Spatiotemporally Consistent Indoor Lighting Estimation with Diffusion Priors

MUTIAN TONG, Columbia University, USA
RUNDI WU, Columbia University, USA
CHANGXI ZHENG, Columbia University, USA

Fig. 1. We present a method that achieves consistent indoor lighting estimation from a video wherein the lighting condition changes spatially and temporally. Here we show four frames of an video. For each frame, we show **(top)** the result of virtual object insertion under the estimated lighting and **(bottom)** estimated environment maps at locations where the two mirror spheres are inserted. In each column, the environment map on the bottom-left corresponds to the mirror sphere above the desk, and the one on the bottom-right corresponds to the sphere under the desk.

Indoor lighting estimation from a single image or video remains a challenge due to its highly ill-posed nature, especially when the lighting condition of the scene varies spatially and temporally. We propose a method that estimates from an input video a continuous light field describing the spatiotemporally varying lighting of the scene. We leverage 2D diffusion priors for optimizing such light field represented as a MLP. To enable zero-shot generalization to in-the-wild scenes, we fine-tune a pre-trained image diffusion model to predict lighting at multiple locations by jointly inpainting multiple chrome balls as light probes. We evaluate our method on indoor lighting estimation from a single image or video and show superior performance over compared baselines. Most importantly, we highlight results on spatiotemporally consistent lighting estimation from in-the-wild videos, which is rarely demonstrated in previous works.

1 INTRODUCTION

High-quality lighting is crucial for virtual object insertion in applications such as augmented reality and video composition. Oftentimes, in these applications, one needs to estimate lighting represented as an environment map from a single image or video of the scene. However, this problem remains challenging due to its highly ill-posed nature—the estimated environment map must have a high dynamic range (HDR) and capture the scene content beyond the narrow field of view of the input low dynamic range (LDR) images. The problem becomes even more challenging when the input video captures an indoor scene where the lighting condition changes spatially and temporally: lighting intensity may vary as one moves around, and some lights may be turned on or off over time.

Several existing works attempt to estimate HDR lighting from LDR image(s) of a *static* scene. Some can only predict a global illumination of the scene, ignoring spatially varying lighting effects [Barron and Malik 2014; Dastjerdi et al. 2023; LeGendre et al. 2019; Phongthawee et al. 2024; Wang et al. 2022; Yu and Smith 2019], while others estimate spatially varying lighting by predicting either an environment map per pixel [Garon et al. 2019; Li et al. 2020, 2021; Zhu et al. 2022] or a 3D lighting volume [Li et al. 2023; Srinivasan et al. 2020; Wang et al. 2024, 2021]. Yet, none of the existing methods can estimate *spatio-temporally consistent* lighting from a video that captures a dynamic lighting condition.

In this work, we present a method that estimates from an input video a continuous, spatiotemporal light field describing the illumination L at each time instance t and each spatial location $\boldsymbol{x}$ and incident direction $\boldsymbol{d}$. This light field is represented by a multilayer perceptron network (MLP) that approximates a six-dimensional (6D) light field function $L(\boldsymbol{x}, t, \boldsymbol{d})$. The challenge lies in effectively training this MLP from a single image or video.

Our overarching idea is to leverage 2D diffusion priors for such a 6D MLP training, inspired by 3D generation using 2D diffusion priors [Gao et al. 2024; Poole et al. 2022; Wu et al. 2024]. Our 2D diffusion model follows DiffusionLight [Phongthawee et al. 2024], which formulates static lighting estimation as an inpainting task: It inserts a chrome ball as a light probe onto the image using a fine-tuned image diffusion model and then unwraps the inpainted chrome ball into an environment map. However, DiffusionLight is trained to inpaint a single chrome ball at the image center, thus only

able to estimate a global environment map at the image's viewpoint. It cannot estimate coherent lighting across multiple spatial locations.

Instead, we directly train a diffusion model for spatially consistent lighting prediction. This is done by jointly inpainting multiple chrome balls conditioned on the relative depths of the balls and the background scene. To this end, we build our training dataset on Infinigen Indoors [Raistrick et al. 2024]—a procedural indoor scene generator based on Blender [Blender 2024]—by rendering ground-truth environment maps at different spatial locations. After obtaining the 2D diffusion model, we use it to train an MLP-represented light field $L(\boldsymbol{x}, t, \boldsymbol{d})$. The training process encourages 2D renderings of chrome balls under the light field to follow image priors encoded in the 2D diffusion model. For a single image input, we simply drop the input time from the MLP and follow the same training procedure to obtain a spatially varying light field.

We evaluate our method on the task of indoor lighting estimation from a single image/video and show superior performance over compared methods, especially when the lighting condition varies spatially and temporally. Moreover, we highlight results on spatiotemporal lighting estimation from in-the-wild videos, which remains elusive in previous works.

2 RELATED WORK

Lighting Estimation. Estimating lighting conditions has been a long-standing problem in computer vision and graphics. Some existing works make use of light probes in the image and perform lighting estimation via inverse rendering [Debevec 2008; Park et al. 2020; Verbin et al. 2024; Yi et al. 2018; Yu et al. 2023]. Others do not require such light probes to appear in images and often rely on a neural network to predict the lighting. Many of these works predict single fixed lighting of the scene, either producing an environment map at one particular location where the virtual object is inserted [Liang et al. 2025] or recovering a full HDR panorama from the input image's limited field of view [Dastjerdi et al. 2023; Gardner et al. 2017; LeGendre et al. 2019; Phongthawee et al. 2024; Somanath and Kurz 2021; Wang et al. 2022; Yu and Smith 2019].

