SELF-SUPERVISED SEPARATE MODELING OF LIGHTING AND APPEARANCE FOR PHOTOREALISTIC PORTRAIT RELIGHTING

Anonymous authors
Paper under double-blind review



Figure 1: **Comparison of portrait relighting.** By separately modeling of lighting and appearance under the supervision of real-world portrait images themselves, our method enables photorealistic portrait relighting, while faithfully preserving the lighting-independent appearance contexts for both the foreground and background. The target lighting is given in the bottom-left corner of the input.

ABSTRACT

Existing portrait relighting methods often rely on a highly ambiguous inverse rendering process to estimate the face geometry, reflectance, and lighting, making them prone to visual artifacts. Moreover, their reliance on synthetic paired training data further leads to performance degradation and insufficient generalizability. In this work, we present a self-supervised portrait relighting framework that allows to produce photorealistic results by separately modeling lighting and appearance. To this end, we design appearance adapter and geometry-aware light modulator to obtain appearance and lighting embeddings that are able to reconstruct the input portrait. To avoid possible entanglement of lighting and appearance, we further refine the separation with a light perturbation based reconstruction optimization. Experiments show that our method outperforms existing methods in visual fidelity and lighting coherence, and generalizes well to diverse real-world and even out-of-domain portraits (e.g., cartoon and animation).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Portrait relighting enables post-capture lighting adjustment without reshooting, supporting applications in portrait enhancement, photographic post-processing, film and advertising production, and virtual human creation. Although recent advances in generative models have made remarkable progress in this field, achieving photorealistic lighting changes while faithfully maintaining the content remains a challenge.

Recent portrait relighting methods can be broadly categorized into two types. Among them, one type of methods use generative models to directly produce relit images conditioned on the input image and target lighting (Chaturvedi et al., 2025; Ren et al., 2024; Zhang et al., 2025; Chadebec et al., 2025; Liu et al., 2026; Mei et al., 2025). However, these methods lack geometry-aware lighting modeling, making it difficult to produce realistic lighting effects consistent with the portrait geometry. Moreover, lighting adjustment remains coupled with the generation of image content, potentially altering identity, local structures, or background content. The other type of methods first solve an inverse rendering problem to obtain intrinsic attributes, such as surface normals, albedo, and material parameters, and then produces the output through rerendering (Kim et al., 2024; Pandey et al., 2021; Liang et al., 2025; 2026; Zeng et al., 2024b). Although they enables explicit control of lighting, inaccurate estimation of intrinsic attributes can lead to disturbing visual artifacts. This limitation motivate us to separately model portrait appearance and geometry-aware lighting without performing inverse rendering.

Training data poses an extra challenge in this field. Most existing methods (Zhang et al., 2025; Chadebec et al., 2025; Chaturvedi et al., 2025; Yeh et al., 2022; Chen et al., 2025; Rao et al., 2024; Liang et al., 2026; Liu et al., 2026; Mei et al., 2023; Chaturvedi et al., 2026; Cha et al., 2025; Ren et al., 2024; Kim et al., 2024; Xue et al., 2026; Pandey et al., 2021; Magar et al., 2025; Liu et al., 2025; Zeng et al., 2024a; Liang et al., 2025) rely on synthetic paired training data, constructed either by linearly combining one-light-at-a-time (OLAT) (Teufel et al., 2025) captures according to environment maps or by rendering virtual subjects under simulated illumination. However, OLAT acquisition requires expensive specialized equipment and precise calibration, limiting data scale and diversity. Virtual rendering though offers a convenient data collection alternative, but will inevitably introduce domain gap between rendered images and real photographs, limiting the generalization capability to real-world portraits.

To address these challenges, we present a self-supervised portrait relighting framework that separately models portrait appearance and geometry-aware lighting, enabling photorealistic relighting. Our key idea is to first establish separate appearance and lighting conditioning pathways through portrait reconstruction, and then refine their separation through a light perturbation strategy. Our framework involves two training stages. In the first stage, we introduce normal conditioning and an Appearance Adapter to provide complementary geometric and global appearance references for reconstruction. We also design a Geometry-Aware Light Modulator that converts lighting conditions into shading hints rendered from coarse facial geometry and combines them with geometric information to guide the reconstruction, associating lighting changes to portrait geometry. In the second stage, we introduce light perturbation based reconstruction optimization to reduce interference from source-lighting cues retained in the appearance representation, encouraging the model to follow the specified lighting condition despite illumination changes in the appearance reference.

Our main contributions are summarized as follows:

- We present a self-supervised framework for photorealistic portrait relighting, which unlike previous method, does not rely on synthetic paired training data.
- We develop appearance adapter and geometry-aware light modulator to separately model lighting and appearance.
- We introduce a light perturbation based reconstruction optimization to reduce interference from source shadows and highlights during target-light adjustment, which helps refine the separation of lighting and appearance.

2 Related Works

Portrait relighting. Early learning-based methods are mostly 2D-based method, while recent approaches have explored both 3D-aware and 2D formulations. Efficient 3D GAN frameworks, such as EG3D (Chan et al., 2022) and GSGAN (Hyun & Heo, 2024), have enabled 3D-aware portrait relighting. NeRFFaceLighting (Jiang et al., 2023) learns an implicit representation for facial illumination disentanglement, while Holo-Relighting (Mei et al., 2024) extracts 3D-aware features and maps them to a relightable volumetric representation. GSHeadRelight (Lv et al., 2025) and LumiGAN (Deng et al., 2024) further propose to incorporate physically motivated light transport to learn relightable 3D representations and support albedo-aware relighting. Despite their advantages in 3D consistency, these methods may compromise high-frequency textures and local details (Yeh et al., 2022). Inverse-rendering-based methods provide another solution for portrait relighting (Kim et al., 2024; Pandey et al., 2021). These methods typically first estimate intrinsic components, such as surface normals, albedo, and lighting, from the input and then rerender it under the target light. This paradigm provides explicit control over the relighting process. However, their results depend heavily on the accuracy of the estimated intrinsic components, and may contain visual artifacts.

Diffusion-based Relighting. The development of diffusion models has stimulated extensive research on image and portrait relighting (Zhang et al., 2025; Chadebec et al., 2025; Ren et al., 2024; Jin et al., 2024; Zeng et al., 2024a; Chaturvedi et al., 2025; Cha et al., 2025; Chen et al., 2025; He et al., 2024; Liu et al., 2025; 2026). Benefiting from the strong visual generative priors of pretrained diffusion models, these methods have made significant progress in the realism and flexibility of relighting. IC-Light (Zhang et al., 2025) introduces light-transport consistency to improve light control. DreamLight (Liu et al., 2026) incorporates position-guided light features to enhance the model's awareness of the target light direction. DiLightNet (Zeng et al., 2024a) uses radiance hints rendered from coarse geometry to guide diffusion-based relighting. HarmLight (Ren et al., 2024) encodes background-conditioned light features to enable background-aware portrait relighting. SynthLight (Chaturvedi et al., 2025) and VLSRelight (Yeh et al., 2022) perform end-to-end relighting training using virtual portrait assets and further improve the realism of their results through fine-tuning on real images. Diffusion models have also been incorporated into inverse-rendering-based relighting pipelines. PI-Light (Liang et al., 2026), DiffusionRenderer (Liang et al., 2025), and RGB-X (Zeng et al., 2024b) extend the conventional inverse-rendering paradigm through diffusion-based intrinsic-component estimation and neural rendering. Although these methods leverage generative priors of diffusion models to improve intrinsic estimation and rendering quality, their results are very sensitive to errors in accurate estimation.

