

Supplementary Material

GeoRelight: Learning Joint Geometrical Relighting and Reconstruction with Flexible Multi-Modal Diffusion Transformers

Abstract

In this supplementary document, we provide further details about the implementation details, training data, evaluation protocol, etc. We also show additional comparison with state-of-the-art approaches. Please refer to our supplementary video for additional visualization of our results. Finally, we discuss limitations of our model and potential future works.

1. Additional Implementation Details

This section provides further details on our network architecture and training procedure, expanding on Sec. 4.1 of the main paper.

1.1. Model Architecture

Base Model Our model’s architecture is a Diffusion Transformer (DiT) [17]. We initialize our weights from the publicly available inverse rendering model of DiffusionRenderer-Cosmos-7B [12], which is built upon the Cosmos-Predict-1 7B DiT [15].

Causal VAE Latent Space We operate in the latent space of the pretrained Cosmos causal VAE [15], which can process both image or videos. The VAE remains frozen during our training. This VAE has a spatial compression factor of 8. Therefore, our input images at a training resolution of 832×1280 pixels are encoded into a latent space of 104×160 . The channel dimension is 16. All five target modalities ($\mathbf{z}^a, \mathbf{z}^n, \mathbf{z}^g, \mathbf{z}^s, \mathbf{z}^{IE}$) and the global image condition (\mathbf{z}^I) and illumination condition (\mathbf{z}^E) share this latent resolution.

1.2. Detailed DiT Conditioning

As illustrated in Fig. 1, our architecture adapts a standard DiT block to accept our multi-modal conditioning. The total input to the DiT for a single modality m is a channel-wise concatenation of several features that are added to the noisy latent z_r^m before the first DiT block.

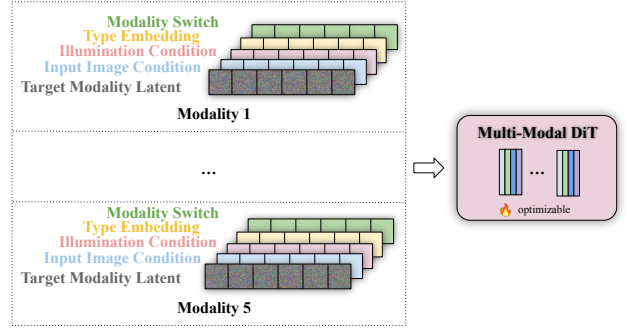


Figure 1. **Conditioning on the modality latent.** Each modality latent after conditioning have the shape $\mathbb{R}^{H \times W \times C(16+16+3*16+3+1)}$. Different modalities are concatenated "temporal-wise" to a sample $\mathbb{R}^{M \times H \times W \times C}$ in one batch.

- **Global Condition \mathbf{z}^I :** The VAE-encoded latent of the input image \mathbf{I} . This is concatenated to all five target modalities to provide shared identity and shape information.
- **Illumination Condition (\mathbf{z}^E):** The VAE-encoded latent of the three LDR light maps ($\mathbf{E}_{ldr}, \mathbf{E}_{log}, \mathbf{E}_{dir}$). This is concatenated only to the relit image latent \mathbf{z}^{IE} and is zero-padded for all other modalities.
- **Modality Type Embedding (\mathbf{c}_{modal}):** A learnable embedding of shape $\mathbb{R}^{M \times C_{type}}$ (where $M=5, C_{type} = 3$) that is broadcast to the latent spatial resolution and concatenated. This informs the model which modality it is currently processing (e.g., albedo vs. normal).
- **Modality Switch Mask (\mathbf{c}_{switch}):** A binary mask of shape $\mathbb{R}^{H \times W \times 1}$ that is broadcast and concatenated. It signals whether the modality is a "clear" condition (1) or a "noisy" target (0).

These concatenated features $\in \mathbb{R}^{H \times W \times (16+16+3*16+3+1)}$ are then processed by the DiT. We replace the original 3D RoPE positional embeddings with a shared 2D RoPE [19] that is applied to all modalities. This ensures that spatial position (x,y) is aligned across all modalities, which is critical for learning cross-modal correlations.

1.3. Training Details

Hardware and Optimizer Our model is trained on 64 NVIDIA A100 (80GB) GPUs. We use the AdamW [9] optimizer with a constant learning rate of 2×10^{-5} , a weight decay of 0.01, and enable bfloat16 mixed-precision training. Our total effective batch size is 128, distributed across all GPUs.

