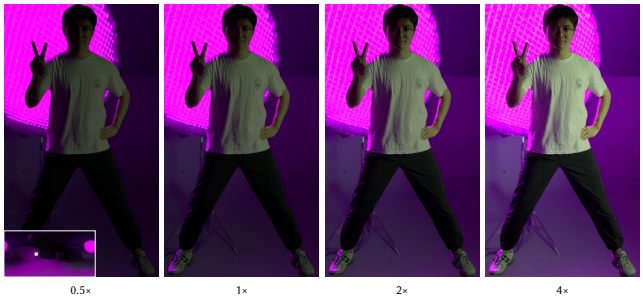


Fig. 23. **Relative intensity control under a fixed environment map.** Scaling the same environment map from 0.5× to 4× produces corresponding changes in the relit appearance.

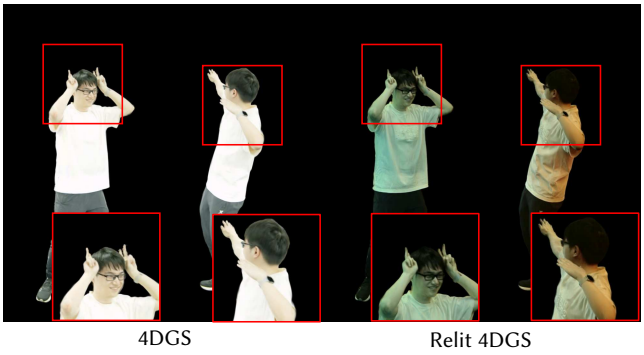


Fig. 24. **4DGS reconstruction from relit videos.** Left: Reconstruction from original flat-lit inputs. Right: Reconstruction from our relit outputs. The plausible novel view synthesis suggests that our method maintains sufficient implicit multi-view consistency to support downstream 3D/4D tasks.

