DiLightNet: Fine-grained Lighting Control for Diffusion-based Image Generation

Chong Zeng*
State Key Lab of CAD and CG,
Zhejiang University
Hangzhou, China
chongzeng2000@gmail.com

Yue Dong
Microsoft Research Asia
Beijing, China
yuedong@microsoft.com

Pieter Peers
College of William & Mary
Williamsburg, USA
ppeers@siggraph.org

Youkang Kong*
Tsinghua University
Beijing, China
kykdqs@gmail.com

Hongzhi Wu
State Key Lab of CAD and CG,
Zhejiang University
Hangzhou, China
hwu@acm.org

Xin Tong
Microsoft Research Asia
Beijing, China
xtong@microsoft.com

"futuristic soldier with advanced armor weaponry and helmet" *"rusty steel toy frog with spatially varying materials with the body diffuse but shinny eyes"*

Figure 1: Examples of generated images specified via a text-prompt (listed below each example) and with fine-grained lighting control. Each prompt is plausibly visualized under two different user-provided lighting environments.

ABSTRACT

This paper presents a novel method for exerting fine-grained lighting control during text-driven diffusion-based image generation. While existing diffusion models already have the ability to generate images under any lighting condition, without additional guidance these models tend to correlate image content and lighting. Moreover, text prompts lack the necessary expressional power to describe detailed lighting setups. To provide the content creator with fine-grained control over the lighting during image generation, we augment the text-prompt with detailed lighting information in the form of radiance hints, i.e., visualizations of the scene geometry with a homogeneous canonical material under the target lighting. However, the scene geometry needed to produce the radiance hints is unknown. Our key observation is that we only need to guide the diffusion process, hence exact radiance hints are not necessary; we only need to point the diffusion model in the right direction. Based on this observation, we introduce a three stage method for controlling the lighting during image generation. In the first stage, we leverage a standard pretrained diffusion model to generate a provisional image under uncontrolled lighting. Next, in the second stage, we resynthesize and refine the foreground object in the generated image by passing the target lighting to a refined diffusion model, named DiLightNet, using radiance hints computed on a coarse shape of the foreground object inferred from the provisional image. To retain the texture details, we multiply the radiance hints with a neural encoding of the provisional synthesized image before passing it to DiLightNet. Finally, in the third stage, we resynthesize the background to be consistent with the lighting on the foreground object. We demonstrate and validate our lighting controlled diffusion model on a variety of text prompts and lighting conditions.

CCS CONCEPTS

• **Computing methodologies** → **Image-based rendering**.

*Work partially done during internship at Microsoft Research Asia.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGGRAPH Conference Papers '24, July 27-August 1, 2024, Denver, CO, USA
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0525-0/24/07…$15.00
https://doi.org/10.1145/3641519.3657396

KEYWORDS

Diffusion, Image Synthesis, Lighting Control, Radiance Hints

ACM Reference Format:
Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. 2024. DiLightNet: Fine-grained Lighting Control for Diffusion-based Image Generation. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24 (SIGGRAPH Conference Papers '24), July 27-August 1, 2024, Denver, CO, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3641519.3657396

1 INTRODUCTION

Text-driven generative machine learning methods, such as diffusion models [Nichol et al. 2022; Ramesh et al. 2022; Rombach et al. 2022; Saharia et al. 2022], can generate fantastically detailed images from a simple text prompt. However, diffusion models also have built in biases. For example, Liu *et al.* [2023] demonstrate that diffusion models tend to prefer certain viewpoints when generating images. As shown in Figure 2, another previously unreported bias is the lighting in the generated images. Moreover, the image content and lighting are highly correlated. While diffusion models have the capability to sample different lighting conditions, there currently does not exist a method to precisely control the lighting and the image content independently in the generated images.

In this paper we aim to exert fine-grained control on the effects of lighting during diffusion-based image generation (Figure 1). While text prompts have been used to provide relative control of non-rigid deformations of objects [Cao et al. 2023; Kawar et al. 2023], the identity and gender of subjects [Kim et al. 2022], and the material properties [Sharma et al. 2023] of objects, it is more difficult to impose precise control over the lighting via a text prompt; language generally offers only qualitative (e.g., warm, cold, cozy, etc.) and coarse positional (e.g., left, right, rim-lighting, etc.) descriptions of lighting. Furthermore, current text embeddings also have difficulty in encoding fine-grained information [Paiss et al. 2023]. However, due to the entanglement of the lighting and text embeddings, simply conditioning the text-to-image model on the lighting (e.g., by passing the light direction) will not allow for independent control of lighting and image content. Moreover, using a lighting representation such as a light direction vector or an environment map limits the types of lighting that can control the image generation.

In this paper we employ an alternative method of passing lighting conditions, namely radiance hints; a rendering of the target scene with a canonical homogeneous material lit by the target lighting. However, this typically requires precise knowledge of the underlying geometry which is unknown in the case of text-driven image generation. A key observation is that even though the diffusion model's sampling of the distribution of images is biased in terms of lighting, the learned distribution does contain the effects of different lighting conditions. Hence, in order to control the lighting during image generation, we need to guide the diffusion sampling process. Armed with this key observation, we revisit radiance hints and note that for guiding the sampling process, we do not need exact radiance hints, only a coarse approximation; we rely on the generative powers of the diffusion model to fill in the details.

We present a novel three stage method for providing fine-grained lighting control for diffusion-based image generation from text prompts. Since the background in an image is part of the lighting condition imposed on the foreground object, we focus primarily on controlling the lighting on the foreground object, allowing the background to change accordingly. In a first stage, we generate a provisional image of the given text prompt under uncontrolled (biased) lighting using a standard pretrained diffusion model. In the second stage, we compute a proxy shape from the provisional image using an off-the-shelf depth estimation network [Bhat et al. 2023] and foreground mask generator [Qin et al. 2020], from which we generate a set of radiance hints. Next, we resynthesize the image that matches both the text-prompt and the radiance hints using a refined diffusion model named *DiLightNet* (**Di**ffusion **Light**ing Control**Net**). To retain the rich texture information, we transform the generated provisional image using a learned encoder and multiply it with the radiance hints before passing it to DiLightNet. In the third stage, we inpaint a new background consistent with the target lighting. As our model is derived from large scale pretrained diffusion models, we can generate multiple replicates of the synthesized image that samples ambiguous interpretations of the materials.

We demonstrate our lighting controlled diffusion model on a variety of text-prompt-generated images and under different types of lighting, ranging from point lights to environment lighting. In addition, we perform an extensive ablation study to demonstrate the efficacy of each of the components that comprise DiLightNet.

Figure 2: Examples of lighting bias in diffusion-based image generation. Left: a batch of 12 images (text prompt: *"a photo of a soccer ball"*). The majority of the images are lit by a flash light; only two exhibit off-center lighting (3rd row, 1st column and 3rd column). Right: a batch of generated images of a robot dominated by light coming from either the front-left or front-right (text prompt: *"a photo of a toy robot standing on a wooden table"*; images are generated with a depth conditioned model to ensure a consistent shape).