Identity-consistent portrait generation. DreamBooth (Ruiz et al., 2023) proposes to learn subject-specific identifiers through model fine-tuning, while PhotoMaker (Li et al., 2024) and IP-Adapter-FaceID (Ye et al., 2023) encode identity information into compact image embeddings for conditioning. InstantID (Wang et al., 2024) further combines identity embeddings with facial landmarks for improved identity preservation. These methods demonstrate that reference-derived identity features can serve as stable anchors during diffusion generation. Inspired by this principle, we extend reference anchoring to portrait relighting, where the reference must preserve identity and appearance while remaining insensitive to its source illumination. Our approach therefore combines appearance and geometric anchors with light-perturbation consistency and local detail preservation to achieve controllable relighting with improved content consistency.

3 Method

We aim to develop a portrait relighting framework that does not rely on synthetic paired training data, while allowing for controllable high-fidelity relighting. To this end, we separately model portrait appearance and light in the diffusion latent space through a two-stage training framework, as illustrated in Fig. 2. In Stage 1 (Sec. 3.1), we establish global appearance anchors to preserve identity, geometry, and background, and also introduce a Geometry-Aware Light Modulator (GALM) to model lighting. These separate conditioning representations enable initial separation of appearance and lighting. In Stage 2 (Sec. 3.2), we perturb the light of appearance-reference inputs and impose consistency constraints to reduce interference from source-illumination leakage during target-lighting modulation, thereby further refining the separation.

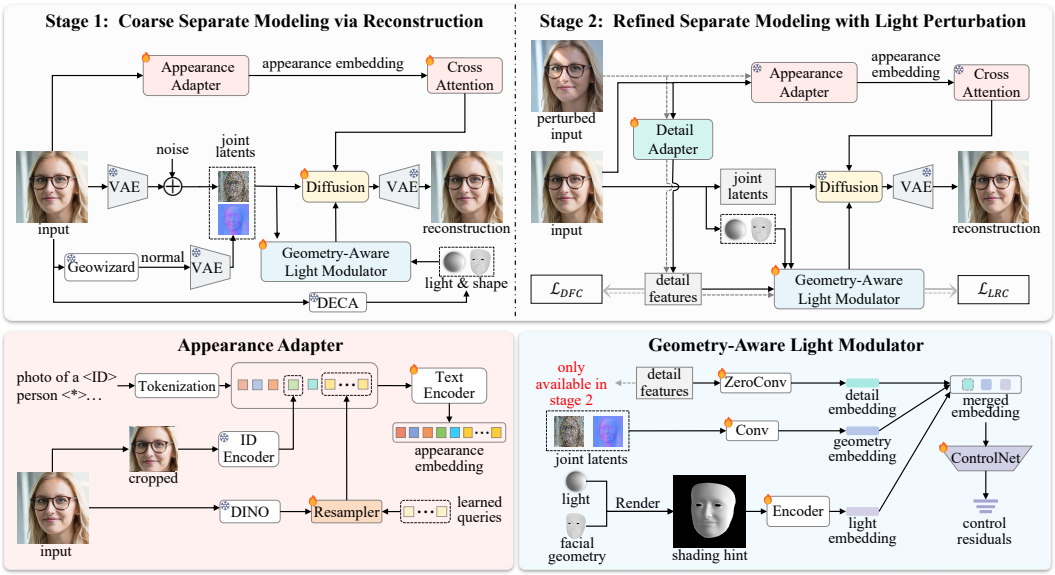

Figure 2: **Overview of our framework.** The training of our framework consists of two stages. Stage 1 combines normal conditioning and an Appearance Adapter to model portrait geometry and appearance, while a Geometry-Aware Light Modulator is utilized to model light and enables spatial illumination control. Stage 2 further refines the separation of lighting and appearance through light perturbation based consistency constraints, and a Detail Adapter is designed to preserve fine-grained details. Note, the inference procedure is given in Fig. 3.

3.1 Stage 1: Coarse Separate Modeling via Reconstruction

In Stage 1, we establish the basic portrait relighting model with separate appearance and light conditioning. Given an image $x_0 \in \mathbb{R}^{H\times W\times 3}$, we estimate its normal map $N \in \mathbb{R}^{H\times W\times 3}$ using pretrained GeoWizard (Fu et al., 2024) and encode both with a VAE encoder as $z_0 = \mathcal{E}(x_0)$ and $z_n = \mathcal{E}(N)$. At timestep t, the noisy latent is $z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon$, where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$, and β_i is the noise variance at step i. We concatenate z_t and z_n along the channel dimension to form the joint latents $u_t = \mathrm{Cat}(z_t, z_n)$, and feed them into the U-Net with correspondingly expanded input channels. The normal condition provides an explicit spatial reference for geometric consistency throughout denoising.

Appearance Adapter. We use a frozen ArcFace encoder (Deng et al., 2019) to extract identity feature $F_{\mathrm{id}} \in \mathbb{R}^{512}$ from the cropped face, and a frozen DINOv2 encoder (Oquab et al., 2023) to extract image features $F_g \in \mathbb{R}^{257\times 768}$ from the input portrait. We normalize and pad the identity feature to obtain $f_{\mathrm{id}} \in \mathbb{R}^{768}$. A trainable resampler with $m = 8$ learned queries compresses F_g into m appearance tokens $f_g \in \mathbb{R}^{m\times 768}$. Next, we tokenize the pseudo-prompt "photo of a $\langle$ID$\rangle$ person $\langle * \rangle \ldots$" to obtain an embedding sequence $s = \{e_1, e_2, e_3, e_{\mathrm{id}}, e_5, e_*, \ldots\}$. We replace the embedding e_{id} associated with $\langle$ID$\rangle$ with f_{id}, and the appearance placeholder embedding e_* with f_g. The resulting sequence s' is processed by the text encoder to obtain the appearance embedding $E_a = \mathcal{A}(x_0) \in \mathbb{R}^{77\times 768}$. The resulting embeddings condition the diffusion U-Net through cross-attention. During Stage 1, we fine-tune the resampler and text encoder to adapt the appearance representation to portrait reconstruction. Together with normal conditioning, the Appearance Adapter provides complementary identity, appearance, and spatial references for the generation process.