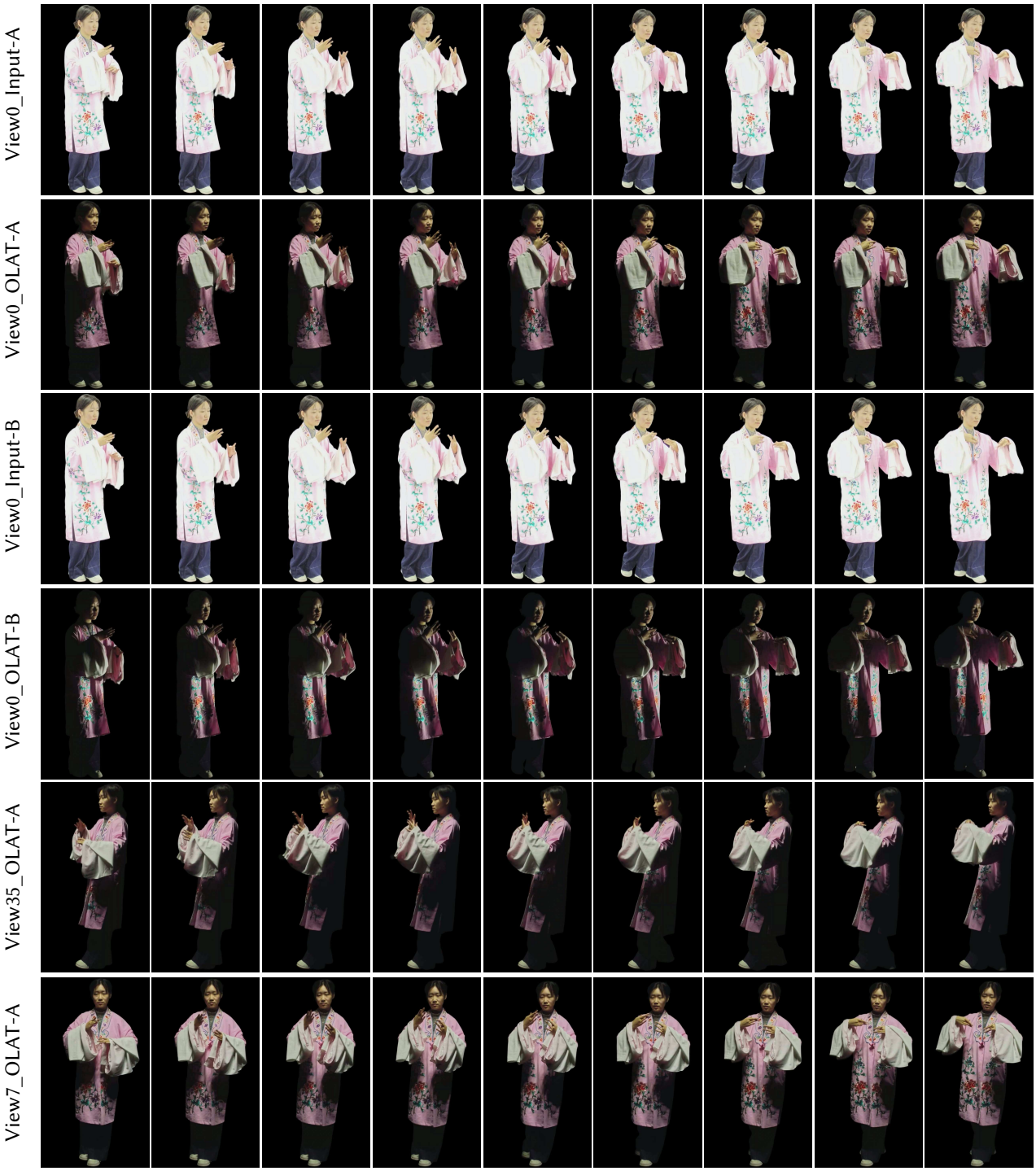


Fig. 25. **Examples of training pairs from the *Dynamic-subset*.** This figure shows the details of video training pairs in one sequence from several views.

