EdgeRelight360: Text-Conditioned 360-Degree HDR Image Generation for Real-Time On-Device Video Portrait Relighting

Min-Hui Lin* Mahesh Reddy* Guillaume Berger Michel Sarkis Fatih Porikli Ning Bi
Qualcomm AI Research†

Abstract

In this paper, we present EdgeRelight360, an approach for real-time video portrait relighting on mobile devices, utilizing text-conditioned generation of 360-degree high dynamic range image (HDRI) maps. Our method proposes a diffusion-based text-to-360-degree image generation in the HDR domain, taking advantage of the HDR10 standard. This technique facilitates the generation of high-quality, realistic lighting conditions from textual descriptions, offering flexibility and control in portrait video relighting task. Unlike the previous relighting frameworks, our proposed system performs video relighting directly on-device, enabling real-time inference with real 360-degree HDRI maps. This on-device processing ensures both privacy and guarantees low runtime, providing an immediate response to changes in lighting conditions or user inputs. Our approach paves the way for new possibilities in real-time video applications, including video conferencing, gaming, and augmented reality, by allowing dynamic, text-based control of lighting conditions.

1. Introduction

On-device video conferencing has emerged as a vital tool in our daily communications. Customizing the background in these platforms enhances user privacy, yet the default backgrounds are usually confined to pre-existing 2D images, and the mismatch in lighting between the subject and the virtual background often compromises the quality of the immersive experience.

In recent years, generative content creation, image/video relighting, and edge computing have witnessed a surge in quality and interest. In this context, we propose an on-device inference pipeline involving the generation of a 360-degree high dynamic range image (HDRI) map from a textual description provided by the user, followed by portrait relighting to seamlessly integrate a streaming user into the newly generated scene.

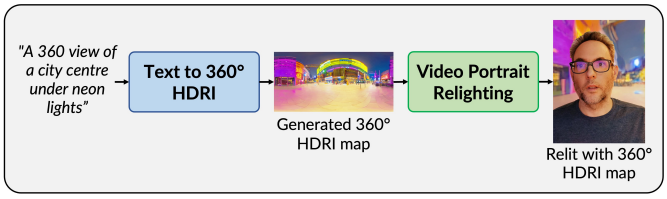

Figure 1. Our proposed method for generating 360-degree environment map from text prompt followed by video portrait relighting in real-time on mobile devices.

To enable efficient and high-quality on-device deployments, there are several challenges in 360-degree HDRI map generation and video relighting, respectively. HDRI maps play a pivotal role in creating backgrounds and lighting for immersive on-device applications. Although an existing approach [10] can generate high-resolution HDRI maps, this method involves a complex two-stage setup which first generates a low dynamic range (LDR) 360-degree image from text, before transforming it into an HDRI map in the second stage. Unfortunately, the complexity of the neural network designs used in this approach hinder its implementation on compute-constrained devices, and the adversarial training framework often results in a lack of diversity in the generated HDRI environment maps [10].

Relighting helps to naturally embed captured subjects into new environments by synthesizing physically consistent lighting effects. To cope with dynamic input lighting and to better synthesize high-frequency lighting effects, recent relighting methods [20, 27–29, 35, 43] first perform learning-based intrinsic decomposition to obtain surface normal and albedo, and then incorporate per-pixel lighting representations as explicit priors into the network design. Despite producing promising relighting results, their complex structures, involving multiple networks or decoders, hinder deployment on compute-constrained devices. Additionally, current state-of-the-art relighting approaches [28] exhibit issues with temporal stability, particularly in clothing regions.

*denotes equal contribution.
†Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

In our work, we address the above challenges and propose a portrait video relighting framework at the edge using text-conditioned generations of 360-degree HDRI maps. As illustrated in Fig. 1, we generate a 360-degree HDRI map with real-world brightness information from a text prompt. To achieve this, we leverage the generative capabilities of a Stable Diffusion [31] model to produce 360-degree HDRI maps by training it on 8-bit quantized HDRI maps following the HDR10 standard [32, 37].

Following this, we use the generated 360-degree image as an omnidirectional illumination source in the proposed portrait video relighting pipeline. A key novelty of our relighting model is the balance between relighting performance and computational efficiency. To enhance the computational efficiency, we only leverage a single network to infer surface normals for computing the diffuse light map and propose a shading equation for relighting. And to ensure better relighting performance, we propose to add diffuse shading to the input camera images for realism, and apply a temporal filter to enhance temporal consistency.

Finally, we compute a diffuse light map from the generated HDRI map for the current view specified by the user, apply lighting effects to the subject and composite the relit portrait with the new background created from the panorama.

In summary, our main contributions include:

- We propose an end-to-end on-device framework to generate 360-degree HDRI maps from text descriptions, and use it to relight video portraits in real-time. By combining HDRI map generation and lightweight video relighting, users can virtually appear anywhere imaginable in video applications at the edge.
- We present a diffusion-based text-conditioned 360-degree HDRI map generation to produce diverse environment maps on-device for relighting video portraits.
- We propose a light-weight video relighting framework combining a normal estimation network and light adding based rendering. Our on-device implementation shows realistic, fast, and stable relighting results for in-the-wild portrait videos, demonstrating the effectiveness, efficiency, temporal consistency, and generalization of the proposed method.