2 RELATED WORK

Diffusion Models for Image Generation. Diffusion models have been shown to excel at the task of generating high quality images by sampling from a learned distribution (e.g., of photographs) [Karras et al. 2022; Song et al. 2021], especially when conditioned on text-prompts [Nichol et al. 2022; Ramesh et al. 2022; Rombach et al. 2022; Saharia et al. 2022]. Follow up work has endeavored to enrich text-driven diffusion models to exert higher level semantic control over the image generation process [Avrahami et al. 2022; Brooks et al. 2023; Ge et al. 2023; Hertz et al. 2022; Liu et al. 2020b; Mokady et al. 2023; Tumanyan et al. 2023; Voynov et al. 2023b], including

non-rigid semantic edits [Cao et al. 2023; Kawar et al. 2023], modifying the identity and gender of subjects [Kim et al. 2022], capturing the data distribution of underrepresented attributes [Cong et al. 2023], and material properties [Sharma et al. 2023]. However, with the exception of Alchemist [Sharma et al. 2023], these methods only offer mid and high level semantic control. Similar to Alchemist, our method aims to empower the user to control low level shading properties. Complementary to Alchemist which offers relative control over material properties such as translucency and gloss, our method provides fine-grained control over the incident lighting in the generated image.

Alternative guidance mechanisms have been introduced to provide spatial control during the synthesis process based on (sketch, depth, or stroke) images [Meng et al. 2022; Voynov et al. 2023a; Ye et al. 2023], identity [Ma et al. 2023; Ruiz et al. 2023b; Xiao et al. 2023], photo-collections [Ruiz et al. 2023a], and by directly manipulating mid-level information [Ho and Salimans 2021; Mou et al. 2023; Zhang et al. 2023b]. However, none of these methods provide control over the incident lighting. We follow a similar process and inject radiance hints modulated by a neural encoded version of the image into the diffusion model via a ControlNet [Zhang et al. 2023b].

2D diffusion models have also been leveraged to change viewpoint or generate 3D models [Liu et al. 2023; Watson et al. 2022; Xiang et al. 2023; Zhang et al. 2023a]. However, these methods do not offer control over incident lighting, nor guarantee consistent lighting between viewpoints. Paint3D [Zeng et al. 2023] directly generates diffuse albedo textures in the UV domain of a given mesh. Fantasia3D [Chen et al. 2023] and MatLaber [Xu et al. 2023] generate a richer set of reflectance properties in the form of shape and spatially-varying BRDFs by leveraging text-to-image 2D diffusion models and score distillation. Diffusion-based SVBRDF estimation [Sartor and Peers 2023; Vecchio et al. 2023] and diffusion-based intrinsic decomposition [Kocsis et al. 2023] also produce rich reflectance properties, albeit from a photograph instead of a text-prompt. However, all these methods require a rendering algorithm to visualize the appearance, including indirect lighting and shadows. In contrast, our method directly controls the lighting during the sampling process, leveraging the space of plausible image appearance embedded by the diffusion model.

Single Image Relighting. While distinct, our method is related to relighting from a single image, which is a highly underconstrained problem. To provide additional constraints, existing single image methods focus exclusively on either outdoor scenes [Griffiths et al. 2022; Liu et al. 2020a; Türe et al. 2021; Wu and Saito 2017; Yu et al. 2020], faces [Han et al. 2023; Nestmeyer et al. 2020; Pandey et al. 2021; Peers et al. 2007; Ranjan et al. 2023; Shu et al. 2017; Sun et al. 2019; Wang et al. 2008], or human bodies [Ji et al. 2022; Kanamori and Endo 2018; Lagunas et al. 2021]. In contrast, our method aims to offer fine-grained lighting control of general objects. Furthermore, existing methods expect a captured photograph of an existing scene as input, whereas, importantly, our method operates on, possibly implausible, generated images. The vast majority of prior single image relighting methods explicitly disentangle the image in various components, that are subsequently recombined after changing the lighting. In contrast, similar to Sun *et al.* [2019], we forego explicit decomposition of the input scene in disentangled components. However, unlike Sun *et al.*, we do not use a specially trained encoder-decoder model, but rely on a general generative diffusion model to produce realistic relit images. Furthermore, the vast majority of prior single image relighting methods represents incident lighting using a Spherical Harmonics encoding [Ramamoorthi 2002]. Notable exceptions are methods that represent the incident lighting by a shading image. Griffiths *et al.* [2022] pass a cosine weighted shadow map (along with normals and the main light direction) to a relighting network for outdoor scenes. Similarly, Kanamori *et al.* [2018] and Ji *et al.* [2022] pass shading and ambient occlusion maps to a neural rendering network. To better model specular reflections, Pandey *et al.* [2021] and Lagunas *et al.* [2021] pass, in addition to a diffuse shading image, also one or more specular shading images for neural relighting of human faces and full bodies respectively. We follow a similar strategy and pass the target lighting as a diffuse and (four) specular radiance hints as conditions to a diffusion model.

Relighting using Diffusion Models. Ding *et al.*[2023] alter lighting, pose, and facial expression by learning a CGI-to-real mapping from surface normals, albedo, and a diffuse shaded 3D morphable model fitted to a single photograph [Feng et al. 2021]. To preserve the identity of the subject in the input photograph, the diffusion model is refined on a small collection (~ 20) of photographs of the subject. Ponglertnapakorn *et al.*[2023] leverage off-the-shelf estimators [Deng et al. 2019; Feng et al. 2021; Yu et al. 2018] for the lighting, a 3D morphable model, the subject's identity, camera parameters, a foreground mask, and cast-shadows to train a conditional diffusion network that takes a diffuse rendered model under the novel lighting (blended on the estimated background), in addition to the identity, camera parameters, and target shadows to generate a relit image of the subject. While we follow a similar overall strategy, our method differs on three critical points. First, our method operates on general scenes which exhibit a broader range of shape and material variations than faces. Second, we provide multiple radiance hints (diffuse and specular) to control the lighting during the diffusion process. Finally, DiLightNet operates purely on an image generated via a text-prompt and our method does not require a real-world captured input photograph.

Lasagna [Bashkirova et al. 2023] also shares the goal of controlling the lighting in diffusion-based image generation. However, instead of radiance hints, Lasagna uses language tokens to control the lighting and thus lacks the fine-grained lighting control of DiLightNet. Furthermore, it only supports a predefined set of 12 directional lights while DiLightNet handles both point and environmental lighting.