Recently, Phongthawee et al. [2024] proposed DiffusionLight, which uses a pre-trained large-scale image diffusion model to inpaint a chrome ball into the center of an input image. The resulting chrome ball image is then unwrapped into an environment map. This method fine-tunes the diffusion model on a dataset of HDR panorama paired with random crops. Our diffusion model follows a similar design but instead jointly inpaints multiple chrome balls, in order to achieve consistent lighting prediction across different spatial locations of the scene. To train the model to reason about spatially varying lighting, we construct a synthetic dataset with ground-truth lighting at different spatial locations.

Apart from estimating a single environment map, several recent works can predict spatially varying lighting from a single image by outputting either per-pixel lighting [Garon et al. 2019; Li et al. 2020, 2021; Zhu et al. 2022] or a 3D lighting volume [Li et al. 2023; Srinivasan et al. 2020; Wang et al. 2024, 2021]. In particular, the work by Li et al. [2023] is among the first to take a video input and use a recurrent neural network (RNN) to improve lighting prediction while preserving spatiotemporal consistency. However, this method assumes the input video to capture a static scene and cannot handle dynamic scenes with time-varying lighting conditions. To our knowledge, our work is the first toward spatiotemporally consistent lighting estimation from a video that captures dynamic lighting.

3D from 2D generative priors. Recent research have attempted to leverage 2D image diffusion models for 3D content generation from a text prompt [Lin et al. 2023; Poole et al. 2022; Wang et al. 2023] or images [Gao et al. 2024; Liu et al. 2023; Shi et al. 2023]. This is motivated by the fact that native 3D data are too limited (in comparison to 2D image data) to train large diffusion models. A pioneer work along this direction, DreamFusion [Poole et al. 2022], proposes score distillation sampling (SDS), wherein a 3D model is optimized with supervision from a 2D image diffusion model. Subsequent works [Tang et al. 2024; Wu et al. 2024; Zhou and Tulsiani 2023] all use a variant of SDS, i.e., sampling an image by running multiple DDIM [Song et al. 2020] sampling steps on a noisy encoding of the current rendered images. Here, the sampled images serves as a pseudo ground-truth for 3D reconstruction, and the 2D image diffusion model is to provide a prior of partial observations (*i.e.*, the rendered images) of the 3D model.

Our framework follows a similar spirit when optimizing the spatiotemporal light field. Our inpainting diffusion model is meant to provide a prior of partial observations—in our case, perfectly reflective chrome balls—of the environment lighting.

3 METHOD

Given an input LDR image (or video) of an indoor scene, we aim to estimate spatially (and temporally) varying HDR lighting represented as environment maps. Our method starts by fine-tuning a pre-trained image diffusion model to jointly predict lighting at multiple positions (Sec. 3.1). Then, taking advantage of the learned priors from the diffusion model, we distill a spatially (and temporally) varying light field $L(\boldsymbol{x}, t, \boldsymbol{d})$ represented as an MLP (Sec. 3.2). Please refer to Fig. 2 for an overview of our framework.

3.1 Diffusion Model for Lighting Prediction

First, we consider a single LDR image as input. We aim to learn a prior for lighting estimation at locations within the image's view frustum. This prior will be used when presented with a video input for estimating spatiotemporally dynamic lighting (in Sec. 3.2).

Inspired by DiffusionLight [Phongthawee et al. 2024], we formulate the lighting estimation problem as an inpainting task, *i.e.*, inserting chrome balls onto the image using an image diffusion model and unwrapping them into equirectangular environment maps. Formally, given an input image I, we estimate the lighting incident to the target locations $\{\boldsymbol{x}_i\}_{i=1}^N$. First, we construct a set of chrome balls $B = \{(\boldsymbol{x}_i, r_i)\}_{i=1}^N$ as light probes at these locations. The radius r_i of each chrome ball is chosen such that the diameter of its projection on the image plane is about 1/4 of the image size. We use a diffusion model to learn the conditional distribution over the image I^B with the chrome balls inserted, *i.e.*, $p(I^B|I, B)$.

Diffusion Model. We build our diffusion model on the pre-trained Stable Diffusion inpainting model [Rombach et al. 2022] with depth-conditioned ControlNet [Zhang et al. 2023]. To this end, we first

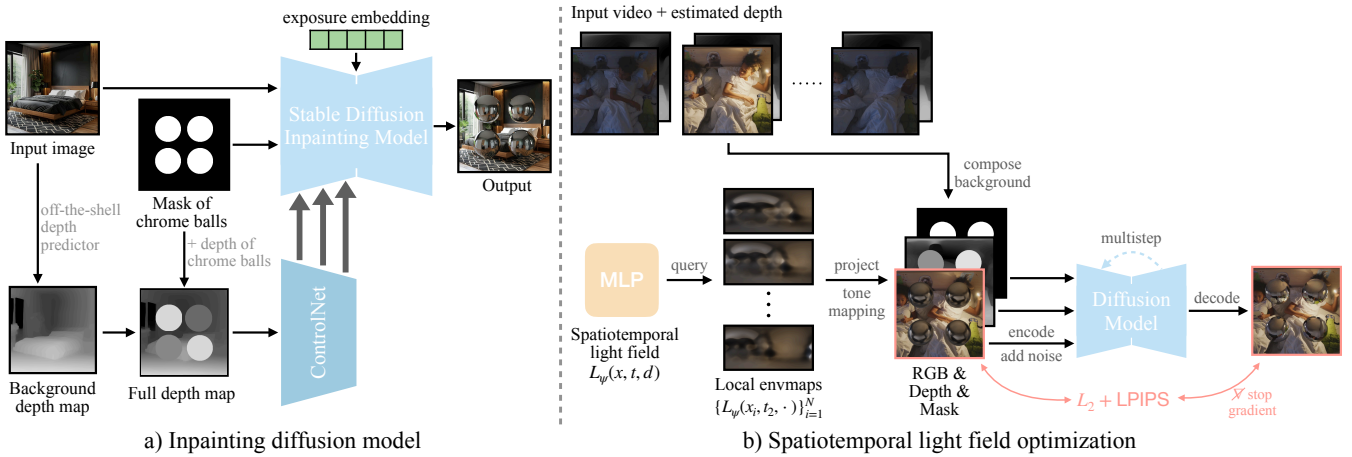

Fig. 2. Method overview. **(a)** We fine-tune a pre-trained Stable Diffusion Inpainting Model [Rombach et al. 2022] for single image lighting estimation by inpainting multiple chrome balls as light probes in the image (Sec. 3.1). **(b)** Given a video of an indoor scene with dynamic lighting changes, we distill a spatiotemporal light field (represented as an MLP) from our diffusion model via optimization (Sec. 3.2).