2. Related work

Text-to-image: Most recent text-conditioned image generative models are based on diffusion [13, 21, 34], a technique which generates new samples via progressive denoising from pure random noise. In order to reduce the computational footprint of such approaches, latent diffusion models, like Stable Diffusion (SD) [31], perform diffusion in the latent space of a variational autoencoder (VAE) [22]. Text-to-image generation methods have unlocked other use cases such as text-to-panorama with applications in virtual reality environments to enable immersive experiences. In particular, several works [7, 9, 40] have extended SD models to generate panorama from text. However, these approaches only consider horizontal (left-to-right) panoramic rotations. Lin et al. [24] use an adversarial setup for unconditioned seamless panoramic generation in a patch-wise manner. More recently, LDM3D-VR [36] fine-tunes a pretrained SD v1.5 model to generate panoramic RGB images with monocular depth. However, the generated image contains a visible border seam due to mismatch in the generated content on both ends of the image. Most text-to-panoramic 360-degree approaches are limited in their applications due to the generation of LDR images. In contrast, HDRI panoramas can assist in synthesizing 360-degree photorealistic lighting and reflections for scene or portrait relighting. Text2Light [10] proposes a text-conditioned panoramic 360-degree HDRI map generation by using a complex dual-stage architecture. Although the generated images are in the HDRI space, the Text2Light framework lacks the image diversity compared to text-to-image models based on diffusion.

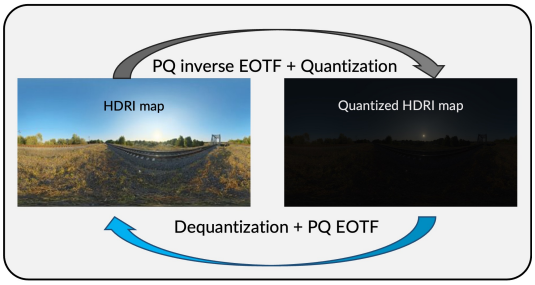

Figure 2. We propose to combine PQ inverse EOTF used in the HDR10 standard with 8-bit quantization to obtain the quantized HDRI maps to generate the training dataset. Similarly, dequantization and the PQ EOTF can be performed to recover the original HDRI map.

Portrait relighting: The pioneering work by Debevec et al. [12] designs a spherical rig called Light Stage to capture a person's reflectance fields as one-light-at-a-time (OLAT) images and uses image-based relighting to relight static faces. Other approaches [16, 25, 44] utilize time-multiplexed illumination or color gradient illumination to relight dynamic subjects, but these methods require expensive custom capture rigs that are known for being hard to set up.

With the advancement in mobile photography, several deep learning approaches to relight portrait images captured in unconstrained environments have emerged. Zhou et al. [47] and Sun et al. [38] are early works to apply deep learning to portrait relighting by employing an encoder-decoder network to take a single image as input, inject

Figure 3. Encoding and decoding (a) quantized HDRI map with a pre-trained VAE leads to significant artifacts such as (b) blue patches and RGB color distortions. However, these issues can be resolved by (c) fine-tuning the VAE on quantized HDRI maps, which reproduces the quantized images close to the (d) dequantized original images.

Figure 4. To augment the perspective HDRI dataset, we generate 20 perspective HDR images for every 360-degree equirectangular HDRI map. The images are tone mapped for visualization purpose.

the target illumination into the bottleneck layer of the network, and output the re-illuminated image. More recently, pixel-aligned components such as normal, albedo, diffuse light map, specular light map, visibility map, and shadow map have been incorporated into network designs to improve performance [20, 27–29, 35, 43]. However, these approaches often involve compute-heavy pipelines with multiple large networks, hindering the portability of these methods to mobile devices. Additionally, these approaches lack temporal consistency for in-the-wild videos.

In video relighting, it is critical to have a good balance between high-quality, video stability, and model complexity. Zhang et al. [46] introduces the flow-based temporal loss supervised on dynamic OLAT dataset for explicit temporal modeling. Despite showing real-time relighting on mobile devices, the light-weight encoder-decoder network cannot produce high-quality relit videos with sufficient facial details. Yeh et al. [45] propose to learn two temporal residual networks for improving consistency of intermediate normal and albedo predictions, but the network size is not designed for consumer device deployment.

Even several commercial relighting solutions do not support on-device inference and video relighting with HDRI environments. Portrait Mode on iPhone [8] only provides 5 studio lights mode, mostly for photo editing. Portrait Light on Google's Pixel [41] and Google Meet [15] only assist in relighting a subject with point light sources, rather than using 360-degree environment light for illumination. Clipdrop [11] offers controllability of point light sources, but cannot relight videos. The recent SwitchLight [39] supports video relighting, but it is a frame-based solution that run on a remote server, hence not a real-time solution. In contrast, our proposed video relighting method combines generalized normal estimation network and light adding based rendering, leveraging mobile computing to achieve convincing and coherent relighting results in real-time.