Geometry-Aware Light Modulator. Global illumination representations provide limited spatial guidance. We therefore render the estimated illumination on a coarse facial mesh and combine the resulting shading hint with normal information for spatially aware illumination control. Specifically, DECA (Feng et al., 2021) estimates facial shape β, pose θ, and expression ϕ, from which FLAME (Li et al., 2017) generates the coarse mesh: $S = \mathrm{FLAME}(\beta, \theta, \phi)$. DECA also estimates nine RGB spherical-harmonics coefficients $L = \{L_i\}_{i=1}^{9}$. We average each coefficient across RGB channels: $\widetilde{L} = \{\frac{L_i^R + L_i^G + L_i^B}{3}\}_{i=1}^{9}$. We render the shading hint $\ell_s = \mathcal{R}\mathrm{SH}(\widetilde{L}, S)$ and encode it as

$h_l = \mathcal{E}_L(\ell_s)$. A lightweight convolutional layer jointly encodes the normal and noisy latents into a geometry embedding h_g. The two embeddings are combined and processed by ControlNet $\mathcal{C}_\varphi$: $R_I = \mathcal{C}_\varphi\left(h_l + h_g, t, \mathcal{A}(x_0)\right)$, where $\mathcal{A}(x_0)$ is the appearance embedding. The resulting multi-scale residuals R_I are injected into the diffusion U-Net to provide geometry-aware illumination guidance with spatially varying effects.

Training Objective. At this stage, we train on large-scale natural images using the standard diffusion noise-prediction loss. We first optimize the U-Net and Appearance Adapter to establish stable, appearance-consistent generation, then unfreeze GALM to learn light-conditioned spatial modulation. The training objective is

$$\mathcal{L}_{\text{stage1}} = \mathbb{E}_{x_0,t,\epsilon}\left[\left\|\epsilon - \epsilon_\theta\left(z_t, z_n, t, \mathcal{A}(x_0), \ell_s\right)\right\|_2^2\right]. \tag{1}$$

3.2 Stage 2: Refined Separate Modeling with Light Perturbation

Stage 1 provides appearance consistency and spatial light control, but its appearance anchors may retain source-illumination cues and lack fine-grained spatial details. Stage 2 addresses these limitations by perturbing illumination in the appearance references and enforcing consistency under a shared reconstruction target. We further introduce a Detail Adapter to complement the global appearance anchors with local spatial features.

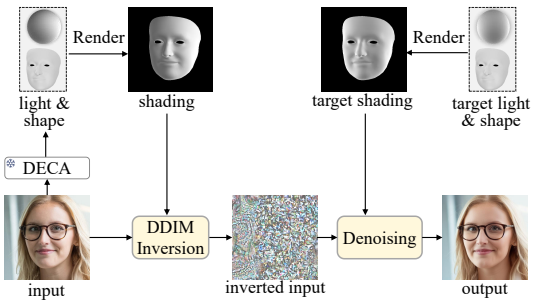

Figure 3: **Inference pipeline.** During inference, we first invert the input under its source illumination and then denoise it using a shading hint rendered under the target illumination, while keeping the geometry and appearance conditions fixed.

Given a natural image x_0, we use Qwen-Image-Edit (Wu et al., 2025) to generate a light-perturbed image x_0^p by randomly modifying the lighting while preserving portrait content and background. We set the original image as $x_0^o = x_0$ and retain it as the reconstruction target, while using x_0^p as an alternative appearance reference. We construct two conditioning paths using x_0^o and x_0^p. Both share the noisy latent z_t derived from x_0, normal latent z_n, timestep t, and shading hint ℓ_s, differing only in the appearance and detail references. This encourages the model to preserve content while remaining invariant to source light in the appearance references.

Detail Adapter. The Appearance Adapter provides a compact global appearance reference but may lose fine-grained spatial details. We therefore introduce a Detail Adapter $\mathcal{A}_d$ to provide local features through the light-modulation pathway. For each reference image x_0^j with $j \in \{o, p\}$, a frozen DINOv2 encoder (Oquab et al., 2023) extracts patch features $F_p^j \in \mathbb{R}^{256\times768}$. We reshape them into a spatial feature map and process them with a detail encoder comprising residual blocks, smooth upsampling, and convolutional layers. A SiLU activation and zero-initialized convolution then project the features into the GALM conditioning space. During training, we randomly drop the Detail Adapter with probability 0.2 to reduce over-reliance on local features.

Reconstruction Consistency. Although the two appearance references have different source light, both conditioning paths reconstruct the same original image. The noise prediction for each path is

$$\widehat{\epsilon}^j = \epsilon_\theta\left(z_t, z_n, t, \mathcal{A}(x_0^j), \mathcal{A}_d(x_0^j), \ell_s\right), \qquad j \in \{o, p\}, \tag{2}$$

where the detail condition is incorporated through GALM. We use the following loss to supervise both paths toward a consistent prediction:

$$\mathcal{L}_{\text{RC}} = \frac{1}{2}\mathbb{E}\left[\|\epsilon - \widehat{\epsilon}^o\|_2^2 + \|\epsilon - \widehat{\epsilon}^p\|_2^2\right]. \tag{3}$$

This objective encourages the model to follow the shared light condition while reducing interference from lighting cues in the appearance reference.

Light-Response Consistency. We further align the ControlNet bottleneck residuals to make light control invariant to changes in the appearance reference. Let $R_{I,\text{mid}}^j$ denote the bottleneck residual

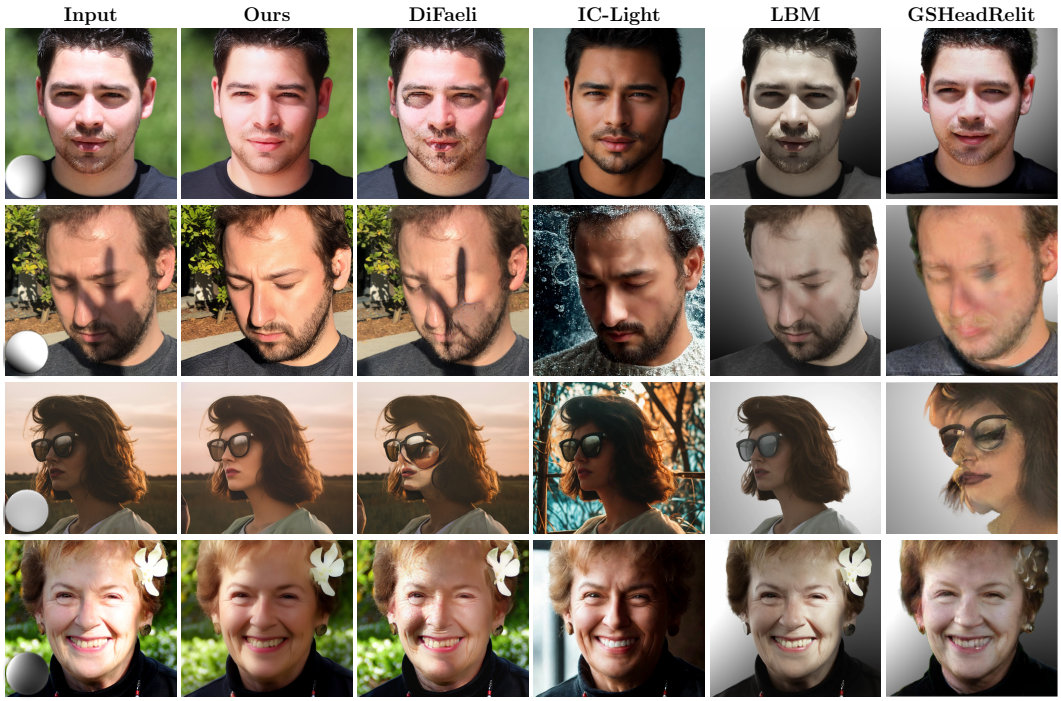

Figure 4: **Qualitative comparison on real-world portraits with complex facial illumination.**

for path j. We define the following light-response consistency loss

$$\mathcal{L}_{\text{LRC}} = \left\| R^o_{I,\text{mid}} - R^p_{I,\text{mid}} \right\|_2^2 . \tag{4}$$

