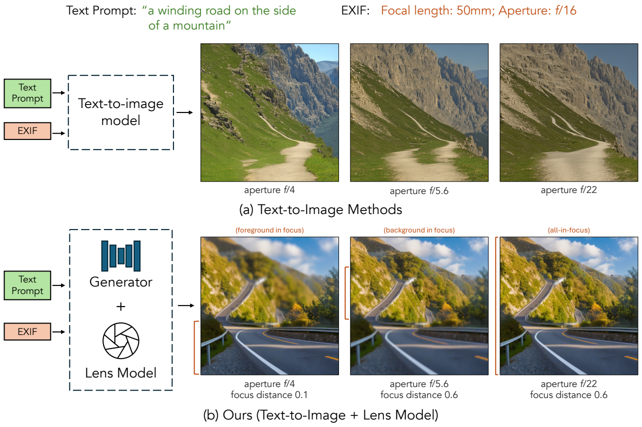
Fine-grained Defocus Blur Control for Generative Image Models

Ayush Shrivastava[1,2*] Connelly Barnes[2] Xuaner Zhang[2] Lingzhi Zhang[2]
Andrew Owens[1] Sohrab Amirghodsi[2] Eli Shechtman[2]
[1]University of Michigan [2]Adobe Research

https://www.ayshrv.com/defocus-blur-gen

Abstract

Current text-to-image diffusion models excel at generating diverse, high-quality images, yet they struggle to incorporate fine-grained camera metadata such as precise aperture settings. In this work, we introduce a novel text-to-image diffusion framework that leverages camera metadata, or EXIF data, which is often embedded in image files, with an emphasis on generating controllable lens blur. Our method mimics the physical image formation process by first generating an all-in-focus image, estimating its monocular depth, predicting a plausible focus distance with a novel focus distance transformer, and then forming a defocused image with an existing differentiable lens blur model [32]. Gradients flow backwards through this whole process, allowing us to learn without explicit supervision to generate defocus effects based on content elements and the provided EXIF data. At inference time, this enables precise interactive user control over defocus effects while preserving scene contents, which is not achievable with existing diffusion models. Experimental results demonstrate that our model enables superior fine-grained control without altering the depicted scene.



(a) Text-to-Image Methods

(b) Ours (Text-to-Image + Lens Model)

Figure 1. **Overview.** We propose a text-to-image model that precisely controls the amount and location of defocus blur in generated images while preserving the scene content. (a) Previous methods like Fang et al. [12] (shown here), struggle to preserve scene content when modifying aperture settings. (b) Our model generates images with varying defocus blur while keeping the scene intact. Each image exhibits a distinct blur effect based on the specified aperture and focus distance. At lower focus distances, nearby objects appear sharper (foreground in focus), whereas at higher focus distances, the focus shifts to distant objects (background in focus). The orange bar [on the left of each image indicates the in-focus region. Higher f-stops result in reduced blur, producing an all-in-focus image for $f/22$.

1. Introduction

Recent advances in text-to-image generative models have shown impressive capabilities in producing realistic, high-quality images [25]. However, in a real photo shoot, a photographer using a DSLR or mirrorless camera might adjust defocus effects (modifying the focal plane and aperture) to direct attention to key subjects and de-emphasize background details. Current generative models lack the ability to make such fine-grained adjustments while preserving scene integrity, as would occur in an actual shoot.

Many real-world images encode details such as focal distance, aperture, lens type, and camera model in EXIF (EXchangeable Image File Format) metadata. A generative model that can *conditionally* and *controllably* adapt to these EXIF parameters would be highly valuable. For example, given a prompt and an aperture value of $f/1.8$, the model should generate an image that accurately reflects both the prompt and the specified aperture. If the user wishes to reduce blur, they could specify a smaller aperture (larger f-stop), prompting the model to generate the same scene with less blur. This task is inherently challenging, as the model must determine the appropriate amount of blur to add or remove, identify which regions should be blurred by recognizing salient areas, and apply blur selectively to create an aesthetically pleasing image while preserving scene content. Achieving a balance between content preservation and

[*] Work done during an internship at Adobe.

depth-aware spatial blur control makes this a complex but crucial problem in generative image modeling.

One of the major challenges is disentangling the image content from the camera parameters. For example, recent work [12] has attempted to address this by training camera parameter embeddings for text-to-image models to guide generation conditioned on EXIF tags. However, that approach struggles to maintain scene consistency when adding or reducing defocus blur across different camera settings. Likewise, camera properties and scene content are closely intertwined in off-the-shelf text-to-image models. As a result, applying camera effects as a postprocessing step is ineffective, since the generated images already contain other camera properties (e.g., existing defocus blur).

We propose a method that decouples scene generation from lens properties, enabling control over EXIF-based image generation while preserving the physicality of image formation. Our model produces plausible focal distances and grants users fine-grained control over lens parameters while maintaining scene content. An example of this control is shown in Figure 1.

Our approach is based on training a few-step generative model conditioned on EXIF information (specifically aperture). This model obtains a supervision signal from a differentiable lens model. We train a fast few-step generative model conditioned on EXIF information to generate deep (or shallow) depth-of-field (DoF) images using weak supervision from simply extracting high-quality deep (and shallow) DoF subsets from the dataset. In particular, the shallow DoF images are generated by our novel focus distance transformer, which predicts a focus distance and scale used by a differentiable lens blur model based on Wang et al. [32]. Because focus distances are typically omitted or unreliable in EXIF tags, we show that a weakly supervised approach can be used to learn these directly from shallow DoF images. By explicitly modeling defocus blur within the lens, our approach enables end-to-end learning of plausible focal distances while manipulating lens properties separately from scene content. We base our generator on improved Distribution Matching Distillation (DMD2) [38] since it can simultaneously learn a fast few-step model to enable user interactivity, and its unpaired DMD and GAN losses are suited for our weak supervision. While these components have been studied in isolation, our key contribution lies in integrating them into a unified, unsupervised training framework that enables precise and intuitive user control over the defocus blur effect — empowering interactive generation of images with controllable DoF.

Our main contribution is a generative framework that enables learning lens properties and allowing precise interactive user control over such lens properties while preserving scene content. Our experiments show clear improvements, enabling such fine-grained user interactivity. Our technical contributions are: (1) A unified generative pipeline that learns without explicit supervision to model depth of field, which is done with the help of (2) A novel focus distance transformer that learns to predict plausible focus distances and scales, which are used in a physically-inspired thin lens blur model [32]. (3) We show that weak supervision from shallow and deep depth-of-field images can successfully supervise a diffusion model with a lens model, and that this form of supervision can easily be obtained from unlabeled image datasets.

2. Related Work

Depth of field rendering. Finite apertures produce defocus effects long studied for light-field rendering [2] and adaptive ray-tracing sampling [4, 5, 47], with recent differentiable depth-of-field-aware rendering [24]. AR-GAN [14] estimates a depth map and all-in-focus RGB image in an unsupervised manner and integrates over a finite aperture to render defocus; unlike it, we adopt a diffusion framework for text-to-image synthesis with learned modules such as the focal-distance model. AR-NeRF [15] uses a similar unsupervised approach based on NeRFs [23], while DOF-GS [33] extends 3D Gaussian splatting [17] with a thin-lens circle-of-confusion for 3D reconstruction and defocus. Wang et al. [32] reconstruct an all-in-focus HDR radiance map and depth map from time-aperture-focus stacks, whose differentiable thin lens model we use, while Xin et al. [36] recover all-in-focus images and defocus maps from dual-pixel images via multiplane optimization. Dr.Bokeh [30] uses a layered scene representation with a differentiable occlusion-aware bokeh model, which we adopt as a plug-and-play lens component during inference. DC^2 [3] enables post-capture defocus control by fusing dual-camera inputs with different fixed apertures but requires multi-camera hardware. DiffCamera [34] performs arbitrary post-capture refocusing using a diffusion transformer trained on simulated multi-focus data, whereas our method learns defocus from in-the-wild data and provides fine-grained control over focus and aperture.

Diffusion models. We use SDXL [25], a text-to-image latent diffusion model [26] as our generator architecture. We estimate depth with frozen diffusion-based monocular metric depth estimation model Metric3Dv2 [11]. Voynov et al. [31] uses a per-pixel coordinate conditioning method to generate diffusion images using different optical systems, such as fisheye and concave lenses, and spherical panoramas. Diffusion-based image restoration methods such as SUPIR [40] have shown strong performance on camera lens-related restoration tasks like recovering from mixed blur, super-resolution, and/or noise degradations.

Learning with metadata. Several models leverage metadata for conditioning. DiffusionSat [18] generates satellite

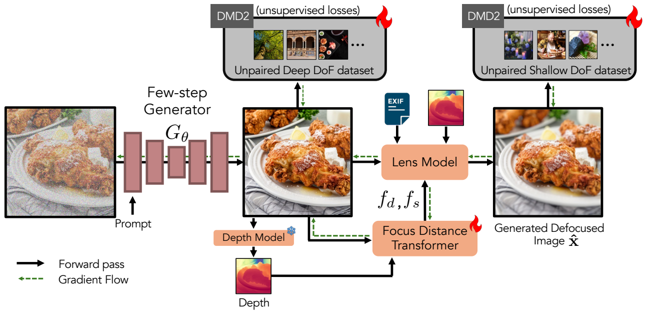


Figure 2. **Model Architecture.** We propose an image generation method that enables precise control over blur intensity and location within an image. This model obtains its supervision from a differentiable lens model and training examples of images with shallow and deep depth-of-field. We train our model to generate an *all-in-focus* image using G_θ. A depth model then predicts depth for this image, which, along with the image itself, is fed into a model that estimates the focus distance and depth scale. Finally, a lens model combines EXIF data with these predictions to apply spatially varying blur, generating the final image. We train the all-in-focus generator using unsupervised DMD2 [38] losses on our unpaired Deep DoF dataset and optimize the entire pipeline with DMD2 losses on the unpaired Shallow DoF dataset.

imagery conditioned on metadata such as GPS, date, cloud cover, some of which also appear in EXIF metadata for cameras. EXIF as Language [45] uses contrastive learning to connect image patches with EXIF data for spliced image detection. Camera Settings as Tokens [12] trains a LoRA [10] adapter and camera parameter embeddings to condition generation on focal length, aperture, ISO, and exposure. but struggles to preserve scene content and consistently control blur. In contrast, our method decouples scene generation from defocus blur for precise control. Generative Photography [41], learns camera embeddings using paired videos with varying EXIF settings, whereas our method is entirely unsupervised, leveraging only unpaired deep and shallow depth-of-field data. A concurrent work, Bokeh Diffusion [8], conditions a diffusion model on a defocus parameter for scene-consistent bokeh control but relies on synthetic blur for supervision.