3. Text-conditioned 360-degree HDRI map generation

Our goal is to enable on-device real-time video portrait relighting by leveraging high dynamic range image (HDRI) maps from a text-to-image generative model for relighting with diverse background environments. To that end, we leverage the generative ability of Stable Diffusion [31] (SD) which we extend to 360-degree HDRI map generation.

3.1. Quantized HDR images

To synthesize 360-degree images from user text prompts, we propose fine-tuning a pre-trained SD v1.5 model on HDRI panoramas. However, since our end goal is to achieve fast inference on the AI accelerator of a device with a Snapdragon® Gen 3 Platform*, we aim to leverage 8-bit model quantization [26] at test time to reduce the computational and memory footprint of our text-to-image models, as previously done in [48]. Yet, it poses challenges for HDR

*Snapdragon branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries.

prediction due to the disparity between the high-dynamic range input and output signals (with luminance range approximately 200,000 using FLOAT32), and the INT8 quantization applied to both weights and internal activations during inference. To mitigate this potential performance drop post-model quantization, we propose to preemptively quantize the raw FLOAT32 HDRI maps using the perceptual quantizer (PQ) based HDRI quantization workflow [32, 37] to obtain 8-bit images while preserving the high-dynamic luminance spectrum.

More specifically, we combine the perceptual quantizer (PQ) used in the HDR10 standard [37] with an 8-bit UINT quantization formula. As shown in Eq. 1, first the PQ inverse electro-optical transfer function (EOTF) is used to convert from linear luminance to non-linear color values. Then, the quantization formula [32] is applied to produce quantized HDR images in UINT8 domain:

$$\begin{aligned} E' &= PQ^{-1}_{EOTF}(F_D) \\ PQ^{-1}_{EOTF}(F_D) &= \left(\frac{c_1 + c_2 \cdot Y^{m_1}}{1 + c_3 \cdot Y^{m_1}} \right)^{m_2} \\ D' &= int(scale \cdot E') \end{aligned} \tag{1}$$

where F_D is the linear luminance in cd/m^2, E' is the non-linear color value, $Y = F_D/10000$, $m_1 = 0.1593017578125$, $m_2 = 78.84375$, $c_1 = 0.8359375$, $c_2 = 18.8515625$, and $c_3 = 18.6875$. For the quantization formula, we set $scale = 198$ to encode luminance in original HDRI maps at a maximum of $200,000\ cd/m^2$. The scale is set based on statistical analysis: among 624 real HDRI panoramas in PolyHaven[6], only 33 panoramas have an average 4 pixels with values greater than $200,000$.

Post fine-tuning the SD v1.5 [31] on UINT8 quantized HDR images (D'), at inference, the generated quantized HDRI map is transformed to the original HDRI space by applying the dequantization followed by PQ EOTF as indicated in Eq. 2. An overview of the images generated from the forward and reverse HDRI quantization are shown in Fig. 2.

$$\begin{aligned} E' &= D'/scale \\ F_D &= PQ_{EOTF}(E') \\ PQ_{EOTF}(E') &= 10000 \left(\frac{max[(E'^{1/m_2} - c_1), 0]}{c_2 - c_3 \cdot E'^{1/m_2}} \right)^{1/m_1} \end{aligned} \tag{2}$$

3.2. Training setup

Training data: We combine PolyHaven [6] and Laval [14, 19], two publicly accessible yet relatively small-scale

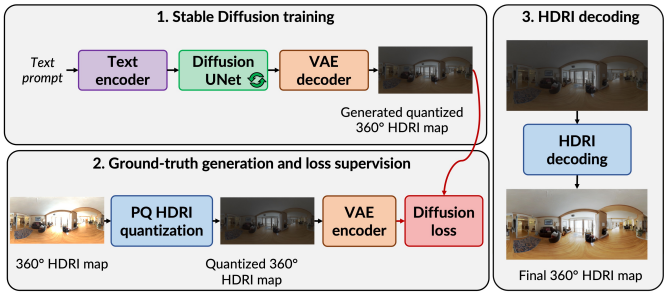

Figure 5. An overview of the (1) text to 360-degree training setup along with (2) the quantized HDRI generation. The generated quantized image can be (3) dequantized with inverse PQ transformation to produce the final HDRI map.