predict the camera intrinsics K and depth map D_I from the image using off-the-shelf estimators [Gantelius 2024; Jin et al. 2023; Ranftl et al. 2021]. We then project the chrome balls $B = \{(\boldsymbol{x}_i, r_i)\}_{i=1}^N$ onto the image plane to obtain an inpainting mask M^B, and compose the projected depth of the chrome balls onto the background depth map D_I to obtain a depth map with chrome balls inserted $D^{I,B}$. Next, we encode the input image I via a VAE encoder while resizing the mask M^B to the same resolution as the encoded image. The encoded I and M^B are then concatenated with the noisy latent vector z_τ and fed to the denoising U-Net ϵ_θ. At the same time, the conditional depth map $D^{I,B}$ is provided as input to the ControlNet. See Fig. 2-a for an illustration of our model architecture.

To allow for prediction of HDR chrome balls, we follow DiffusionLight [Phongthawee et al. 2024] to condition the diffusion model on an exposure level ev: the illuminance of the inpainted chrome balls in output image I^B is scaled by 2^{ev} before being tone-mapped to an LDR image. Specifically, we condition the diffusion model on an exposure embedding, which is an interpolation of two text CLIP embeddings [Radford et al. 2021] as a function of ev:

$$\zeta^{ev} = \zeta^{\max} + \frac{ev}{ev^{\min}} \cdot (\zeta^{\min} - \zeta^{\max}) \tag{1}$$

where $ev \in [ev^{\min}, 0]$. $\zeta^{\max}$ is the embedding of the prompt "perfect mirrored reflective chrome ball spheres" while $\zeta^{\min}$ is the embedding of the prompt "perfect black dark mirrored reflective chrome ball spheres".

We train the model with the standard diffusion training loss on masked pixels:

$$\mathcal{L} = \mathbb{E}_{z_0,t,\epsilon,M^B,D^{I,B},ev} \left\| M^B \odot (\epsilon_\theta(z_\tau, t, M^B, D^{I,B}, ev) - \epsilon) \right\|_2^2 \tag{2}$$

where z_τ is the latent vector of inpainted image I^B at diffusion timestep τ. Note that instead of using LoRA as in DiffusionLight [Phongthawee et al. 2024], we fine-tune all weights of the denoising U-Net and the ControlNet, which we found empirically gives better results.

Dataset Curation. Training our lighting prediction model requires ground-truth lighting at different spatial locations. Therefore, we construct our training dataset using synthetic indoor scenes. Among existing datasets of synthetic 3D scenes [Fu et al. 2021; Li et al. 2018, 2022, 2021; Raistrick et al. 2024], we choose the recently released Infinigen Indoors [Raistrick et al. 2024]—a Blender-based procedural scene generator—to leverage its diverse lighting conditions of the generated scenes and realistic rendering provided by Blender.

First, we use Infinigen Indoors to randomly generate 500 indoor scenes, and sample 5 distinct viewpoints within each scene. Each viewpoint is randomly sampled so that the minimum depth from this viewpoint is greater than a certain threshold. This is to ensure that the sampled camera (viewpoint) is not blocked by nearby objects.

Then, for each selected viewpoint, we place $N \sim \mathcal{U}[1, 9]$ perfectly reflective chrome balls inside the view frustum while ensuring that these balls have no intersection with the scene geometry. After constructing the scenes, we render HDR images I^B with chrome balls inserted into the scenes, and also render chrome-ball masks M^B and depth maps $D^{I,B}$. At the same time, we hide the chrome balls and render the background images I that serve as paired input.

3.2 Distilling Spatiotemporal Light Field

The trained diffusion model can predict HDR lighting at multiple spatial locations from a single image. Our ultimate goal is to distill from this 2D diffusion model a spatiotemporally consistent light field $L_\psi(\boldsymbol{x}, t, \boldsymbol{d})$, where $\boldsymbol{x}$ is a spatial location, $t \in [1, T]$ is the frame index in time, and $\boldsymbol{d} = (\theta, \phi)$ represents an incoming light direction.

Here, we consider a video input $\{I_t\}_{t=1}^T$ (with T frames) that may capture spatially and temporally varying lighting. We represent the light field $L_\psi(\boldsymbol{x}, t, \boldsymbol{d})$ using an MLP, and denote $L_\psi(\boldsymbol{x}, t, \cdot)$ as the HDR environment map at location $\boldsymbol{x}$ and time t, one that is obtained by querying the MLP with all incoming light directions.