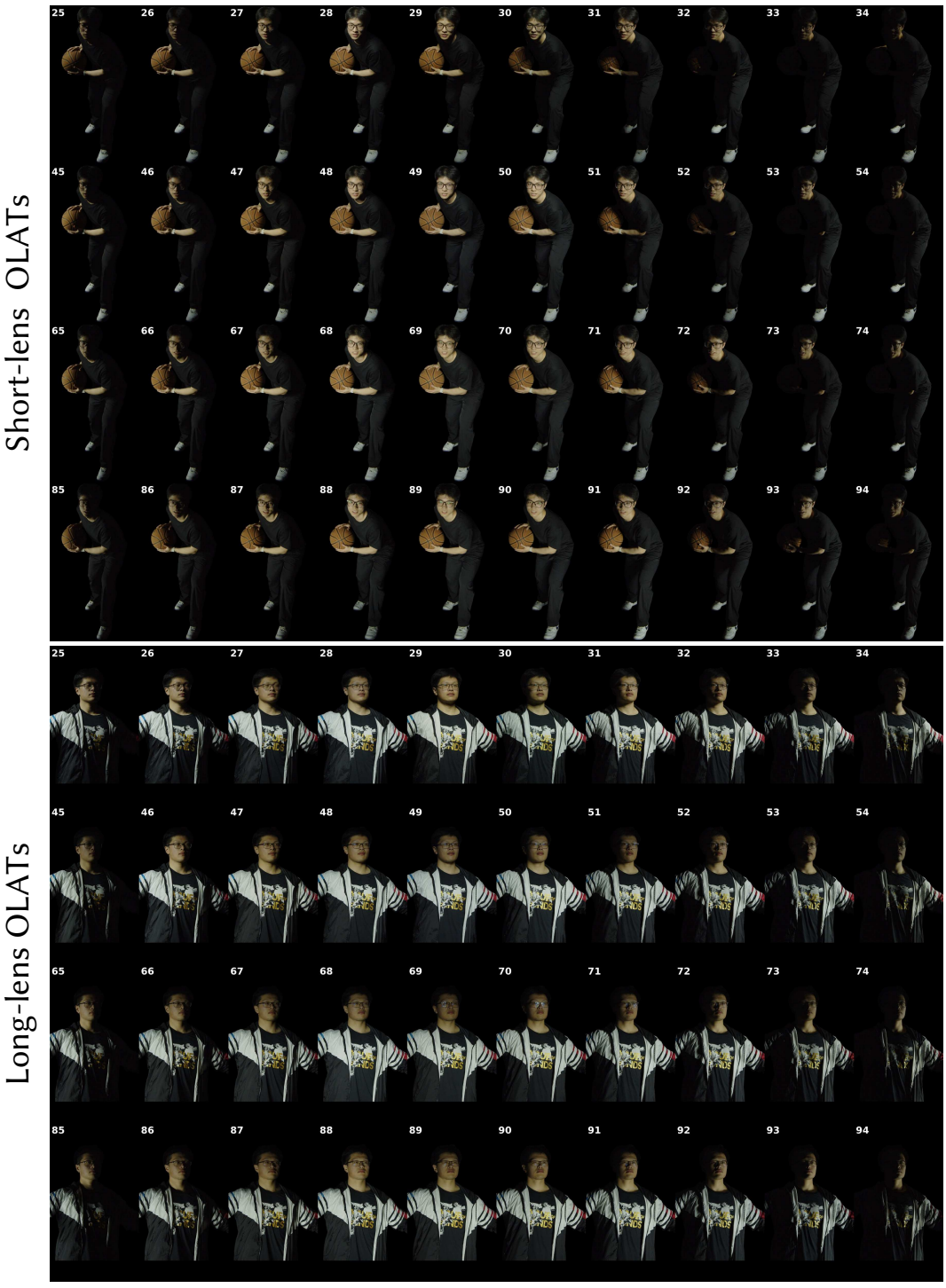


Fig. 26. **OLAT images from short-lens and telephoto cameras in the *Static-subset*.** The left-up corner number represent lighting direction index.

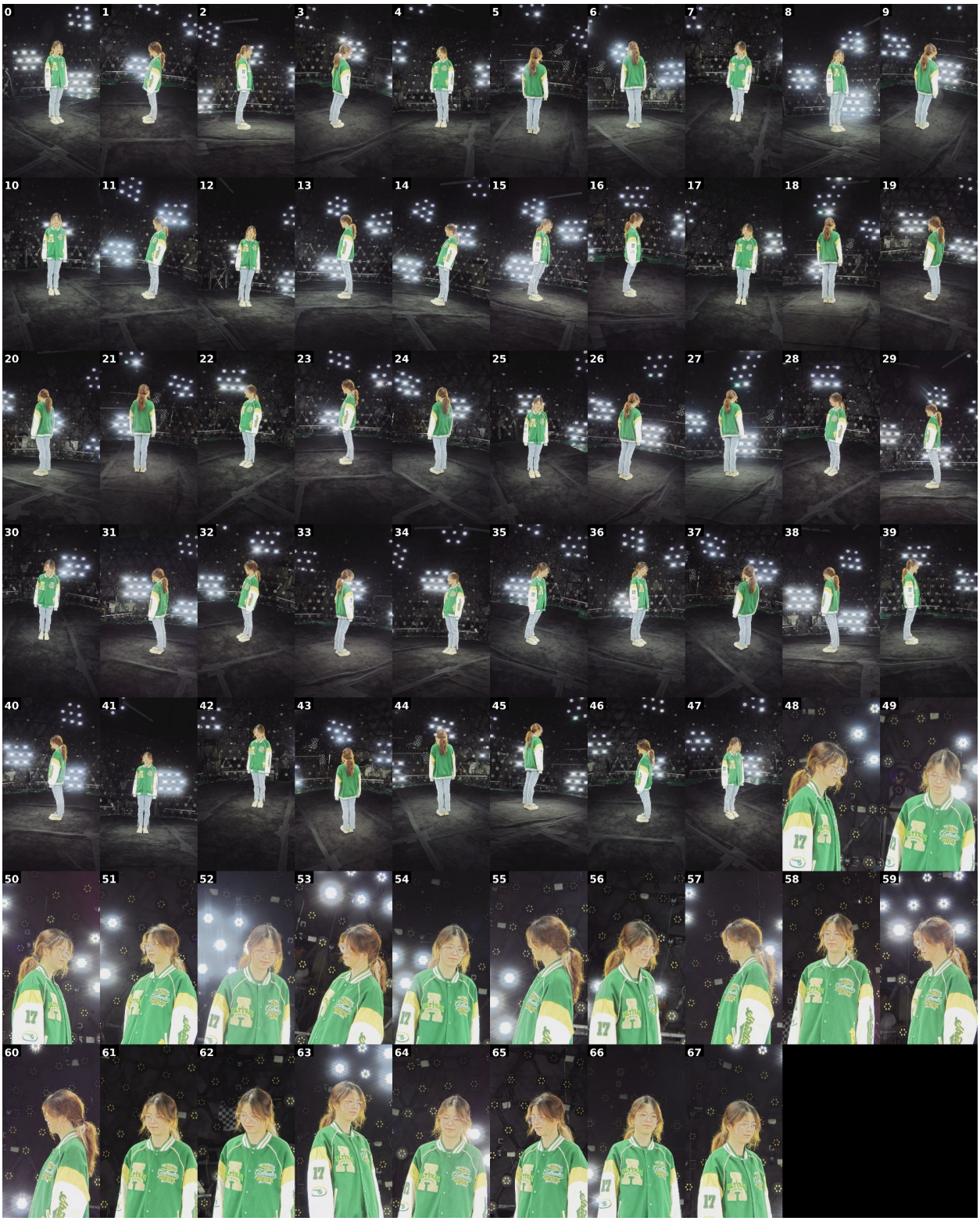


Fig. 27. Snapshots from all views of an example subject in the *Dynamic-subset*.