Two Stage Training As described in the main paper (Sec 4.1), our training is conducted in two distinct stages:

1. Stage 1 (Synthetic Pre-training): We first train the model for 30,000 steps exclusively on our synthetic *Synth* dataset. This stage teaches the model the fundamental disentanglement of intrinsics, geometry, and lighting. The resulting model from this stage is then used as our "auto-labeler" for the real-world datasets.
2. Stage 2 (Mixed Post-Training): We then finetune the model from Stage 1 for an additional 10,000 steps. In this stage, we train on a combined hybrid dataset of synthetic, light stage, and in-the-wild data, applying the strategic tasks outlined in our main paper’s Table 1. The batch is composed by sampling from our datasets with the following probabilities: 33% *Synth*, 35% *Dome*, 32% *ITW*.

1.4. Inference Runtime

We report our method’s inference speed on a single NVIDIA A100 80GB GPU. For a single input image at our test resolution of 832×1280 , the full joint-generation process takes approximately 35 seconds. This time include 35 diffusion sampling steps, VAE decoding time, and I/O time for saving the results.

1.5. Ablation Details

We provide more details about implementation of ablation study in main paper section 4.2. The quantitative ablation study was reported on full-body subset of our synthetic evaluation dataset because full-body images contain more wrinkles compared to face-only subset.

Geometry is essential for Relighting We compare three distinct modes to validate this claim:

- Joint Modeling: The model generates all five modalities (relit image, albedo, normal, segmentation, and geometry) simultaneously from noise.
- w/ GT Geometry: We replace the initial noise of the geometry modalities (iNOD, normal, segmentation) with the ground-truth latent representations. We set their corresponding switch masks to $c_{\text{switch}} = 1$ (condition) and keep them fixed (clear) during the reverse sampling process, forcing the model to generate the relit image conditioned on perfect geometry.

- w/o Geometry: We disable the generation of geometry modalities. To keep the input tensor dimensions consistent with our architecture, we replace the geometry latents with zero-tensors. This formulation is analogous to the dropped-condition scenario in classifier-free guidance [5], effectively removing geometric guidance from the generation process.

Relighting helps Shape-from-Shading. To validate the synergistic effect of appearance on geometry, we compare the following modes:

- **w/o Appearance:** We isolate the geometry generation by suppressing the relighting modality. Similar to the “w/o Geometry” setting, we replace the relit image latent z^{IE} and the illumination condition z^E with zero-tensors, effectively forcing the model to hallucinate geometry from the input image I alone, without any shading cues from the new lighting context.
- **w/ GT Appearance:** We provide the ground-truth relit image as a strong conditioning signal. We encode the GT relit image into latent space and set its switch mask to $c_{\text{switch}} = 1$. This allows the model to explicitly use the shading and shadows present in the target relit image to refine the surface normals and iNOD geometry.

2. Detailed Data Creation and Sources

In this section, we provide a detailed breakdown of the three data sources used in our hybrid dataset, as first introduced in Sec. 3.4 of the main paper and visually summarized in our main paper’s Figure 4.

2.1. Synthetic Data

The synthetic dataset forms the foundation of our method. Its primary purpose is to provide a large corpus of physically-accurate, fully-labeled data to train our Stage 1 model. We create 1000 high-quality, artist-created 3D human meshes, which contain detailed facial components (eyes, teeth, hair). We augment it to 8000 identities by rigging for pose variation. We pair these with a library of high-quality PBR texture maps for clothing (albedo, normal, roughness).

Our system is built with Blender, and images are rendered with the Cycles renderer. For each of the 8000 appearance samples, we render it under 400 random HDR environment maps, creating a large-scale training set. For each rendered scene, we save all five of our target modalities: the relit image ($\mathbf{I_E}$), albedo (\mathbf{a}), segmentation mask (\mathbf{s}), surface normals (\mathbf{n}), and our novel iNOD geometry (\mathbf{g}), which is computed from the ground-truth metric depth and camera intrinsics.

As noted in our main paper, this synthetic-only model yields satisfactory intrinsics but lacks photorealism in the final relit image, which motivates our hybrid-data approach.