We view the trained diffusion network as a model that provides partial observations of the scene lighting. From this perspective, those partial observations can be used to supervise the training of

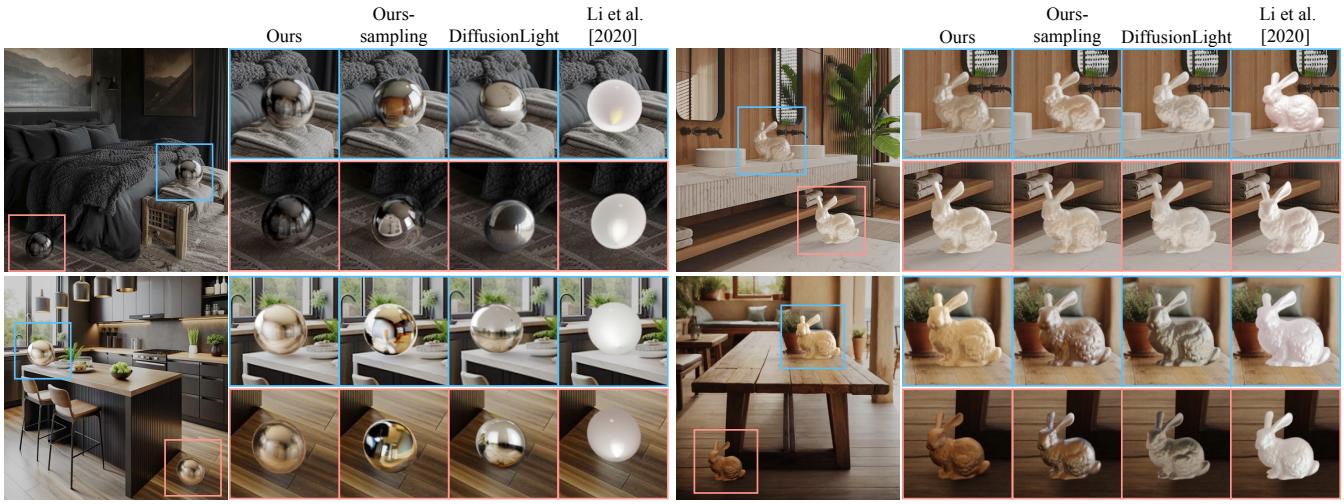

Fig. 3. Qualitative comparison of single image lighting prediction for object insertion on in-the-wild scenes. For each example, we insert the object at two different locations. We show the full image of our result on the left and zoom-in crops for each method on the right.

the light-field MLP. A similar view has been taken in training text-to-3D generation: a 3D scene can be distilled from a 2D diffusion model using score distillation sampling (SDS) [Poole et al. 2022].

In our case, we iteratively improve the MLP. In each iteration, we randomly pick a frame I_t and randomly sample N locations within the view frustum to insert chrome balls $B = \{(\boldsymbol{x}_i, r_i)\}_{i=1}^{N}$. With these chrome balls, we query the MLP to obtain the HDR environment maps at their locations $\{L_\psi(\boldsymbol{x}_i, t, \cdot)\}_{i=1}^{N}$. We then warp the environment maps onto the chrome balls and project them onto the selected image frame I_t. This process results in a LDR image I_t^B (which is tone mapped using a sampled exposure value ev). The goal of each training iteration is to improve the MLP so that the resulting image I_t^B better align with the image priors provided by the 2D diffusion model.

In particular, we encode each image in I_t^B into a latent vector and perturb it into a noisy latent z_τ with a noise level $\tau \sim \mathcal{U}[\tau_{\min}, \tau_{\max}]$. For each z_τ, we generate a sample using the latent diffusion model by running DDIM sampling [Song et al. 2020] for k denoising steps and denote the resulting clean latent sample as z_0. Finally, z_0 is decoded into an image $\hat{I}_t^B$, one that carries the image prior of our previously trained diffusion model. Therefore, we treat $\hat{I}_t^B$ as a pseudo ground truth for supervision:

$$\psi = \text{argmin}_\psi \, \mathbb{E}_{t,B,ev,\tau}\left[\|I_t^B - [\hat{I}_t^B]_{\not\nabla}\|_2^2 + \mathcal{L}_p(I_t^B, [\hat{I}_t^B]_{\not\nabla})\right], \quad (3)$$

where $\mathcal{L}_p$ is the perceptual distance LPIPS [Zhang et al. 2018], and $[\cdot]_{\not\nabla}$ is the stop gradient operator, which prevents $\hat{I}_t^B$ from being affected by the training process, as it is treated as a pseduo ground truth for providing image priors. This loss function and optimization strategy resemble the one used in several sparse-view 3D reconstruction methods [Tang et al. 2024; Wu et al. 2024; Zhou and Tulsiani 2023], and we empirically found it to work better than score distillation sampling [Poole et al. 2022].

In practice, we sample the chrome ball locations by sampling $N = 9$ pixel locations and then unprojecting each to 3D space with a randomly chosen depth smaller than the background scene depth at that pixel. During optimization, we linearly decrease ev from 0 to $ev^{\min}$ instead of randomly sampling it from $[ev^{\min}, 0]$. We empirically found that this linear sweep of ev value helps to synthesize overexposed regions more accurately.

In the case of a single image input, we drop the MLP's input time dimension (*i.e.*, using $T = 1$) and follow the above distillation procedure as is.

3.3 Implementation Details

We fine-tune the diffusion model for 15000 iterations with a batch size of 16 and a learning rate of 10^{-5}. To enable classifier-free guidance (CFG), we randomly dropout the exposure embedding with a probability of 0.1. During fine-tuning, we concatenate the latent background image, the noisy latent image, and the chrome ball mask as input to the diffusion model. This setup allows the model to leverage global context while inpainting the masked regions, maintaining high image quality even when multiple chrome balls occlude a large portion of the background. Our light field MLP shares a similar architecture as [Mildenhall et al. 2020] and has 6 hidden layers with a hidden dimension of 256 and positional encoding for the input. We also adopt the skip-connection design as [Mildenhall et al. 2020] and choose the third layer as the skip layer. We use a positional encoding frequency of 6 for the position $\boldsymbol{x}$, 4 for the time t, and 4 for the direction $\boldsymbol{d}$. During optimization, we fix $\tau_{\max} = 1.0$ for all steps, and linearly anneal $\tau_{\min}$ from 1.0 to 0.0. Given $\tau \sim \mathcal{U}[\tau_{\min}, \tau_{\max}]$, we sample the denoised image with $k = \lceil 10 \cdot \tau \rceil$ steps and a classifier-free guidance scale of 12.5. Within each optimization step, we firstly adjust the queried HDR environment map from MLP using the sampled exposure level, and then tone-map it into a LDR environment map using a fixed gamma of 2.4, following [Wang et al. 2022] and [Phongthawee et al. 2024]. The diffusion model training takes 14 hours on an NVIDIA A6000, and the distillation procedure typically takes 40 minutes for a video of 100 frames.

