

Generative Relightable Avatars

Kunwar Maheep Singh, Christian Theobalt, and Rishabh Dabral

Max Planck Institute for Informatics,
Saarland Informatics Campus
{ksingh,theobalt,rdabral}@mpi-inf.mpg.de

Abstract. We present Generative Relightable Avatars (GRA), a person-specific method for photorealistic free-view rendering and environment-map relighting of full-body humans. We postulate that modeling fine-grained appearance details is inherently a one-to-many problem that can benefit from a generative formulation. In contrast to fully regressive relightable avatar methods, GRA follows a hybrid approach that combines controllable, physics-grounded relighting with probabilistic refinement. Starting from a tracked animated mesh, we optimize material parameters in UV-space and render a coarse relit appearance under a target HDR environment map. Next, we refine the textures with a feed-forward model to capture pose-dependent texture dynamics and illumination effects beyond simplified reflectance assumptions. Finally, a fine-tuned video-to-video diffusion model transforms the physically grounded renderings into temporally coherent, high-detail videos while preserving 3D control, with an error-recycling strategy for generating long videos. Experimental evaluations demonstrate our method’s improved perceptual quality over prior relightable avatar baselines. We urge the readers to watch the supplementary video. See the project page for more details.

Keywords: Full body avatars · Relighting · Generative

1 Introduction

Developing photorealistic digital human avatars that can be realistically relit under arbitrary illumination remains a long-standing goal in computer vision and graphics, central to applications in telepresence, virtual production, and AR/VR. The challenge is particularly acute for full-body, clothed humans under dynamic motions, where minute appearance cues like wrinkles and shading critically influence realism.

Prior works on relightable 3D avatars largely fall into two categories. One line of work comprises deterministic pipelines that reconstruct a controllable 3D human representation, estimate explicit reflectance factors, and render novel views under novel illumination via physically based rendering or a feed-forward model [2, 3, 10, 19, 29, 30, 36]. While such methods provide strong 3D controllability, their inherent determinism and limited reflectance modeling restrict their ability to capture the one-to-many relationship between skeletal motion and fine appearance details, leading to over-smooth renderings. In contrast, a second line of methods adopts generative relighting in the image or video domain, producing convincing lighting changes [8, 20] but typically lacking physical consistency and explicit 3D control, limiting their applicability for free-viewpoint rendering.

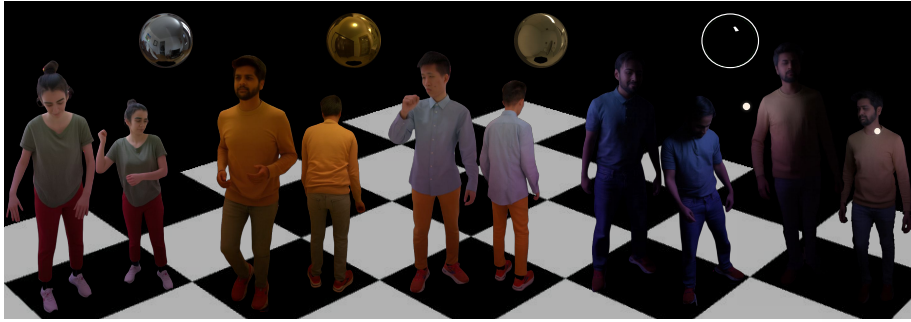


Fig. 1: We present *Generative Relightable Avatars*, a novel method that enables photorealistic relighting of full-body humans in motion. From pose, viewpoint, and illumination as input, our method generates temporally consistent, physically grounded renderings under novel lighting, which also qualitatively generalizes to out-of-distribution one-light-at-a-time (OLAT) setups and near-field lighting effects. This enables the realistic placement of dynamic humans into arbitrary virtual environments.

To address these challenges, we introduce Generative Relightable Avatars (GRA), a relightable framework for generating pose-driven photorealistic renderings under arbitrary viewpoints and illumination. Rather than asking a single generative model to jointly infer geometry, motion, viewpoint, and lighting, we adopt a coarse-to-fine decomposition: (i) an explicit 3D avatar provides controllable, physically grounded intermediate renderings under the target environment map, and (ii) a conditional video generative model probabilistically synthesizes the fine-scale details that are inherently stochastic and ambiguous under imperfect tracking, as shown in Fig. 1.

Concretely, we first train a person-specific parametric avatar that maps tracked 3D motion and camera parameters to a temporally consistent animated mesh. We optimize person-specific material parameters in UV space and render the mesh under a target environment map using a microfacet reflectance model [4]. A lightweight neural model, RelightNet, then captures illumination effects beyond the simplified reflectance assumptions and models pose-dependent texture dynamics in UV space. By construction, this stage supports novel viewpoint synthesis, yielding physically grounded renderings that capture the overall lighting response but remain visually coarse.

In the second stage, we finetune a video generative model to resolve the one-to-many ambiguity in high-frequency appearance by formulating it as a video-to-video task over the coarse physically-based renderings. We introduce an Atemporal Cross-Attention block to inject illumination conditioning by cross-attending per-frame environment maps with the corresponding video frames. This encodes per-frame lighting correspondence as an inductive bias, motivated by our training data where every frame is captured under a distinct illumination. To support long dynamic motions beyond the limited generation horizon of current video diffusion models, we integrate an error-recycling strategy [15] that maintains temporal consistency over substantially longer sequences.

In summary, our contributions are threefold.

- We propose a system for relightable human avatars that combines physical plausibility with photorealism through a hybrid explicit and generative pipeline, enabling the sampling of diverse wrinkle patterns.
- A video relighting generative model that introduces novel lighting controls to a video-to-video generative model, while ensuring long-form video generation
- We achieve state-of-the-art results with qualitative out-of-distribution performance on unseen environment illumination, including OLAT sequences.

2 Related Work

2.1 3D Full-body Relightable Avatars.

Recovering a relightable full-body human avatar from images has been widely studied, with approaches differing primarily in their assumptions about illumination calibration, sensor setup, and appearance representation. Most methods formulate the problem as inverse rendering, where geometry, material, and illumination parameters are optimized through a differentiable rendering pipeline to match observed images.

Inverse Rendering under Uncalibrated Illumination. A large body of work targets relightable avatar reconstruction from casually captured images under unknown lighting [2, 3, 9, 10, 29, 35, 36, 41], where single-view or multi-view observations are used to jointly estimate environment illumination along with dynamic geometry and material properties. However, recovering illumination jointly with material and geometry introduces severe ambiguities, often necessitating strong priors or simplified reflectance models. To make the problem tractable, most approaches adopt analytic BRDFs such as microfacet [3, 19], Disney [29], or Lambertian models [9, 34], and represent appearance using implicit neural fields defined in a canonical pose and deformed into observation space via skinning [3], enabling differentiable rendering but complicating accurate visibility reasoning. Subsequent works improve shading fidelity by combining explicit meshes with implicit material fields [2], allowing more reliable surface normals and view-dependent effects. Significant effort has also been devoted to accelerating visibility and light transport computation; for example, Xu et al. [36] introduce hierarchical distance queries to efficiently compute ray-surface intersections in deforming implicit fields, while Zhao et al. [41] use surfel-based representations to speed up visibility estimation. Despite these advances, uncalibrated inverse rendering methods often suffer from residual lighting-material ambiguity, and render over-smoothed appearance.