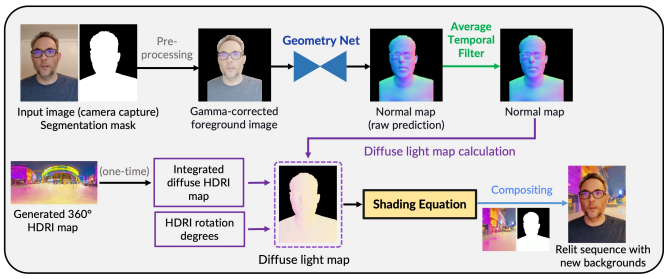

Figure 6. Overview of the proposed relighting pipeline. Given an input image and segmentation mask from the camera, we generate a gamma-corrected foreground image to obtain the surface normals from Geometry Net. The raw normal map prediction is temporally stabilized with an average temporal filter. To perform relighting, we first use the generated 360-degree HDRI map and the filtered normal map to compute the diffuse light map, and then use the shading equation to relight the portrait.

datasets offering 360-degree HDR images. PolyHaven contains around 660 high-resolution equirectangular images depicting a variety of indoor and outdoor scenes. Laval, on the other hand, provides a collection of 2100 indoor and 206 outdoor 360-degree images. To augment our training data, we also utilize Hypersim [30], which consists of standard HDR images, particularly during VAE fine-tuning. We apply the transfer function specified in Eq. 1 to all three datasets to generate quantized HDRI maps.

VAE: Utilizing the original VAE from the SD v1.5 [31] pipeline to encode quantized HDR images leads to notable color distortions and artifacts in the decoded images, as shown in Fig. 3. To address this, we fine-tune the VAE on quantized HDRI data. Given that the VAE does not require to specialize on 360-degree images, we can substantially expand our dataset for this phase. Rather than directly using the full 360-degree equirectangular images from our combined PolyHaven-Lavel dataset, we construct a data augmentation pipeline which entails rendering perspective views from these 360-degree images, as shown in

Fig 4. Moreover, to further expand our training dataset, we incorporate HDR images from the Hypersim [30] dataset. Post-training, as depicted in Fig. 3, our fine-tuned VAE successfully reconstructs HDR images devoid of color distortions or other artifacts initially observed with the original model.

Diffusion UNet: To fine-tune the original SDv1.5 diffusion U-Net for text-conditioned 360-degree HDR generation, we adopt a two-stage training strategy to address the limited availability of 360-degree HDRI data. In the first stage, the objective is to adapt the diffusion U-Net to the fine-tuned latents and learn a prior on the luminance range (HDR domain adaptation). For this, we do not use 360-degree images and instead reuse the augmented, perspective-based, quantized HDRI dataset constructed during VAE fine-tuning. Text prompts are generated using the BLIP [23] image captioning model, applied to tone mapped low dynamic range (LDR) versions of the images.

In the second stage, we further fine-tune the latent diffusion U-Net on 360-degree images this time. We use the quantized images from PolyHaven and Laval along with their BLIP-based captions. Remarkably, even with a relatively limited dataset comprising only a few thousand 360-degree images, the model effectively learns the capability to generate consistently accurate 360-degree imagery. During inference, we revert the generated quantized images back to the original HDRI format using the inverse PQ function outlined in Eq. 2, thereby readying them for downstream tasks such as portrait relighting. To ensure 360-degree consistency at the edges of the generated equirectangular image, we adopt the circular latent padding technique proposed in [42] throughout all denoising timesteps. An overview of the training, quantized data preparation and inference is illustrated in Fig. 5.

4. Video portrait relighting

We propose a relighting pipeline to generate relit portrait videos with temporal stability, high-quality and with light-weight network architecture. Our proposed setup involves a Geometry Net for surface normal map prediction, a average temporal filter for enhancing temporal consistency, diffuse light map calculation, a shading equation for rendering relit foreground, and background composition as shown in Fig. 6.

4.1. Geometry Net

To estimate the surface normals, inspired by the Total Relighting [28] framework, we develop a light-weight Geometry Net with a similar but smaller UNet architecture with 13 layers. In addition, the size of network input is set to 512×512 to reduce computational complexity and to handle both portrait and landscape camera captures. The Geometry Net aims to produce per-pixel surface normal map N_t as a

Method	**Runtime** (seconds)	**Memory** (GB)
Text2Light [10] (2048×4096)	61.25	3.1
Ours (512×512)	5.2	$\sim$3.4
Ours (512×1024)	14.8	$\sim$3.4

Table 1. We measure the runtime and memory requirements of Text2Light [10] and our proposed 360-degree generative approach on a single NVIDIA A100 GPU as Text2Light does not support on-device inference.

Figure 7. Our Stable Diffusion [31] based text-conditioned 360-degree HDRI generative model can generate realistic and diverse environments compared to existing Text2Light [10] model. The images are tone mapped for visualization purpose.

geometrical cue to calculate physically correct diffuse light map. Given an input camera capture I, we leverage our proprietary segmentation network with on-device support to extract the foreground segmentation, but we can leverage any off-the-shelf segmentation network for this task. To prepare the input to the Geometry Net, the camera capture I is first gamma corrected and masked using the foreground segmentation, then resized and padded to 512×512. As for training, due to the challenges of setting up a light stage, inspired by [43], we use head meshes captured by the 3DMD system [2], HDRI panoramas in PolyHaven[6], and 3D software Blender [3] to create the synthetic normal dataset. The synthetic normal dataset contains paired rendered images and normal ground truths in camera space coordinates for 60 identities, each with 31 expressions. Please refer to the supplementary material for more details on the network architecture and the creation of synthetic normal dataset.