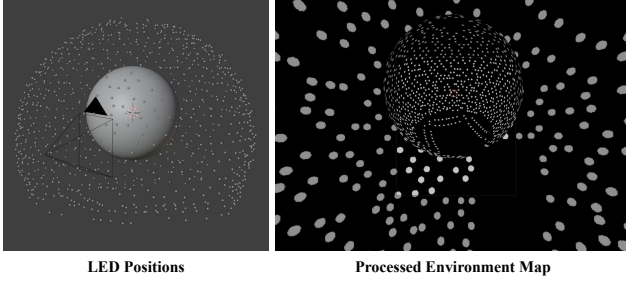


Figure 2. **Processed Environment Illumination from Light-Stage.** From the 3-dimensional LED positions, we project it to a latlong image to model the environment map.

2.2. Light Stage Data

The Dome dataset is our high-quality, real-world dataset that acts as a bridge between synthetic data and in-the-wild images. It is sourced from the Codec Avatar Studio [13] and Relightable Full-Body Gaussian Codec Avatars [24] datasets.

Paired Data Creation This dataset is critical because its light stage capture setup, which employs 1024 individually controllable white LED light sources with known locations, allows us to create paired real-world training data. The captured videos consist of alternating fully-lit frames and partially-lit frames (random 10-20 lights) at 60 Hz.

Due to fast motion, we cannot reliably warp frames to perform traditional image-based relighting as in [8, 14, 16]. Hence, we treat neighboring frames as the paired input and relit image. To generate the corresponding illumination condition **E**, we map the 3D LED light locations and intensities from the partially-lit frame to an equirectangular image (Fig. 2, simulating a sparse environment map). This process provides us with realistic, photometrically-aligned pairs of (Input, Target Light, Target Relit Image), which we use for the "Default" and "Rendering" training modes (Table 1 in main paper).

2.3. In-the-wild Data

To further enhance diversity and photorealism, we use a large-scale in-the-wild dataset. We sample 10K human images from the CosmicMan [11] and ReLaion [10] datasets as our data pool.

Label Scarcity This dataset is defined by what it lacks. Unlike *Synth* and *Dome*, *ITW* images have no ground-truth intrinsics, no paired relit image, and no known illumination condition **E**. It is used exclusively for our "Intrinsic→Relit" task.

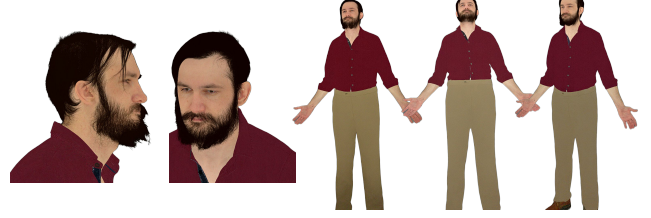


Figure 3. **Robustness of our Auto-Labeler.** Our auto-labeled albedo (shown) and other intrinsics are consistent across multiple views of the same subject from our Dome dataset, demonstrating the high quality of our pseudo-ground-truth.

2.4. Auto-Labeling Process

The "auto-labeling" process is the important component of our strategic mixed-data training, inspired by [12, 14]. We use the model trained exclusively on our *Synth* dataset (Stage 1) as our "auto-labeler."

We run this model in inference mode over our entire Dome and ITW datasets. For each real-world image, we generate and save the four corresponding pseudo-ground-truth intrinsic maps: albedo (z^a), segmentation mask (z^s), surface normals (z^n), and our iNOD geometry (z^g). As shown in Fig 3, this auto-labeler is surprisingly robust, generating high-quality, view-consistent intrinsics even for real-world subjects. This process is what enables us to use the Dome and ITW datasets for our advanced training tasks in Stage 2.

3. Additional Details on iNOD

Our novel geometry representation, isotropic Normalized Orthographic Depth (iNOD), is a core contribution of our work. Unlike point map which is unsuitable for latent diffusion due to the noise introduced after encoding and decoding (Fig. 4), our iNOD is naturally suitable for being applied in latent diffusion model. We provide a more detailed algorithm for its creation and a justification for our dilation-based boundary refinement.

3.1. iNOD Creation Algorithm

As described in the main paper (Sec. 3.3), iNOD is generated from a source 3D point cloud. For our *Synth* data, this is computed from the ground-truth metric depth and known camera intrinsics. The process, detailed in Algorithm 1, is designed to be simple, fast, and distortion-free.