Table 1. Quantitative comparison of single image lighting prediction on synthetic datasets Infinigen Indoor (in-distribution) and 3D-FRONT (out-of-distribution), and and real-world Laval Indoor Spatially Varying (out-of-distribution, real-world).

Dataset	Method	Scale-invariant RMSE ↓			Angular Error ↓			Normalized RMSE ↓		
		Diffuse	Matte	Mirror	Diffuse	Matte	Mirror	Diffuse	Matte	Mirror
Infinigen Indoor [Raistrick et al. 2024]	DiffusionLight	0.50	0.52	0.58	6.21	6.39	6.62	0.47	0.46	0.46
	DiffusionLight-Distilled	0.60	0.61	0.65	5.85	5.96	6.11	0.57	0.49	0.48
	Ours-Sampling	0.48	0.49	0.56	5.63	5.86	6.23	0.46	0.42	0.36
	Li et al. [2020]	0.47	0.50	0.56	6.50	6.68	6.88	0.62	0.54	0.52
	Ours	0.41	0.43	0.48	3.72	4.35	4.55	0.45	0.44	0.45
3D-FRONT [Fu et al. 2021]	DiffusionLight	0.32	0.33	0.36	4.95	5.13	5.36	0.47	0.47	0.48
	DiffusionLight-Distilled	0.49	0.50	0.52	4.84	5.00	5.17	0.57	0.52	0.52
	Ours-Sampling	0.29	0.30	0.34	4.73	4.88	5.16	0.46	0.48	0.44
	Li et al. [2020]	0.30	0.32	0.35	5.73	5.89	6.07	0.62	0.51	0.43
	Ours	0.25	0.27	0.31	3.49	3.84	4.04	0.43	0.45	0.46
Laval Indoor-Spatially Varying [Garon et al. 2019]	DiffusionLight	0.40	0.42	0.43	7.51	7.86	8.03	0.53	0.60	0.65
	Ours-Sampling	0.34	0.34	0.39	7.73	7.75	7.96	0.56	0.58	0.65
	Li et al. [2020]	0.37	0.34	0.42	8.27	8.42	8.76	0.63	0.56	0.56
	Ours	0.26	0.28	0.31	5.45	5.57	6.02	0.48	0.49	0.53

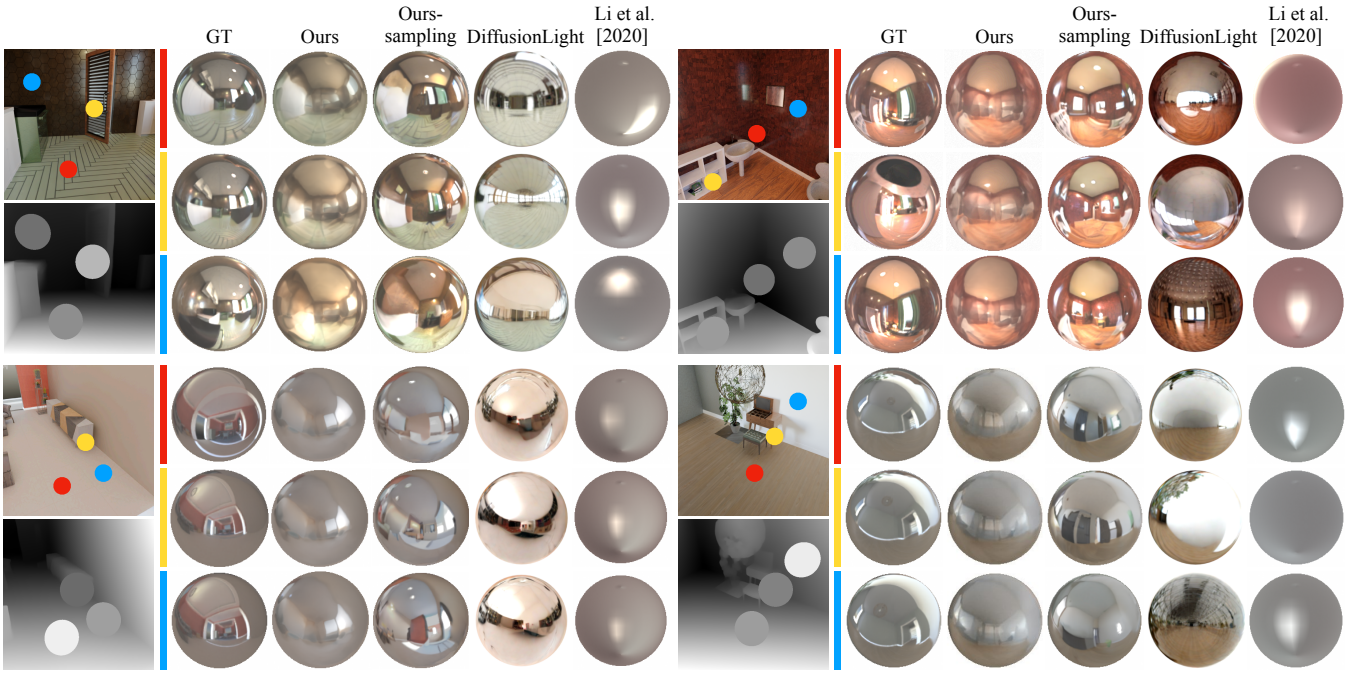

Fig. 4. Qualitative comparison of single image indoor lighting prediction on Infinigen Indoor (top) and 3D-FRONT (bottom). For each example, we show on the left the input image and corresponding depth map. Three different locations are marked on the image with color red, yellow and blue. On the right, we show estimated lighting (represented as chrome ball) from each method at the marked locations correspondingly.