4.2. Average Temporal Filter

When applying the Geometry Net to video sequences, minor flickering occurs between consecutive normal estimations because the network is frame-based. In addition, since the head meshes in our synthetic normal dataset lack hair and cloth geometry, the instability issue around hair and cloth areas are severer. To further improve the temporal

Method	Image relighting	Video relighting	Run on mobile device	Video consistency	HQ relighting	Target light	Controllability
Apple iPhone [8]	✔	✗	✗	✗	✔	Studio light	Limited lighting options & no env. rotation
Google Pixel [41]	✔	✗	✗	✗	✔	Single point light	Light direction & intensity
Google Meet [15]	✔	✔	✗	✔	✔	Multiple point lights	Light direction, color & intensity
Clipdrop [11]	✔	✗	✗	✗	✔	Multiple point lights	Light direction, intensity, color, distance & radius
SwitchLight [39]	✔	✔	✗	✗	✔	360° HDR image	HDR rotation degrees
Zhang et al. [46]	✔	✔	✔	✔	✗	360° HDR image	HDR rotation degrees
Yeh et al. [45]	✔	✔	✗	✔	✔	360° HDR image	HDR rotation degrees
Ours	✔	✔	✔	✔	✔	360° HDR image	HDR rotation degrees

Table 2. Unlike other existing approaches, our proposed relighting method supports all key features such as image and video relighting, on-device inference, video consistency, high-quality relighting, relight with 360-degree HDR environment maps, and fine-grained control over the HDR rotation.

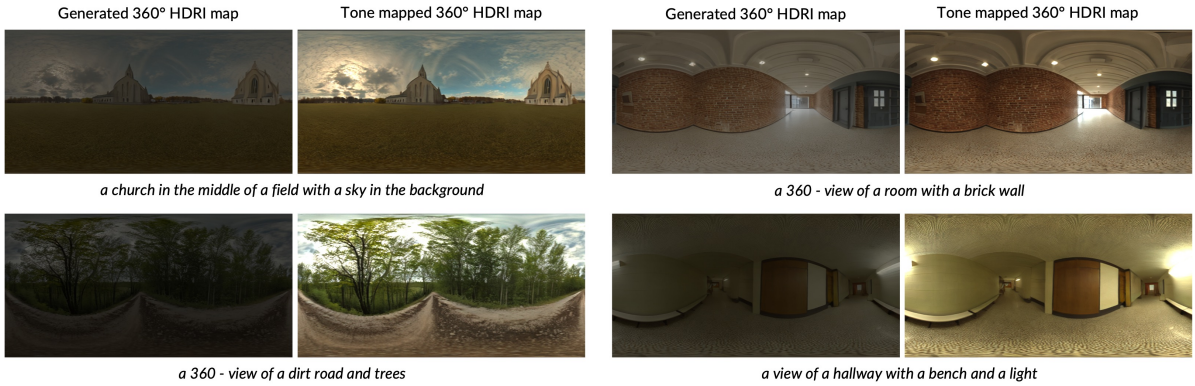

Figure 8. Our text to 360-degree HDRI generative model can produce environment maps for diverse text prompts. We show both the generated quantized HDR image and it's dequantized with inverse PQ, and tone mapped image.

Method	Runtime (seconds)	Model Size (MB)
Text-to-360-degree image	~ 5	~ 1100
Video Relighting	~ 0.04	~ 20.5

Table 3. Overview of the on-device runtime and model size for both the 360-degree image generation and video relighting pipelines on a Snapdragon Gen 3 platform. Note that the model size for video relighting is the sum of the size of the video segmentation network ($\sim$ 17.3 MB) and the Geometry Net ($\sim$ 3.2 MB).

consistency while maintaining low computation complexity at edge, we do not add an additional temporal refinement network to the architecture as [45], but adopt an effective solution to apply the average temporal filter on three consecutive normal maps N_{t-2}, N_{t-1}, and N_t predicted by the Geometry Net. The operation is formulated in Eq. 3:

$$\widetilde{N_t} = \frac{1}{3}(N_{t-2} + N_{t-1} + N_t) \tag{3}$$

where $\widetilde{N_t}$ is the average normal map used to calculate the diffuse light map, which helps in mitigating flickering issues in final relit sequences.

Figure 9. To demonstrate the strong generative capabilities of our proposed text to 360-degree image generation, we compare the results from an LDR variant of the model with LDM3D-VR [36]. For more qualitative comparison, please refer to the supplementary material.