Few-step diffusion model distillation. Teacher models can be distilled to few-step students via progressive distillation [21, 27], GAN-based approaches (ADD [29], LADD [28], Diffusion2GAN [16]), and distribution-matching methods DMD [39] and DMD2 [38]. We adopt DMD2 pipeline for efficiency and—crucially—to match the shallow and deep DoF data distributions (which are unpaired) within a single pipeline.

3. Approach

Our objective is to develop a text-to-image model conditioned on EXIF metadata, enabling precise, fine-grained control over the degree and location of image defocus blur. To achieve this, we decouple image generation from the blurring process by introducing a lens model that applies controlled blur based on depth information. Our pipeline is shown in Figure 2 and begins with a fast, few-step generator that produces an all-in-focus image. This is followed by a focal distance model that identifies regions of interest and guides those areas to stay in focus. Finally, the lens model applies a spatially-varying blur according to a differentiable physically-inspired thin lens model. The entire pipeline is differentiable and learned end-to-end.

3.1. Generator

Our generator builds on the few-step generator architecture from DMD2 [38], which distills a Stable Diffusion-XL (SDXL) teacher model into a fast few-step generator by aligning the generator's output distribution with that of the teacher model. The process involves two key components:

Distribution Matching Distillation (DMD). This component distills the teacher diffusion model (SDXL) into a few-step generator G by minimizing the Kullback-Liebler (KL) divergence between the diffused target distribution $p_{\text{real},t}$ and the diffused generator output distribution $p_{\text{fake},t}$ for a timestep t. The loss is derived from the difference between two score functions and is computed as:

$$\begin{aligned}\nabla\mathcal{L}_{\text{DMD}} &= \mathbb{E}_t\left(\nabla_\theta \text{KL}(p_{\text{fake},t} \| p_{\text{real},t})\right)\\ &= \mathbb{E}_t\left(\int \left(s_{\text{fake}}(x_t,t) - s_{\text{real}}(x_t,t)\right)\frac{dG_\theta(z)}{d\theta}\, dz\right),\end{aligned} \tag{1}$$

where s_{real} and s_{fake} are the score functions approximated using diffusion models μ_{real} and μ_{fake}. Here, $t \sim \mathcal{U}[0,T]$, $x_t = F(G_\theta(z),t)$ where F is the forward diffusion process (adding noise), $z \sim \mathcal{N}(0,\mathbf{I})$ is random Gaussian noise, θ denotes the generator parameters. μ_{real} is the pre-trained frozen diffusion model (SDXL) used as the teacher and μ_{fake} is a dynamically trained diffusion model that is optimized alongside G using the denoising score-matching loss ($\mathcal{L}_{\text{denoise}}$), conditioned on the output of G.

GAN Loss. A discriminator D is trained to distinguish between real images and those generated by G. A classification branch is added to the bottleneck features of the fake diffusion denoiser μ_{fake}, with the loss given by:

$$\begin{aligned}\mathcal{L}_{\text{GAN}} &= \mathbb{E}_{x\sim p_{\text{real}}}[\log D(F(x,t))]\\ &\quad + \mathbb{E}_{z\sim p_{\text{noise}}}[-\log(D(F(G_\theta(z),t)))],\end{aligned} \tag{2}$$

where $t \sim \mathcal{U}[0,T]$ and D represents the discriminator.

Training. The generator G and μ_{fake} are initialized with a pre-trained diffusion model. During training, G minimizes $\mathcal{L}_{\text{DMD}} + \mathcal{L}_{\text{GAN}}$, while μ_{fake} minimizes $\mathcal{L}_{\text{denoise}} + \mathcal{L}_{\text{GAN}}$. This setup can be summarized as DMD2 (G, μ_{fake}) where G is a few-step generator conditioned on a text prompt t and generates an all-in-focus image $\mathbf{x}$ from a noise sample z:

$$\mathbf{x} = G_\theta(z, t) \tag{3}$$

3.2. Focus Distance Transformer

To achieve aesthetically pleasing blur without rendering images entirely out of focus, the model must determine the salient regions in the image and decide which areas should remain sharp. In photographic terms, this requires selecting a depth for the focal plane to ensure key objects appear in focus. To this end, we develop a focus distance prediction model that outputs the focus distance (in depth units) at which the most important objects are located. We first extract a monocular depth map $\mathbf{d}$ from the generated all-in-focus image using a frozen Metric3Dv2 [11] model. Next, the focus distance transformer takes as input the generated all-in-focus image and depth map, and produces the focus distance f_d. Additionally, it produces a scale factor f_s, which allows the generative model to align the metric depth and focal length and still generate plausible results when either is inaccurate. Our model fine-tunes the Visual Saliency Transformer (VST) [20], which we adapt for this specific task.

Focus Distance. To predict the focus distance, we compute a saliency map from our VST network decoder and take its weighted average with the depth map, yielding a focus distance prediction within the range of the depth map.

$$f_d = \|\mathbf{d} \odot \text{VST}(\mathbf{x}, \mathbf{d})\|_1 / \|\text{VST}(\mathbf{x}, \mathbf{d})\|_1 \tag{4}$$

To supervise learning the focal distance, we use a frozen pre-trained copy of the VST network to calculate a reference focus distance from this weighted average. We apply a weak supervision for f_d with a Huber loss between the reference focus distance from the pretrained VST network.

Focus Distance Scale. For the focus distance scale, we extract the saliency token VST_{SAL} from the first layer of the VST decoder and use a linear head to predict its value. This is learned through end-to-end training with DMD2 losses.

$$f_s = \text{MLP}(\text{VST}(\mathbf{x}, \mathbf{d})_{\text{SAL}}) \tag{5}$$

3.3. Lens Model

To achieve precise control over defocus blur within the image generation pipeline, we require a differentiable lens blur model that can be trained jointly with the generator, allowing gradients to flow back from the defocused image. For this, we use the thin lens model from [32], which produces defocus blur in the image through differentiable kernels. The lens model is parameterized by focal length f, focus distance f_d and aperture N. We extend this model to include the focus distance scale f_s mentioned above. The circle of confusion (CoC) disk diameter is computed as:

$$\textbf{coc} = \frac{|\mathbf{d} - f_d|}{\mathbf{d}} \frac{f^2}{N(f_s \cdot f_d - f)} \tag{6}$$

After computing the **coc** for each pixel, we simulate defocus blur using a spatially-varying convolution W as $\hat{\mathbf{x}} = W * \mathbf{x}$ as implemented in [32]. The kernel W is a unit-energy disk with **coc** as its diameter and a differentiable soft boundary based on pixel distance from the kernel center. We denote the combination of the lens model and focus distance model as $\hat{G}$ such that $\hat{G}(\mathbf{x}) = \hat{\mathbf{x}}$. For convenience, we refer to this lens model as the *TAF Lens* [32].

Swapping the Lens Model at Inference. Our framework allows swapping out the lens model at inference time. For example, we use Dr.Bokeh [30] as an alternative lens model to demonstrate results with our generator during inference. The TAF lens model is fast and therefore suitable for training, whereas the Dr.Bokeh lens model is slower but produces higher-quality results due to its layered representation and inpainting.

3.4. Deep and Shallow Depth-of-Field Datasets

We desire a generator G that produces all-in-focus, or deep depth-of-field (DoF) images $\mathbf{x}$, and a lens model that generates shallow DoF images $\hat{\mathbf{x}}$. We train these models using only a weak form of supervision. We construct datasets of deep and shallow DoF images from a large pool of uncurated images. We provide details about dataset curation in Supplemental Section D.

3.5. Putting Everything Together

In summary, our generator G produces deep depth-of-field (DoF) images, $\mathbf{x}$, while the lens model $\hat{G}$ generates shallow DoF images, $\hat{\mathbf{x}}$. We aim for $\mathbf{x}$ to have a deep DoF and $\hat{\mathbf{x}}$ to have a shallow DoF. To achieve this, we train G on the Deep DoF dataset using DMD2 (G, μ_{fake}) and train $\hat{G}$ on the Shallow DoF dataset using DMD2 ($\hat{G}$, μ_{fake}). Our overall loss is:

$$\lambda_1 \text{DMD2}(G, \mu_{\text{fake}}) + \lambda_2 \text{DMD2}(\hat{G}, \mu_{\text{fake}}) + \lambda_3 L_{\text{Huber}} \tag{7}$$

4. Experiments

We use a commercially available stock-photography dataset for training. We extract 1.5M samples for the deep DoF and shallow DoF datasets. We pretrain the generator G on the deep DoF dataset. Then we jointly fine-tune G on the deep DoF dataset and $\hat{G}$ on the shallow DoF dataset.

4.1. Evaluation

To assess the efficacy of our model, we evaluate two key properties: (1) how the blur amount changes in response to EXIF information, particularly aperture value, and (2)

Table 1. **Results.** We compare our method to teacher and distilled-SDXL models, variants of these with lens models, and other baselines.

	#	Method	Lens Model	EXIF Conditioning	Blur ↑ Monotonicity	Content ↑ Consistency	LPIPS ↓	$\mathbf{FID_{DDoF}}$ ↓	$\mathbf{FID_{SDoF}}$ ↓
Gen.	1	SDXL [25]	-	EXIF as text	48.47	82.91	0.1398	17.88	18.17
	2	SDXL [25]	-	DoF as text	53.80	81.93	0.0563	17.87	18.17
	3	4-step SDXL (Distilled)	-	DoF as text	52.85	79.67	0.0829	14.19	18.06
	4	4-step SDXL (Distilled)	-	EXIF embedding	53.76	88.30	0.0344	15.38	16.92
	5	Camera Settings as Tokens [12]	-	EXIF embedding	56.23	66.97	0.2311	28.07	28.55
Gen. + Blur	6	SDXL [25]	TAF [32]	EXIF as text	67.04	79.54	0.1409	26.02	30.23
	7	SDXL [25]	Dr.Bokeh [30]	EXIF as text	79.40	81.07	0.1506	18.04	19.5
	8	SDXL (EXIF-Fixed) [25]	Dr.Bokeh [30]	EXIF as text	81.10	87.04	0.0356	19.01	18.8
	9	Deep-DoF Gen	TAF [32]	Aperture, Focal Length	82.12	87.62	0.0338	18.54	24.04
	10	Ours	TAF [32]	Aperture, Focal Length	93.91	**92.34**	**0.0064**	**13.24**	**16.69**
	11	Ours	Dr.Bokeh [30]	Aperture, Focal Length	**96.89**	91.42	0.0144	13.67	17.51

Table 2. **Ablations.** Every quantitative metric becomes worse as components are removed from our full model (Ours + TAF [32]).