Fig. 28. Snapshots from all views of an example subject in the *Static-subset*.

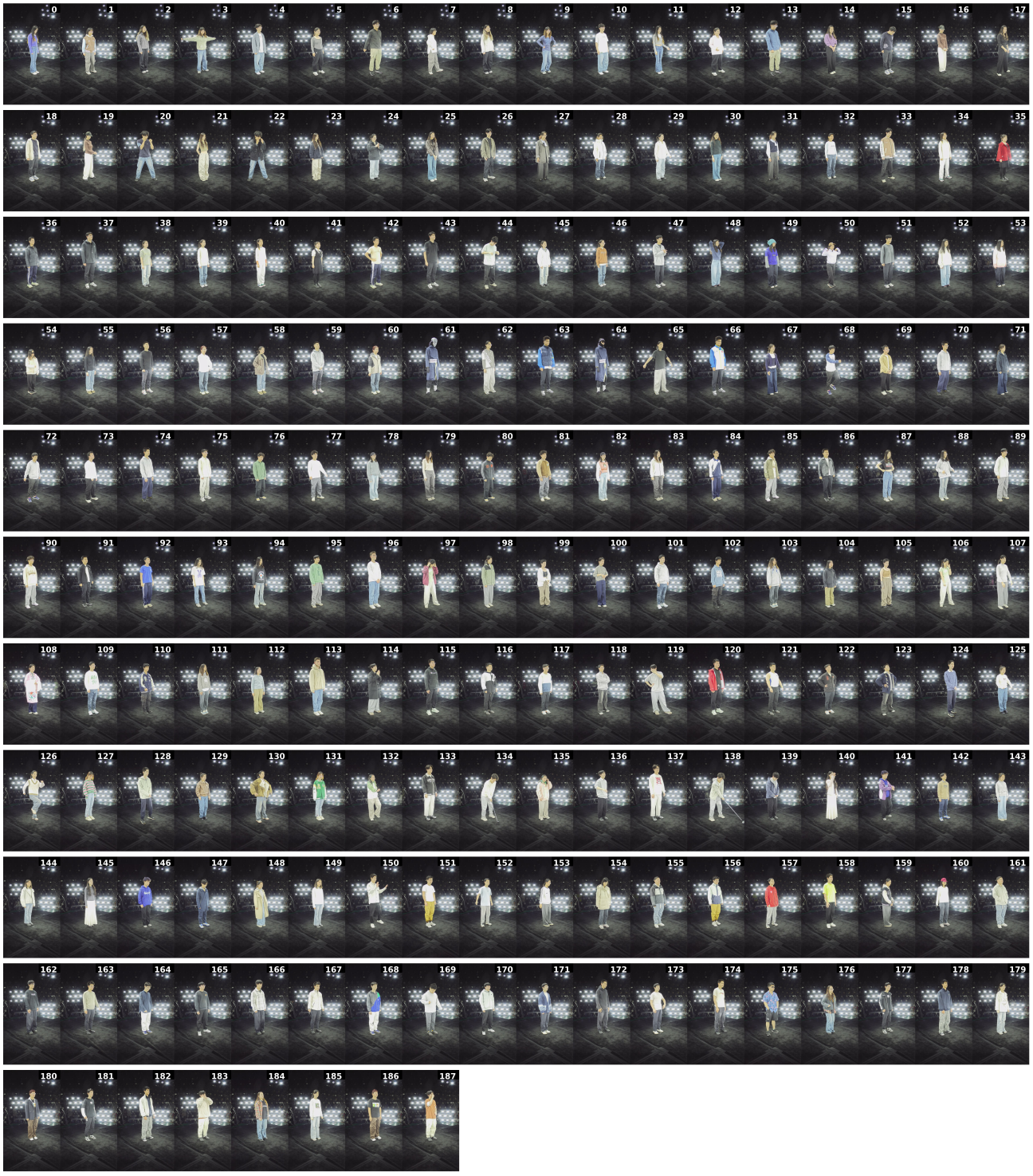


Fig. 29. Overview of all identity-outfit combinations in the *Dynamic-subset*.



Fig. 30. Overview of all identity-outfit combinations in the *Static-subset*.