4 EXPERIMENTS

In this section, we first evaluate our framework on spatially varying lighting estimation from a single image (Sec. 4.1), and then demonstrate results on spatiotemporally varying lighting estimation from a video (Sec. 4.2). Lastly, we present ablation studies on several design choices (Sec. 4.3). We strongly recommend the readers to view the videos in our supplementary material that best demonstrate the spatiotemporal lighting effects.

4.1 Single Image Indoor Lighting Estimation

Evaluation Datasets. For quantitative comparison, we test our method on both synthetic and real-world datasets. For the synthetic datasets, we use both in-distribution and out-of-distribution ones, *i.e.*, Infinigen Indoors [Raistrick et al. 2024] and 3D-FRONT[Fu et al. 2021]. From each dataset, we sample 25 distinct scenes and for each scene render an input image from a randomly sampled viewpoint. Within the frustum of the input image, we sample 10 different

Table 2. Ablation study. 'w/o MLP' uses a $3 \times 3 \times 3$ discrete grid with trilinear interpolation between them as the lighting representation for optimization. 'w/o multiple probes' trains the diffusion model to inpaint only a single sphere probe. 'w/o decreasing exposure' uniformly samples the exposure value during optimization instead of gradually decreasing it.

Dataset	Method	Scale-invariant RMSE ↓			Angular Error ↓			Normalized RMSE ↓		
		Diffuse	Matte	Mirror	Diffuse	Matte	Mirror	Diffuse	Matte	Mirror
Infinigen Indoor [Raistrick et al. 2024]	w/o MLP	0.47	0.49	0.55	8.47	8.63	8.99	0.39	0.40	0.44
	w/o multiple probes	0.45	0.47	0.53	4.74	4.91	5.12	0.49	0.52	0.55
	w/o decreasing exposure	0.45	0.46	0.53	4.69	4.87	5.11	0.45	0.44	0.45
	Ours	0.45	0.47	0.53	4.54	4.71	4.95	0.47	0.46	0.47
3D-FRONT [Fu et al. 2021]	w/o MLP	0.36	0.38	0.40	8.04	8.18	8.32	0.48	0.45	0.47
	w/o multiple probes	0.31	0.33	0.35	3.59	3.67	3.83	0.53	0.57	0.61
	w/o decreasing exposure	0.31	0.33	0.36	3.49	3.62	3.79	0.49	0.52	0.51
	Ours	0.30	0.32	0.35	3.49	3.61	3.79	0.46	0.49	0.49

locations and obtain the ground truth lighting by rendering a HDR environment map from a panorama camera placed at each location. For real-world testing, we note that the InteriorNet [Li et al. 2018] testing set used in prior work [Srinivasan et al. 2020; Wang et al. 2024] has become inaccessible. As an alternative, we evaluate our method on the Laval Indoor Spatially Varying HDR Dataset [Garon et al. 2019], which provides four spatially varying ground-truth lighting conditions per scene, captured using DSLR cameras with exposure bracketing, across 20 different testing scenes.

Evaluation Metrics. We follow the three-sphere evaluation protocol used in previous work [Gardner et al. 2019, 2017; Phongthawee et al. 2024; Wang et al. 2022]. Specifically, we use the estimated HDR environment map of size 128×256 pixels to render three spheres with different materials (gray-diffuse, silver-matte and silver-mirror). To compare the rendered sphere images to those under the ground truth HDR environment map, we follow the literature and adopt three scale-invariant metrics: scale-invariant Root Mean Square Error (si-RMSE) [Grosse et al. 2009], Angular Error [LeGendre et al. 2019] and normalized RMSE (mapping the 0.1st and 99.9th percentiles to 0 and 1 [Marnerides et al. 2018]). We report the numbers averaged over all samples locations of all scenes.

Compared Baselines. Since most of recent methods for predicting spatially varying lighting from a single image do not have open-source code [Li et al. 2023; Liang et al. 2025; Wang et al. 2024, 2021], we compare against the only one we found available [Li et al. 2020]. Li et al. [2020] trained a network to regress an environment map of size 16×32 for each pixel, *i.e.*, lighting at each visible surface point of the scene. Yet, in our evaluation dataset, the evaluated location could be arbitrary in the 3D view frustum. Therefore, we take the predicted lighting at the nearest surface point as their prediction. In addition, we compare our method with three baselines: 1) direct sampling with DiffusionLight [Phongthawee et al. 2024] 2) optimization with our distillation with DiffusionLight and 3) direct sampling with our diffusion model (*Ours-sampling*), *i.e.*, independently generating environment maps at each evaluated spatial location. For the two direct sampling baselines, we follow the HDR merging algorithm in DiffusionLight [Phongthawee et al. 2024] to obtain HDR environment maps.

Results. We present quantitative comparison results in Table 1 and qualitative comparison results in Fig. 4 and Fig. 5. Our method outperforms compared baselines on almost all evaluation metrics. While the estimated lighting from *Ours-sampling* and DiffusionLight have sharper texture, they are not consistent across different spatial locations due to independent sampling. The distilled version of DiffusionLight produces spatially smooth lighting estimates but lacks fine-grained detail. This limitation stems from the fact that DiffusionLight samples only a single chrome ball during each optimization step, and its base model is fine-tuned specifically to inpaint chrome balls located at the image center. As a result, it yields increasingly distorted predictions toward the image periphery.

Li et al. [2020] does predict spatially varying lighting in a consistent manner, yet their estimated environment map has a limited resolution of 16×32 pixels and thus is overly blurred. In contrast, our method estimates the spatially varying lighting with sufficient details. Additionally, we show the comparison results for in-the-wild scenes in Fig. 3 and refer the readers to the videos in the supplementary material that best demonstrate the spatiotemporally varying lighting effects.