The key steps are (1) Unprojection to a metrically-accurate 3D point cloud, (2) Isotropic Normalization where the entire point cloud is scaled to fit a $[-1, 1]$ cube based on its longest axis, and (3) Orthographic Projection where we simply take the z -value of the normalized points and write it into a 2D map. This process preserves the relative 3D geometry and aspect ratio of the subject while ensuring the

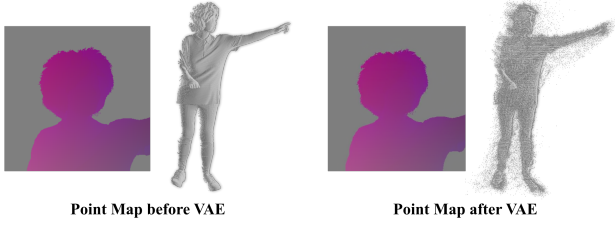


Figure 4. **Limitation of Point Map in Latent Space.** As a popular geometry representation [21, 23] in image space, point map shows strong limitation in latent space. Although visually the point map looks similar before and after VAE, the boundary lost huge precision (please zoom in) and it contains much noise after VAE.

Algorithm 1 iNOD Creation from a Metric Depth Map

```

1: Input: metric_depth (H, W), cam_intrinsics (K)
2: Output: inod_map (H, W)
3:
4: Step 1: Unproject to 3D point cloud
5: pixels_uv  $\leftarrow$  create_pixel_grid(H, W)
6: cam_points  $\leftarrow$  unproject(pixels_uv, metric_depth, K)
7:
8: Step 2: Find scaling factor for isotropic normalization
9: min_bound  $\leftarrow$  min_axis(cam_points.xyz)
10: max_bound  $\leftarrow$  max_axis(cam_points.xyz)
11: max_edge  $\leftarrow$  max(max_bound - min_bound)
12:
13: Step 3: Apply isotropic normalization
14: center  $\leftarrow$  (min_bound + max_bound) / 2
15: norm_points  $\leftarrow$  (cam_points.xyz - center) / max_edge
16:
17: Step 4: Create map via orthographic projection
18: inod  $\leftarrow$  create_empty_map(H, W, value = 0)
19: for  $i \leftarrow 1$  to num_points do
20:    $u, v \leftarrow$  pixels_uv[ $i$ ]
21:   inod_map[v, u]  $\leftarrow$  norm_points.z[ $i$ ]
22: end for
23:
24: return inod

```

final 2D map is in a VAE-friendly $[-1, 1]$ range.

3.2. Dilation for VAE Artifacts

As noted in the main paper (Sec. 3.3) and visually demonstrated in our main paper’s Fig. 2, the pretrained Cosmos VAE [15], like most VAEs, can struggle with the extremely sharp boundaries of a foreground silhouette against a blank background.

When a standard iNOD map is passed through the VAE (encoded and then decoded), this compression can intro-



Figure 5. **iNOD Generalization Beyond Humans.** We validate iNOD on subjects with increasing Depth-to-Height (D:H) ratios: humans (~ 0.1), a Grim Reaper [18] (0.5), a Capybara [4] (1.0), and a Building [1] (3.0). The 3D shape is faithfully recovered in all cases from the ground-truth iNOD.

duce minor "ringing" or noise artifacts at the very edge of the geometry. To solve this and ensure a clean, robust geometric boundary after decoding, we apply a simple 2D morphological *dilation* operation to the iNOD map before it is passed to the VAE. We use the assign neighbor iNOD value to dilated pixels. The background pixels are assigned to 0. This slightly thickens the silhouette’s boundary, providing a small buffer that is more robust to the VAE’s lossy compression. After decoding, we use original mask to "cut off" the dilated region which contains noise. Then we obtain noise-free iNOD geometry.

The effectiveness of this simple step is shown in Fig.2 in main paper. The non-dilated iNOD produces noticeable boundary artifacts after being decoded, while the dilated version results in a clean, sharp, and artifact-free geometric boundary.

3.3. Generalizing iNOD Beyond Humans

While we focus on human subjects in this work, iNOD is a general-purpose representation that is not inherently limited to any specific category. In Fig. 5, we validate that iNOD faithfully encodes and decodes 3D geometry for subjects with increasing Depth-to-Height (D:H) ratios, from humans (~ 0.1) to objects like buildings (3.0). The representation remains effective as long as the D:H ratio is moderate; extremely elongated scenes (e.g., tunnels with $D:H > 10$) would compress the depth range, potentially degrading reconstruction quality. Consequently, the generalizability of iNOD unlocks the potential of our framework for joint relighting, intrinsic decomposition, and 3D reconstruction in broader object-centric domains.