#	DoF datasets	Deep DoF Pretraining	Lens Model	Focus Distance Transformer	Blur ↑ Monotonicity	Content ↑ Consistency	LPIPS ↓	$\mathbf{FID_{DDoF}}$ ↓	$\mathbf{FID_{SDoF}}$ ↓
1	✓				56.34	86.45	0.1432	15.67	16.95
2	✓	✓			57.51	88.51	0.0862	14.23	17.31
3	✓	✓	✓		73.21	90.65	0.0138	14.10	17.54
4	✓	✓	✓	✓	93.91	92.34	0.0064	13.24	16.69
5	SDoF only	✓	✓	✓	82.50	90.04	0.0031	14.05	16.99

whether scene content remains unchanged as blur is adjusted. The lens models TAF [32] and Dr.Bokeh [30] that we use have been previously validated in isolation for paired data and metrics. We define these metrics:

Blur Monotonicity. To verify that the model appropriately reflects changes in aperture, we check if decreasing the aperture value leads to increased defocus in the generated images. This is tested by evaluating whether the signal energy decreases as the aperture value decreases. Formally, for $\text{ap} \in \{\text{ap}_i\}_{i=1}^{N}$ with $\text{ap}_i < \text{ap}_{i+1}$ and I_{ap_i} representing the image generated for each aperture value, we count the percentage of instances where $E(I_{\text{ap}_i}) < E(I_{\text{ap}_{i+1}})$. Here, $E(\cdot)$ denotes the signal energy, which can be computed either as the sum of squared magnitudes of the 2D Fourier spectrum as $\sum_{\vec{k}} \left|\text{FFT2}(\cdot)_{\vec{k}}\right|^2$, where the sum is over all frequencies $\vec{k}$, or, by Parseval's theorem, equivalently as the sum of squared magnitudes in the spatial domain, $\sum_p |(\cdot)_p|^2$ over pixels p. We include in Supplemental Sec. F.2 a statistical analysis on real photos and a proof (for a simple case of a scene with uniform depth) that this metric decreases from all-in-focus images to defocused images. A higher value for this metric indicates better model performance in controlling blur based on aperture. We compute this metric given the 8 aperture values: [1.8, 2.8, 4, 5.6, 8, 11, 16, 22].

Content Consistency. This metric assesses whether the scene content remains unchanged as the amount of defocus varies in the image. For a set of aperture values [4, 5.6, 8, 11, 16, 22], we generate an image at each aperture, compute its semantic segmentation [6], and compare each pixel's segmentation class across images generated at different apertures. If a pixel's class remains the same, we count its contribution as 1; otherwise, as 0. We calculate the mean across all samples. Higher values for this metric indicate a stronger ability to separate defocus effects from scene-content changes.

LPIPS. Our content consistency metric uses semantic segmentation, which has a benefit of being less sensitive to changes in blur, but might not capture certain fine-grained textural changes. Thus, we also evaluate LPIPS [43]. We report the mean of the $\text{LPIPS}(I_{\text{ap}_i}, I_{\text{ap}_{i+1}})$ metrics between adjacent apertures across $i = 0, \ldots, 4$ for the same aperture values ap_i as in content consistency.

FID. We compute the Fréchet Inception Distance (FID) [9] scores on 10,000 samples from both the Deep DoF and Shallow DoF datasets to assess image quality. A low $\mathbf{FID_{DDoF}}$ indicates that the model generates sharp, blur-free images, while a low $\mathbf{FID_{SDoF}}$ suggests that the model effectively reproduces images with regional blur.

5. Results

Baselines. We categorize baselines into two groups: (1) methods that function purely as image generators and (2) methods that integrate a lens blur model with the genera-

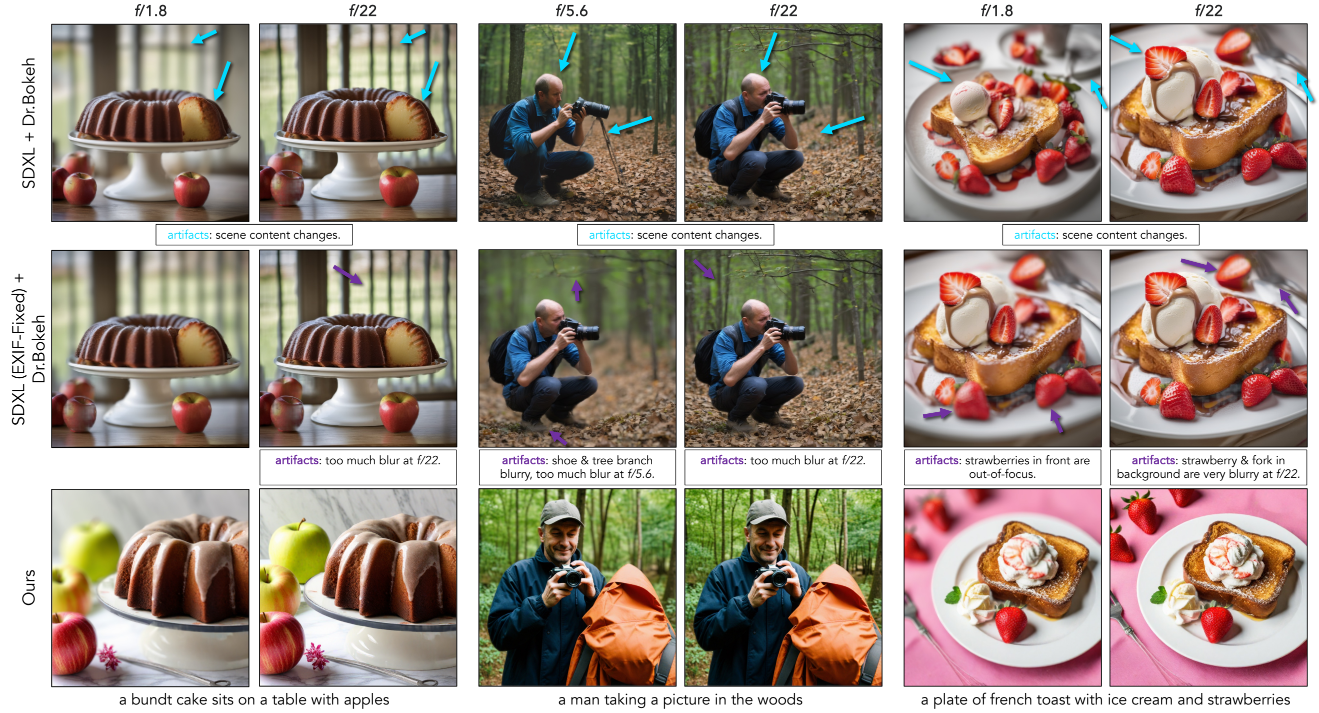

f/1.8
f/22
f/5.6
f/22
f/1.8
f/22
SDXL + Dr.Bokeh
artifacts: scene content changes.
artifacts: scene content changes.
artifacts: scene content changes.
SDXL (EXIF-Fixed) + Dr.Bokeh
artifacts: too much blur at f/22.
artifacts: shoe & tree branch blurry, too much blur at f/5.6.
artifacts: too much blur at f/22.
artifacts: strawberries in front are out-of-focus.
artifacts: strawberry & fork in background are very blurry at f/22.
Ours
a bundt cake sits on a table with apples
a man taking a picture in the woods
a plate of french toast with ice cream and strawberries

Figure 3. **Qualitative Comparisons with SDXL + Dr.Bokeh.** SDXL + Dr.Bokeh struggles to preserve scene content across aperture settings, with unrealistic defocus even when EXIF input to the generator is fixed. In contrast, our method effectively reduces blur while maintaining scene structure as aperture increases. At $f/22$, we produce sharp images as expected, while SDXL + Dr.Bokeh introduces noticeable background blur. Blue arrows: undesired content changes, purple arrows: unrealistic defocus effects.

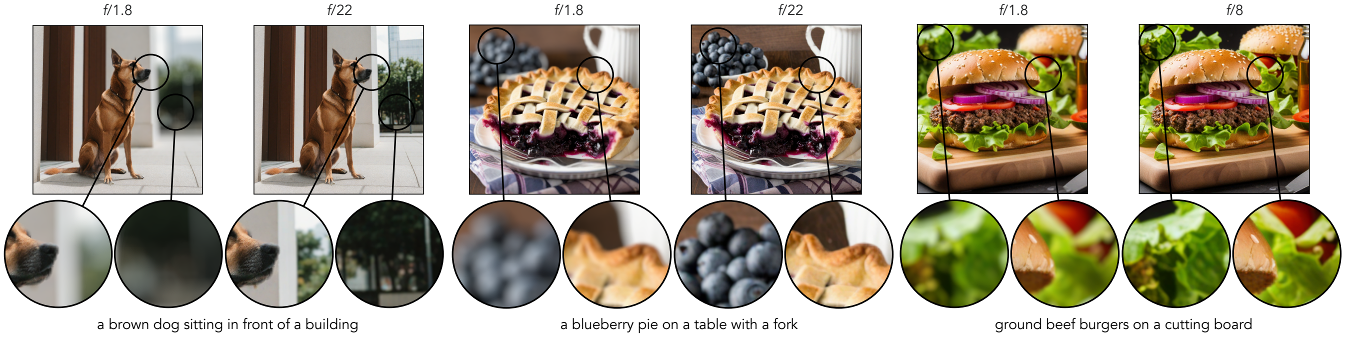

f/1.8
f/22
f/1.8
f/22
f/1.8
f/8
a brown dog sitting in front of a building
a blueberry pie on a table with a fork
ground beef burgers on a cutting board

Figure 4. **Qualitative Results.** We show our model's performance across different apertures and prompts. As the aperture increases, background blur decreases while the overall scene remains consistent. *Please see the website videos for more results.*

tor ("Gen + Blur"), which are inspired by our framework but use more pretrained models rather than fully end-to-end training as we do. We evaluate our method against the following baselines.