4.3. Light Adding based Rendering

Unlike state-of-the-art relighting methods [28, 45] which require at least four sequential networks, to save calculations, we explore combining a neural network with physically based rendering (PBR). Additionally, when the subjects are illuminated with physically reasonable diffuse light, we observe that they appear to be in the virtual environment. Therefore, we propose a light adding based rendering to produce a relit foreground by adding diffuse shad-

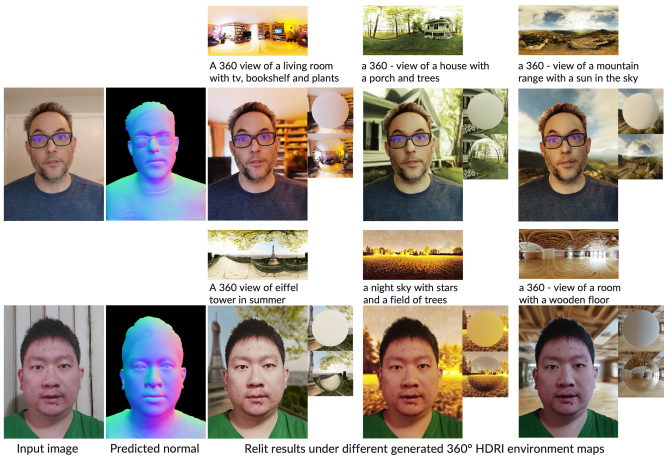

Figure 10. Our proposed framework shows realistic relighting of different portraits on diverse generative HDRI environment maps. Our results can preserve facial details such as wrinkles and beards and generate lighting effects that are consistent with diffuse sphere references.

ing to the camera capture. Specifically, given a generated 360-degree HDRI map, we pre-compute the diffuse cubemap by summing all diffuse light from the surrounding environment along each sample direction. For each frame, a diffuse light map can be calculated by sampling the pre-integrated diffuse cubemap using per-pixel normal vectors, which is a popular technique in real-time PBR implementation [5]. To further render relit foregrounds, the shading equation Eq. 4 is formulated as a scaled addition of raw camera capture and diffuse shading, expressed as the multiplication of the low saturation camera capture I_{lowS} and the diffuse light map D:

$$R = s_1 \cdot I + s_2 \cdot I_{lowS} \odot D \tag{4}$$

where $\odot$ denotes the element-wise multiplication, and s_1,s_2 are scaling constants. Note that generally the diffuse shading is formulated as the multiplication of albedo and diffuse light map [5, 27, 28]. Considering that adding an sequential network after Geometry Net increases the operation time by $\sim$ 24 milliseconds for a 1024×768 albedo estimation, it introduces a computational bottleneck for on-device real-time video relighting. Therefore, we compute a low saturation camera capture I_{lowS} using Eq. 5 instead:

$$I_{lowS} = 0.6 \cdot I + 0.4 \cdot I_{gray} + 0.05 \tag{5}$$

where I_{gray} is the grayscale camera capture. Specifically, I_{lowS} is designed to be computationally efficient without requiring color space conversion from RGB to HSV, and the low saturation is designed to reduce the impact of the input lighting's hue.

Method	CLIP ↑	FID ↓
LDM3D-VR [36]	28.73	42.04
Ours-LDR	**29.94**	**39.79**

Table 4. We measure the CLIP and FID metrics on the LDR images generated by LDM3D-VR and our LDR 360-degree model to demonstrate the better generative capabilities.

5. On-device inference

To enable inference on Snapdragon Gen 3 platform, we leverage network quantization and real-time rendering. For on-device network inference, we use the AIMET [33] library to conduct post-training quantization from FP32 to INT8 for both the 360-degree image generation and video relighting models. Additionally, we implement the light adding based rendering module in OpenGL Shading Language [4] to save expensive floating-point operations on the CPU.

6. Results

Our proposed framework achieves real-time video portrait relighting based on text-prompted 360 degree image generations. We now present quantitative and qualitative analyses to showcase the capabilities of our proposed pipeline.

6.1. Text to 360-degree HDRI map generation

We compare images generated by our text-conditioned 360-degree HDRI generation against the Text2Light [10] approach in Fig. 7 and report the corresponding runtime on a single A100 GPU for both approaches in Tab. 1. Although the Text2Light model is capable of generating high-resolution 360-degree HDRI maps from text, it exhibits a lack of diversity and realism while requiring a significantly higher runtime. Our approach, on the other hand, can generate high-quality 360-degree images in $\sim$ 5 seconds on an A100 . More text-conditioned HDRI generations for both indoor and outdoor scenes are shown in Fig. 8. Note that, for the sake of visualization, we show the corresponding tone mapped images.

We also conduct a study to compare with the recent LDM3D-VR [36] method, which proposes a similar approach to ours but is limited to low-dynamic range prediction. For this study, we adapt our training setup and fine-tune the SDv 1.5 U-Net on 360-degree LDR images from PolyHaven [6] and Matterport360 [1] with an image resolution of 512×1024 to match LDM3D-VR. We report FID [18] and CLIP [17] scores for both models in Tab. 4. On both metrics, our fine-tuned SD v1.5 model shows better performance. This can be attributed to (i) our model's capacity to generate consistent 360-degree image thanks to circular padding [42], and (ii) the fact that we only use real-world images for fine-tuning unlike LDM3D-VR. Further-

more, Fig. 9 provides a qualitative comparison to demonstrate the superior quality of our generated images. For an extended qualitative analysis of the generated LDR images, we refer readers to the supplementary materials.