Relighting with Calibrated Illumination. To alleviate illumination ambiguity, several methods assume partially or fully calibrated lighting during capture. Relightable Neural Actor [19] leverages a small set of known illumination conditions, simplifying optimization and improving shading consistency. More recently, Relightable Full-body Gaussian Codec Avatars [30] exploit a lightstage setup with hundreds of illumination conditions, obtained via randomized groupings of OLAT measurements, to densely sample the illumination basis space and tightly constrain material estimation. While calibrated lighting significantly improves physical correctness, these methods still rely on deterministic appearance models and remain sensitive to tracking errors and limited pose coverage.

2.2 Generative Relighting Methods.

Generative priors have been transformative in estimating the geometry [25, 33], and material [32, 38] properties of objects from input images. Image-space diffusion models have been successfully adapted to perform scene relighting [5, 12, 39], portrait relighting [1, 14, 21, 22] and associated tasks such as shadow editing and harmonization [23, 37]. More recent works like Comprehensive Relighting address the broader human setting with temporal input and background harmonization [27]. Closely related to our setting are the works that perform video relighting via video diffusion models or diffusion-based neural rendering [16, 20, 28]. While these approaches produce highly photorealistic lighting effects, they typically operate in the image plane and, therefore, offer limited explicit 3D control. In contrast, our work couples an explicit, pose- and view-controllable full-body relightable avatar with a conditional video generative refinement model.

3 Method

Our goal is to perform physics-guided, photorealistic relighting of a human character performing motion θ , rendered from arbitrary viewpoint κ under arbitrary environment lighting $\mathbf{E}$. As motion is a sparse input signal, rendering fine-grained appearance details such as wrinkles is inherently a one-to-many task: the same skeletal motion can explain a variety of wrinkle patterns and appearance effects. This motivates us to seek a generative solution to the task, where the character’s appearance, $\mathbf{z}$, can be modeled as a conditional distribution $p(\mathbf{z}|\theta, \kappa, \mathbf{E})$ over the skeletal motion, camera viewpoint and the environment illumination.

However, jointly learning the conditional probability distribution above is challenging as the model needs to precisely disentangle the effects of illumination, viewpoint and motion. Therefore, we follow a divide-and-conquer approach and reformulate the task as $p(\mathbf{z}|\theta, \kappa, \mathbf{E}) = p(\mathbf{z}|\mathbf{z}', \mathbf{E})$, where $\mathbf{z}' = q(\theta, \kappa, \mathbf{E})$ is a deterministic function representing the intermediate, coarse renderings of a mesh avatar illuminated with $\mathbf{E}$ and rendered from the target viewpoint κ . This simplifies the generative task to that of modeling the fine-grained details like wrinkles, and generating realistic illumination-dependent effects like shadows and specularities from coarse guidance. In practice, we model $p(\mathbf{z}|\mathbf{z}', \mathbf{E})$ by adapting a video-to-video generative model to accept illumination conditioning. The intermediate representation $\mathbf{z}' = q(\theta, \kappa, \mathbf{E})$ is modeled by taking a person-specific 3D avatar [7] and making it relightable by optimizing material properties under approximate reflectance assumptions, followed by learning pose and illumination-dependent effects with a feed-forward neural network.

Fig. 2 outlines our overall schema. Following a coarse-to-fine strategy, we first use a mesh-based 3D avatar that predicts a coarse surface from skeletal motion. On this surface, an explicit physics-based texture produces an approximate relighting of the character under novel illumination (Sec. 3.1). A subsequent feed-forward network, RelightNet, then approximates illumination-dependent effects beyond the coarse microfacet approximation, yielding a coarse, but physically grounded relit appearance (Sec. 3.2). Fine-scale details are modeled probabilistically using a generative model for refinement (Sec. 3.3), capturing the inherent ambiguity in mapping pose to high-frequency appearance variations.

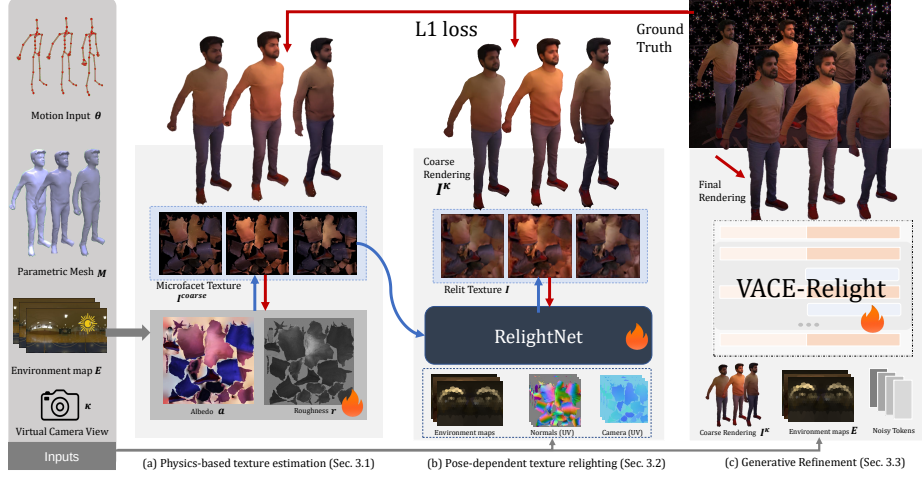


Fig. 2: Overview. Given skeletal motion θ , viewpoint κ , and environment illumination E , GRA outputs a photorealistic rendering of the character avatar in three steps: (a) A person-specific animatable mesh predicts pose-dependent geometry M , relit under a microfacet BRDF to produce a coarse appearance (Sec. 3.1), which is refined by (b) RelightNet to capture illumination- and pose-dependent effects (Sec. 3.2). (c) Finally, a finetuned video diffusion model synthesizes fine-scale details conditioned on the coarse renderings and per-frame environment maps (Sec. 3.3). Blue arrows indicate the inference path; red arrows indicate gradient propagation.

3.1 Physics-based Texture

We begin with a person-specific parametric body model [7] that is trained to produce a temporally consistent character mesh M_t , given the input skeletal motion θ_t at time t . The parametric model is also trained to produce dynamic, view-dependent textures, but is agnostic to the target illumination E . Therefore, our first goal is to obtain a physics-guided approximation of the character’s appearance under novel lighting that can effectively guide subsequent *learned* relighting. To this end, we introduce an explicit, physically motivated texture that approximates the rendering equation for a deforming human surface.