4. Evaluation Details

This section provides additional specifics on the baseline configurations and metric implementations used for the comparisons in our main paper (Sec 4.3).

To ensure a fair and reproducible comparison, we used publicly available code and pretrained models for all baselines wherever possible. We also tried our best effort for evaluating close-sourced human-centric relighting methods



Figure 6. **Artifacts in HumanOLAT ground-truth relit images.** Due to motion blur of captured OLAT images, the linear combination leads to blurry ground-truth which may affect quantitative evaluation results. Please zoom in for details in hair and eyes.

by contacting their authors. Due to license of our evaluation data [20], we are prohibited to redistribute any of them to conduct quantitative comparison. Hence, we share our in-the-wild evaluation images and focus on qualitative comparison and user study with them.

4.1. Baseline Configurations

Appearance Baselines For methods taking square images [6, 25, 26] as input, we pad our vertical image horizontally and resize to the desired input resolution. For others we directly resize our image into their desired resolution.

For DiLightNet [25] which takes additional text prompt as input, we use the unified prompt "A real photo of a human". For IC-Light [27] which takes background image instead of environment map as input, we render background image from conditioning environment map using Blender and use as its input.

For closed-source baselines, we get reply from authors of LuxPostFacto [14] and deeply appreciate them for sharing inference results on *face cropped* in-the-wild images with us. LuxPostFacto is a face-centric relighting methods and can have downgrading performance when the input face crop is small or blurry. For TotalRelighting [16], SynthLight [3], and SwitchLight [8], we cannot get reply from authors on time and will include the comparison at first moment when we obtain their inference results from authors.

Geometry Baselines For VGGT [21], we use the 1B model and pad our input image into square format and resize to the desired input resolution. For MoGe2 [22] we use the largest model which has 331M parameters and supports metric-scale reconstruction and normal estimation. For Sapiens [7], we use the normal normal estimator with 2B parameters.

4.2. Metric Implementation Details

Relighting and Albedo Evaluation To mitigate the factor of scale ambiguity in lighting and albedo, we follow the protocol from DiffusionRenderer [12] and apply chromatic alignment (a scale-invariant evaluation) between the ground-truth and predicted images for all evaluated methods. All metrics are computed on the foreground pixels.

Geometry (Point Cloud) Evaluation As described in the main paper, we compare against feed-forward estimators (VGGT, MoGe2) that predict point maps. Our evaluation process is as follows:

1. Normalize both the predicted and ground-truth point clouds to fit within a shared $[-1, 1]^3$ bounding box.
2. Align the predicted shape to the ground-truth shape using the Iterative Closest Point (ICP) algorithm [2].
3. Calculate all geometry metrics (CD, F-Score) based on these normalized and aligned shapes.

Normal Evaluation For surface normal evaluation, all metrics (Angular Error, RMSE) are computed on the foreground pixels using the ground-truth segmentation mask

5. Additional Results

5.1. User Study

GeoRelight is capable of jointly generating all five modalities in parallel from a single input image: a high-fidelity relit image under novel illumination, a clean albedo map, a detailed surface normal map, a segmentation mask, and a robust 3D reconstruction. The model generalizes well to out-of-distribution subjects, robustly handling a wide variety of ages, accessories, complex clothing, and non-standard poses. Furthermore, GeoRelight supports controllable generation: rotating the target environment map produces physically-plausible relit images with dynamic shadows, while the estimated albedo and normal maps remain stable, and scaling the light intensity correctly modulates the brightness and shadow depth. We refer readers to our supplementary video for extensive visualization of joint generation results and controllable relighting.

As mentioned in 4.1, a direct quantitative comparison with closed-source methods like LuxPostFacto [14] is not feasible due to data license restrictions. Hence, we conduct a user study to evaluate the performance.

We asked 32 participants to evaluate a set of 20 randomly selected subjects from our in-the-wild test set. The study was divided into two tasks. The first task was a two-alternative forced-choice (2AFC) task for relighting. Participants were shown the input image and environment map, followed by the relit results from our GeoRelight and LuxPostFacto [14], and were asked, "Which relit image is better? Please focus on details such as hair and face". The second task was a three-alternative forced-choice (3AFC) task for geometry. Participants were shown the results from GeoRelight, MoGe2 [22], and VGGT [21]. To fairly evaluate the 3D shape, we provided rotating animations of each reconstruction and asked, "Which 3D reconstruction is better? Please focus on details such as eyes and hair". We randomize the order of each method in each question.