In Fig. 5, we show qualitative comparison to LightOctree [Wang et al. 2024] on real world scenes [Garon et al. 2019]. Our method achieves comparable results for virtual object insertion.

4.2 Dynamic Lighting Estimation from a Single Video

Our proposed framework enables lighting estimation from a single video of dynamic lighting conditions, which has not been demonstrated in previous work. As it is not easy to build a synthetic evaluation dataset that reflects realistic lighting changes, we instead choose to present qualitative results on diverse in-the-wild videos in Fig. 6 (and in supplementary material). We include a baseline where we predict lighting by sampling our diffusion model independently for each frame (*Ours-sampling*). In comparison, our full pipeline can estimate the scene lighting in a spatiotemporally consistent manner. We also show object insertion results under dynamic lighting in Fig. 1 and Fig. 7. The visual results are best appreciated in the supplementary videos.

Fig. 5. Qualitative evaluation on virtual object insertion on Laval Indoor Spatially Varying HDR dataset [Garon et al. 2019]. Results of LightOctree [Wang et al. 2024] are from their paper.

4.3 Ablation Studies

We perform ablation studies by comparing our full pipeline with several variants: 1) *w/o decreasing exposure*, where we uniformly sample the exposure value during optimization instead of gradually decreasing it; 2) *w/o multiple probes*, where our diffusion model is trained to inpaint a single sphere probe instead of jointly inpainting multiple probes; 3) *w/o MLP*, where instead of a MLP we use a 3×3×3 discrete grid with trilinear interpolation between them as the lighting representation for optimization. Table 2 shows the comparison results on a subset of our evaluation dataset for single image indoor lighting estimation. Our full method surpasses these ablated versions, therefore validating our design choices. We further compare our diffusion model against DiffusionLight to evaluate the performance of the respective base models. Specifically, we test on the Laval Indoor HDR dataset [Gardner et al. 2017], which differs from the Laval Indoor Spatially Varying dataset in that it provides only a single center lighting ground truth per scene. Following the same one-time central sampling setup used in DiffusionLight, our model achieves si-RMSE values of 0.41/0.43/0.49 for diffuse/matte/mirror spheres, angular errors of 3.61/3.83/4.42, and normalized RMSE values of 0.36/0.34/0.33. These results are comparable to those reported in DiffusionLight (Table 1, line 5, SDXL+LoRA), even though our model is not explicitly trained for this setting, where lighting is captured from the camera viewpoint using a chrome sphere.

5 DISCUSSION AND LIMITATIONS

We present a method that can estimate spatiotemporally consistent lighting from a video of an indoor scene where the illumination varies spatially and temporally. We highlight our lighting estimation results on challenging real-world videos in the wild, which is rarely demonstrated in the literature. Our method can also estimate spatially varying lighting from a single image, and we show competitive results compared to prior work.

While our diffusion model generalizes well to in-the-wild indoor scenes thanks to the pre-trained weights, it struggles on outdoor scenes where the sunlight dominates. Including synthetic outdoor scenes [Raistrick et al. 2023] or captured real-world panoramas [Hold-Geoffroy et al. 2019] in the training data may help to address this issue. Our optimization also encounters over-smoothing issues in both the appearance of the reflective spheres and the temporal evolution of lighting, a challenge common to many distillation-based methods [Poole et al. 2022]. To alleviate this, we experimented with increasing the degree of positional encoding for time t from 4 to 6 and further to 8. While increasing to 6 showed some reduction in over-smoothing, raising the degree to 8 offered only marginal additional improvement and introduced subtle flickering effects temporally. These results suggest diminishing returns beyond a certain encoding complexity. A promising direction for future work is to explore volumetric lighting representations [Srinivasan et al. 2020; Wang et al. 2024], as volumetric rendering can propagate loss gradients along entire rays, potentially preserving more spatial detail. Lastly, our method also shares the limitation with many other lighting estimation methods when used for virtual object insertion. The appearance of rendered object may not seamlessly blend with the image since the graphics renderer is only an approximation of real world light transport. Adding supervision on the final composed image [Liang et al. 2025] could potentially alleviate this problem.

REFERENCES

Jonathan T Barron and Jitendra Malik. 2014. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence* 37, 8 (2014), 1670–1687.

Blender. 2024. *Blender - a 3D modelling and rendering package.* Blender Foundation. http://www.blender.org

Mohammad Reza Karimi Dastjerdi, Jonathan Eisenmann, Yannick Hold-Geoffroy, and Jean-François Lalonde. 2023. EverLight: Indoor-outdoor editable HDR lighting estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7420–7429.

Paul Debevec. 2008. Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *Acm siggraph 2008 classes*. 1–10.

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 2021. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10933–10942.

Per Gantelius. 2024. *fSpy - an open source, cross platform app for still image camera matching.* https://fspy.io/

Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. 2024. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314* (2024).

Marc-André Gardner, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Christian Gagné, and Jean-François Lalonde. 2019. Deep parametric indoor lighting estimation. In