Detail-Feature Consistency. Local features may also encode source shadows and highlights. To reduce the transmission of these lighting cues from the Detail Adapter to GALM and decrease its sensitivity to source light, we align the detail features before the SiLU activation and zero-convolution projection, and define the following loss

$$\mathcal{L}_{\text{DFC}} = \| f^o_d - f^p_d \|_2^2 . \tag{5}$$

Training Objective. We initialize Stage 2 from the Stage 1 model and freeze the diffusion U-Net and Appearance Adapter. We then optimize the GALM and newly introduced Detail Adapter using the following objective:

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{RC}} + \lambda_{\text{LRC}}\mathcal{L}_{\text{LRC}} + \lambda_{\text{DFC}}\mathcal{L}_{\text{DFC}}, \tag{6}$$

where $\lambda_{\text{LRC}} = \lambda_{\text{DFC}} = 0.1$. Together, these objectives promote reconstruction consistency and light-invariant control while preserving local details.

4 Experiments

4.1 Implementation Details

Training Details. We train our model on the FFHQ portrait dataset (Karras et al., 2019), preprocessed according to the EG3D pipeline (Chan et al., 2022). For Stage 2, we apply Qwen-Image-Edit (Wu et al., 2025) to a subset of 10,000 FFHQ images to introduce Light perturbations to the appearance-reference inputs. To leverage robust facial priors, the Stable Diffusion backbone (Rombach et al., 2022) is initialized with pretrained Arc2Face weights (Papantoniou et al., 2024). More implementation details are provided in Appendix A.

Evaluation Datasets and Metrics. We evaluate on natural portraits and a synthetic benchmark with paired ground truth. For natural portraits, we use CelebA-HQ (Lee et al., 2020) and randomly



Figure 5: **Qualitative comparison on out-of-domain portraits.**

Table 1: **Quantitative comparison on the CelebA-HQ and synthetic dataset.** "All" and "Foreground" indicate the evaluation regions. Best results are shown in bold.

Method	*CelebA-HQ*				*Synth. All*			*Synth. Foreground*			Resolution	Time (s)
	ID ↓	Light ↓	Depth Error ↓	FID ↓	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑		
DiFaReli	0.200	**0.861**	0.043	34.824	21.936	0.186	0.880	19.237	0.151	0.867	256	113
IC-Light	0.255	1.924	0.112	40.771	11.963	0.364	0.736	14.053	0.162	0.812	512	2.89
LBM	0.086	2.749	0.072	33.897	12.119	0.353	0.678	13.804	0.161	0.785	512	0.96
GSHeadRelit	0.504	1.610	0.180	54.968	8.336	0.426	0.564	17.810	0.240	0.726	512	62.65
DiffusionRender	0.469	2.838	0.097	90.308	10.829	0.467	0.547	14.887	0.197	0.739	512	18.64
Ours	**0.079**	1.124	**0.037**	**21.946**	**23.176**	**0.149**	**0.884**	**20.570**	**0.118**	**0.878**	512	2.68

sample target illumination references from the dataset. Following Jiang et al. (2023), we use the off-the-shelf DECA estimator (Feng et al., 2021) to compute Lighting Error (Light). We measure identity discrepancy (ID) using ArcFace (Deng et al., 2019), structural consistency using Depth Anything V2 (Yang et al., 2024), and image realism and fidelity using Fréchet Inception Distance (FID). For paired evaluation, we use MetaHuman in Unreal Engine 5 (ue5, 2025) to create 20 digital human subjects and render 1,000 image pairs across 20 scenes with varying illumination directions and intensities. We compare relit images with their corresponding targets using PSNR, SSIM, and LPIPS (Zhang et al., 2018), computed over both the full image and the foreground region, with foreground masks estimated by BiRefNet (Zheng et al., 2024).

4.2 Comparison with State-of-the-Art Methods

Baselines. We compare our method with DiFaReli (Ponglertnapakorn et al., 2023), IC-Light (Zhang et al., 2025), LBM (Chadebec et al., 2025), DiffusionRenderer (Liang et al., 2025), and GSHeadRelit (Lv et al., 2025), covering self-supervised, diffusion-based, intrinsic-decomposition-based, and 3D Gaussian Splatting-based relighting methods. All methods except GSHeadRelit are diffusion-based. For IC-Light and LBM, we use a proxy background from a sphere rendered under the target spherical-harmonics illumination. For DiffusionRenderer, we use DiffusionLight (Lv et al., 2025) to estimate the target HDR map.

Qualitative Comparison. Figs. 1 and 4 compare our method with existing approaches on unseen portraits, including real-world images, portraits with strong self-shadows or external cast shadows, and underexposed images. Our method faithfully follows the target light, effectively removes shadows and highlights caused by the source light, and better preserves identity, background appearance, and local details. In contrast, existing methods may introduce visual artifacts, retain source-light effects, or cause unintended changes to the face, background, clothing, and other appearance details. Our method also handles challenging light configurations, such as overhead lighting, while producing natural cast shadows from accessories such as eyeglasses that are consistent with the target light. Fig. 5 further demonstrates the robust out-of-domain generalization of our method on artistic portraits and cartoon images. More visual comparisons are provided in Appendix C.

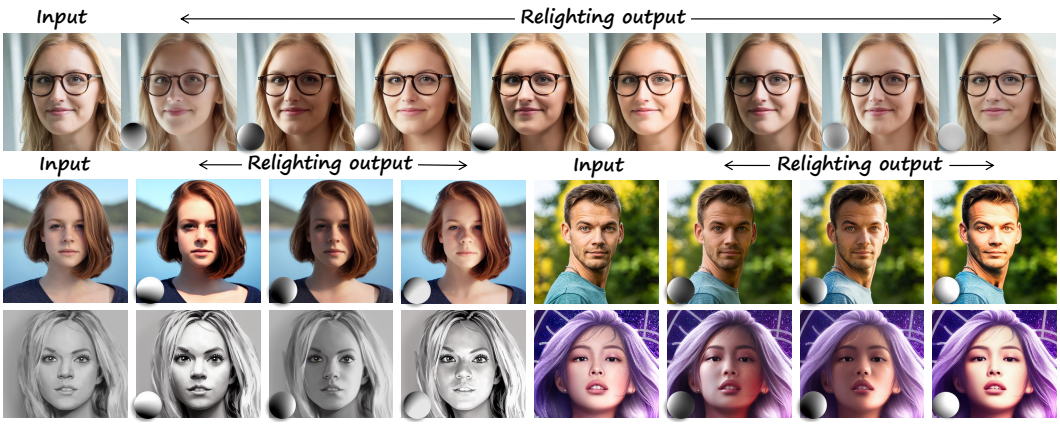

Figure 6: **Controllable lighting editing.**

Quantitative Comparison. Tab. 1 reports quantitative results on CelebA-HQ and our synthetic benchmark. On CelebA-HQ, our method achieves the lowest ID Error, Depth Error, and FID, while maintaining competitive relighting quality. DiFaReli achieves a comparable Light score, but operates at a lower input resolution, reflecting a different trade-off between relighting quality and appearance preservation. Overall, these results demonstrate that our method provides effective illumination control while better preserving identity and facial structure. On the synthetic benchmark, our method achieves the best PSNR, SSIM, and LPIPS across both full-image and foreground evaluations, indicating improved reconstruction fidelity under controlled illumination changes.