We collected a total of 640 votes (32 participants \times 20

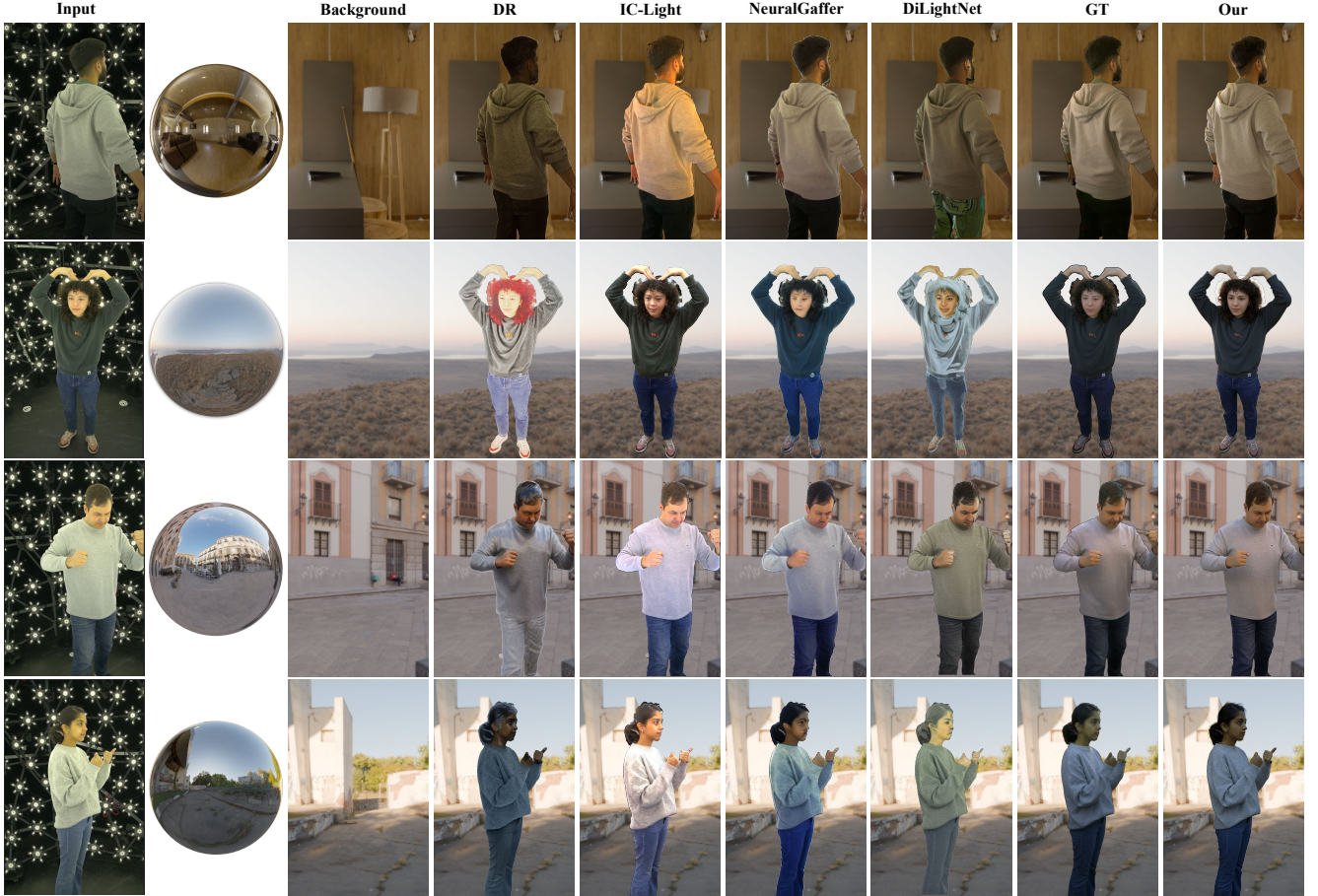


Figure 7. **Qualitative comparison on relighting on HumanOLAT.** Our model (right) produces more physically-plausible results compared to open-source baselines.

comparisons) for each task. The results demonstrate an overwhelming preference for our method. For relighting, in the comparison against LuxPostFacto, GeoRelight was preferred in 93.8% (596.5) of votes, while LuxPostFacto was preferred in only 6.2% (43.5) of votes. For geometry, in the reconstruction task, GeoRelight was preferred in 93.4% (598.0) of votes, decisively outperforming both MoGe2 at 5.2% (33.0) and VGGT at 1.4% (9.0).