3 OVERVIEW

Our method takes as input a text prompt (describing the image content), the target lighting, a content-seed that controls variations in shape and texture, and an appearance-seed that controls variations in light-material interactions. The resulting output is a generated image corresponding to the text prompt that is consistent with the target lighting. We assume that the image contains an isolated foreground object, and that the background content is implicitly described by the target lighting. We make no assumption on the

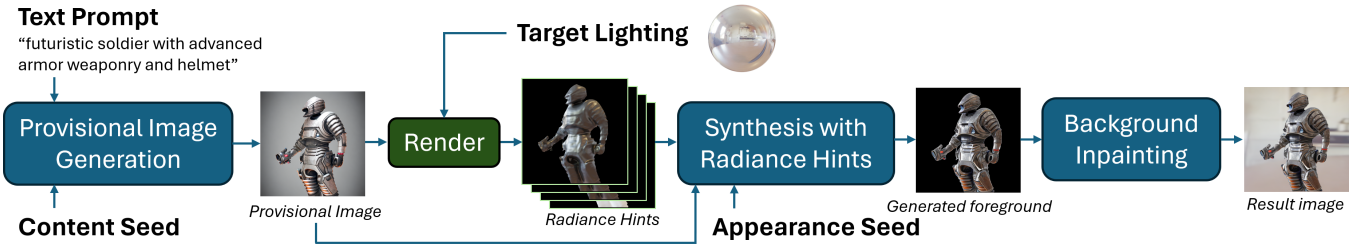

Figure 3: Overview of our pipeline for lighting-controlled prompt-driven image synthesis: (1) We start by generating a *provisional image* using a pretrained diffusion model under uncontrolled lighting given a text prompt and a content-seed. (2) Next, we pass an appearance-seed, the provisional image, and a set of radiance hints (computed from the target lighting and a coarse estimate of the depth) to DiLightNet that will resynthesize the image such that becomes consistent with the target lighting while retaining the content of the provisional image. (3) Finally, we inpaint the background to be consistent with foreground object and the target lighting.

target lighting, and support arbitrary lighting conditions. Finally, while we do not impose any constraint on the realism of the synthesized content (e.g., fantastic beasts), we assume an image style that depicts physically-based light-matter interactions (e.g., we do not support artistic styles such as cell-shading or surrealistic images).

Our pipeline for lighting-controlled prompt-driven image synthesis consists of three separate stages (Figure 3):

(1) *Provisional Image Generation:* In the first stage, we generate a provisional image with uncontrolled lighting given the text-prompt and the content-seed using a pretrained diffusion model [Stability AI 2022b]. The goal of this stage is to determine the shape and texture of the foreground object. Optionally, we add *"white background"* to the text-prompt to facilitate foreground detection.
(2) *Synthesis with Radiance Hints:* In the second stage (section 4), we first generate radiance hints given the provisional image and target lighting. Next, the radiance hints are multiplied with a neural encoded version of the provisional image, and passed to DiLightNet together with the text-prompt and appearance-seed. The result of this second stage is the foreground object with consistent lighting.
(3) *Background Inpainting:* In the third stage (section 5), we inpaint the background consistent with the target lighting.

4 SYNTHESIS WITH RADIANCE HINTS

Our goal is to synthesize an image with the same foreground object as in the provisional image, but with its appearance consistent with the given target lighting. We will finetune the same diffusion model used to generate the provisional image to take in account the target lighting via a ControlNet [Zhang et al. 2023b]. A ControlNet assumes a control signal per pixel, and thus we cannot directly guide the diffusion model using a direct representation of the lighting such as an environment map or a spherical harmonics encoding. Instead, we encode the *effect* of the target lighting on each pixel's outgoing radiance using radiance hints.

4.1 Radiance Hint Generation

A radiance hint is a visualization of the target shape under the target illumination, where the material of the object is replaced by a homogeneous proxy material (e.g., uniform diffuse). However, we do not have access to the shape of the foreground object. To circumvent this challenge, we observe that ControlNet typically does not require very precise information and it has been shown to work well on sparse signals such as sketches. Hence, we argue that an approximate radiance hint computed from a coarse estimate of the shape suffices.

To estimate the shape of the foreground object, we first segment the foreground object from the provisional image using an off-the-shelf salient object detection network. Practically, we use U2Net [Qin et al. 2020] as it offers a good trade-off between speed and accuracy; we revert to SAM [Kirillov et al. 2023] for the rare cases where U2Net fails to provide a clean foreground segmentation. Next, we apply another off-the-shelf depth estimation network (ZoeDepth [Bhat et al. 2023]) on the segmented foreground object. The estimated depth map is subsequently triangulated in a mesh and rendered under the target lighting with the proxy materials. However, single-image depth estimation is a challenging problem, and the resulting triangulated depth maps are far from perfect. Empirically we find that ControlNet is less sensitive to low-frequency errors in the resulting shading, while high-frequency errors in the shading can lead to artifacts. We therefore apply a Laplace smoothing filter over the mesh to reduce the impact of high-frequency discontinuities.

Inspired by the positional encoding in NeRFs [Mildenhall et al. 2020], we also encode the impact of different frequencies in the target lighting on the appearance of the foreground shape in separate radiance hints. Leveraging the fact that a BRDF acts as a band-pass filter on the incident lighting, we generate 4 radiance hints, each rendered with a different material modeled with the Disney BRDF model [Burley 2012] (one pure diffuse material and three specular materials with roughness set to 0.34, 0.13, and 0.05 respectively). We render the radiance hints, inclusive of shadows and indirect lighting, with Blender's Cycles path tracer.

4.2 Lighting Conditioned ControlNet

As noted before, we finetune a diffusion model to incorporate the radiance hint images using ControlNet, as well as the original text prompt used to generate the provisional image, and the appearance-seed. However, as we finetune the model, there is no guarantee

that it will generate a foreground object with the same shape and texture as in the provisional image. Therefore, we want to include the provisional image into the diffusion process. However, the texture and shape information in the provisional image is entangled with the unknown lighting from the first stage. We disentangle the relevant texture and shape information by first encoding the provisional image (with the alpha channel set to the segmentation mask). Our encoder follows Gao *et al.*'s [2020] deferred neural relighting architecture, but with a reduced number of channels to limit memory usage (see supplemental material). In addition, we include a channel-wise multiplication between the 12-channel encoded feature map of the provisional image and the 4×3-channel radiance hints which are subsequently passed to ControlNet.

4.3 Training

To train DiLightNet, we opt for a synthetic 3D training set that allows us to precisely control the lighting, geometry, and the material distributions. It is critical that the synthetic training set contains a wide variety of shapes, materials, and lighting.

Shape and Material Diversity. We select synthetic objects from the LVIS category in the Objaverse dataset [Deitke et al. 2022] that also have either a roughness map, a normal map, or both, yielding an initial subset of $13K$ objects. In addition, we select $4K$ objects from the Objaverse dataset (from the LVIS category) that only contain a diffuse texture map and assign a homogeneous specular BRDF with a roughness log-uniformly selected in $[0.02, 0.5]$ and specular tint set to 1.0. To ensure that the refined diffusion model has seen objects with homogeneous materials, we select an additional $4K$ objects (from the LVIS category) and randomly assign a homogeneous diffuse albedo and specular roughness sampled as before.