SDXL: We use Stable Diffusion XL to evaluate this dataset, running it for 50 denoising steps. Since it does not natively support EXIF encoding, we test two methods for incorporating EXIF data. In the first variant, we convert Aperture and Focal Length to strings and add them as additional text input. In the second variant, we use a Depth-of-Field (DoF) prompt based on the aperture value: for apertures greater than 10, we set the prompt to "Deep Depth-of-Field", and for apertures less than 10, we set it to "Shallow Depth-of-Field" and add it to the text.

4-step SDXL (Distilled): We distill the SDXL teacher model into a 4-step generator and train it with DMD2 losses. We train two variants: one that includes a DoF prompt in the text input and another that uses an EXIF projection layer for improved EXIF data processing where we add sinusoidal positional encoding for Aperture and Focal Length and process them using two projection layers and add it to the time embedding. See the supplemental section F.3 for details regarding the EXIF projection.

Camera Settings as Tokens: We evaluate [12] as a baseline, where camera parameter embeddings and a LoRA are trained to guide image generation based on EXIF metadata. The model takes aperture, focal length, ISO rating, and exposure time as inputs in addition to the text prompt and uses Stable Diffusion 2 [26] as the base generator.

SDXL + TAF Lens: This baseline is inspired by our framework, but utilizes pretrained models in a plug-and-play manner. We use SDXL as the image generator and add the TAF lens model [32] to simulate defocus effects. Depth is predicted using the Metric3Dv2 model [11]. To determine an optimal focus distance, we use a frozen Visual Saliency Transformer [20] model to generate a saliency mask, and we use the mean depth value of the salient region as the focal distance prediction.

SDXL + Dr.Bokeh: This baseline is like SDXL + TAF Lens, except we replace the lens model with Dr.Bokeh [30].

SDXL (EXIF-Fixed) + Dr.Bokeh: As in SDXL + Dr.Bokeh, but the EXIF metadata fed to SDXL is fixed across aperture settings to preserve scene content, closely following the strategy of our framework.

Deep-DoF Gen. + TAF Lens: Inspired by our framework, we train a model similar to ours, but instead of training a model to predict focus distance, we again use the mean depath of the salient region from the frozen VST model [20].

Quantitative Results. We evaluate all baselines in Table 1. The SDXL baselines without lens models, both with and without distillation, and with and without EXIF conditioning (rows 1-4) all struggle to generate different defocus effects as aperture and focal distance are changed. Camera Settings as Tokens (row 5), despite being explicitly conditioned on EXIF metadata, struggles to adjust blur effectively based on aperture changes and struggles to maintain scene integrity. This is seen in its low Blur Monotonicity and Content Consistency scores (see also Figure 5).

The SDXL + TAF Lens baseline (row 6), a fully pretrained model, outperforms earlier methods by leveraging additional information such as depth and focus distance, enabling more consistent blur application as the aperture decreases. Replacing the lens model with Dr.Bokeh (row 7) further improves blur rendering. Further fixing the EXIF input to SDXL (row 8) preserves the scene content (following our framework), but often generates out-of-focus blur (Fig. 3). The baseline in row 9 resembles our approach, except it does not have focus distance prediction and only trains on the DDoF dataset. Without learning an appropriate scale for focus distance (as our model does), it struggles, often producing out-of-focus blur effects, and also obtains a poor FID score on the SDoF dataset. Our model with TAF Lens (row 10) outperforms all baselines, consistently increasing blur as the aperture decreases while preserving scene content across aperture changes. Replacing TAF Lens with Dr.Bokeh during inference (row 11) further improves the blur metric, due to Dr.Bokeh's superior blur rendering capabilities.

Ablations. We ablate key components of our model in Table 2. All ablations are trained on our Shallow and Deep DoF datasets (except row 5). Rows 1 and 2 exclude the lens model and instead use aperture and focal length as additional text inputs to control blur behavior. These configurations show reduced Blur Consistency, indicating that representing aperture solely as text does not provide sufficient information to consistently simulate blur changes. Notably, row 2 outperforms row 1, demonstrating the benefit of pretraining on the Deep DoF dataset. This pretraining provides our generator $\mathbf{G}$ with a strong prior for generating all-in-focus images during fine-tuning.

In row 3, the addition of the lens model improves Content Consistency when changing aperture and enhances the model's ability to apply blur consistently at a given aperture. Row 3 removes the focal distance model and replaces it with mean depth, which compared to the full model gives worse visual quality (FID) and worse blur monotonicity as the focus distance scale is not learned, causing unintended out-of-focus blur and limiting the lens model's ability to increase blur beyond a certain point. Finally, in row 4, adding the focus distance model allows the lens model to better focus on salient regions, reducing out-of-focus artifacts. Row 5 shows pretraining the generator on Deep-DoF and freezing it, followed by finetuning only on Shallow DoF, leads to worse all-in-focus image generation.

Qualitative Results. We show qualitative results for our method and baselines. In Figure 3, we compare our method, using Dr.Bokeh as the lens model against an SDXL generator with a Dr.Bokeh lens model (Table 1, row 7). Our method effectively preserves scene content while adjusting the blur effect based on the specified aperture value, enabling precise control over defocus blur during image generation. In contrast, SDXL + Dr.Bokeh struggles to increase blur as the aperture changes and alters the scene content. Even when we fix the EXIF input to SDXL, the defocus blur is not realistic. Figure 4 illustrates the blurring effects of our method at different aperture values. It keeps the foreground (or salient region) in focus while modifying the blur in non-salient regions without altering the overall scene.

In Figure 5, we compare against the Camera Settings as Tokens [12] baseline. This method fails to maintain scene content, significantly altering the generated images while providing limited control over blur. Figure 6 further compares our method with the Deep-DoF Gen + TAF baseline, which struggles to produce aesthetically pleasing focal planes, because it neither learns an appropriate focus-distance scale nor is fine-tuned on a shallow DoF dataset. In Supplemental Section B, we extend the analysis: (1) evaluating a ControlNet-based variant conditioned on scene depth to improve the Camera Settings as Tokens approach, and (2) adding additional comparisons with real photographs. Even with conditional depth, camera-setting

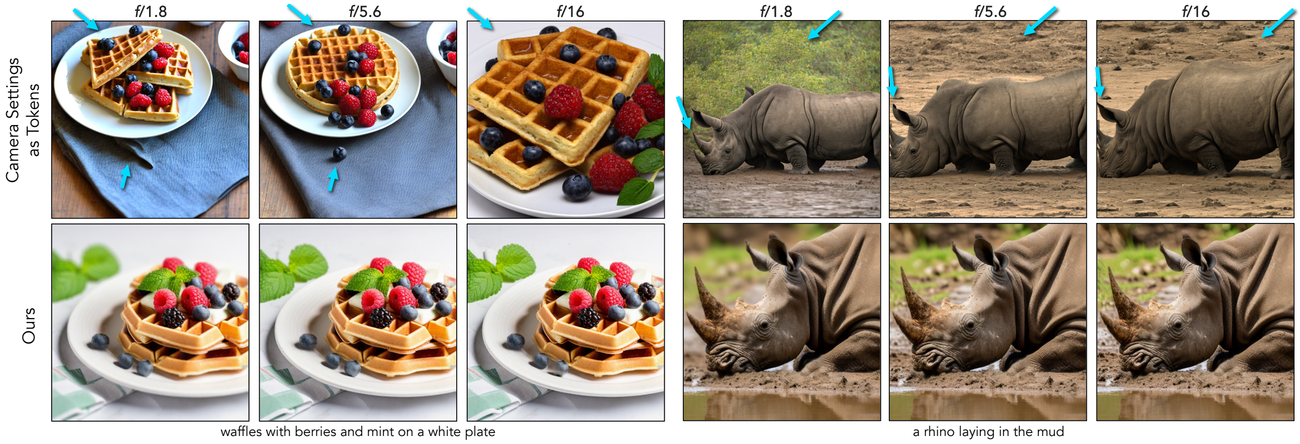

f/1.8
f/5.6
f/16
f/1.8
f/5.6
f/16
Camera Settings as Tokens
Ours
waffles with berries and mint on a white plate
a rhino laying in the mud

Figure 5. **Comparison to Camera Settings as Tokens [12].** Despite being trained with EXIF conditioning, [12] struggles to control blur effectively and often alters the image significantly, making it entirely different. Whereas, our model preserves scene content while successfully decoupling defocus blur from the scene, modifying only the blur without changing the overall composition. Blue arrows: undesired content changes.

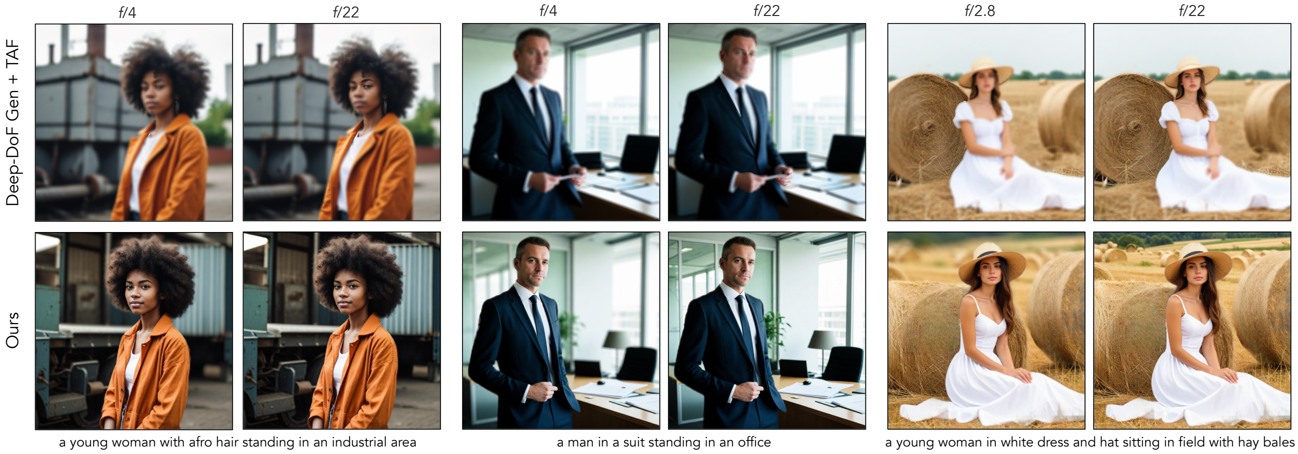

f/4
f/22
f/4
f/22
f/2.8
f/22
Deep-DoF Gen + TAF
Ours
a young woman with afro hair standing in an industrial area
a man in a suit standing in an office
a young woman in white dress and hat sitting in field with hay bales

Figure 6. **Comparison to Deep-DoF Gen. + TAF.** We compare our method with Deep-DoF Gen + TAF model (main paper Table 1, row 8). Our method is able to consistently vary blur amount while preserving the scene, whereas the generated images from the baseline appear out-of-focus without focusing on salient regions in the image.

embeddings, and a text prompt, the ControlNet variant still fails to preserve scene content consistently.