The outgoing radiance L_o at a surface point x with outgoing direction ω_o is given by the rendering equation [13]:

$$L_o(x, \omega_o) = \int_{\omega_i} f_r(x, \omega_i, \omega_o) L_i(x, \omega_i) V(x, \omega_i) \langle \omega_i, \mathbf{n} \rangle d\omega_i, \quad (1)$$

where ω_i is the incoming light direction, $\mathbf{n}$ is the normal at x , f_r is the surface BRDF, L_i is the incoming light intensity and V indicates visibility. We model the surface reflectance using a simplified Cook–Torrance microfacet BRDF [4], parameterized by spatially varying albedo $\mathbf{a}$ and roughness $\mathbf{r}$ maps:

$$f_r(x, \omega_i, \omega_o) = \text{BRDF}(\omega_i, \omega_o; \mathbf{a}(x), \mathbf{r}(x)). \quad (2)$$

As the albedo and roughness maps are unknown, we optimize for them by evaluating Eq. 1 in the UV space of M_t under a one-bounce assumption using the known mesh geometry, camera parameters κ , and environment map $\mathbf{E}$:

$$\mathbf{a}^*, \mathbf{r}^* = \arg \min_{\mathbf{a}, \mathbf{r}} \sum_t \|\mathcal{P}(M_t, \kappa, \mathbf{E}; \mathbf{a}, \mathbf{r}) - \mathbf{I}_t^{\text{gt}}\|_1 \quad (3)$$

where $\mathcal{P}(\cdot)$ denotes the differentiable rendering of the mesh under the microfacet BRDF, and $\mathbf{I}_t^{\text{gt}}$ is the ground-truth image. This optimization yields a coarse UV-space texture $\mathbf{I}_t^{\text{coarse}}$ under arbitrary illumination (see Fig. 2).

While approximate, $\mathbf{I}_t^{\text{coarse}}$ encodes physically meaningful dependencies on geometry, view direction, and illumination, and serves as a strong physics-guided inductive bias for subsequent learned relighting.

3.2 RelightNet

The explicit physics-based texture $\mathbf{I}_t^{\text{coarse}}$ estimated above provides a physically grounded but coarse approximation of surface appearance as can be seen in Fig. 2. This approximation has two key limitations: (i) due to the restricted expressiveness of the microfacet BRDF and the single-bounce assumption, it does not span the full space of relit appearances observed in real data; and (ii) the use of a static UV texture on an imperfectly tracked deforming mesh leads to pose-dependent smoothing of high-frequency appearance.

To address these limitations, we introduce *RelightNet* $\mathcal{R}$, a feed-forward neural model that approximates the complete rendering equation in the UV space of the tracked mesh. RelightNet learns a mapping from illumination, geometry, and view configuration to the final relit mesh texture $\mathbf{I}_t$ as:

$$\mathbf{I}_t = \mathcal{R}(\mathbf{E}_t, \mathbf{I}_t^{\text{coarse}}, \mathbf{N}_{t-k:t}, \mathbf{C}_t), \quad (4)$$

where $\mathbf{E}_t$ denotes the environment illumination, $\mathbf{I}_t^{\text{coarse}}$ is the coarse relit texture from Sec. 3.1, $\mathbf{N}_{t-k:t}$ is the set of surface normal maps with a history of k frames, and $\mathbf{C}_t$ encodes the per-texel view direction to the camera center. The relit texture $\mathbf{I}_t$ is projected back onto the mesh and rasterized into camera κ to get the relit rendering $\mathbf{I}_t^\kappa$.

RelightNet follows the UNet-based architecture proposed in RHC [24] and operates entirely in the UV texture space. It uses cross-attention to condition the texture prediction on the environment map $\mathbf{E}$, enabling efficient relighting of dynamic human characters under arbitrary illumination. The predicted texture is trained using L1 loss with ground truth images. Unlike RHC [24], which leverages image observations to transfer details in texture space, we rely only on the coarse geometry and derived features, recovering fine details generatively (Sec. 3.3).

3.3 Generative Refinement

While RelightNet produces physically grounded appearance, its deterministic output $\mathbf{I}_t^\kappa$ tends to regress to the mean for stochastic high-frequency details such as dynamic wrinkles. To recover these details, we formulate a probabilistic

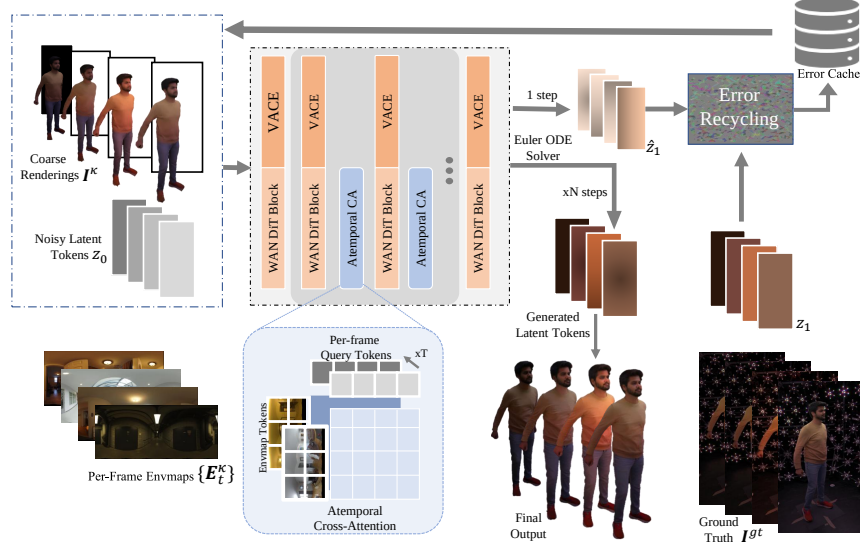


Fig. 3: Generative Refinement. Given the coarse renderings of the character from a target viewpoint, we finetune the WAN2.1-VACE architecture to produce fine-grained renderings. Environment illumination is introduced through Atemporal Cross-Attention that performs per-frame attention between Envmap tokens and video tokens.

solution by introducing a generative refinement stage that treats the final appearance as a conditional distribution. We leverage a pretrained video generation backbone based on *Latent Flow Matching (LFM)* [17], adapted to incorporate lighting controls. Importantly, we do so in the image space instead of the UV texture space to best leverage the learned priors of the pre-trained video models.