4.3 More Analysis

Controllable lighting editing. To evaluate Lighting controllability, we fix the input portrait and vary the target light conditions. For each condition, we render the corresponding target shading and relight the same portrait while keeping the appearance and geometry conditions fixed. Fig. 6 shows the resulting relit portraits. Our method consistently adapts the shadows and highlights to the target light while preserving facial identity, structure, and background appearance across different lighting conditions. These results demonstrate that the proposed disentangled appearance and lighting representations enable consistent and controllable lighting editing. Appendix C presents more light editing results, as well as reference-based light transfer results.

Comparison with powerful multimodal models. Fig. 7 compares our method with Qwen3, ChatGPT6, and Nano Banana. In the illustrated examples, our method follows the target illumination more closely and reduces residual source shadows and highlights. In contrast, the compared models may retain the original shading or introduce unintended changes to portrait details. These results demonstrate that our method better preserves portrait content while achieving the desired illumination changes.

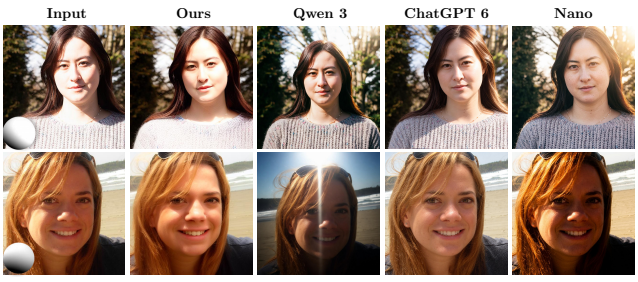

Figure 7: **Comparison with powerful multimodal models.**

Inference time. Tab. 1 reports the inference time. As shown, DiFaReli and GSHeadRelit require more time because of relying on iterative inversion procedures. In comparison, our method achieves a runtime comparable to IC-Light at a resolution of 512×512.

4.4 Ablation Study

We conduct ablation studies on CelebA-HQ to evaluate the effectiveness of normal conditioning (NM), Geometry-Aware Light Modulator (GALM), the Appearance Adapter (AA), Light-Perturbation refinement (LP), and the Detail Adapter (DA). Quantitative results are reported in

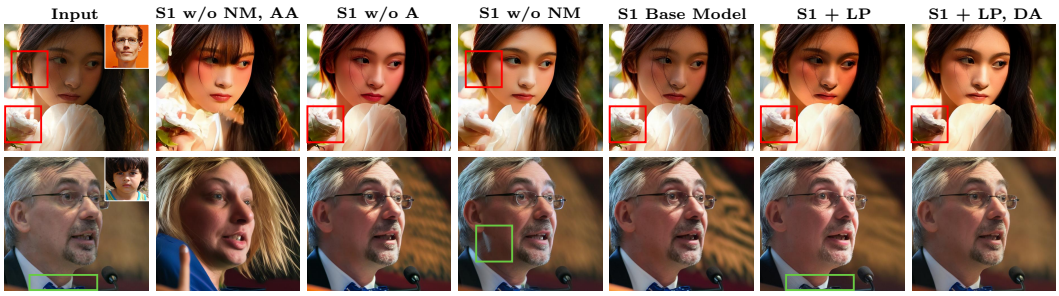

Figure 8: **Visual ablation study on different components in our method.** S1 Base Model denotes the basic portrait relighting model trained in Stage 1. NM, AA, LP, and DA denote the normal map condition, Appearance Adapter, Light perturbation, and Detail Adapter, respectively.

Tab. 2, while visual comparison results are shown in Fig. 8. Additional ablation analyses are provided in Appendix B.

Effect of GALM and appearance adapter. We first ablate GALM and adopt a lighting-injection scheme similar to IC-Light, using only a proxy background rendered under the target illumination as the lighting condition. As shown in the fourth row of Tab. 2, without GALM, the model cannot effectively separate lighting from appearance, leading to lighting cues to be fully captured by the appearance embedding. We then evaluate the Appearance Adapter. As shown in Fig. 8, removing AA results in noticeable identity and appearance changes, whereas incorporating AA better preserves skin tone, hair color, and background content.

Effect of light-perturbation refinement. Although the complete Stage 1 model preserves portrait appearance well, its illumination error remains relatively high, indicating that source-illumination information retained by the appearance conditions can interfere with target-light editing. We therefore compare the Stage 1 base model with a variant trained with Light perturbations and the corresponding consistency constraints. As shown in the fifth and sixth rows of Tab. 2, introducing LP substantially reduces illumination error. The sixth column of Fig. 8 (S1 + LP) further shows improved adaptation to the target illumination.

Table 2: **Ablation study on CelebA-HQ.** ✓ and ✗ indicate with and without an component.

Stage 1			Stage 2		ID ↓	Light ↓	Depth Error ↓
NM	AA	GALM	LP	DA			
✗	✗	✓	✗	✗	0.416	**0.985**	0.104
✓	✗	✓	✗	✗	0.212	1.278	0.061
✗	✓	✓	✗	✗	0.138	1.388	0.057
✓	✓	✗	✗	✗	**0.036**	5.650	**0.030**
✓	✓	✓	✗	✗	0.091	1.569	0.038
✓	✓	✓	✓	✗	0.118	1.120	0.044
✓	✓	✓	✓	✓	0.079	1.124	0.037

Effect of detail adapter. The global appearance representation learned in Stage 1 does not fully preserve local details, and Light-perturbation refinement alone cannot prevent content changes during relighting. As shown in the fifth and sixth columns of Fig. 8 (S1 Base Model and S1 + LP), details such as background, clothing, and skin tone remain susceptible to unintended changes. Adding DA alleviates these artifacts by better preserving local content. The last row of Tab. 2 further shows that DA reduces unintended content changes while maintaining effective illumination control.

5 Conclusion

We have presented a self-supervised portrait relighting framework that enables controllable lighting editing while preserving lighting-invariant portrait appearance. Our method combines normal conditioning, an Appearance Adapter, and a Geometry-Aware Light Modulator to achieve spatially coherent lighting control, while light-perturbed references, a Detail Adapter, and multi-level consistency constraints reduce source-illumination leakage and preserve fine-grained details. Together with DDIM inversion, our method produces high-fidelity full-image relighting with diverse illumination changes while maintaining identity, structure, background, and local details. Experiments on natural portraits and 3D-rendered virtual humans demonstrate improved relighting quality and generalization over existing methods.