6. Limitations, Future Work, Conclusion

Due to strong priors in the SDXL model that we distill from, the generated images from the all-in-focus generator G, on rare occasions, have a background blur already present, which limits how in focus of an image can be obtained. This could likely be mitigated by further increasing the DDoF dataset size. The focus distance scale is learned from weak signals from DMD2 losses and does not have a direct supervisory signal, which can in some cases result in blurry photos without a good focal plane. Our unsupervised method has the advantage of not needing explicit collection of focus distance values, which are typically not present or accurate in EXIF tags. However, future work might consider the task of capturing RGBD photos with metric depth along with precise measurement of focus distances to enable supervised learning. Even with the high-quality Dr.Bokeh renderer, high-frequency details along occlusion boundaries such as the dog fur in Figure 4 can occasionally be blurry; this is due to limited depth-map resolution and could be resolved by using a depth estimator that resolves such high-frequency details [22]. We learn only disc-shaped bokehs, but a benefit of our framework is that the lens model can be swapped out at inference time (Sec. 3.3), so in future work, other lens models or stylized bokehs such as hexagons or hearts could be used [37].

In conclusion, our approach is a flexible framework that decouples the image generator from a lens model. This approach is useful for obtaining explicit and fine-grained control of aperture and focus properties in diffusion generative models. Our framework shows the benefits of explicit focus distance estimation and our joint end-to-end learning on both DDoF and SDoF datasets.

References

[1] Huggingface spaces. https://huggingface.co/spaces. 14

[2] Andrew Adams and Marc Levoy. General linear cameras with finite aperture. In *Rendering Techniques*, pages 121–126. Citeseer, 2007. 2

[3] Hadi Alzayer, Abdullah Abuolaim, Leung Chun Chan, Yang Yang, Ying Chen Lou, Jia-Bin Huang, and Abhishek Kar. Dc2: Dual-camera defocus control by learning to refocus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages –, 2023. 2

[4] Laurent Belcour, Cyril Soler, Kartic Subr, Nicolas Holzschuch, and Fredo Durand. 5d covariance tracing for efficient defocus and motion blur. *ACM Transactions on Graphics (TOG)*, 32(3):1–18, 2013. 2

[5] Jiating Chen, Bin Wang, Yuxiang Wang, Ryan S Overbeck, Jun-Hai Yong, and Wenping Wang. Efficient depth-of-field rendering with adaptive sampling and multiscale reconstruction. In *Computer Graphics Forum*, pages 1667–1680. Wiley Online Library, 2011. 2

[6] Jiaqi Chen, Zeyu Yang, and Li Zhang. Semantic segment anything. https://github.com/fudan-zvg/Semantic-Segment-Anything, 2023. 5, 14

[7] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023. 13

[8] Armando Fortes, Tianyi Wei, Shangchen Zhou, and Xingang Pan. Bokeh diffusion: Defocus blur control in text-to-image diffusion models. *arXiv preprint arXiv:2503.08434*, 2025. 3

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[11] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2, 4, 7

[12] Jun-Cheng Chen I-Sheng Fang, Yue-Hua Han. Camera settings as tokens: Modeling photography on latent diffusion models. In *Proc. SIGGRAPH Asia 2024*. ACM, 2024. 1, 2, 3, 5, 6, 7, 8, 12

[13] Andrey Ignatov, Jagruti Patel, and Radu Timofte. Rendering natural camera bokeh effect with deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 418–419, 2020. 11, 12, 14

[14] Takuhiro Kaneko. Unsupervised learning of depth and depth-of-field effect from natural images with aperture rendering generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15679–15688, 2021. 2

[15] Takuhiro Kaneko. Ar-nerf: Unsupervised learning of depth and defocus effects from natural images with aperture rendering neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18387–18397, 2022. 2

[16] Minguk Kang, Richard Zhang, Connelly Barnes, Sylvain Paris, Suha Kwak, Jaesik Park, Eli Shechtman, Jun-Yan Zhu, and Taesung Park. Distilling diffusion models into conditional gans. *ECCV 2024*, 2024. 3

[17] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2

[18] Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David Lobell, and Stefano Ermon. Diffusionsat: A generative foundation model for satellite imagery. *arXiv preprint arXiv:2312.03606*, 2023. 2

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 11

[20] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4722–4732, 2021. 4, 7

[21] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023. 3

[22] S. Mahdi H. Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yagiz Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9685–9694, 2021. 8

[23] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[24] Stanislav Pidhorskyi, Timur Bagautdinov, Shugao Ma, Jason Saragih, Gabriel Schwartz, Yaser Sheikh, and Tomas Simon. Depth of field aware differentiable rendering. *ACM Transactions on Graphics (TOG)*, 41(6):1–18, 2022. 2

[25] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 2, 5, 14

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2, 7

[27] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 3

[28] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. *arXiv preprint arXiv:2403.12015*, 2024. 3

[29] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2025. 3

[30] Yichen Sheng, Zixun Yu, Lu Ling, Zhiwen Cao, Xuaner Zhang, Xin Lu, Ke Xian, Haiting Lin, and Bedrich Benes. Dr. bokeh: Differentiable occlusion-aware bokeh rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4515–4525, 2024. 2, 4, 5, 7

[31] Andrey Voynov, Amir Hertz, Moab Arar, Shlomi Fruchter, and Daniel Cohen-Or. Curved diffusion: A generative model with optical geometry control. In *European Conference on Computer Vision*, pages 149–164. Springer, 2024. 2

[32] Chao Wang, Krzysztof Wolski, Xingang Pan, Thomas Leimkühler, Bin Chen, Christian Theobalt, Karol Myszkowski, Hans-Peter Seidel, and Ana Serrano. An implicit neural representation for the image stack: Depth, all in focus, and high dynamic range. Technical report, 2023. 1, 2, 4, 5, 7

[33] Yujie Wang, Praneeth Chakravarthula, and Baoquan Chen. Dof-gs: Adjustable depth-of-field 3d gaussian splatting for refocusing, defocus rendering and blur removal. *arXiv preprint arXiv:2405.17351*, 2024. 2

[34] Yiyang Wang, Xi Chen, Xiaogang Xu, Yu Liu, and Hengshuang Zhao. Diffcamera: Arbitrary refocusing on images. *SIGGRAPH Asia 2025 Conference Papers*, 2025. 2

[35] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016. 11

[36] Shumian Xin, Neal Wadhwa, Tianfan Xue, Jonathan T Barron, Pratul P Srinivasan, Jiawen Chen, Ioannis Gkioulekas, and Rahul Garg. Defocus map estimation and deblurring from a single dual-pixel image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2228–2238, 2021. 2

[37] Yang Yang, Haiting Lin, Zhan Yu, Sylvain Paris, and Jingyi Yu. Virtual dslr: High quality dynamic depth-of-field synthesis on mobile platforms. *Electronic Imaging*, 28:1–9, 2016. 8

[38] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *arXiv preprint arXiv:2405.14867*, 2024. 2, 3

[39] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6613–6623, 2024. 3

[40] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25669–25680, 2024. 2

[41] Yu Yuan, Xijun Wang, Yichen Sheng, Prateek Chennuri, Xingguang Zhang, and Stanley Chan. Generative photography: Scene-consistent camera control for realistic text-to-image synthesis. *arXiv preprint arXiv:2412.02168*, 2024. 3

[42] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 11, 12

[43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[44] Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023. 14

[45] Chenhao Zheng, Ayush Shrivastava, and Andrew Owens. Exif as language: Learning cross-modal associations between images and camera metadata. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6945–6956, 2023. 3

[46] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 11, 12

[47] Matthias Zwicker, Wojciech Jarosz, Jaakko Lehtinen, Bochang Moon, Ravi Ramamoorthi, Fabrice Rousselle, Pradeep Sen, Cyril Soler, and S-E Yoon. Recent advances in adaptive sampling and reconstruction for monte carlo rendering. In *Computer graphics forum*, pages 667–681. Wiley Online Library, 2015. 2

Fine-grained Defocus Blur Control for Generative Image Models

Supplementary Material

A. Video

We provide videos on our webpage (`link`) showing the controllability of defocus blur using our model. We also include qualitative examples demonstrating defocus control in generated images for several prompts.

B. Comparisons to our model

Comparison with the ControlNet baseline. Figure A1 examines an alternative approach where a ControlNet [42] is conditioned on depth to improve scene preservation for Camera Settings as Tokens. Despite using conditional depth, camera embeddings, and a text prompt, this approach still struggles to maintain scene content. As seen in Figure A1 (right), the baseline preserves the dog's pose but changes its identity and the background scene. In contrast, our method maintains both the subject and background while effectively adjusting the blur.

Qualitative Comparison with Real Images. To assess our model's qualitative performance against real photographs, we compare its outputs with the Everything is Better with Bokeh! (EBB!) dataset [13]. This dataset contains pairs of images of the same scene captured at two aperture settings: $f/1.8$ (shallow depth-of-field) and $f/16$ (all-in-focus).

To generate comparable results without bias toward either aperture, we first caption each $f/16$ image using the InternVL3 model [46]. These captions serve as neutral text prompts. We then provide the caption along with the target aperture ($f/1.8$ or $f/16$) to our model to synthesize shallow depth-of-field and all-in-focus images, respectively. Representative outputs are shown in Figure A2.

The figure demonstrates that our method faithfully reproduces the expected optical characteristics of each aperture. When conditioned on $f/1.8$, our model produces images with pronounced background blur and smooth bokeh, closely matching the shallow-focus ground truth and showing sharp foreground with naturally defocused backgrounds. When conditioned on $f/16$, it generates images with crisp details across the full depth of field, consistent with the all-in-focus reference photographs.

These results confirm that our approach not only captures the semantic content of a scene but also accurately models the physical effects of aperture on defocus, validating the effectiveness of our aperture-aware image generation framework.

C. Controllability of defocus blur in image generation

Our model takes EXIF metadata (e.g., aperture, focal length) and a text prompt as input, generating an image that faithfully reflects both. A trained focus distance model predicts the scene's focus distance during generation, which the lens model uses to apply defocus blur consistent with the metadata.