We operate in the latent space $\mathcal{Z}$ of a pre-trained 3D VAE, where a video sequence $\mathcal{V}$ is encoded as latent $z = \mathcal{E}(\mathcal{V})$ via the VAE encoder. LFM defines a velocity field v that describes the probability path $p_s(z)$ between an initial noise distribution $p_0 = \mathcal{N}(0, \mathbf{I})$ at flow-step $s = 0$ and the target data distribution p_1 at $s = 1$. Following LFM [17], we define a linear probability path: $z_s = (1 - s)z_0 + sz_1$, where z_1 is the encoded clean latent and $z_0 \sim p_0$ is Gaussian noise. The velocity field $v(z_s, s, \mathbf{c})$ is modeled by a neural network trained to regress the field by minimizing:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_s \|v(z_s, s, \mathbf{c}) - (z_1 - z_0)\|^2, \quad (5)$$

where $s \sim \mathcal{U}[0, 1]$ and $\mathbf{c}$ is the conditional context. During inference, we solve the Ordinary Differential Equation (ODE) from $s = 0$ to $s = 1$ using an ODE solver (e.g., Euler). The final conditionally generated video is recovered via the VAE decoder $\hat{\mathcal{V}} = \mathcal{D}(z_0)$.

In our case, the conditioning set $\mathbf{c} = \{\mathbf{I}^K, \mathbf{E}\}$ comprises of the coarse image-space renderings of the RelightNet output and the target illumination. Since the coarse renderings $\mathbf{I}^K$ are pixel-aligned with the target photorealistic output, this can be formulated as a video-to-video generation task, allowing us to use

WAN2.1-VACE [11] as the pretrained backbone. Built on top of WAN2.1 [26], WAN2.1-VACE is a video-to-video model that allows video generation with spatio-temporal control derived from a given reference video. In addition, we propose to use atemporal cross-attention to introduce the per-frame environment maps $\mathbf{E} \in \mathbb{R}^{T \times 256 \times 512 \times 3}$ into the WAN2.1 architecture.

VACE Control. To introduce precise spatio-temporal control from the RelightNet output, the WAN DiT backbone is augmented using VACE [11], which extends the base model with adapter layers that accept a reference video as a spatio-temporal conditioning signal.

Central to VACE is a *concept decoupling* mechanism that separates the conditioning signal into two complementary components via a Video Control Unit (VCU). Given the coarse RelightNet rendering $\mathbf{I}^\kappa \in \mathbb{R}^{T \times H \times W \times 3}$ and a per-frame binary mask $\mathbb{M} \in \mathbb{R}^T$, frames are partitioned into *reactive* frames, which the model actively refines, and *inactive* frames, which serve as fixed temporal anchors:

$$F_{\text{reac}} = \mathbf{I}^\kappa \odot \mathbb{M}, \quad F_{\text{inac}} = \mathbf{I}^\kappa \odot (1 - \mathbb{M}). \quad (6)$$

The reactive component captures the frames to be refined under the target illumination, while the inactive component provides stable spatial and temporal context that anchors the generation and prevents content drift. Both components, together with the mask $\mathbb{M}$, are encoded via the 3DVAE encoder and concatenated channel-wise to form the VCU input. This is processed by dedicated attention blocks within the VACE adapter and fused into each DiT block of the WAN2.1 backbone as an additive residual, propagating the spatio-temporal conditioning signal throughout the full depth of the network.

In addition to the coarse renderings, we provide a reference image $\mathbf{I}^{\text{ref}}$ of the subject captured under neutral illumination from the training set, which is passed to the VACE VCU alongside the coarse renderings. We append one additional token to the DiT corresponding to this reference image, with a velocity target defined from noise to its encoded latent. Since the model must reproduce the reference image at this token, it is encouraged to attend to the subject’s identity-specific appearance independently of the target lighting condition.

During training, we leverage this decoupling to teach the model temporal continuation. Specifically, we replace the first 5 frames of $\mathbf{I}^\kappa$ with ground-truth frames with probability 40%, marking them as inactive with $\mathbb{M}_{1:5} = 0$ and setting $\mathbb{M}_{6:T} = 1$ for the remaining frames. This trains the model to treat the seed frames as a fixed spatio-temporal reference and coherently continue the sequence from them, which is the key capability exploited during infinite rollout at inference time (Sec. 3.3).

Atemporal Lighting Conditioning. We train our model on the RHC dataset [24], which captures human performance under a distinct environment map in each frame. While the coarse RelightNet renderings $\mathbf{I}^\kappa$ encode approximate lighting, they lack explicit global illumination context, forcing the generative backbone to infer target lighting from limited cues.

To provide direct per-frame lighting information, we introduce *atemporal lighting conditioning*, which encodes each environment map $\{\mathbf{E}_t\}$ (transformed to camera space $\{\mathbf{E}_t^\kappa\}$) into a latent space using the WAN2.1 3DVAE encoder:

$$\ell = \mathcal{E}(\{\mathbf{E}_t^\kappa\}), \quad (7)$$

yielding temporally-aligned embeddings $\ell \in \mathbb{R}^{T' \times H' \times W' \times d}$. Each latent ℓ_i is injected into the corresponding DiT feature $\mathbf{f}_i$ via cross-attention, with $\tilde{\mathbf{f}} = \text{Concat}_i(\text{CrossAttn}(\mathbf{f}_i, \ell_i))$. No temporal cross-attention is applied, ensuring each frame is conditioned only on its own environment map.

As shown in the OLAT results of our supplementary video, this explicit per-frame illumination context enables the model to faithfully reproduce fine-grained, temporally varying lighting while preserving motion priors.

Error Recycling. Although history-anchored conditioning (Sec. 3.3) improves temporal continuation, long-horizon generation remains susceptible to drift due to the distribution mismatch between ground-truth context observed during training and self-generated context at inference.

To mitigate this, we adopt *Error Recycling*, recently introduced in Stable Video Infinity [15]. The key idea is to re-inject reconstruction residuals from previous model predictions as stochastic perturbations during subsequent training iterations. This explicitly trains the model to recover from errors induced by its own predictions. Concretely, given potentially perturbed clean and noisy latents $\tilde{\mathbf{z}}_1$ and $\tilde{\mathbf{z}}_0$, we construct the noisy latent $\tilde{\mathbf{z}}_s = (1-s)\tilde{\mathbf{z}}_0 + s\tilde{\mathbf{z}}_1$, and define the target velocity $v_s^{\text{rcy}} = \mathbf{z}_1 - \tilde{\mathbf{z}}_0$, where the target always points to the clean latent $\mathbf{z}_1$ regardless of the perturbation applied to $\tilde{\mathbf{z}}_0$, encouraging the model to correct corrupted trajectories back toward clean outputs.