Empirically, we found that the diversity of detailed spatially varying materials in the Objaverse dataset is limited. Therefore, we further augment the dataset with the shapes with the most "likes" (a statistic provided by the Objaverse dataset) from each LVIS category. For each of these selected shapes we automatically generate UV coordinates using Blender (we eliminate the shapes (17) for which this step failed), and create 4 synthetic objects per shape by assigning a randomly selected spatially varying material from the INRIA-Highres SVBRDF dataset [Deschaintre et al. 2020], yielding a total of $4K$ additional objects with enhanced materials.

In total, our training set contains $25K$ synthetic objects with a wide variety of shapes and materials. We scale and translate each object such that its bounding sphere is centered at the origin with a radius of 0.5m.

Lighting Diversity. We consider five different lighting categories:

(1) *Point Light Source* random uniformly sampled on the upper hemisphere (with $0 \leq \theta \leq 60^\circ$) surrounding the object with radius sampled in $[4m, 5m]$, and with the power uniformly chosen in $[500W, 1500W]$. To avoid completely black images when the point light is positioned behind the object, we also add a $1W$ uniform white environment light.
(2) *Multiple Point Light Sources:* three light sources sampled in the same manner as the single light source case, including the uniform environment lighting.
(3) *Environment Lighting* sampled from a collection of 679 environment maps from Polyhaven.com.
(4) *Monochrome Environment Lighting* are the luminance only versions of the environment lighting category. Including this category combats potential inherent biases in the overall color distribution in the environment lighting.
(5) *Area Light Source* simulates studio setups with large light boxes. We achieve this by randomly placing an area light source on the hemisphere surrounding the object (similar to point light sources) aimed at the object, with a size randomly chosen in the range $[5m, 10m]$ and total power sampled in $[500W, 1500W]$. Similar to the point lighting, we add a uniform white environment light of $1W$.

Rendering. We render each of the $25K$ synthetic objects from four viewpoints uniformly sampled on the hemisphere with radius uniformly sampled from $[0.8m, 1.1m]$ and $10^\circ \leq \theta \leq 90^\circ$, aimed at the object with a field of view sampled from $[25^\circ, 30^\circ]$, and lit with 12 different lighting conditions, selected with a relative ratio of $3:1:3:2:3$ for point source lighting, multiple point sources, environment maps, monochrome environment maps, and area light sources respectively. For each rendered viewpoint, we also require corresponding radiance hints. However, at *evaluation* time, the radiance hints will be constructed from estimated depth maps; using the ground truth geometry and normals during *training* would therefore introduce a domain gap. We observe that depth-derived radiance hints include two types of approximations. First, due to the smoothed normals, the resulting shading will also be smoothed and shading effects due to intricate geometrical details are lost; i.e., it locally affects the radiance hints. Second, due to the ambiguities in estimating depth from a single image, missing geometry and global deformations cause incorrect shadows; i.e., a non-local effect. We argue that diffusion models can plausibly correct the former, whereas the latter is more ambiguous and difficult to correct. Therefore, we would like the training radiance hints to only introduce approximations on the local shading. This is achieved by using the ground truth geometry with modified shading normals. We consider two different approximations for the shading normals, and randomly select at training time which one to use: (1) we use the geometric normals and ignore any shading normals from the object's material model, or (2) we use the corresponding normals from the smoothed triangulated depth (to reduce computational costs, we estimate the depth for each synthetic object for each viewpoint under uniform white lighting instead for each of the 9 sampled lighting conditions).

Training Dataset. At training time we dynamically compose the input-output pairs. We first select a synthetic object and view uniformly. Next, we select the lighting for the input and output image. To select the lighting condition for the input training image, we note that images generated with diffusion models tend to be carefully white balanced. Therefore, we exclude the input images rendered under (colored) environment lighting. For the output image, we randomly select any of the 12 precomputed renders (including those rendered with colored environment lighting). We select the radiance hints corresponding to the output with a 1:9 ratio for the radiance hints with smoothed depth-estimated normals versus geometric normals. To further improve robustness with respect to colored

lighting, we apply an additional color augmentation to the output images by randomly shuffling their RGB color channels; we use the same color channel permutation for the output image and its corresponding radiance hints.

5 BACKGROUND INPAINTING

Environment-based Inpainting. When the target lighting is specified by an environment map, we can directly render the background image using the same camera configuration as for the radiance hints. We composite the foreground on the background using the previously computed segmentation mask filtered with a 3×3 average filter to smooth the mask edges.

Diffusion-based Inpainting. For all other lighting conditions, we use a pretrained diffusion-based inpainting model [Rombach et al. 2022] (i.e., the *stable-diffusion-2-inpainting* model [Stability AI 2022a]). We input the synthesized foreground image along with the (inverse) segmentation mask, as well as the original text prompt, to complete the foreground image with a consistent background.

6 RESULTS

We implemented DiLightNet in PyTorch [Paszke et al. 2019] and use *stable diffusion v2.1* [Stability AI 2022b] as the base pretrained diffusion model to refine. We jointly train the provisional image encoder as well as the ControlNet using AdamW [Loshchilov and Hutter 2018] with a 10^{-5} learning rate (all other hyper-parameter are kept at the default values) for $150K$ iterations using a batch size of 64. Training took approximately 30 hours using 8× NVidia V100 GPUs. The training data is rendered using Blender's Cycles path tracer [Blender Foundation 2011] at 512×512 resolution with 4096 samples per pixel.

Consistent Lighting Control. Figure 9 shows five generated scenes (the provisional image is shown in the first column for reference) under 5 different lighting conditions (point light (2nd and 3rd column), and 3 different environment maps from [Debevec 1998]: Eucalyptus Grove (4th column), Kitchen (5th column), and Grace Cathedral (last column)) for five different prompts. Each prompt was chosen to demonstrate our method's ability to handle different material and geometric properties such high specular materials (1st row), rich geometrical details (2nd row), objects with multiple homogeneous materials (3rd row), non-realistic geometry (4th row), and spatially-varying materials (last row). The provisional image in the last two rows are generated with *DALL-E3* instead of *stable diffusion v2.1* to better model the more complex prompt. We observe that DiLightNet produces plausible results and that the appearance is consistent under the same target lighting for different prompts. Furthermore, the lighting changes are plausible over each prompt. Please refer to the supplemental material for additional results.

Additional User Control. One advantage of our three step solution is that the user can alter the appearance-seed in the second stage to modify the interpretation of the materials in the provisional image. Figure 4 showcases how different appearance-seeds affect the generated results. Altering the appearance-seed yields alternative explanations of the appearance in the provisional image. Conversely, using the same appearance-seed produces a consistent appearance under different controlled lighting conditions (as demonstrated in Figure 9).