AI USE STATEMENT

Generative AI tools were used in this work for data generation, experimental comparison, language polishing, and parts of the code implementation. Specifically, Qwen-Image-Edit was used to generate illumination-perturbed appearance-reference images for 10,000 FFHQ portraits during Stage 2. It was also used to generate randomly relit images for the paired-supervision ablation presented in Appendix B. Qwen3, ChatGPT 6, and Nano Banana were used as comparison systems to produce the qualitative results reported in Fig. 7. OpenAI Codex was used to assist with language polishing and parts of the code implementation. All AI-assisted text and code were reviewed by the authors, and the code was tested and verified for correctness. The authors also examined the AI-generated data and comparison outputs used in the experiments. The research methodology, experimental validation, and scientific conclusions were determined and verified by the authors. The authors take full responsibility for the final content of this work, including all text, claims, results, code, and artifacts produced with the assistance of generative AI.

REFERENCES

Ue5. https://www.unrealengine.com/, 2025.

Junuk Cha, Mengwei Ren, Krishna Kumar Singh, He Zhang, Yannick Hold-Geoffroy, Seunghyun Yoon, HyunJoon Jung, Jae Shin Yoon, and Seungryul Baek. Text2relight: Creative portrait relighting with text guidance. In *AAAI*, volume 39, pp. 1980–1988, 2025.

Clément Chadebec, Onur Tasar, Sanjeev Sreetharan, and Benjamin Aubin. Lbm: Latent bridge matching for fast image-to-image translation. In *ICCV*, pp. 29086–29098, 2025.

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2022.

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. In *CVPR*, pp. 369–379, 2025.

Sumit Chaturvedi, Yannick Hold-Geoffroy, Mengwei Ren, Jingyuan Liu, He Zhang, Yiqun Mei, Julie Dorsey, and Zhixin Shu. Tokenlight: Precise lighting control in images using attribute tokens. In *CVPR*, pp. 19737–19748, 2026.

Zhuo Chen, Chengqun Yang, Zhuo Su, Zheng Lv, Jingnan Gao, Xiaoyuan Zhang, Xiaokang Yang, and Yichao Yan. Polar: A portrait olat dataset and generative framework for illumination-aware face modeling. *arXiv preprint arXiv:2512.13192*, 2025.

Boyang Deng, Yifan Wang, and Gordon Wetzstein. Lumigan: Unconditional generation of relightable 3d human faces. In *3DV*, pp. 302–312. IEEE, 2024.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pp. 4690–4699, 2019.

Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021.

Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *ECCV*, pp. 241–258. Springer, 2024.

Mingming He, Pascal Clausen, Ahmet Levent Taşel, Li Ma, Oliver Pilarski, Wenqi Xian, Laszlo Rikker, Xueming Yu, Ryan Burgert, Ning Yu, et al. Diffrelight: Diffusion-based facial performance relighting. In *SIGGRAPH Asia*, pp. 1–12, 2024.

Sangeek Hyun and Jae-Pil Heo. Gsgan: Adversarial learning for hierarchical generation of 3d gaussian splats. *NeurIPS*, 37:67987–68012, 2024.

Kaiwen Jiang, Shu-Yu Chen, Hongbo Fu, and Lin Gao. Nerffacelighting: Implicit and disentangled face lighting representation leveraging generative prior in neural radiance fields. *ACM Transactions on Graphics*, 42(3):1–18, 2023.

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. *NeurIPS*, 37:141129–141152, 2024.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pp. 4401–4410, 2019.

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. In *CVPR*, pp. 25096–25106, 2024.

Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, pp. 5549–5558, 2020.

Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *CVPR*, pp. 8640–8650. IEEE, 2024.

Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Chih-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, et al. Diffusionrenderer: Neural inverse and forward rendering with video diffusion models. In *CVPR*, pp. 26069–26080. IEEE, 2025.

Zhexin Liang, Zhaoxi Chen, Yongwei Chen, Tianyi Wei, Tengfei Wang, and Xingang Pan. Pi-light: Physics-inspired diffusion for full-image relighting. *arXiv preprint arXiv:2601.22135*, 2026.

Ropeway Liu, Hangjie Yuan, Bo Dong, Jiazheng Xing, Jinwang Wang, Rui Zhao, Yan Xing, Weihua Chen, and Fan Wang. 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. Dreamlight: Towards harmonious and consistent image relighting. *NeurIPS*, 38:36041–36060, 2026.

Henglei Lv, Bailin Deng, Jianzhu Guo, Xiaoqiang Liu, Pengfei Wan, Di Zhang, and Lin Gao. Gsheadrelight: Fast relightability for 3d gaussian head synthesis. In *SIGGRAPH*, pp. 1–12, 2025.

Nadav Magar, Amir Hertz, Eric Tabellion, Yael Pritch, Alex Rav-Acha, Ariel Shamir, and Yedid Hoshen. Lightlab: Controlling light sources in images with diffusion models. In *SIGGRAPH*, pp. 1–11, 2025.

Yiqun Mei, He Zhang, Xuaner Zhang, Jianming Zhang, Zhixin Shu, Yilin Wang, Zijun Wei, Shi Yan, HyunJoon Jung, and Vishal M Patel. Lightpainter: Interactive portrait relighting with freehand scribble. In *CVPR*, pp. 195–205, 2023.

Yiqun Mei, Yu Zeng, He Zhang, Zhixin Shu, Xuaner Zhang, Sai Bi, Jianming Zhang, HyunJoon Jung, and Vishal M Patel. Holo-relighting: Controllable volumetric portrait relighting from a single image. In *CVPR*, pp. 4263–4273, 2024.

Yiqun Mei, Mingming He, Li Ma, Julien Philip, Wenqi Xian, David M George, Xueming Yu, Gabriel Dedic, Ahmet Levent Taşel, Ning Yu, et al. Lux post facto: Learning portrait performance relighting with conditional video diffusion and a hybrid dataset. In *CVPR*, pp. 5510–5522. IEEE, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Rohit Pandey, Sergio Orts-Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul E Debevec, and Sean Ryan Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Trans. Graph.*, 40(4):43–1, 2021.

Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2face: A foundation model for id-consistent human faces. In *ECCV*, pp. 241–261. Springer, 2024.

Puntawat Ponglertnapakorn, Nontawat Tritrong, and Supasorn Suwajanakorn. Difareli: Diffusion face relighting. In *ICCV*, pp. 22646–22657, 2023.

Pramod Rao, Gereon Fox, Abhimitra Meka, Mallikarjun BR, Fangneng Zhan, Tim Weyrich, Bernd Bickel, Hanspeter Pfister, Wojciech Matusik, Mohamed Elgharib, et al. Lite2relight: 3d-aware single image portrait relighting. In *SIGGRAPH*, pp. 1–12, 2024.

Mengwei Ren, Wei Xiong, Jae Shin Yoon, Zhixin Shu, Jianming Zhang, HyunJoon Jung, Guido Gerig, and He Zhang. Relightful harmonization: Lighting-aware portrait background replacement. In *CVPR*, pp. 6452–6462, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pp. 22500–22510. IEEE, 2023.