The model predicts velocity $v(\tilde{\mathbf{z}}_s, s, \mathbf{c})$, from which we obtain first-order estimates of the clean and noisy latents:

$$\hat{\mathbf{z}}_0 = \tilde{\mathbf{z}}_s - s v(\tilde{\mathbf{z}}_s, s, \mathbf{c}), \quad \hat{\mathbf{z}}_1 = \tilde{\mathbf{z}}_s + (1-s) v(\tilde{\mathbf{z}}_s, s, \mathbf{c}). \quad (8)$$

Analogously, using the target velocity v_s^{rcy} , we define reference reconstructions to obtain a consistent reference under the ground-truth velocity:

$$\mathbf{z}_0^{\text{rcy}} = \tilde{\mathbf{z}}_s - s v_s^{\text{rcy}}, \quad \mathbf{z}_1^{\text{rcy}} = \tilde{\mathbf{z}}_s + (1-s) v_s^{\text{rcy}}. \quad (9)$$

The reconstruction residuals can then be computed as: $\mathbf{e}_0 = \hat{\mathbf{z}}_0 - \mathbf{z}_0^{\text{rcy}}$, and $\mathbf{e}_1 = \hat{\mathbf{z}}_1 - \mathbf{z}_1^{\text{rcy}}$. These residuals are stored in an *Error Cache* and reused in subsequent iterations to perturb all latents involved in history-conditioned training:

$$\tilde{\mathbf{z}}_0 \leftarrow \mathbf{z}_0 + \mathbf{e}_0, \quad \tilde{\mathbf{z}}_1 \leftarrow \mathbf{z}_1 + \mathbf{e}_1, \quad \tilde{\mathbf{z}}_{\text{inac}} \leftarrow \mathbf{z}_{\text{inac}} + \mathbf{e}_1, \quad (10)$$

where $\mathbf{z}_{\text{inac}} = \mathcal{E}(F_{\text{inac}})$ denotes the VAE-encoded inactive conditioning frames. Each perturbation is applied independently with probabilities 0.01, 0.9, and 0.9, respectively. Training proceeds with the standard velocity regression objective under the perturbed latent distribution:

$$\mathcal{L}_{ER} = \mathbb{E}_s \|v(\tilde{\mathbf{z}}_s, s, \mathbf{c}) - (\mathbf{z}_1 - \tilde{\mathbf{z}}_0)\|^2. \quad (11)$$

By exposing the model to trajectories corrupted by its own reconstruction residuals, error recycling explicitly regularizes against accumulation of temporal drift

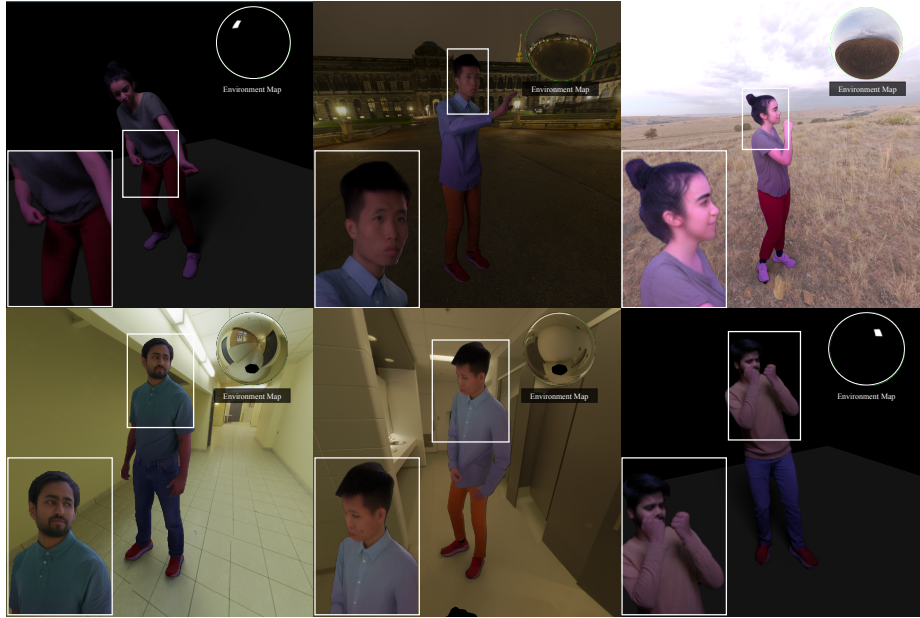


Fig. 4: Qualitative Results. GRA generates photorealistic free-view renderings of dynamic humans under novel poses, viewpoints, and illumination. Characters are seamlessly placed into diverse real-world environments, demonstrating realistic relighting, coherent shading, and consistent appearance under challenging lighting conditions. Notably, our physics-grounded model is able to generalize to OLAT lighting conditions despite not having seen these at training time.

and improves stability in long-form generation. As demonstrated in the supplementary video, it plays a key role in preserving high-frequency motion fidelity over extended sequences. For training details, see the supplementary material.

4 Experiments

We demonstrate that our method produces photorealistic renderings under novel poses, viewpoints, and illumination conditions in Fig. 4, across both indoor and outdoor environments. Owing to the physically grounded nature of our relighting formulation, our approach generalizes to one-light-at-a-time lighting setups even though such conditions are not observed during training.

Dataset. We use the Relightable Holoported Characters (RHC) dataset, providing 40 synchronized camera viewpoints of five subjects performing diverse motions, with interleaved uniformly lit and relit frames. Training uses 1,015 HDR environment maps from the Laval Indoor dataset [6]; testing uses 8 unseen environments. Test sequences consist of 128 consecutive relit frames under a fixed environment. Each subject has 28,420 training and 14,336 testing frames per camera, captured at 60 Hz. We evaluate on four subjects, training all components per-subject, with 37 cameras for training and 3 held out for testing.

Metrics. We report PSNR, SSIM [31], and LPIPS [40] averaged over every 10th frame across held-out test views, evaluated on the middle frame of a 17-frame

Table 1: Quantitative Evaluation. We compare our method to prior methods for human performance relighting, including Relighting4D (R4D) [3]. Following RHC [24], we provide high-quality DDC-tracked meshes and calibrated ground-truth environment maps for multi-view training. This training protocol is further extended to IntrinsicAvatar (IA) [29], MeshAvatar (MA) [2], and DNF-Avatar (DNFA) [10]. Our method achieves the best perceptual quality across all subjects.