Our user study measures perceived photorealism and geometric accuracy, which strongly validates our model’s superior performance, aligning with the quantitative metrics in the main paper.

5.2. Qualitative Comparison on Relighting

In Figure 7, we provide a qualitative comparison for the task of relighting on the HumanOLAT dataset. We compare GeoRelight against state-of-the-art methods, including DiffusionRenderer (DR) [12], IC-Light [27], NeuralGaffer [6], and DiLightNet [25].

Our method consistently demonstrates superior photorealism and physical plausibility. We observe that many baselines struggle with these examples. DiffusionRenderer of-

ten produces results that are overly dark or lack contrast. DiLightNet shows significant instability, frequently introducing severe color artifacts. Methods like IC-Light and NeuralGaffer can struggle with correct exposure, color balance, or preserving fine details.

In contrast, GeoRelight produces relit images that are free of artifacts and well-harmonized with the target illumination. Our method properly handles complex interactions between light, hair, and clothing, and accurately renders shadows consistent with the target environment map.

5.3. Qualitative Comparison on Reconstruction

We compare our 3D reconstruction against the leading feed-forward geometry estimators, VGGT [21] and MoGe2 [22]. We also provide qualitative comparison on geometry reconstruction in the main paper (Fig. 7).

We observe that VGGT often produces overly smooth or “blobby” reconstructions. It captures the general human form but fails to preserve fine-grained details such as clothing folds, hair, or accessories.

While MoGe2 is able to capture more high-frequency detail, it frequently introduces significant noise, “floater” arti-

facts, and non-manifold holes in the geometry. This results in a brittle and visually unappealing surface.

GeoRelight successfully combines the strengths of both approaches. Our reconstructions are both robustly complete and rich in sharp detail. Our method accurately captures complex surfaces like clothing wrinkles, accessories, and hair, all while maintaining a coherent and clean 3D mesh. We refer readers to our supplementary video, which shows these 3D reconstructions as rotating animations for a full multi-view comparison.

6. Extension to Video Relighting

GeoRelight is primarily designed for single-image relighting and reconstruction by repurposing the temporal dimension T of a pretrained video DiT as a modality dimension M . However, the framework can be naturally extended to multi-modal video relighting with minimal architectural modification.

Architecture Modification The key change is in the positional encoding. In our static model, we apply a shared 2D RoPE [19] ($H \times W$) identically across all M modalities to ensure spatial alignment. For the video extension, we replace this with a shared 3D RoPE ($T \times H \times W$) across all M modalities. This allows the model to reason about both spatial correlations across modalities and temporal consistency across frames, while still maintaining per-modality spatial alignment.

Preliminary Experiment To validate this approach, we generated 4,000 synthetic human videos and trained a video variant of GeoRelight. Due to computational constraints, we use 17-frame videos at 1280×832 resolution, which are compressed to $3 \times 160 \times 104$ latents using the Cosmos $8 \times 8 \times 8$ causal VAE [15]. The model is initialized from our trained static GeoRelight weights. Our preliminary results show that this video extension achieves meaningful temporal consistency for multi-modal video relighting, confirming that the architecture supports this generalization. A full exploration of video relighting and reconstruction is an exciting direction for future work.

7. Limitations and Failure Cases

While GeoRelight demonstrates state-of-the-art performance and high generalization, our model shares limitations common to generative frameworks and also has unique failure modes related to its joint-generation task.

Texture-Geometry Ambiguity: As a single-image model, GeoRelight can struggle with the ambiguity between a 3D object and a 2D texture. For example, a small object held by the subject (e.g., a flower) may be incorrectly interpreted as a 2D texture pattern. In such cases, the object is “baked”

into the albedo map but is entirely absent from the surface normal and geometry (iNOD) outputs. This leads to a physically implausible relit image where the object appears as a flat pattern rather than a 3D object that should cast its own shadows.

Temporal Inconsistency: GeoRelight is a single-image framework. Generating a sequence of images (e.g., by rotating the light source) requires multiple independent forward passes. Because the generative process is not strictly deterministic, this can result in small inconsistencies between frames, leading to “temporal flickering” when viewed as a video. Enforcing temporal consistency for video relighting is a significant challenge and a clear direction for future work.

