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

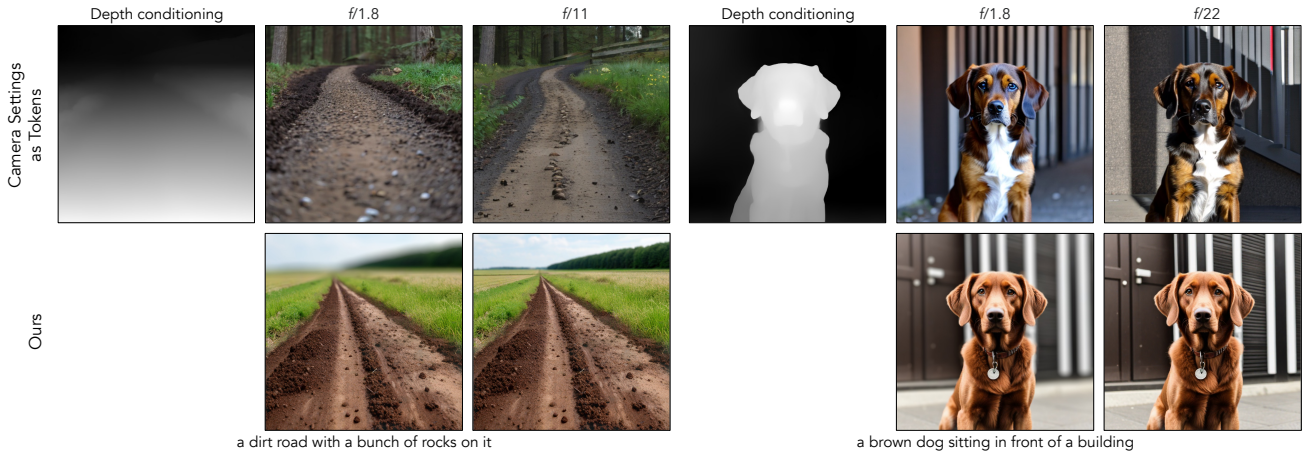

Figure A1. **Comparison with Camera Settings as Tokens [12] based ControlNet [42].** We compare our method to a depth-conditioned ControlNet that uses Camera Settings as Tokens embeddings. While the ControlNet effectively adheres to scene depth, it alters scene content within those depth planes. Notably,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