In addition to the appearance-seed, we can further specialize the text prompt between the first and second stage to provide additional guidance on the material properties. Figure 5 shows four specializations of an initial prompt (*"toy robot"*) by adding: *"paper made"*, *"plastic"*, *"specular shinny metallic"*, and *"mirror polished metallic"*. From these results we can see that all variants are consistent under the same lighting, but with a more constrained material appearance (i.e., diffuse without a highlight, a mixture of diffuse and specular, and two metallic surfaces with a different roughness).

User Study. We perform two user studies to measure the perceptual lighting accuracy and the consistency of the resulting appearance under varying lighting; i.e., how well changes induced by the target lighting are disentangled from the appearance-seed.

In the first study, participants rate the lighting similarity of the foreground objects in image pairs (four-level rating range where 0 means least similar and 3 means most similar) selected from three groups of image pairings (10 pairs in each group):

(1) a synthetic object rendered under the target lighting is paired with any of the generated images shown in this paper and the supplemental material under identical lighting;
(2) a pair of synthetic objects rendered under identical target lighting (this serves as the positive baseline); and
(3) a synthetic image paired with a generated image without lighting control (the negative baseline). To avoid that the background affects the judgment, we replace the background with the target environment lighting.

The average total rating over 20 non-expert participants with images shown in randomized order for each of the three classes is: 19.61/19.85/12.25, showing that DiLightNet scores similar to the positive reference.

In a second study, participants rate the appearance consistency of the foreground objects in image pairs generated with rotated environment lighting. We opt for rotating the lighting to retain the overall color balance and frequency of lighting. The three groups of pairings under rotated lighting are:

(1) image pairs generated with the same prompt and seeds;
(2) image pairs rendered with the same synthetic object (positive baseline); and
(3) pairs generated without lighting control with the same text prompt but different content-seeds (negative baseline).

The average total rating was 25.75/25.05/11.35, confirming appearance consistency on par with the positive baseline.

7 ABLATION STUDY

We perform a series of qualitative and quantitative ablation studies to better understand the impact of the different components that comprise our method. For quantitative evaluation, we create a synthetic test set by selecting objects from the Objaverse dataset that have the 'Staff Picked' label and *no* LVIS label, ensuring that there is no overlap between the training and test set. To ensure high quality synthetic objects, we manually remove scenes that are not limited to a single object and/or objects with low quality

Figure 4: Impact of changing the appearance-seed. If not sufficiently constrained by the text prompt, the generated provisional image (left) might not provide sufficient information for DiLightNet to determine the exact materials of the object. Altering the appearance-seed directs DiLightNet to sample a different interpretation of light-matter interaction in the provisional image. In this example, altering the appearance-seed induces changes in the interpretation of the glossiness and smoothness of the leather gloves.

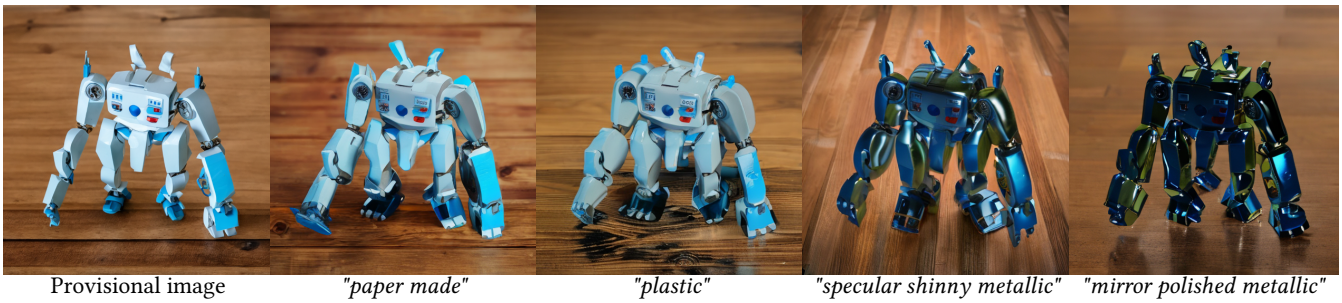

Figure 5: Impact of prompt specialization in DiLightNet. Instead of altering the appearance-seed, the user can also specialize the prompt with additional material information in the 2nd stage. In this example the initial prompt (*"toy robot"*) is augmented with additional material descriptions while keeping the (point lighting) fixed.

Table 1: Quantitative comparison of different variants of passing radiance hints to the DiLightNet (rows 1-3), the number of radiance hints (rows 4-6), impact of including the segmentation mask (row 7-8) and different training data augmentation schemes (rows 9-12).

Variant	PSNR	SSIM	LPIPS
Our Network	**22.97**	**0.8249**	**0.1165**
Direct ControlNet	22.82	0.8216	0.1212
Non-Encoded Multiplication	22.40	0.8174	0.1232
3 Radiance Hints	22.92	0.8197	0.1188
4 Radiance Hints	**22.97**	**0.8249**	**0.1165**
5 Radiance Hints	22.79	0.8200	0.1176
w/ Mask	**22.97**	**0.8249**	**0.1165**
w/o Mask	22.23	0.8148	0.1184
Full Augmentation	**22.97**	**0.8249**	**0.1165**
w/o Material Augmentation	22.90	0.8235	0.1178
w/o Smoothed Normal	21.88	0.7974	0.1314
w/o Color Augmentation	22.54	0.8161	0.1223

scanned textures with baked in lighting effects, yielding a test set of 50 high quality synthetic objects. We render each test scene for 3 viewpoints and 6 lighting conditions. We quantify errors with the PSNR, SSIM, and LPIPS [Zhang et al. 2018] metrics. Because the appearance-seed is a user controlled parameter, we assume that the user would select the appearance-seed that produces the most plausible result. To simulate this process, we report the errors for each scene/view/lighting combination that produces the lowest LPIPS errors on renders generated with 4 different appearance-seeds.

Provisional Image Encoding. DiLightNet multiplies the (encoded) provisional image with the radiance hints. We found that both the encoding, as well as the multiplication is critical for obtaining good results. Figure 6 shows a comparison of DiLightNet versus two alternate architectures:

(1) *Direct ControlNet* passes the provisional image directly as an additional channel (in addition to the radiance hints) instead of multiplying, yielding 16 channels input for ControlNet (3-channels for the provisional image, plus (4×3)-channels for the radiance hints, and 1 channel for the mask); and
(2) *Non-encoded Multiplication* of the provisional image (without encoding) with the radiance hints.

Neither of the variants generates satisfactory results. This qualitative result is further quantitatively confirmed in Table 1 (rows 1-3).