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 *ICCV*, pp. 29131–29141. IEEE, 2025.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Yuxuan Xue, Ruofan Liang, Egor Zakharov, Timur Bagautdinov, Chen Cao, Giljoo Nam, Shunsuke Saito, Gerard Pons-Moll, and Javier Romero. Georelight: Learning joint geometrical relighting and reconstruction with flexible multi-modal diffusion transformers. *arXiv preprint arXiv:2604.20715*, 2026.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *NeurIPS*, 37:21875–21911, 2024.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. Learning to relight portrait images via a virtual light stage and synthetic-to-real adaptation. *ACM Transactions on Graphics (TOG)*, 41(6):1–21, 2022.

Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. Dilightnet: Fine-grained lighting control for diffusion-based image generation. In *SIGGRAPH*, pp. 1–12, 2024a.

Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. Rgb↔x: Image decomposition and synthesis using material-and lighting-aware diffusion models. In *SIGGRAPH*, pp. 1–11, 2024b.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *ICLR*, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595, 2018.

Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *arXiv preprint arXiv:2401.03407*, 2024.

A Implementation Details

Training is conducted sequentially: in Stage 1, we optimize the U-Net and Appearance Adapter for 40,000 steps at a learning rate of 5×10^{-6}, and then unfreeze the Geometry-Aware Light Modulator for an additional 200,000 steps at 1×10^{-5}; in Stage 2, we initialize from the Stage 1 checkpoint, freeze the U-Net and Appearance Adapter, and optimize GALM and the Detail Adapter for 200,000 steps at 1×10^{-5}. All experiments are executed on a single NVIDIA RTX 4090 GPU (24 GB) using the AdamW optimizer with a batch size of 4.

Portrait Relighting at Inference. Given an input portrait x_0, we first estimate its source illumination and render the shading hint ℓ_s^{src}. We then perform DDIM inversion conditioned on the source shading hint, normal latent, and appearance features:

$$z_T^{\text{src}} = \text{DDIMInv}\left(z_0, z_n, \mathcal{A}(x_0), \mathcal{A}_d(x_0), \ell_s^{\text{src}}\right), \tag{7}$$

where $z_0 = \mathcal{E}(x_0)$. Next, we render the same facial mesh under the target illumination to obtain the target shading hint ℓ_s^{tar}. Keeping all other conditions unchanged, we denoise from the inverted latent:

$$z_0^{\text{tar}} = \text{Denoising}\left(z_T^{\text{src}}, z_n, \mathcal{A}(x_0), \mathcal{A}_d(x_0), \ell_s^{\text{tar}}\right). \tag{8}$$

Finally, the VAE decoder produces the relit portrait $x_0^{\text{tar}} = \mathcal{D}(z_0^{\text{tar}})$.

B More Ablation Studies

Visualization of the impact of illumination perturbation To examine whether illumination perturbation (LP) enhances GALM's sensitivity to the target illumination, we visualize differences in the control residuals produced by GALM. For each variant, we fix the noisy latent, facial geometry, appearance features, and diffusion timestep, and evaluate GALM separately using the source and target shading conditions. We then compute $\Delta R = R_{\text{target}} - R_{\text{source}}$, where R denotes the multi-scale control residuals, and visualize the channel-wise Root-Mean-Square (RMS) magnitude of ΔR during early and late denoising. As shown in Figure 9, GALM with LP produces stronger and more persistent target-lighting responses, particularly in regions requiring illumination correction, leading to more effective removal of the original shadows in the final output.

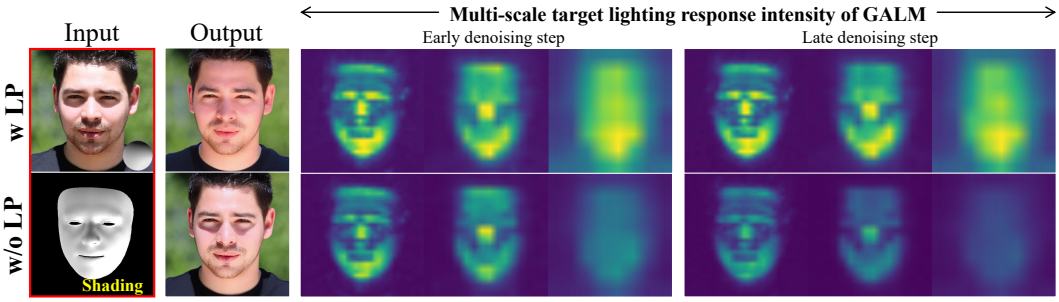

Figure 9: Visualization of GALM with and without illumination perturbation (LP). The heatmaps show the channel-wise RMS intensity of light-specific residuals at three scales during early and late denoising. LP yields stronger and more persistent target-lighting responses, enabling effective removal of source shadows and more faithful relighting.

Effect of Self-Supervised Illumination Disentanglement To validate the advantages of our self-supervised disentanglement strategy, we construct a counterpart trained with paired supervision. Specifically, we use Qwen to generate randomly relit versions of in-the-wild FFHQ images (Karras

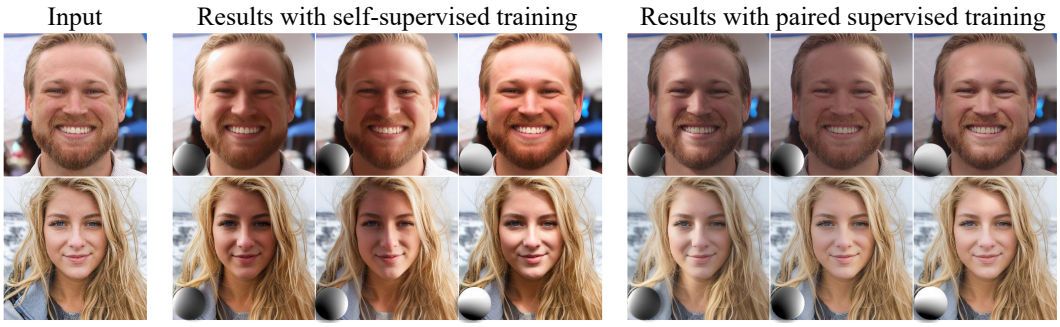

Figure 10: Effect of Self-Supervised Illumination Disentanglement. We compare two variants of our framework trained with self-supervision on natural images and paired supervision on synthetic relighting data, respectively. When applied to unseen real portraits, the paired-supervision model produces less photorealistic relighting results and responds weakly to the target illumination, particularly for lighting conditions that are rare in the training data, such as illumination from top. In contrast, self-supervised training produces more realistic and pronounced illumination changes while preserving portrait appearance, demonstrating better illumination disentanglement and generalization to real-world inputs.

et al., 2019) and pair them with the corresponding original images. Since generative relighting may not strictly preserve physical illumination and geometric consistency, we further employ the Unreal Engine 5 (ue5, 2025) rendering engine to generate 10,000 paired samples with diverse lighting conditions across 50 virtual scenes. To mitigate the domain gap between synthetic and real images, we first pretrain the model on natural images and subsequently fine-tune it using the paired data. As shown in Fig. 10, the paired-supervised model largely preserves facial identity and structure, but tends to produce weak or perceptually implausible illumination changes. Its generalization is particularly limited for less frequently observed lighting configurations, such as overhead illumination. In contrast, our self-supervised method produces more pronounced and realistic lighting effects and generalizes better across different lighting directions. These results demonstrate that self-supervised disentanglement reduces reliance on the distribution of synthetic training pairs and enables more generalizable relighting.