To enable controllability over the focus in the generated image, users can intercept the predicted focus distance and provide their own focus distance value. The lens model applies spatial blur based on this user-defined focus distance, allowing precise control over where the generated image should focus. The focus distance is represented on the depth output scale of the Metric3Dv2 depth model. For instance, a focus distance of 0.1 corresponds to the depth plane with a value of 0.1 in the depth map. We demonstrate the effects of varying focus distance in Figure A3, where low and high focus distance values result in noticeable shifts in the focal plane within the image.

In addition to the prompt, our model provides the ability to manipulate focus distance and aperture, offering fine-grained control over image generation. By leveraging this information, the model determines where and how much defocus blur to apply. This controllability is illustrated in a video attached in the supplementary material, along with several qualitative examples.

D. Deep and Shallow Depth-of-Field Datasets

Our generator G produces all-in-focus (deep depth-of-field, or Deep DoF) images $\mathbf{x}$, while the lens model renders shallow DoF images $\hat{\mathbf{x}}$, trained with only weak supervision ($\mathbf{x}$, $\hat{\mathbf{x}}$ shown in Fig. 2). To supervise this pipeline, we curate large-scale datasets of deep and shallow DoF images from roughly 300 million uncurated photographs drawn from a commercially available stock-photography dataset. We discard photos with no EXIF data. For photos without captions, we generate captions using BLIP2 [19]. A ResNeXt–FPN classifier [35] is used to identify whether each image contains no blur, desirable blur, or undesirable blur.

Filtering Criteria. From the classifier outputs and EXIF metadata, we apply the following filters to construct the two datasets:

- **Aperture range:** Shallow DoF images retain aperture values below 10, producing a narrow depth of field and

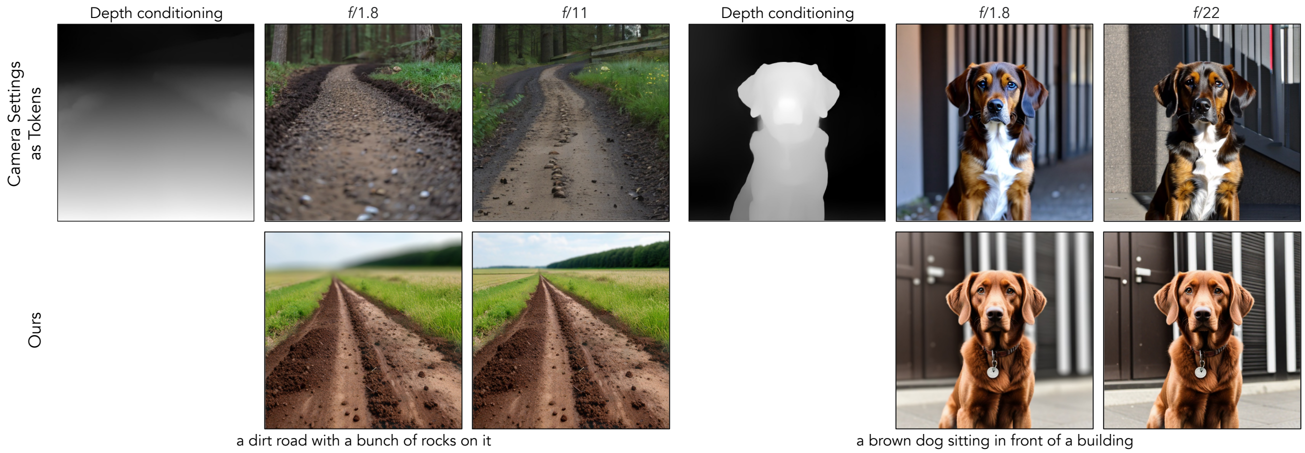

Depth conditioning
f/1.8
f/11
Depth conditioning
f/1.8
f/22
Camera Settings as Tokens
Ours
a dirt road with a bunch of rocks on it
a brown dog sitting in front of a building

Figure A1. **Comparison with Camera Settings as Tokens [12] based ControlNet [42].** We compare our method to a depth-conditioned ControlNet that uses Camera Settings as Tokens embeddings. While the ControlNet effectively adheres to scene depth, it alters scene content within those depth planes. Notably, depth is used as a conditioning input for ControlNet but not for our generator.

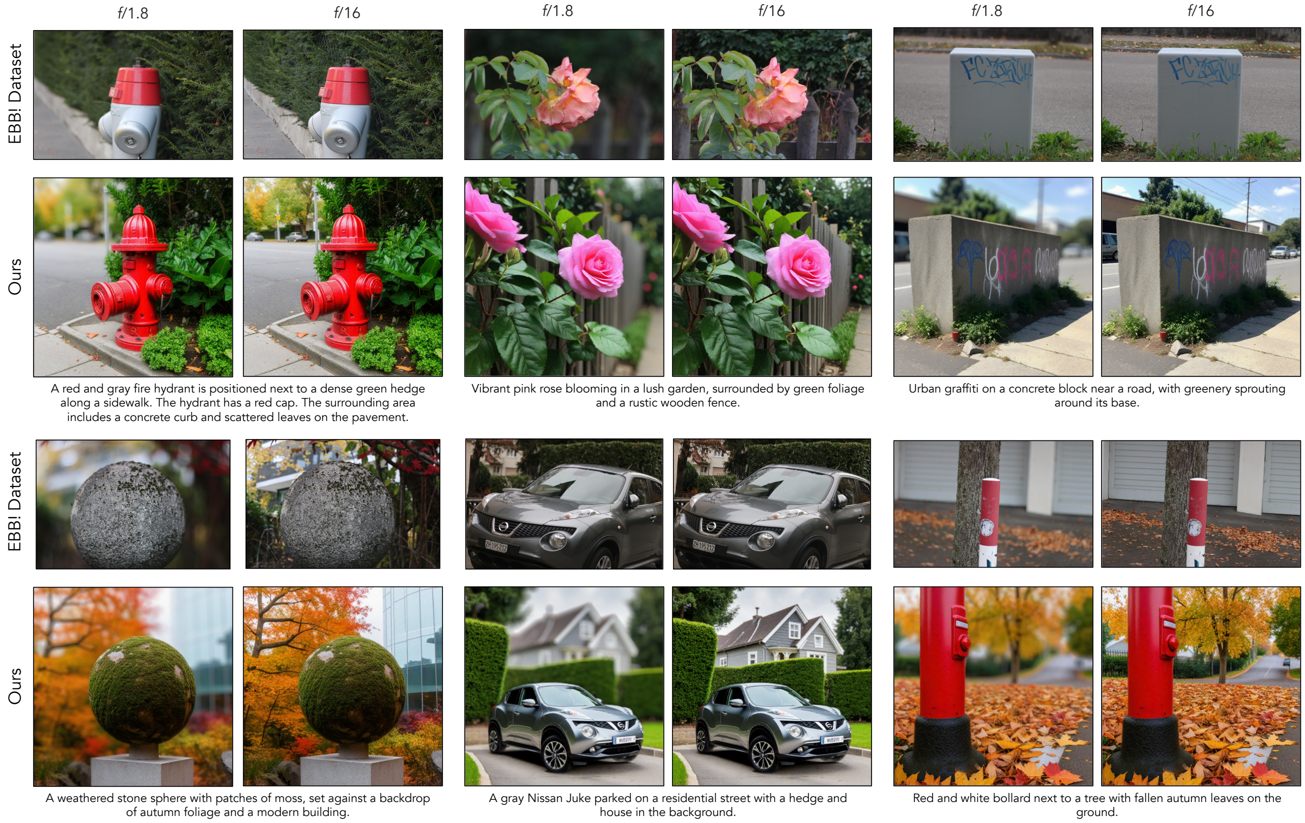

f/1.8
f/16
f/1.8
f/16
f/1.8
f/16
EBB! Dataset
Ours
A red and gray fire hydrant is positioned next to a dense green hedge along a sidewalk. The hydrant has a red cap. The surrounding area includes a concrete curb and scattered leaves on the pavement.
Vibrant pink rose blooming in a lush garden, surrounded by green foliage and a rustic wooden fence.
Urban graffiti on a concrete block near a road, with greenery sprouting around its base.
EBB! Dataset
Ours
A weathered stone sphere with patches of moss, set against a backdrop of autumn foliage and a modern building.
A gray Nissan Juke parked on a residential street with a hedge and house in the background.
Red and white bollard next to a tree with fallen autumn leaves on the ground.

Figure A2. **Comparison to EBB! Dataset.** Using the EBB! dataset [13], which provides image pairs captured at two apertures: $f/16$ (all-in-focus) and $f/1.8$ (shallow depth of field), we first caption the $f/16$ image with the InternVL3 [46] model (shown below the "Ours" images). We then use that caption as the text prompt, along with the specified aperture ($f/16$ or $f/1.8$), to generate images with our method. Our results closely match the expected defocus characteristics, producing pronounced blur at $f/1.8$ and sharp, well-focused images at $f/16$.

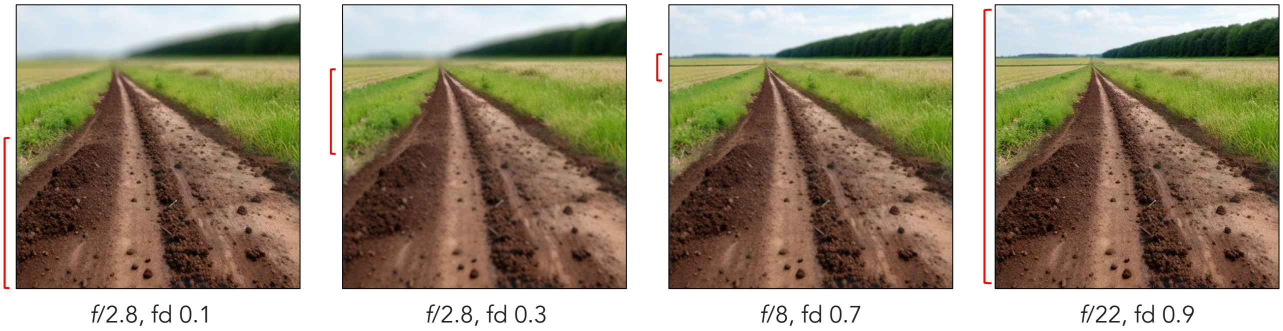


Figure A3. **Varying focus distances in the generation process.** We show that varying the focus distance in our model produces images with focus shifting across different focal planes. As the focus distance increases from low to high values, the focal plane transitions from the near to the far plane. The red bar [highlights the region of the image that is in focus.