Method	Subject 1 (S1)				Subject 2 (S2)				Subject 3 (S3)				Subject 4 (S4)			
	PSNR $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$	FVD $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$	FVD $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$	FVD $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$	FVD $\downarrow$
R4D [3]	29.89	10.31	87.15	247	31.13	8.04	87.08	161	29.32	10.62	87.67	186	31.98	10.07	87.90	168
IA [29]	27.25	18.25	81.39	408	28.87	15.43	82.07	345	26.14	22.52	79.07	621	29.50	18.46	82.91	354
MA [2]	29.71	<u>9.09</u>	84.69	<u>77</u>	30.36	<u>8.03</u>	84.99	<u>74</u>	28.35	<u>10.16</u>	84.60	<u>64</u>	31.13	<u>8.50</u>	84.59	<u>73</u>
DNFA [10]	26.32	11.57	82.50	212	28.57	9.40	82.92	131	26.37	12.04	82.26	150	29.31	10.82	84.12	124
Ours	30.49	6.42	87.52	49	31.15	5.64	87.50	48	<u>29.11</u>	7.40	<u>87.16</u>	52	32.09	6.55	<u>87.35</u>	63

generation window to avoid boundary effects. We also report FVD [18], computed on 17-frame clips sampled every 200 frames, to capture temporal coherence and overall video realism.

Baselines. We compare against state-of-the-art relightable 3D avatar methods. For R4D [3], we extend it to a motion-driven setting with multi-view supervision and ground-truth environment maps, replacing its SMPL mesh with our higher-quality DDC-tracked mesh for a fair comparison. We also compare against IntrinsicAvatar (IA) [29], MeshAvatar (MA) [2], and DNF-Avatar (DNFA) [10], trained with multi-view supervision and ground-truth envmaps using their standard SMPL-based representations, as their tight coupling with SMPL/SMPL-X prevents substitution with our DDC mesh. We exclude Relightable Full-Body Gaussian Codec Avatars [30] as its implementation is not publicly available.

4.1 Comparisons on Novel Sequences, Viewpoints, and Lighting

Quantitative Comparisons. Tab. 1 reports results on novel sequences under novel viewpoints and unseen lighting. Our method outperforms all baselines in perceptual quality across all subjects. R4D with DDC tracking surpasses IA and MA on PSNR and SSIM, confirming the importance of accurate geometry for pixel-wise metrics. MA achieves sharper results than IA and R4D via its hybrid explicit-implicit representation, but remains limited by pose-only conditioning and restrictive BRDF assumptions. DNFA improves perceptual sharpness over its teacher IA via a more expressive Gaussian representation, but inherits its tracking limitations and remains less robust than MA. Our method achieves the best perceptual and video performance across all subjects.

We note that PSNR and SSIM are imperfect proxies for relighting quality: both favor blurry, deterministic outputs, and fine-scale details like wrinkles are inherently ambiguous at test time. LPIPS and FVD better reflect the perceptual gains of our approach.

Qualitative Comparisons. Fig. 7 shows our qualitative comparisons. Our method produces sharp, physically plausible relighting with dynamic wrinkles, while all baselines yield noticeably blurrier results due to deterministic pose conditioning. R4D produces over-smooth results. IA additionally suffers from SMPL tracking inaccuracies, leading to further loss of detail. MA generates comparatively sharper textures but fails to reproduce fine-scale material responses. DNFA

Table 2: Ablation Study. We ablate different design choices of our method. Note the progressive improvement by the addition of individual components. When removing individual components from the pipeline, the results degrade confirming the importance of our individual design choices.

Method	PSNR $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$	FVD $\downarrow$
Microfacet Rendering	29.34	12.26	86.85	178.42
+ RelightNet	30.90	8.92	89.25	118.24
+ Generative Refinement (Ours)	30.49	6.42	87.52	48.79
Ours w/ Temporal Lighting Conditioning	30.78	6.53	88.00	58.18
Ours w/o Atemporal Lighting Conditioning	30.53	6.48	87.76	58.21
Ours w/o VACE	30.35	6.93	87.12	61.84
Ours w/o RelightNet	29.56	6.66	87.26	63.81
Ours w/o VACE Finetuning	24.92	22.43	70.76	230.94
Ours w/o Prior	30.02	9.60	86.67	394.91
Ours w/ Neural Gaffer [12]	29.33	9.25	86.62	102.10
Ours w/ Few-shot Adaptation	29.82	7.01	87.30	65.25

is sharper than IA but exhibits visible high-frequency artifacts. Additional results are in the supplementary material.

4.2 Ablation Studies

We conduct quantitative (Tab. 2) and qualitative (Suppl. Mat.) ablation studies to validate each component. Full details are provided in the supplementary material. The microfacet rendering achieves reasonable PSNR from physically correct low-frequency shading but lacks fine-scale detail. RelightNet improves PSNR and SSIM via learned pose-correlated appearance, but its determinism limits perceptual fidelity, motivating the generative refinement stage. Atemporal cross-attention outperforms temporal cross-attention by inducing per-frame lighting correspondence as an inductive bias rather than requiring the model to learn it implicitly. Without VACE finetuning, the model fails to treat coarse renderings as a shading target, producing physically inconsistent results. Training without the pretrained video prior fails to converge within the same budget, and replacing our video model with an image-based alternative (Neural Gaffer [12]) yields severe temporal inconsistency (as shown by the degraded FVD), confirming the necessity of video-level temporal modeling.

4.3 Discussion

We now discuss the applications and broader implications of our design choices. **Physics-grounded generalization.** Our physics-grounded formulation enables generalization to lighting conditions not seen during training. We demonstrate strong generalization to one-light-at-a-time (OLAT) environment maps (Fig. 1, Fig. 4, and Suppl. Vid.), where the model produces stable, physically plausible relighting under sparse, highly directional illumination despite being trained primarily on natural indoor environments. This generalization stems from the explicit microfacet rendering, which provides a physically consistent inductive

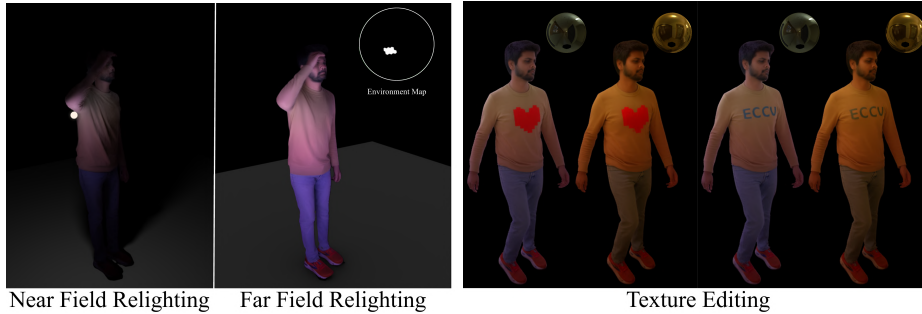


Fig. 6: Physics Grounded Applications. We demonstrate two capabilities of our *w/o RelightNet* model enabled by its explicit physics-grounded design. (Left) A local near-field point light source produces physically plausible shading, including arm self-shadowing. For comparison, our full model approximates the same light as far-field illumination by projecting it onto the lightstage. (Right) Relightable texture editing is achieved by directly modifying the albedo map in UV space.

bias that guides the generative model toward correct shading even under unseen distributions.