Impact of Number of Radiance Hints. Table 1 (rows 4-6) compares the impact of changing the number of (specular) radiance hints;

all variants include a diffuse radiance hint. The 3 radiance hints variant includes 2 specular radiance hints with roughness 0.13, and 0.34. The 4 radiance hints variant includes one additional specular radiance hint with roughness 0.05. Finally, the 5 radiance hints variant includes an additional (sharp specular) hint with roughness 0.02. From the quantitative results in Table 1 we can see that 4 radiance hints perform best. Upon closer inspection of the results, we observe that there is little difference for scenes that exhibit a simple shape with simple materials. However, for scenes with a more complex shape we find that the 3 radiance hints are insufficient to accurately model the light-matter interactions. For scenes with complex materials, we found that providing too many radiance hints can also be detrimental due to the limited quality of the (smoothed) depth-estimated normals.

Foreground Masking. DiLightNet takes the foreground mask as additional input. To better understand the impact of including the mask, we also train a variant without taking the mask as an additional channel. Instead we fill the background with black pixels in the provisional image. During training we also remove the background in the reference images. As a consequence, DiLightNet will learn to generate a black background. For the ablation, we only compute the errors over the foreground pixels. As shown in Table 1 (rows 7-8), the variant trained without a mask produces larger errors especially on cases with either complex shape or materials.

Training Augmentation. We eliminate each of the three augmentations from the training set to better gauge their impact (Table 1, rows 9-12):

- *Without Normal Augmentation:* This variant is trained using radiance hints rendered with the ground truth shading normals, instead of the smoothed depth-estimated normals or the geometric normals;
- *Without Color Augmentation:* This variant is trained on the full training set without swapping the RGB color channels; and
- *Without Material Augmentation:* This model is trained with the basic 13K dataset without material augmentations.

From Table 1, we observe that all three augmentations improve the robustness of DiLightNet. Of all augmentations, the normal augmentation has the largest impact as it helps to bridge the domain gap between perfect shading normals (in the training) and the smoothed estimated depth normals. The color augmentation also improves the quality for all test scenes, albeit to lesser degree. The benefits of the material augmentation are most noticeable for objects with smooth shapes (i.e., low geometrical complexity) as errors in the normal estimation can help to mask inaccuracies in representing complex materials.

8 DISCUSSION

Relation to Single Image Relighting. By skipping the first stage and directly inputing a captured photograph as the provisional image into DiLightNet, we can perform approximate single image relighting (Figure 7). However, due to the lack of a text prompt, the relighting results might not be ideal. Furthermore, unlike existing single image relighting methods that are trained for a more narrow class of scenes, DiLightNet is trained to handle any type of synthesized image for which there might not exists a 'real' reference under novel lighting (e.g., the 'giraffe-turtle' in Figure 9), DiLightNet only aims to produce *plausible* images. Nevertheless, the relighting results generated by DiLightNet are plausible for scenes from which a reasonably accurate depth and mask can be extracted. Further refining DiLightNet to be more robust for relighting photographs is a promising avenue for future research.

Limitations. Our method is not without limitations. Due to the limitations of specifying the image content with text prompts, the user only has limited control over the materials in the scene. Consequently, the material-light interactions might not follow the intention of the prompt-engineer. DiLightNet enables some indirect control, beyond text prompts, through the appearance-seed. Integrating material aware diffusion models, such as Alchemist [Sharma et al. 2023], could potentially lead to better control over the material-light interactions. Furthermore, our method relies on a number of off-the-shelf solutions for estimating a rough depth map and segmentation mask of the foreground object. While our method is robust to some errors in the depth map, some types of errors (e.g., the bass-relief ambiguity) can result in non-satisfactory results. An interesting alternative pipeline takes a reference depth map as input (e.g., using a depth conditioned diffusion model such as *"stable-diffusion-2-depth"*), thereby bypassing the need to estimate the depth and mask. As demonstrated in Figure 8, augmenting the input with a reference depth map, further increases the quality of the results. Finally, animating/altering the lighting using a fixed content-seed can result in some minor structural shape changes because the images are generated independently (see supplemental video). Incorporating cross-frame consistency to improve temporal stability is an interesting avenue for future research.

9 CONCLUSION

In this paper we introduced a novel method for controlling the lighting in diffusion-based text-to-image generation. Our method consists of three stages: (1) provisional image synthesis under uncontrolled lighting using existing text-to-image methods, (2) resynthesis of the foreground object using our novel DiLightNet conditioned by the radiance hints of the foreground object, and finally (3) inpainting of the background consistent with the target lighting. Key to our method is DiLightNet, a variant of ControlNet that takes an encoded version of the provisional image (to retain the shape and texture information) multiplied with the radiance hints. Our method is able to generate images that match both the text prompt and the target lighting. For future work we would like to apply DiLightNet to estimate reflectance properties from a single photograph and for text-to-3D generation with rich material properties.

ACKNOWLEDGMENTS

Pieter Peers was supported in part by NSF grant IIS-1909028. Chong Zeng and Hongzhi Wu were partially supported by NSF China (62332015 & 62227806), the Fundamental Research Funds for the Central Universities (226-2023-00145), and Information Technology Center and State Key Lab of CAD&CG, Zhejiang University.

REFERENCES

Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended diffusion for text-driven editing of natural images. In *CVPR*. 18208–18218.

Dina Bashkirova, Arijit Ray, Rupayan Mallick, Sarah Adel Bargal, Jianming Zhang, Ranjay Krishna, and Kate Saenko. 2023. Lasagna: Layered Score Distillation for Disentangled Object Relighting.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. 2023. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288* (2023).

Blender Foundation. 2011. Blender Cycles. https://github.com/blender/cycles.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*. 18392–18402.

Brent Burley. 2012. Physically-based shading at disney. In *ACM Siggraph Courses*, Vol. 2012.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. 2023. MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing. *arXiv preprint arXiv:2304.08465* (2023).

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. 2023. Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation. In *ICCV*.

Yuren Cong, Martin Renqiang Min, Li Erran Li, Bodo Rosenhahn, and Michael Ying Yang. 2023. Attribute-centric compositional text-to-image generation. *arXiv preprint arXiv:2301.01413* (2023).

Paul Debevec. 1998. Rendering synthetic objects into real scenes: bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '98)*. 189–198.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2022. Objaverse: A Universe of Annotated 3D Objects. *arXiv preprint arXiv:2212.08051* (2022).

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

Valentin Deschaintre, George Drettakis, and Adrien Bousseau. 2020. Guided fine-tuning for large-scale material transfer. In *Comp. Graph. Forum*, Vol. 39. 91–105.

Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. 2023. DiffusionRig: Learning Personalized Priors for Facial Appearance Editing. In *CVPR*. 12736–12746.

Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. 2021. Learning an Animatable Detailed 3D Face Model from In-the-Wild Images. *ACM Trans. Graph.* 40, 4, Article 88 (2021).