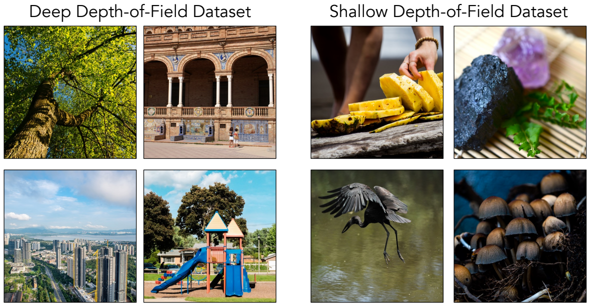


Figure A4. **Deep and Shallow DoF Datasets.** Images shown are selected using our dataset filtering approach mentioned in Sec. D. After filtering, the Deep DoF dataset primarily consists of all-in-focus images, while the Shallow DoF dataset includes images with defocus blur, emphasizing a specific object of interest. We use these datasets to train our model.

aesthetically pleasing background blur. Deep DoF images retain aperture values between 10 and 50 to ensure sharp focus across the scene.

- **Device type:** Smartphone photographs are removed from the shallow DoF set to avoid synthetic blur introduced by computational photography.
- **Exposure time:** Images with exposure times longer than 0.1 seconds are excluded from the deep DoF set to prevent motion blur.
- **Photographic validity:** We discard non-photographic content (e.g., AI-generated or illustrated images) by verifying that each candidate is a real photograph using the vision–language model InternVL [7].
- **Blur classifier output:** Shallow DoF images are those labeled as exhibiting the desired blur, while deep DoF images are those labeled as having no blur.

Dataset Scale. Applying these criteria gives us roughly 1.5 million (image, EXIF, prompt) pairs for each of the Deep DoF and Shallow DoF datasets. Representative samples are shown in Figure A4.

E. Human Study

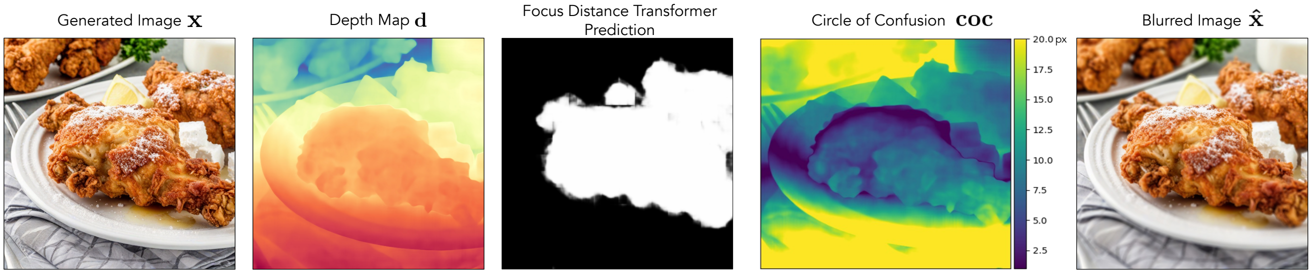
We conducted a human study to validate our method, aiming to evaluate whether the model can preserve the scene and reduce defocus blur when the aperture value in the camera metadata increases. For the study, we used 25 prompts from the validation split of the deep and shallow depth-of-field datasets we created. For each prompt, we generated images corresponding to aperture values in the set [1.8, 2.8, 4, 5.6, 8, 11, 16, 22] across all methods.

The study involved six baseline methods in addition to our approach:

- SDXL (Table 2, Row 2),
- 4-step SDXL (Distilled) (Table 2, Row 4),
- Camera Settings as Tokens (Table 2, Row 5),
- SDXL (EXIF-Fixed) + Dr.Bokeh Lens (Table 2, Row 8),
- SDXL + TAF Lens (Table 2, Row 6),
- Deep-DoF Gen + TAF Lens (Table 2, Row 8).

We created a video for each method per prompt, sequentially increasing the aperture value to illustrate its effect in the video. During the study, participants were shown paired videos—one generated by our model and the other by a baseline—for the same prompt. Participants were instructed to select the video that better preserved scene content, reduced blur as aperture increased, and kept the salient object in focus.

Each participant answered 20 comparison questions, with video pairs randomly assigned. The study involved 25 participants, and their aggregated preferences are presented in Figure A6. Results show that users preferred our method over the baselines in over at least 83% of cases, consistent with the performance metrics in Table 2, further demon-


Generated Image $\mathbf{x}$
Depth Map $\mathbf{d}$
Focus Distance Transformer Prediction
Circle of Confusion $\mathbf{coc}$
20.0 px
17.5
15.0
12.5
10.0
7.5
5.0
2.5
Blurred Image $\hat{\mathbf{x}}$

Prompt: fried chicken breasts on a white plate with powdered sugar

Figure A5. **Image Generation Pipeline.** The pipeline begins with an image generated by the model (left), followed by depth prediction from the depth model. A saliency map is then predicted and used to compute the focus distance as a weighted sum of depth and saliency. The lens model calculates the circle of confusion (CoC) based on depth, focus distance, and other EXIF parameters. Finally, a spatially varying blur kernel, derived from the CoC, is applied to the generated image. The entire pipeline is trained end-to-end.

strating our method's superiority over baselines.

The user study was conducted on the Hugging Face Spaces [1] platform, and the interface used is shown in Figure A7.

F. Implementation Details

Here, we provide more information about training, hyperparameters, and evaluation metrics.

F.1. Training and Hyperparameters

We train the few-step generator by distilling it from the SDXL model [25] over 4 steps. We train our network on 2 nodes, each equipped with 8 A100 GPUs, for a total of 16 GPUs. The Deep DoF generator is trained for one day, followed by an additional day of fine-tuning using the full setup, which includes the lens model, depth estimation module, and focus distance predictor. To scale the training efficiently across 8 nodes, we use the Fully-Shared Data Parallel framework [44].

The training images have a resolution of 1024 × 1024, and the model is optimized using the AdamW optimizer with a learning rate of 5×10^{-7}, a weight decay of 0.01, and beta parameters of (0.9, 0.999). The batch size is set to 1 to fit the entire model in GPU memory. The fake diffusion model μ_{fake} is updated 5 times for each generator update and during generator updates, we alternate between Shallow and Deep DoF. The focus distance model (optimized with L_{Huber}) is updated at every iteration. The guidance scale for the real diffusion model μ_{real} is to be 8. The loss weights are set to $\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 200$.

F.2. Evaluation Metrics

Content Consistency. To evaluate this metric, we compute the segmentation maps using Semantic Segment Anything [6]. This is an open-set segmentation method which means it does not have a predefined set of prediction classes. Due to this, sometimes the top-1 predicted class could be different for the same object. So, we compare the top-3 predicted classes. To check if the semantic class remains the same, we check if any of top-3 classes matches remain the same for the image pixels instead of comparing just top-1.

Blur Monotonicity. We introduced Blur Monotonicity to quantify whether image blur decreases as the aperture value increases. To measure the efficacy of this metric, we use the Everything is Better with Bokeh! dataset [13], which provides approximately 5,000 pairs of all-in-focus images captured at $f/16$ and corresponding shallow-depth images at $f/1.8$. In 96% of the pairs, the signal energy of the all-in-focus image exceeds that of its bokeh counterpart, supporting the premise of our metric. Visual inspection of the remaining 4% reveals negligible defocus differences, making the energy comparison less informative in those specific cases.

We now present a theoretical justification for the validity of our metric. As a reminder, we defined the signal energy $E(\cdot)$ as the sum of squared magnitudes of the 2D Fourier spectrum, computed as $\sum_{\vec{k}} \left|\text{FFT2}(\cdot)_{\vec{k}}\right|^2$, where the sum is over all frequencies $\vec{k}$. For a simple scene of uniform depth (i.e., depth is constant across pixels), we show that the energy of an image formed from that scene is greater than the energy of the image after blurring via convolution with a blur kernel.

Theorem F.1. *Let f, h be d-dimensional tensors with $f, h \in \mathbb{R}^{N_1 \times N_2 \times \cdots \times N_d}$, where $\vec{N} = (N_1, N_2, \ldots, N_d)$. Define the discrete Fourier transform (DFT) of f as*

$$F_{\vec{k}} = \sum_{\vec{n}=\vec{0}}^{\vec{N}-1} f_{\vec{n}}\, e^{-2\pi i\, \vec{k}\cdot\left(\frac{\vec{n}}{\vec{N}}\right)}, \tag{8}$$

where division is element-wise. Define $H_{\vec{k}}$ analogously for h. The multi-index summation is defined as

$$\sum_{\vec{n}=\vec{0}}^{\vec{N}-1} := \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} \cdots \sum_{n_d=0}^{N_d-1} = \sum_{\vec{n}\in\{0,\ldots,N_1-1\}\times\cdots\times\{0,\ldots,N_d-1\}} .$$

Assume $h_{\vec{n}} \geq 0$ *for all* $\vec{n} \in \{0, \ldots, \vec{N}-1\}$ *and* $\sum_{\vec{n}=\vec{0}}^{\vec{N}-1} h_{\vec{n}} = 1$. *If* $g = f * h$ *and* $G_{\vec{k}}$ *is the DFT of* g, *then:*

$$\sum_{\vec{k}=\vec{0}}^{\vec{N}-1} \left|G_{\vec{k}}\right|^2 = \sum_{\vec{k}=\vec{0}}^{\vec{N}-1} \left|F_{\vec{k}}\, H_{\vec{k}}\right|^2 \leq \sum_{\vec{k}=\vec{0}}^{\vec{N}-1} \left|F_{\vec{k}}\right|^2. \tag{9}$$

Proof. The equality in the above equation follows from the convolution theorem, which states that $G_{\vec{k}} = F_{\vec{k}}\, H_{\vec{k}}$. Now we can analyze the inequality. For each frequency $\vec{k}$,

$$H_{\vec{k}} = \sum_{\vec{n}=\vec{0}}^{\vec{N}-1} h_{\vec{n}}\, e^{-2\pi i\, \vec{k}\cdot\left(\frac{\vec{n}}{\vec{N}}\right)} \tag{10}$$

$$\left|H_{\vec{k}}\right| \leq \sum_{\vec{n}=\vec{0}}^{\vec{N}-1} \left|h_{\vec{n}}\right| \left|e^{-2\pi i\, \vec{k}\cdot\left(\frac{\vec{n}}{\vec{N}}\right)}\right| = \sum_{\vec{n}=\vec{0}}^{\vec{N}-1} \left|h_{\vec{n}}\right| = 1 \tag{11}$$