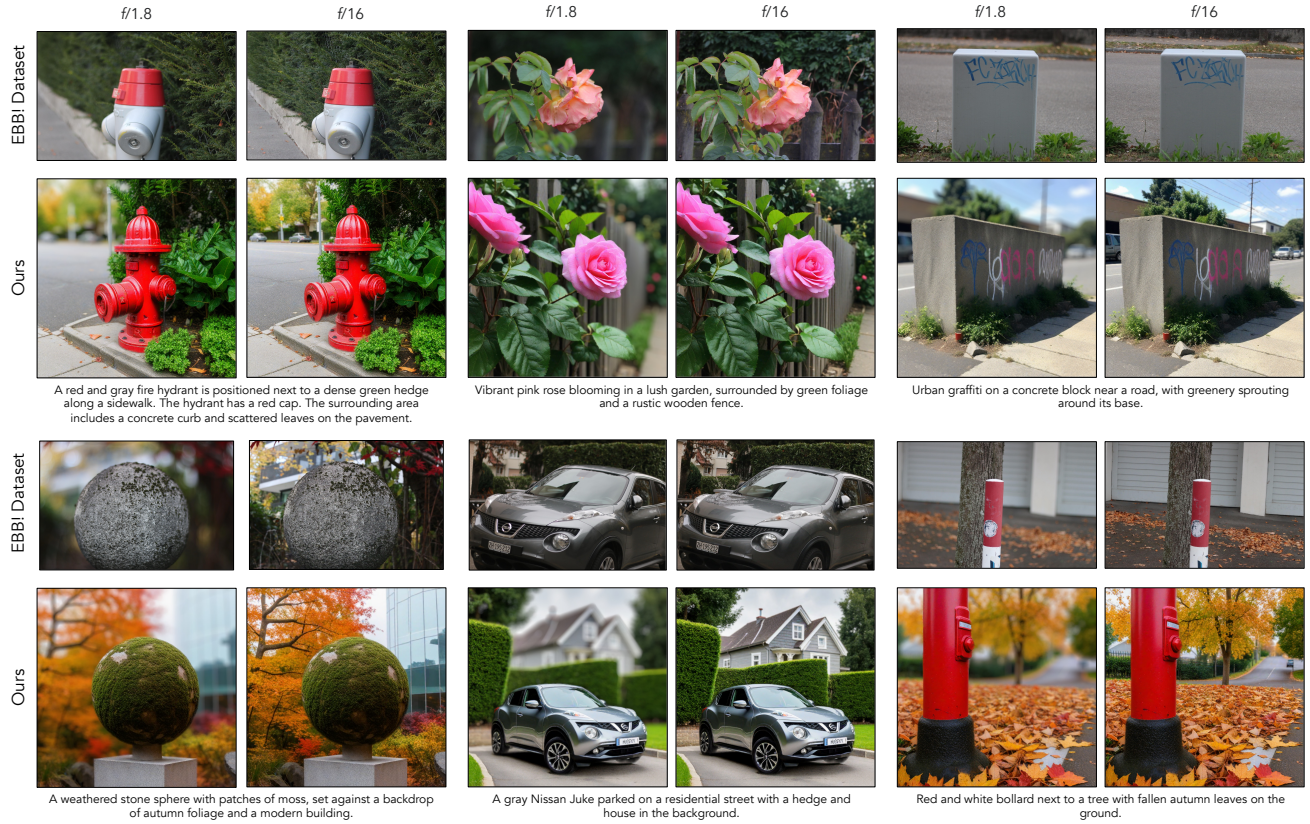

Figure A2. **Comparison to EBB! Dataset.** Using the EBB! dataset [13], which provides image pairs captured at two apertures: $f/16$ (all-in-focus) and $f/1.8$ (shallow depth of field), we first caption the $f/16$ image with the InternVL3 [46] model (shown below the "Ours" images). We then use that caption 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.

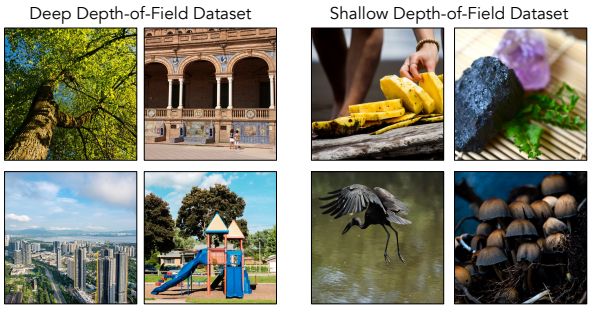

Figure A4. **Deep and Shallow DoF Datasets.** Images shown are selected using our dataset filtering approach mentioned in Sec. D. After filtering, the Deep DoF dataset primarily consists of all-in-focus images, while the Shallow DoF dataset includes images with defocus blur, emphasizing a specific object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

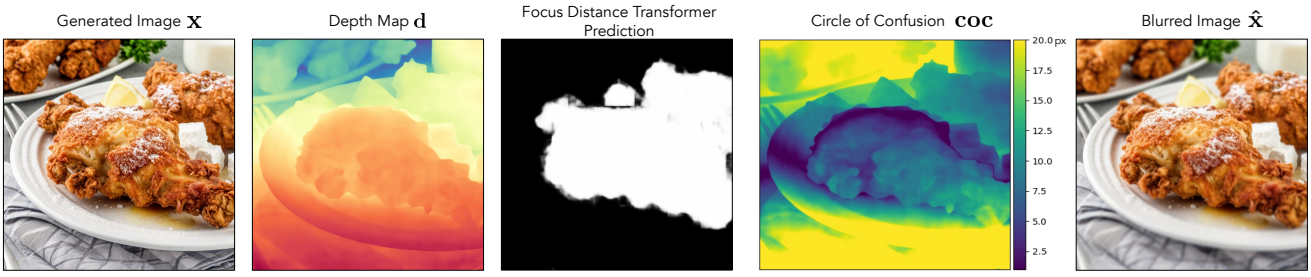

Figure A5. **Image Generation Pipeline.** The pipeline begins with an image generated by the model (left), followed by depth prediction from the depth model. A saliency map is then predicted and used to compute the focus distance as a weighted sum of depth and saliency. The lens model calculates the 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 $\vec{k}$,

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

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

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.