Near-field relighting. Interestingly, our method also supports near-field relighting (Fig. 1, Fig. 6, and Suppl. Vid.) via the *w/o RelightNet* variant. We render microfacet-based shading under a point light source, and pass the resulting coarse renderings directly to the video model (Suppl. for details). The fact that this approximation produces plausible results demonstrates that the physics-based design, combined with the generative prior, is expressive enough to handle localized, unseen lighting configurations without any dedicated training.

Few-shot identity adaptation. The subject-specific finetuning required by our pipeline raises a natural question about generalization across identities. Our few-shot adaptation experiment (Fig. 5) shows that despite prior subject-specific adaptation, the video model retains sufficient generative capacity to transfer to a new identity from as little as 6 seconds of capture data. This suggests that subject-specific finetuning primarily anchors identity-specific appearance without catastrophically forgetting the broader motion and illumination priors learned during pre-training.

Relightable texture editing, as shown in (Fig. 6), is similarly enabled by the explicit material decomposition of the *w/o RelightNet* model. Modifying the optimized UV albedo map and re-rendering with the microfacet model produces edited coarse renderings that the video model refines into photorealistic output. The generative stage faithfully preserves the intended texture modifications

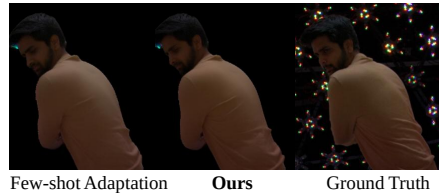


Fig. 5: Starting from a Subject 4 checkpoint, we finetune our video model on 170 frames of Subject 1. The adapted model produces photorealistic, temporally consistent relighting for the new identity despite minimal supervision.

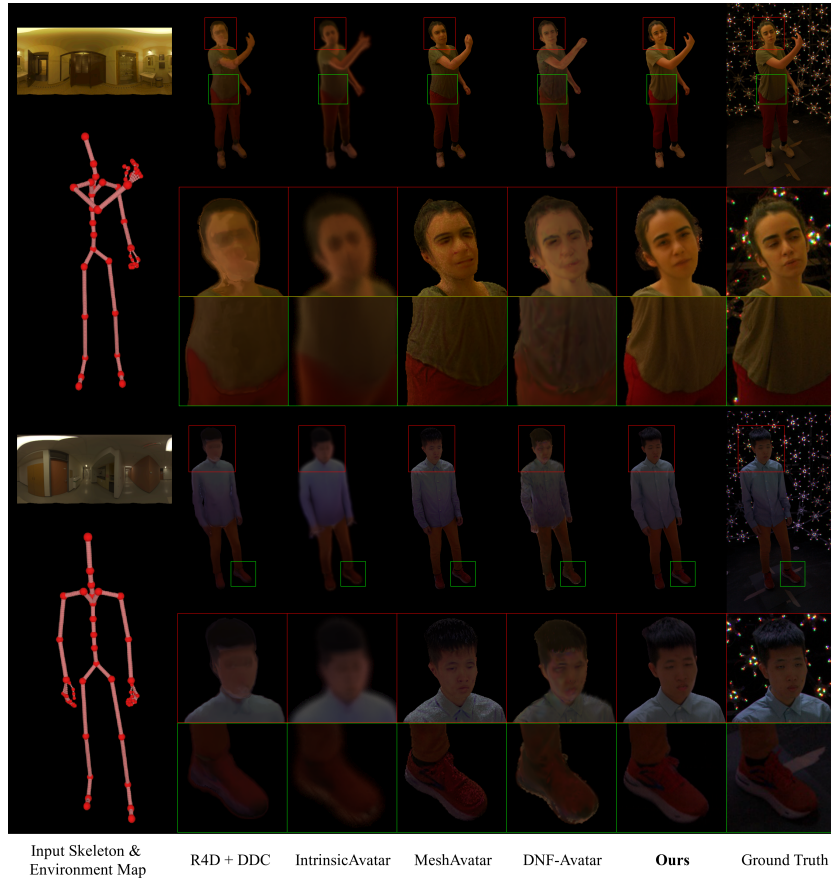


Fig. 7: Qualitative Comparison. We compare our method with state-of-the-art relightable avatar methods: Relighting4D (R4D) [3] augmented with a DDC-tracked template, IntrinsicAvatar [29], MeshAvatar [2], and DNF-Avatar [10]. Our approach consistently improves visual realism and relighting quality.

while synthesizing consistent shading, demonstrating that explicit material control naturally extends to appearance editing with no additional supervision.

5 Conclusion

In this paper, we proposed GRA as a hybrid method to build generative 3D relightable avatars. We showed that integrating explicit, physics-grounded texture appearance with video generative models allows us to have 3D control as well as photorealism at the same time. Interestingly, we showed that our approach can also generalize to out-of-distribution target illuminations, such as OLAT illumination without any retraining. As it is grounded in explicit 3D representation, GRA can model global illumination effects like self-shadows. Finally, we demonstrated the ability of our system to generate long-form videos, a capability not easily afforded by video generative models. In the future, we intend to model the face and hands in high-resolution, while also introducing near-field relighting effects within this framework.

6 Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK 2853/1 “Neuroexplicit Models of Language, Vision, and Action” - project number 471607914.

References

1. Chaturvedi, S., Ren, M., Hold-Geoffroy, Y., Liu, J., Dorsey, J., Shu, Z.: Synthlight: Portrait relighting with diffusion model by learning to re-render synthetic faces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
2. Chen, Y., Zheng, Z., Li, Z., Xu, C., Liu, Y.: Meshavatar: Learning high-quality triangular human avatars from multi-view videos. In: European Conference on Computer Vision. pp. 250–269. Springer (2024)
3. Chen, Z., Liu, Z.: Relighting4d: Neural relightable human from videos. In: European Conference on Computer Vision. pp. 606–623. Springer (2022)
4. Cook, R.L., Torrance, K.E.: A reflectance model for computer graphics. *ACM Trans. Graph.* **1**(1), 7–24 (Jan 1982). <https://doi.org/10.1145/357290.357293>, <https://doi.org/10.1145/357290.357293>
5. Erel, Y., Dabral, R., Golyanik, V., H. Bermano, A., Theobalt, C.: Practilight: Practical light control using foundational diffusion models. *ACM Transactions on Graphics (TOG)* **44**(6), 1–11 (2025)
6. Gardner, M.A., Sunkavalli, K., Yumer, E., Shen, X., Gambaretto, E., Gagné, C., Lalonde, J.F.: Learning to predict indoor illumination from a single image. *ACM Transactions on Graphics (TOG)* **36**(6), 1–14 (2017)
7. Habermann, M., Liu, L., Xu, W., Zollhoefer, M., Pons-Moll, G., Theobalt, C.: Real-time deep dynamic characters. *ACM Transactions on Graphics (ToG)* **40**(4), 1–16 (2021)
8. He, M., Clausen, P., Taşel, A.L., Ma, L., Pilarski, O., Xian, W., Rikker, L., Yu, X., Burgert, R., Yu, N., et al.: Diffrelight: Diffusion-based facial performance relighting. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–12 (2024)
9. Iqbal, U., Caliskan, A., Nagano, K., Khamis, S., Molchanov, P., Kautz, J.: Rana: Relightable articulated neural avatars. In: ICCV (2023)
10. Jiang, Z., Wang, S., Tang, S.: Dnf-avatar: Distilling neural fields for real-time animatable avatar relighting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6383–6394 (2025)
11. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17191–17202 (2025)
12. Jin, H., Li, Y., Luan, F., Xiangli, Y., Bi, S., Zhang, K., Xu, Z., Sun, J., Snavely, N.: Neural gaffer: Relighting any object via diffusion. *Advances in Neural Information Processing Systems* **37**, 141129–141152 (2024)
13. Kajiya, J.T.: The rendering equation. In: Proceedings of the 13th annual conference on Computer graphics and interactive techniques. pp. 143–150 (1986)
14. Kim, H., Jang, M., Yoon, W., Lee, J., Na, D., Woo, S.: 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. pp. 25096–25106 (2024)