Duan Gao, Guojun Chen, Yue Dong, Pieter Peers, Kun Xu, and Xin Tong. 2020. Deferred neural lighting: free-viewpoint relighting from unstructured photographs. *ACM Trans. Graph.* 39, 6, Article 258 (2020).

Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. 2023. Expressive text-to-image generation with rich text. In *CVPR*. 7545–7556.

David Griffiths, Tobias Ritschel, and Julien Philip. 2022. OutCast: Outdoor Single-image Relighting with Cast Shadows. *Computer Graphics Forum* 41, 2 (2022), 179–193.

Yuxuan Han, Zhibo Wang, and Feng Xu. 2023. Learning a 3D Morphable Face Reflectance Model From Low-Cost Data. In *CVPR*. 8598–8608.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626* (2022).

Jonathan Ho and Tim Salimans. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS*.

Chaonan Ji, Tao Yu, Kaiwen Guo, Jingxin Liu, and Yebin Liu. 2022. Geometry-Aware Single-Image Full-Body Human Relighting. In *ECCV*. 388–405.

Yoshihiro Kanamori and Yuki Endo. 2018. Relighting humans: occlusion-aware inverse rendering for full-body human images. *ACM Trans. Graph.* 37, 6 (2018).

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. 2022. Elucidating the Design Space of Diffusion-Based Generative Models. In *NeurIPS*.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In *CVPR*. 6007–6017.

Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *CVPR*. 2426–2435.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. 2023. Segment Anything. In *ICCV*. 4015–4026.

Peter Kocsis, Vincent Sitzmann, and Matthias Nießner. 2023. Intrinsic Image Diffusion for Single-view Material Estimation. *arXiv preprint arXiv:2312.12274* (2023).

Manuel Lagunas, Xin Sun, Jimei Yang, Ruben Villegas, Jianming Zhang, Zhixin Shu, Belen Masia, and Diego Gutierrez. 2021. Single-image Full-body Human Relighting. In *EGSR - DL-only Track*.

Andrew Liu, Shiry Ginosar, Tinghui Zhou, Alexei A Efros, and Noah Snavely. 2020a. Learning to factorize and relight a city. In *ECCV*. Springer, 544–561.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*. 9298–9309.

Xihui Liu, Zhe Lin, Jianming Zhang, Handong Zhao, Quan Tran, Xiaogang Wang, and Hongsheng Li. 2020b. Open-edit: Open-domain image manipulation with open-vocabulary instructions. In *ECCV*. Springer, 89–106.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *ICLR*.

Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. 2023. Subject-Diffusion:Open Domain Personalized Text-to-Image Generation without Test-time Fine-tuning. *arXiv preprint arXiv:2307.11410* (2023).

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *ICLR*.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *ECCV* (2020), 405–421.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text inversion for editing real images using guided diffusion models. In *CVPR*. 6038–6047.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. 2023. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453* (2023).

Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas Lehrmann. 2020. Learning physics-guided face relighting under directional light. In *CVPR*. 5124–5133.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *ICML*. 16784–16804.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. 2023. Teaching clip to count to ten. *arXiv preprint arXiv:2302.12066* (2023).

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS* 32 (2019).

Pieter Peers, Naoki Tamura, Wojciech Matusik, and Paul Debevec. 2007. Post-production facial performance relighting using reflectance transfer. *ACM Trans. Graph.* 26, 3 (2007).

Puntawat Ponglertnapakorn, Nontawat Tritrong, and Supasorn Suwajanakorn. 2023. DiFaReli: Diffusion Face Relighting. *arXiv preprint arXiv:2304.09479* (2023).

Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. 2020. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern recognition* 106 (2020), 107404.

Ravi Ramamoorthi. 2002. *A signal-processing framework for forward and inverse rendering*. Stanford University.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125* (2022).

Anurag Ranjan, Kwang Moo Yi, Jen-Hao Rick Chang, and Oncel Tuzel. 2023. FaceLit: Neural 3D Relightable Faces. In *CVPR*. 8619–8628.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *CVPR*. 10684–10695.

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

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. 2023b. HyperDreamBooth: HyperNetworks for Fast Personalization of Text-to-Image Models. *arXiv preprint arXiv:2307.06949* (2023).

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS* 35 (2022), 36479–36494.

Sam Sartor and Pieter Peers. 2023. MatFusion: A Generative Diffusion Model for SVBRDF Capture. In *SIGGRAPH Asia 2023 Conference Papers*. 1–10.

Prafull Sharma, Varun Jampani, Yuanzhen Li, Xuhui Jia, Dmitry Lagun, Fredo Durand, William T. Freeman, and Mark Matthews. 2023. Alchemist: Parametric Control of Material Properties with Diffusion Models. *arXiv preprint arXiv:2312.02970* (2023).

Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. 2017. Neural face editing with intrinsic image disentangling. In *CVPR*. 5541–5550.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *ICLR*.

Figure 7: A demonstration of single image relighting obtained by bypassing the first stage and directly injecting a captured photograph as the provisional image (left). The resulting generated images (middle and right) represent a plausible relighting of the given photograph.

Figure 8: Lighting control results for a depth-controlled text-to-image diffusion model improves the quality of the results by providing a depth map as additional input.

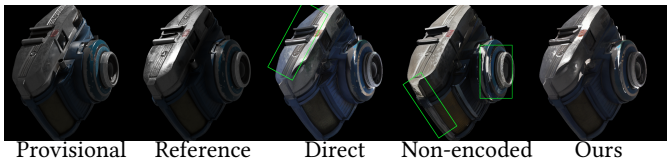

Figure 6: Ablation comparison of different architecture variants that: (1) *direct*ly pass the radiance hints and provisional image (without multiplication) to ControlNet, and (2) multiply the radiance hints with the *non-encoded* provisional image. DiLightNet's encoded multiplication generates visually more plausible results.