Since $\left|e^{i\theta}\right| = 1$ for all $\theta \in \mathbb{R}$. We want to show that

$$0 \geq \sum_{\vec{k}=\vec{0}}^{\vec{N}-1} \left|F_{\vec{k}}\right|^2 \left(\left|H_{\vec{k}}\right|^2 - 1\right). \tag{12}$$

Now $\forall \vec{k}$, $\left|F_{\vec{k}}\right|^2 \geq 0$ and $\left|H_{\vec{k}}\right|^2 - 1 \leq 0$, so

$$\sum_{\vec{k}=\vec{0}}^{\vec{N}-1} \left|F_{\vec{k}}\right|^2 \left(\left|H_{\vec{k}}\right|^2 - 1\right) \leq 0. \tag{13}$$

Further, the inequality in Eq. 9 is strict if there exists some $\vec{k}$ such that $\left|F_{\vec{k}}\right| > 0$ and $\left|H_{\vec{k}}\right| < 1$. □

Application. Assume the Circle of Confusion (CoC) is not spatially varying (i.e., the scene has uniform depth) and let f be the image and h the blur kernel. By the theorem, the signal energy satisfies

$$E(h * f) \leq E(f).$$

Further, the inequality is strict if there exists some $\vec{k}$ such that $\left|F_{\vec{k}}\right| > 0$ and $\left|H_{\vec{k}}\right| < 1$. Assume the image is formed from the scene by a process that adds i.i.d. Gaussian noise to each pixel. Then, almost surely, $\left|F_{\vec{k}}\right| > 0$ for all $\vec{k}$, since the DFT is a linear transformation and each DFT coefficient is the sum of a deterministic component and a Gaussian-distributed random variable. Thus, it suffices to analyze the kernel h and determine whether there exists some $\vec{k}$ with $\left|H_{\vec{k}}\right| < 1$. In particular, any blur kernel h that is non-negative, sums to 1, and is not a delta function (i.e. has at least two non-zero entries) guarantees a strict inequality. This includes, as special cases, disc-shaped and polygonal bokeh corresponding to blur kernels with such shapes. We emphasize again that since this analysis uses the convolution theorem, it applies only to the simplified setting of a scene with uniform depth.

□

Although the theoretical guarantee holds under the assumption of uniform depth and spatially invariant blur, our empirical results on real-world images with spatially varying blur strongly suggest that the blur monotonicity metric remains a reliable indicator of relative blur. This combination of theory and empirical validation supports the practical utility of our metric.

F.3. 4-step SDXL (distilled) with EXIF

To incorporate camera metadata into the distilled 4-step SDXL generator, we design an EXIF projection module that encodes numerical tags — specifically, Aperture and Focal Length. These values are first transformed using sinusoidal positional embeddings, then concatenated and passed through two projection layers to produce a single EXIF embedding, which is added to the diffusion timestep embedding.

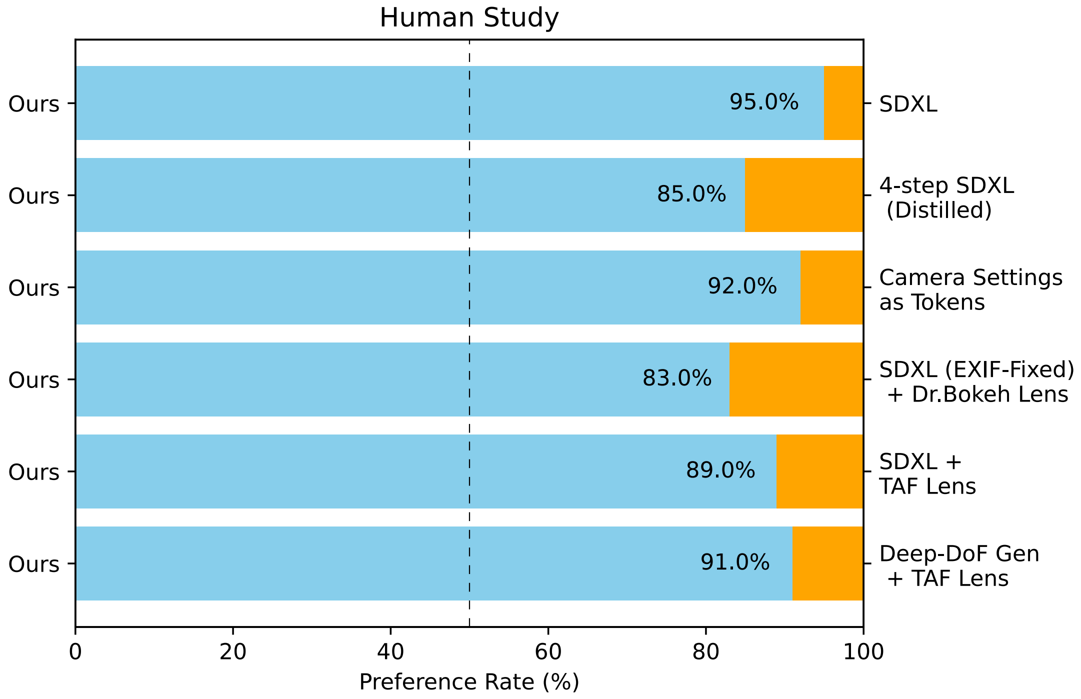
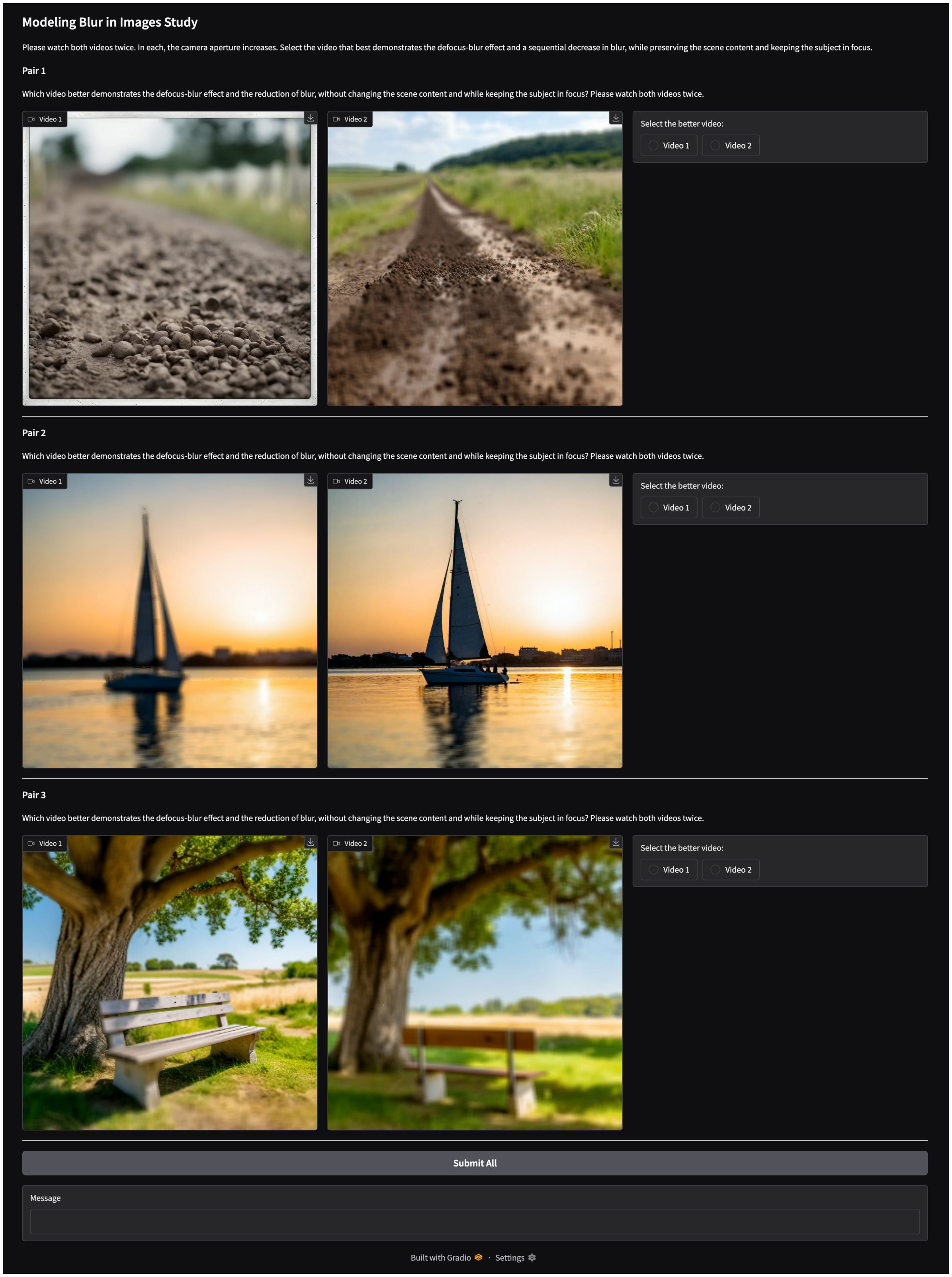


Figure A6. **Human Studies.** Preference rate for selecting our method over the baselines (Section E).



Modeling Blur in Images Study

Please watch both videos twice. In each, the camera aperture increases. Select the video that best demonstrates the defocus-blur effect and a sequential decrease in blur, while preserving the scene content and keeping the subject in focus.

Pair 1

Which video better demonstrates the defocus-blur effect and the reduction of blur, without changing the scene content and while keeping the subject in focus? Please watch both videos twice.

Select the better video: ○ Video 1 ○ Video 2

Pair 2

Which video better demonstrates the defocus-blur effect and the reduction of blur, without changing the scene content and while keeping the subject in focus? Please watch both videos twice.

Select the better video: ○ Video 1 ○ Video 2

Pair 3

Which video better demonstrates the defocus-blur effect and the reduction of blur, without changing the scene content and while keeping the subject in focus? Please watch both videos twice.

Select the better video: ○ Video 1 ○ Video 2

Figure A7. **Human study interface.** We show the Hugging Face Spaces interface used to conduct the user studies. In each question, one video is generated by our method, while the other is randomly selected from one of the baselines for the same prompt. The participants are tasked with selecting the video that shows the realistic defocus-effect and the least amount of scene content change with decrease in blur as the aperture increases in the video.