6.2. Video portrait relighting

To evaluate the proposed video portrait relighting method, first, we relight in-the-wild portrait sequences under different HDRI environments generated by our 360-degree image generative model. Due to the lack of relight ground truth for in-the-wild sequence, we render diffuse spheres to provide a lighting reference in the target HDRI environment. And mirror spheres are also provided to indicate where the main light source locates. As shown in Fig. 10, we are able to produce physically-correct relighting results while preserve facial details and can embed subjects naturally into a variety of generated environments. Additionally, to show temporal-consistency, we provide the recorded screenshots of our on-device EdgeRelight360 application in the supplementary material, demonstrating that our approach is stable and flicker-free.

Second, we compare the runtime and relighting quality of publicly available SwitchLight [39], which also use HDRI maps to relight images. The web version of SwitchLight takes 10 more seconds to run a single image on their remote server, while our method runs locally on the phone and takes only 0.04 seconds per image. Their high computation cost origins from a complete intrinsic decomposition pipeline (including normal, albedo, specular, and roughness estimation), a neural renderer, and fine-grained foreground matting. Fig. 11 shows that we can produce promising results and make subjects blend into the PolyHaven [6] HDRI map naturally, while the default lighting effects of SwitchLight is too strong and less consistent with the lighting reference. Also, consecutive frames from our talking test sequence is put in the supplementary material to demonstrate that results from the web version of SwitchLight produce relit images with flickering, while ours are more stable and flicker-free.

6.3. On-device inference

The primary goal of our proposed approach is to enable end-to-end on-device inference. To the best of our knowledge, we are the first to run both 360-degree HDRI map generation and real-time video portrait relighting. We compare some prior approaches that address image/video relighting with and without on-device support in Tab. 2. In comparison to all the approaches, our proposed framework supports: (a) image and video relighting, (b) runs on device, (c) ensure smooth temporal consistency, (d) enables high-quality relighting, (e) leverages 360-degree HDRI maps, and (f) offers fine-grained rotation control of the HDRI maps. In contrast, all the prior approaches only handle a subset of the features.

Figure 11. Compared with SwitchLight [39], our relighting pipeline is more lightweight and can generate more natural, physically correct, and temporally consistent results.

In Tab. 3 we show the runtime and model size of our proposed framework deployed on a mobile device with Snapdragon Gen 3 processor. The time to generate a single 360-degree HDRI map is $\sim$ 5 seconds. For the end-to-end process which performs face detection, video segmentation, and video relighting concurrently, it runs at around 25 fps using the neural signal processor (NSP) and GPU. Note that we generate the 360-degree images at 480×480 resolution with 20 denoising steps and run the Geometry Net at 512×512, while running rendering and produce relighting results at 1024×768.

7. Discussion

EdgeRelight360 supports real-time video applications by introducing an innovative, on-device video portrait relighting technique. By harnessing the power of text-conditioned generated 360-degree HDRI maps, it offers high-quality, realistic lighting conditions derived from textual descriptions. This not only ensures privacy and low runtime but also provides an immediate response to changes in lighting conditions or user input. A potential improvement to the existing 360-degree generation is to generate higher resolution images to support different edge screen resolutions with high-quality background images.

References

[1] Matterport3d 360o rgbd dataset, 2022. 7

[2] https://3dmd.com/products/, 2024. [Online; accessed March-6-2024]. 5

[3] https://www.blender.org/, 2024. [Online; accessed March-6-2024]. 5

[4] https://www.khronos.org/opengl/wiki/OpenGL_Shading_Language, 2024. [Online; accessed March-6-2024]. 7

[5] https://learnopengl.com/PBR/IBL/Diffuse-irradiance, 2024. [Online; accessed March-6-2024]. 7

[6] https://polyhaven.com/, 2024. [Online; accessed 22-February-2024]. 4, 5, 7, 8

[7] https://skybox.blockadelabs.com/, 2024. [Online; accessed 22-February-2024]. 2

[8] Apple. Use portrait mode on your iphone. https://support.apple.com/en-us/102398. 3, 6

[9] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023. 2

[10] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Text2light: Zero-shot text-driven hdr panorama generation. ACM Transactions on Graphics (TOG), 41(6):1–16, 2022. 1, 2, 5, 7

[11] ClipDrop. https://clipdrop.co/relight. [Online; accessed 22-February-2024]. 3, 6

[12] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In Proceedings of the 27th annual conference on Computer graphics and interactive techniques, pages 145–156, 2000. 2

[13] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. Advances in neural information processing systems, 34:8780–8794, 2021. 2

[14] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. arXiv preprint arXiv:1704.00090, 2017. 4

[15] Google. Enhance your video audio with gemini in google meet. https://support.google.com/meet/answer/14441737. [Online; accessed March-2-2024]. 3, 6

[16] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. ACM Transactions on Graphics (ToG), 38(6):1–19, 2019. 2