Stability AI. 2022a. Stable Diffusion V2 - Inpainting. https://huggingface.co/stabilityai/stable-diffusion-2-inpainting.
Stability AI. 2022b. Stable Diffusion V2.1. https://huggingface.co/stabilityai/stable-diffusion-2-1.
Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. 2019. Single image portrait relighting. *ACM Trans. Graph.* 38, 4 (2019).
Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2023. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. In *CVPR*. 1921–1930.
Murat Türe, Mustafa Ege Çıklabakkal, Aykut Erdem, Erkut Erdem, Pinar Satılmış, and Ahmet Oguz Akyüz. 2021. From Noon to Sunset: Interactive Rendering, Relighting, and Recolouring of Landscape Photographs by Modifying Solar Position. In *Comp. Graph. Forum*, Vol. 40. 500–515.
Giuseppe Vecchio, Rosalie Martin, Arthur Roullier, Adrien Kaiser, Romain Rouffet, Valentin Deschaintre, and Tamy Boubekeur. 2023. ControlMat: A Controlled Generative Approach to Material Capture. *arXiv preprint arXiv:2309.01700* (2023).
Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. 2023a. Sketch-Guided Text-to-Image Diffusion Models. In *ACM SIGGRAPH 2023 Conference Proceedings*. Article 55, 11 pages.
Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. 2023b. P+: Extended Textual Conditioning in Text-to-Image Generation. *arXiv preprint arXiv:2303.09522* (2023).
Yang Wang, Lei Zhang, Zicheng Liu, Gang Hua, Zhen Wen, Zhengyou Zhang, and Dimitris Samaras. 2008. Face relighting from a single image under arbitrary unknown lighting conditions. *IEEE PAMI* 31, 11 (2008), 1968–1984.
Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. 2022. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628* (2022).
Jung-Hsuan Wu and Suguru Saito. 2017. Interactive relighting in single low-dynamic range images. *ACM Trans. Graph.* 36, 2 (2017).
Jianfeng Xiang, Jiaolong Yang, Binbin Huang, and Xin Tong. 2023. 3D-aware Image Generation using 2D Diffusion Models. *arXiv preprint arXiv:2303.17905* (2023).
Guangxuan Xiao, Tianwei Yin, William T. Freeman, Frédo Durand, and Song Han. 2023. FastComposer: Tuning-Free Multi-Subject Image Generation with Localized Attention. *arXiv preprint arXiv:2305.10431* (2023).
Xudong Xu, Zhaoyang Lyu, Xingang Pan, and Bo Dai. 2023. Matlaber: Material-aware text-to-3d via latent brdf auto-encoder. *arXiv preprint arXiv:2308.09278* (2023).
Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. *arXiv preprint arXiv:2308.06721* (2023).
Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*. 325–341.
Ye Yu, Abhimitra Meka, Mohamed Elgharib, Hans-Peter Seidel, Christian Theobalt, and William AP Smith. 2020. Self-supervised outdoor scene relighting. In *ECCV*. 84–101.
Xianfang Zeng, Xin Chen, Zhongqi Qi, Wen Liu, Zibo Zhao, Zhibin Wang, Bin Fu, Yong Liu, and Gang Yu. 2023. Paint3D: Paint Anything 3D with Lighting-Less Texture Diffusion Models. *arXiv preprint arXiv:2312.13913* (2023).
Longwen Zhang, Qiwei Qiu, Hongyang Lin, Qixuan Zhang, Cheng Shi, Wei Yang, Ye Shi, Sibei Yang, Lan Xu, and Jingyi Yu. 2023a. DreamFace: Progressive Generation of Animatable 3D Faces under Text Guidance. *ACM Trans. Graph.* 42, 4, Article 138 (2023).
Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023b. Adding conditional control to text-to-image diffusion models. In *CVPR*. 3836–3847.
Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*. 586–595.

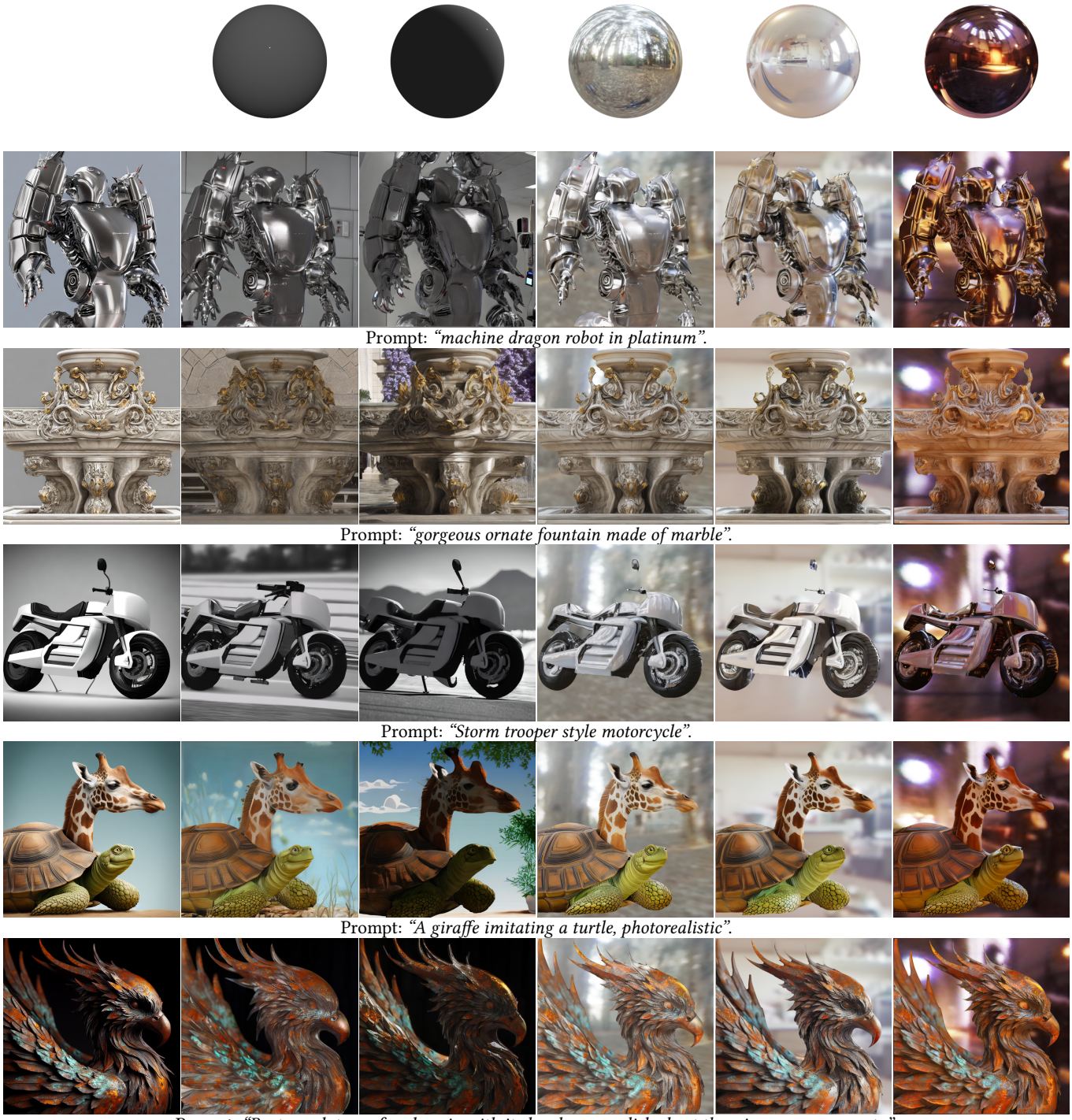

Figure 9: Text-to-image generated results with lighting control. The first column shows the provisional image as a reference, whereas the last five columns are generated under different user-specified lighting conditions (point lighting (columns 2-3) and environment lighting (columns 4-6)). The provisional images for the last two examples are generated with *DALL-E3* instead of *stable diffusion v2.1* to better handle the more complex prompt.