Out-of-Distribution Subjects: Though our model is robust, it can be challenged by subjects with extreme, out-of-distribution (OOD) poses, severe occlusions, or highly complex, non-Lambertian materials (e.g., metallic or specular clothing) that are rare in our hybrid training data.

References

- [1] Abhayexe. Historic european brick building, roebuck. [Link](#), 2026. [License: Free Standard Sketchfab](#). 4
- [2] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3-d point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(5):698–700, 1987. 5
- [3] Sumit Chaturvedi, Mengwei Ren, Yannick Hold-Geoffroy, Jingyuan Liu, Julie Dorsey, and Zhixin Shu. 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. 5
- [4] Grodan. Cappybara 3d model. [Link](#), 2026. [License: CC Attribution \(CC BY 4.0\)](#). 4
- [5] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 2
- [6] Haian Jin, Yuan Li, Fujun Luan, Yuanbo Xiangli, Sai Bi, Kai Zhang, Zexiang Xu, Jin Sun, and Noah Snavely. Neural gaffer: Relighting any object via diffusion. In *Advances in Neural Information Processing Systems*, 2024. 5, 6
- [7] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. *arXiv preprint arXiv:2408.12569*, 2024. 5
- [8] Hoon Kim, Minje Jang, Wonjun Yoon, Jisoo Lee, Donghyun Na, and Sanghyun Woo. Switchlight: Co-design of physics-driven architecture and pre-training framework for human portrait relighting, 2024. 3, 5
- [9] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 2
- [10] LAION. Releasing re-laion-5b: transparent iteration on laion-5b with additional safety fixes. <https://laion.ai/blog/relaion-5b/>, 2024. Accessed: 30 aug, 2024. 3

- [11] Shikai Li, Jianglin Fu, Kaiyuan Liu, Wentao Wang, Kwan-Yee Lin, and Wayne Wu. Cosmicman: A text-to-image foundation model for humans. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 3
- [12] Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Zhi-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, and Zian Wang. Diffusion-renderer: Neural inverse and forward rendering with video diffusion models. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1, 3, 5, 6
- [13] 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. Codec Avatar Studio: Paired Human Captures for Complete, Driveable, and Generalizable Avatars. *NeurIPS Track on Datasets and Benchmarks*, 2024. 3
- [14] Yiqun Mei, Mingming He, Li Ma, Julien Philip, Wenqi Xian, David M George, Xueming Yu, Gabriel Dedic, Ahmet Levant Tasel, Ning Yu, Vishal M Patel, and Paul Debevec. 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. 3, 5
- [15] NVIDIA, :, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Afsan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezani, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefanik, Shitao Tang, Lyne Tchampi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qingsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai, 2025. 1, 4, 7
- [16] Rohit Pandey, Sergio Orts-Escolano, Chloe LeGendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. Total relighting: Learning to relight portraits for background replacement. 2021. 3, 5
- [17] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 1
- [18] Pigsell. Grim reaper faceless death black wings poly. [Link](#), 2026. [License: Standard Fab](#). 4
- [19] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. 1, 7
- [20] Timo Teufel, Pulkit Gera, Xilong Zhou, Umar Iqbal, Pramod Rao, Jan Kautz, Vladislav Golyanik, and Christian Theobalt. Humanolat: A large-scale dataset for full-body human relighting and novel-view synthesis. In *International Conference on Computer Vision (ICCV)*, 2025. 5
- [21] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 4, 5, 6
- [22] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details, 2025. 5, 6
- [23] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 4
- [24] Shaofei Wang, Tomas Simon, Igor Santesteban, Timur Bagautdinov, Junxuan Li, Vasu Agrawal, Fabian Prada, Shoou-I Yu, Pace Nalbony, Matt Gramlich, Roman Lubachersky, Chenglei Wu, Javier Romero, Jason Saragih, Michael Zollhoefer, Andreas Geiger, Siyu Tang, and Shunsuke Saito. Relightable full-body gaussian codec avatars. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, New York, NY, USA, 2025. Association for Computing Machinery. 3
- [25] Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. Dilightnet: Fine-grained lighting control for diffusion-based image generation. In *ACM SIGGRAPH 2024 Conference Papers*, 2024. 5, 6
- [26] Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. Rgbx: Image decomposition and synthesis using material- and lighting-aware diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery. 5
- [27] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *The Thirteenth International Conference on Learning Representations*, 2025. 5, 6