C More Analysis

Additional Visual Results. Figs. 12 and 13 provide further comparisons with state-of-the-art portrait relighting methods. Our method achieves photorealistic relighting while preserving portrait appearance and background content, including under strong external cast shadows, extreme exposure, and domain shifts. DiFaReli produces natural-looking illumination but may introduce artifacts in challenging cases. LBM and GSHeadRelit often alter the background and respond weakly to target lighting, leaving residual source shadows and highlights. IC-Light can generate backgrounds reflecting the target lighting, but may produce mismatched illumination on the subject and alter hair color, accessories, or clothing. Fig. 14 showcases our results on more diverse portraits, while Fig. 15 demonstrates lighting transfer consistent with the reference illumination.

D Limitation and Future Work.

Our method may not fully remove source shadows or highlights located far from the face, as shown in Fig. 11. In these regions, residual source illumination can persist even when the facial lighting follows the target condition, resulting in inconsistent illumination across the image. This limitation may stem from the use of shading hints rendered from coarse facial geometry, which provide limited spatial guidance for regions beyond the face. Future work will improve its robustness under such challenging conditions and extend the framework to temporally consistent portrait video relighting.



Figure 12: Qualitative comparison on arbitrary portraits.

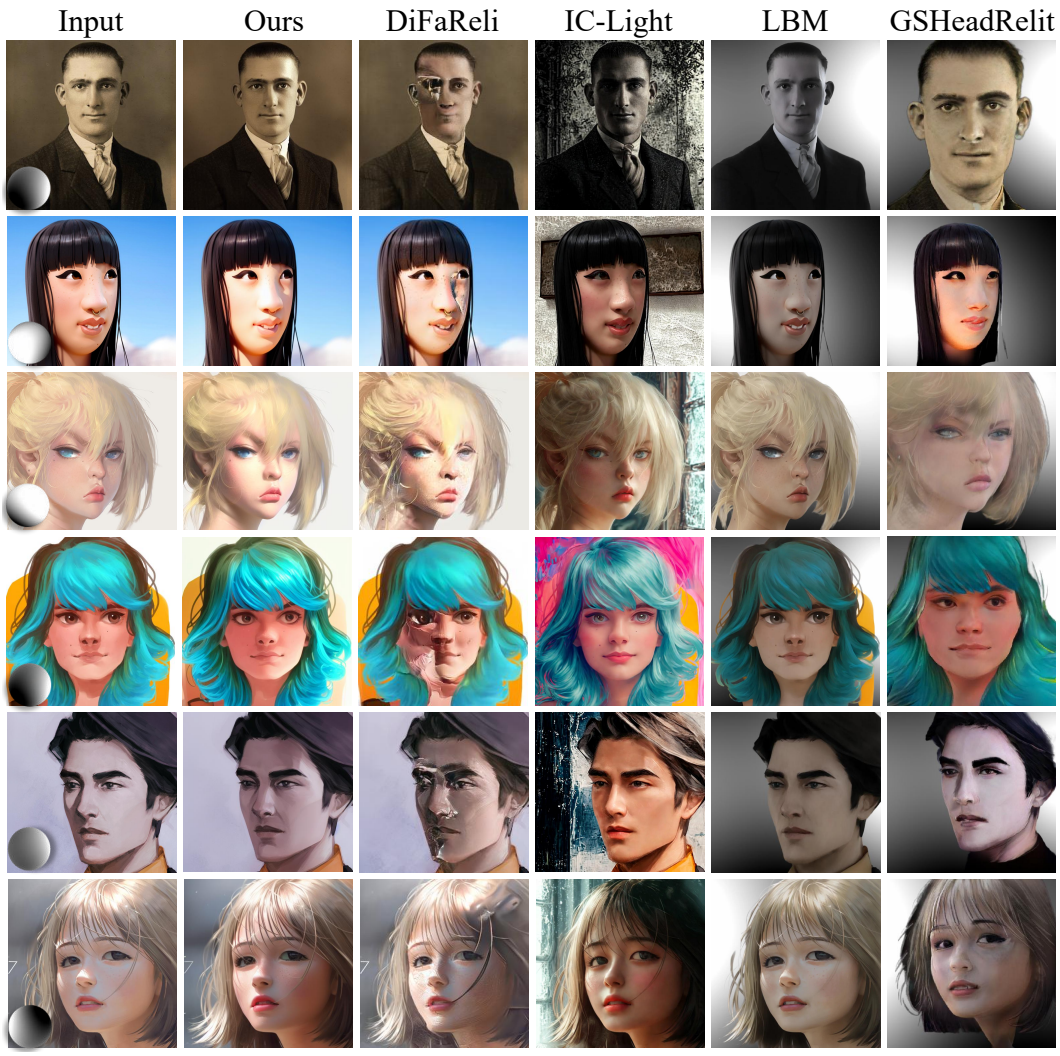

Figure 13: Qualitative comparison on out-of-domain images.



Figure 14: Results on more unseen portraits. In each image pair, the first image shows the input image and the lighting condition.

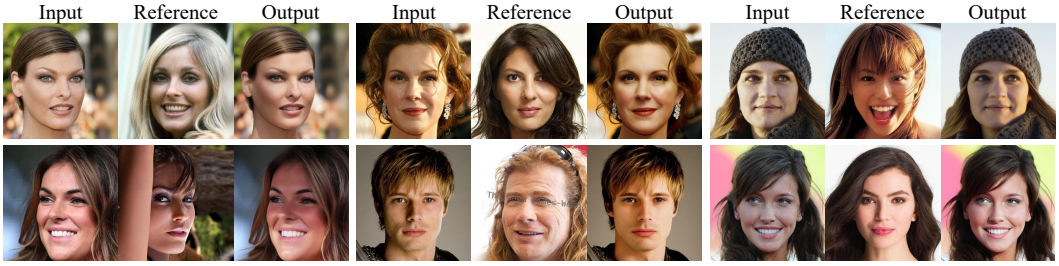

Figure 15: Lighting transfer results.

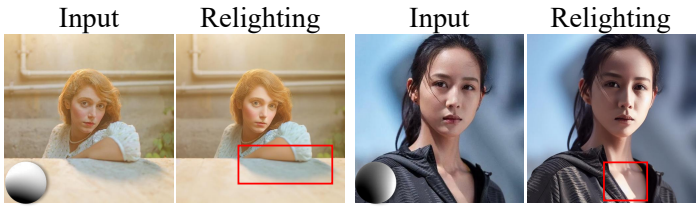

Figure 11: Failure Case. Our method sometimes fails to completely remove shadows or highlights when they are far from the face in the original illumination.