[17] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718, 2021. 7

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 30, 2017. 7

[19] Yannick Hold-Geoffroy, Akshaya Athawale, and Jean-François Lalonde. Deep sky modeling for single image outdoor lighting estimation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6927–6935, 2019. 4

[20] Andrew Hou, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. Face relighting with geometrically consistent shadows. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4217–4226, 2022. 1, 3

[21] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. Advances in neural information processing systems, 34:21696–21707, 2021. 2

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013. 2

[23] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International Conference on Machine Learning, pages 12888–12900. PMLR, 2022. 5

[24] Chieh Hubert Lin, Hsin-Ying Lee, Yen-Chi Cheng, Sergey Tulyakov, and Ming-Hsuan Yang. Infinitygan: Towards infinite-pixel image synthesis. arXiv preprint arXiv:2104.03963, 2021. 2

[25] Abhimitra Meka, Christian Haene, Rohit Pandey, Michael Zollhöfer, Sean Fanello, Graham Fyffe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, et al. Deep reflectance fields: high-quality facial reflectance field inference from color gradient illumination. ACM Transactions on Graphics (TOG), 38(4):1–12, 2019. 2

[26] Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart Van Baalen, and Tijmen Blankevoort. A white paper on neural network quantization. arXiv preprint arXiv:2106.08295, 2021. 3

[27] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas Lehrmann. Learning physics-guided face relighting under directional light. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5124–5133, 2020. 1, 3, 7

[28] 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. ACM Transactions on Graphics (TOG), 40(4):1–21, 2021. 1, 5, 6, 7

[29] Haonan Qiu, Zhaoxi Chen, Yuming Jiang, Hang Zhou, Xiangyu Fan, Lei Yang, Wayne Wu, and Ziwei Liu. Relitalk: Relightable talking portrait generation from a single video. International Journal of Computer Vision, pages 1–16, 2024. 1, 3

[30] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In

Proceedings of the IEEE/CVF international conference on computer vision, pages 10912–10922, 2021. 4, 5

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. 2022 ieee. In CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 10674–10685, 2021. 2, 3, 4, 5

[32] BT Series. Parameter values for ultra-high definition television systems for production and international programme exchange. In Proc. ITU-T, Bt. 2020, pages 1–7, 2012. 2, 4

[33] Sangeetha Siddegowda, Marios Fournarakis, Markus Nagel, Tijmen Blankevoort, Chirag Patel, and Abhijit Khobare. Neural network quantization with ai model efficiency toolkit (aimet). arXiv preprint arXiv:2201.08442, 2022. 7

[34] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In International conference on machine learning, pages 2256–2265. PMLR, 2015. 2

[35] Guoxian Song, Tat-Jen Cham, Jianfei Cai, and Jianmin Zheng. Real-time shadow-aware portrait relighting in virtual backgrounds for realistic telepresence. In 2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), pages 729–738. IEEE, 2022. 1, 3

[36] Gabriela Ben Melech Stan, Diana Wofk, Estelle Aflalo, Shao-Yen Tseng, Zhipeng Cai, Michael Paulitsch, and Vasudev Lal. Ldm3d-vr: Latent diffusion model for 3d vr. arXiv preprint arXiv:2311.03226, 2023. 2, 6, 7

[37] SMPTE Standard. High dynamic range electro-optical transfer function of mastering reference displays. SMPTE ST, 2084(2014):11, 2014. 2, 4

[38] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. Single image portrait relighting. ACM Transactions on Graphics (TOG), 38(4):1–12, 2019. 2

[39] SwitchLight. https://www.switchlight.beeble.ai/. [Online; accessed 22-February-2024]. 3, 6, 8

[40] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. arXiv, 2023. 2

[41] Yun-Ta Tsai and Rohit Pandey. Portrait light: Enhancing portrait lighting with machine learning. https://ai.googleblog.com/2020/12/portrait-light-enhancing-portrait.html, 2020. 3, 6

[42] Jionghao Wang, Ziyu Chen, Jun Ling, Rong Xie, and Li Song. 360-degree panorama generation from few unregistered nfov images. arXiv preprint arXiv:2308.14686, 2023. 5, 7

[43] Zhibo Wang, Xin Yu, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. Single image portrait relighting via explicit multiple reflectance channel modeling. ACM Transactions on Graphics (TOG), 39(6):1–13, 2020. 1, 3, 5

[44] Andreas Wenger, Andrew Gardner, Chris Tchou, Jonas Unger, Tim Hawkins, and Paul Debevec. Performance relighting and reflectance transformation with time-multiplexed illumination. ACM Transactions on Graphics (TOG), 24(3):756–764, 2005. 2

[45] 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. 3, 6

[46] Longwen Zhang, Qixuan Zhang, Minye Wu, Jingyi Yu, and Lan Xu. Neural video portrait relighting in real-time via consistency modeling. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 802–812, 2021. 3, 6

[47] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In Proceedings of the IEEE/CVF international conference on computer vision, pages 7194–7202, 2019. 2

[48] Jilei Hou Ziad Asghar. World's first on-device demonstration of stable diffusion on an android phone, 2023. 3