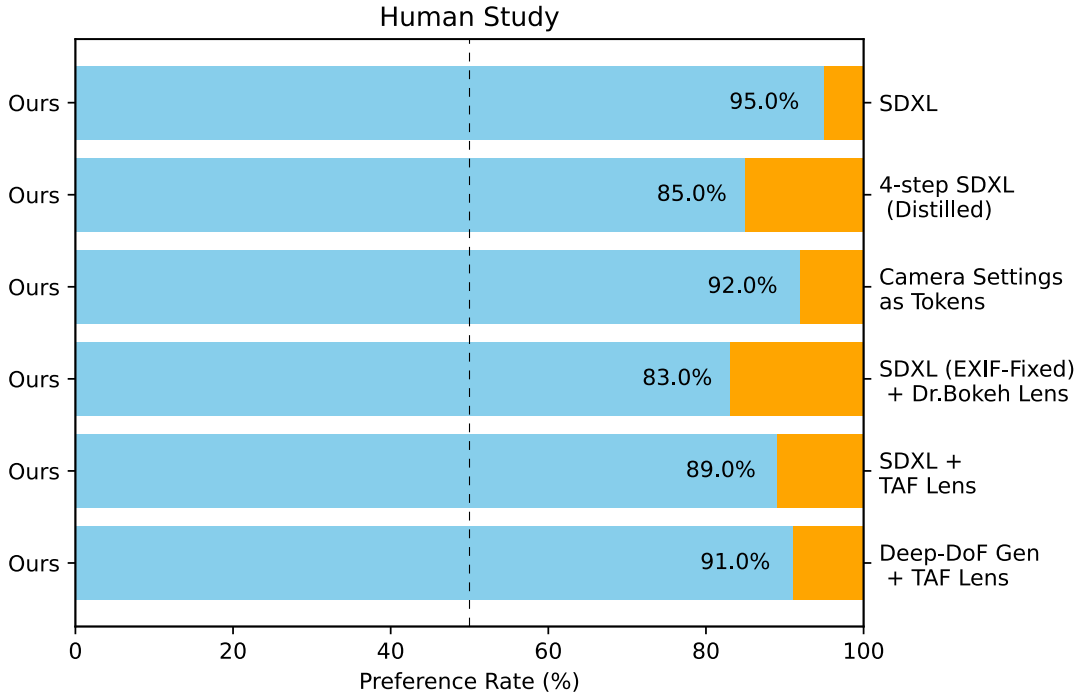

Figure A6. **Human Studies.** Preference rate for selecting our method over the baselines (Section E).

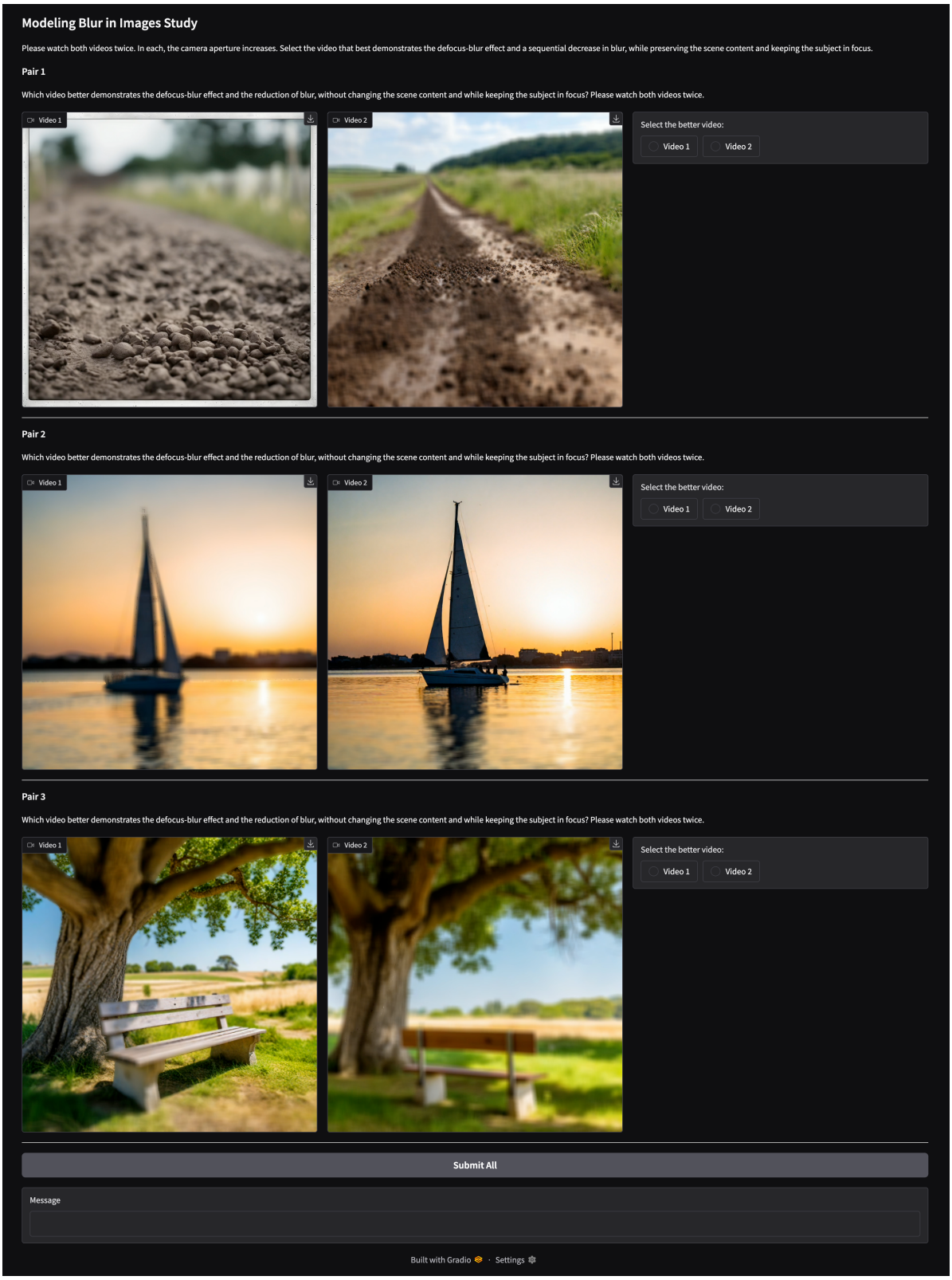

Figure A7. **Human study interface.** We show the Hugging Face Spaces interface used to conduct the user studies. In each question, one video is generated by our method, while the other is randomly selected from one of the baselines for the same prompt. The participants are tasked with selecting the 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.