15. Li, W., Pan, W., Luan, P.C., Gao, Y., Alahi, A.: Stable video infinity: Infinite-length video generation with error recycling. In: International Conference on Learning Representations 2025 (ICLR 2025) (2026)
16. Liang, R., Gojcic, Z., Ling, H., Munkberg, J., Hasselgren, J., Lin, C.H., Gao, J., Keller, A., Vijaykumar, N., Fidler, S., Wang, Z.: Diffusionrenderer: Neural inverse and forward rendering with video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
17. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations
18. Liu, J., Qu, Y., Yan, Q., Zeng, X., Wang, L., Liao, R.: Fréchet video motion distance: A metric for evaluating motion consistency in videos. In: First Workshop on Controllable Video Generation@ ICML24
19. Luvizon, D., Golyanik, V., Kortylewski, A., Habermann, M., Theobalt, C.: Relightable neural actor with intrinsic decomposition and pose control. In: European Conference on Computer Vision (ECCV) (2024)
20. Mei, Y., He, M., Ma, L., Philip, J., Xian, W., George, D.M., Yu, X., Dedic, G., Taşel, A.L., Yu, N., et al.: Lux post facto: Learning portrait performance relighting with conditional video diffusion and a hybrid dataset. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5510–5522 (2025)
21. Mei, Y., Zeng, Y., Zhang, H., Shu, Z., Zhang, X., Bi, S., Zhang, J., Jung, H., Patel, V.M.: Holo-relighting: Controllable volumetric portrait relighting from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4263–4273 (2024)
22. Pandey, R., Orts-Escolano, S., Legendre, C., Haene, C., Bouaziz, S., Rhemann, C., Debevec, P.E., Fanello, S.R.: Total relighting: learning to relight portraits for background replacement. *ACM Trans. Graph.* **40**(4), 43–1 (2021)
23. Ren, M., Xiong, W., Yoon, J.S., Shu, Z., Zhang, J., Jung, H., Gerig, G., Zhang, H.: Relightful harmonization: Lighting-aware portrait background replacement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
24. Singh, K.M., Chen, J., Golyanik, V., Garbin, S.J., Beeler, T., Dabral, R., Habermann, M., Theobalt, C.: Relightable holoported characters: Capturing and relighting dynamic human performance from sparse views. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28882–28892 (2026)
25. Team, T.H.: Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation (2025)
26. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models (2025), <https://arxiv.org/abs/2503.20314>
27. Wang, J., Liu, J., Sun, X., Singh, K.K., Shu, Z., Zhang, H., Yang, J., Zhao, N., Wang, T.Y., Chen, S.S., Neumann, U., Yoon, J.S.: Comprehensive relighting: Gen-

- eralizable and consistent monocular human relighting and harmonization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
28. Wang, L., Jin, S., Cui, R., Dahl, A.B., Frisvad, J.R., Bigdeli, S.: Relumix: Extending image relighting to video via video diffusion models. arXiv preprint arXiv:2509.23769 (2025)
 29. Wang, S., Antic, B., Geiger, A., Tang, S.: Intrinsicavatar: Physically based inverse rendering of dynamic humans from monocular videos via explicit ray tracing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1877–1888 (2024)
 30. Wang, S., Simon, T., Santesteban, I., Bagautdinov, T., Li, J., Agrawal, V., Prada, F., Yu, S.I., Nalbone, P., Gramlich, M., et al.: Relightable full-body gaussian codec avatars. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–12 (2025)
 31. Wang, Z.: Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13**(4), 600–612 (2004)
 32. Xiang, J., Chen, X., Xu, S., Wang, R., Lv, Z., Deng, Y., Zhu, H., Dong, Y., Zhao, H., Yuan, N.J., et al.: Native and compact structured latents for 3d generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14419–14429 (2026)
 33. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21469–21480 (2025)
 34. Xiao, J., Zhang, Q., Xu, Z., Zheng, W.S.: Neca: Neural customizable human avatar. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20091–20101 (2024)
 35. Xie, R., Huang, K., Cho, I.Y., Yang, S., Chen, W., Bao, H., Zheng, W., Li, R., Huo, Y.: Ren human: Learning relightable neural implicit surfaces for animatable human rendering. *ACM Transactions on Graphics* **43**(5), 1–22 (2024)
 36. Xu, Z., Peng, S., Geng, C., Mou, L., Yan, Z., Sun, J., Bao, H., Zhou, X.: Relightable and animatable neural avatar from sparse-view video. In: CVPR (2024)
 37. Yoon, J.S., Shu, Z., Ren, M., Zhang, C., Hold-Geoffroy, Y., Singh, K.k., Zhang, H.: Generative portrait shadow removal. *ACM Transactions on Graphics (TOG)* **43**(6), 1–13 (2024)
 38. Zeng, Z., Deschaintre, V., Georgiev, I., Hold-Geoffroy, Y., Hu, Y., Luan, F., Yan, L.Q., Hašan, M.: Rgb-x: Image decomposition and synthesis using material- and lighting-aware diffusion models. In: ACM SIGGRAPH 2024 Conference Papers (2024)
 39. Zhang, L., Rao, A., Agrawala, M.: Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In: International Conference on Learning Representations (ICLR) (2025)
 40. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
 41. Zhao, Y., Wu, C., Huang, B., Zhi, Y., Zhao, C., Wang, J., Gao, S.: Surfel-based gaussian inverse rendering for fast and relightable dynamic human reconstruction from monocular videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025)