Proceedings of the IEEE/CVF International Conference on Computer Vision. 7175–7183.
Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. 2017. Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090* (2017).
Mathieu Garon, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, and Jean-François Lalonde. 2019. Fast spatially-varying indoor lighting estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6908–6917.
Roger Grosse, Micah K. Johnson, Edward H. Adelson, and William T. Freeman. 2009. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *2009 IEEE 12th International Conference on Computer Vision*. 2335–2342. https://doi.org/10.1109/ICCV.2009.5459428
Yannick Hold-Geoffroy, Akshaya Athawale, and Jean-François Lalonde. 2019. Deep sky modeling for single image outdoor lighting estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6927–6935.
Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Blackburn-Matzen, Matthew Sticha, and David F Fouhey. 2023. Perspective fields for single image camera calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17307–17316.
Chloe LeGendre, Wan-Chun Ma, Graham Fyffe, John Flynn, Laurent Charbonnel, Jay Busch, and Paul Debevec. 2019. Deeplight: Learning illumination for unconstrained mobile mixed reality. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5918–5928.
Wenbin Li, Sajad Saeedi, John McCormac, Ronald Clark, Dimos Tzoumanikas, Qing Ye, Yuzhong Huang, Rui Tang, and Stefan Leutenegger. 2018. Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset. *arXiv preprint arXiv:1809.00716* (2018).
Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. 2020. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2475–2484.
Zhen Li, Lingli Wang, Xiang Huang, Cihui Pan, and Jiaqi Yang. 2022. Phyir: Physics-based inverse rendering for panoramic indoor images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12713–12723.
Zhengqin Li, Li Yu, Mikhail Okunev, Manmohan Chandraker, and Zhao Dong. 2023. Spatiotemporally consistent hdr indoor lighting estimation. *ACM Transactions on Graphics* 42, 3 (2023), 1–15.
Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Meng Song, Yuhan Liu, Yu-Ying Yeh, Rui Zhu, Nitesh Gundavarapu, Jia Shi, et al. 2021. Openrooms: An open framework for photorealistic indoor scene datasets. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7190–7199.
Ruofan Liang, Zan Gojcic, Merlin Nimier-David, David Acuna, Nandita Vijaykumar, Sanja Fidler, and Zian Wang. 2025. Photorealistic object insertion with diffusion-guided inverse rendering. In *European Conference on Computer Vision*. Springer, 446–465.
Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. 2023. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 300–309.
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 *Proceedings of the IEEE/CVF international conference on computer vision*. 9298–9309.
Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista. 2018. Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. In *Computer Graphics Forum*, Vol. 37. Wiley Online Library, 37–49.
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. In *European Conference on Computer Vision*. Springer, 405–421.
Jeong Joon Park, Aleksander Holynski, and Steven M Seitz. 2020. Seeing the world in a bag of chips. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1417–1427.
Pakkapon Phongthawee, Worameth Chinchuthakun, Nontaphat Sinsunthithet, Varun Jampani, Amit Raj, Pramook Khungurn, and Supasorn Suwajanakorn. 2024. Diffusionlight: Light probes for free by painting a chrome ball. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 98–108.
Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988* (2022).
Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, et al. 2023. Infinite photorealistic worlds using procedural generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12630–12641.
Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, Zeyu Ma, and Jia Deng. 2024. Infinigen Indoors: Photorealistic Indoor Scenes using Procedural Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 21783–21794.
René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. 2021. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*. 12179–12188.
Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.
Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. 2023. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512* (2023).
Gowri Somanath and Daniel Kurz. 2021. HDR environment map estimation for real-time augmented reality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11298–11306.
Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).
Pratul P Srinivasan, Ben Mildenhall, Matthew Tancik, Jonathan T Barron, Richard Tucker, and Noah Snavely. 2020. Lighthouse: Predicting lighting volumes for spatially-coherent illumination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8080–8089.
Jiapeng Tang, Davide Davoli, Tobias Kirschstein, Liam Schoneveld, and Matthias Niessner. 2024. GAF: Gaussian Avatar Reconstruction from Monocular Videos via Multi-view Diffusion. *arXiv preprint arXiv:2412.10209* (2024).
Dor Verbin, Ben Mildenhall, Peter Hedman, Jonathan T Barron, Todd Zickler, and Pratul P Srinivasan. 2024. Eclipse: Disambiguating illumination and materials using unintended shadows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 77–86.
Guangcong Wang, Yinuo Yang, Chen Change Loy, and Ziwei Liu. 2022. Stylelight: Hdr panorama generation for lighting estimation and editing. In *European Conference on Computer Vision*. Springer, 477–492.
Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. 2023. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12619–12629.
Xuecan Wang, Shibang Xiao, and Xiaohui Liang. 2024. LightOctree: Lightweight 3D Spatially-Coherent Indoor Lighting Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4536–4545.
Zian Wang, Jonah Philion, Sanja Fidler, and Jan Kautz. 2021. Learning indoor inverse rendering with 3d spatially-varying lighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12538–12547.
Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. 2024. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21551–21561.
Renjiao Yi, Chenyang Zhu, Ping Tan, and Stephen Lin. 2018. Faces as lighting probes via unsupervised deep highlight extraction. In *Proceedings of the European Conference on computer vision (ECCV)*. 317–333.
Hong-Xing Yu, Samir Agarwala, Charles Herrmann, Richard Szeliski, Noah Snavely, Jiajun Wu, and Deqing Sun. 2023. Accidental light probes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12521–12530.
Ye Yu and William AP Smith. 2019. Inverserendernet: Learning single image inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3155–3164.
Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.
Zhizhuo Zhou and Shubham Tulsiani. 2023. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12588–12597.
Rui Zhu, Zhengqin Li, Janarbek Matai, Fatih Porikli, and Manmohan Chandraker. 2022. Irisformer: Dense vision transformers for single-image inverse rendering in indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2822–2831.

Fig. 6. Qualitative results of dynamic lighting estimation from single video on *in-the-wild* scenes. For each example video, we show 4 frames with results from our full pipeline (*Ours*) and our diffusion model samples (*Ours-sampling*). For each frame, we show estimated lighting at two different locations, depicted as chrome balls on the image with corresponding environment maps below. Depth map of the scene is also shown on top left of the first frame. The estimated lighting from our full pipeline has better temporal consistency (see the red box crops). Please refer to the supplement for video results of more examples.

Fig. 7. More examples of object insertion under dynamic lighting. Here we show four frames of each input video. Here we show four frames of an video. For each frame, we show **(top)** the result of virtual object insertion under the estimated lighting and **(bottom)** estimated environment maps at locations where the two mirror spheres are inserted. In each column, the environment map on the bottom-left corresponds to the top mirror sphere, and the one on the bottom-right corresponds to the bottom mirror sphere.