GENERATIVE REFOCUSING: FLEXIBLE DEFOCUS CONTROL FROM A SINGLE IMAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Depth-of-field control is essential in photography, but achieving perfect focus often requires multiple attempts or specialized equipment. Single-image refocusing is still difficult, requiring both sharp content recovery and realistic bokeh synthesis. Current methods have significant drawbacks: they require all-in-focus inputs, rely on synthetic data from simulators, and offer limited control over the aperture. We introduce Generative Refocusing, built around a training scheme that completes missing signals, converting synthetic renderings, single real bokeh photographs, and paired real captures into a unified training format. This allows us to scale training jointly across heterogeneous real-world sources that prior methods could not combine, letting our model learn authentic optical characteristics beyond what simulators capture while enabling precise control over focus and bokeh intensity. Our experiments show we achieve state-of-the-art performance in defocus deblurring, bokeh synthesis, and refocusing benchmarks.

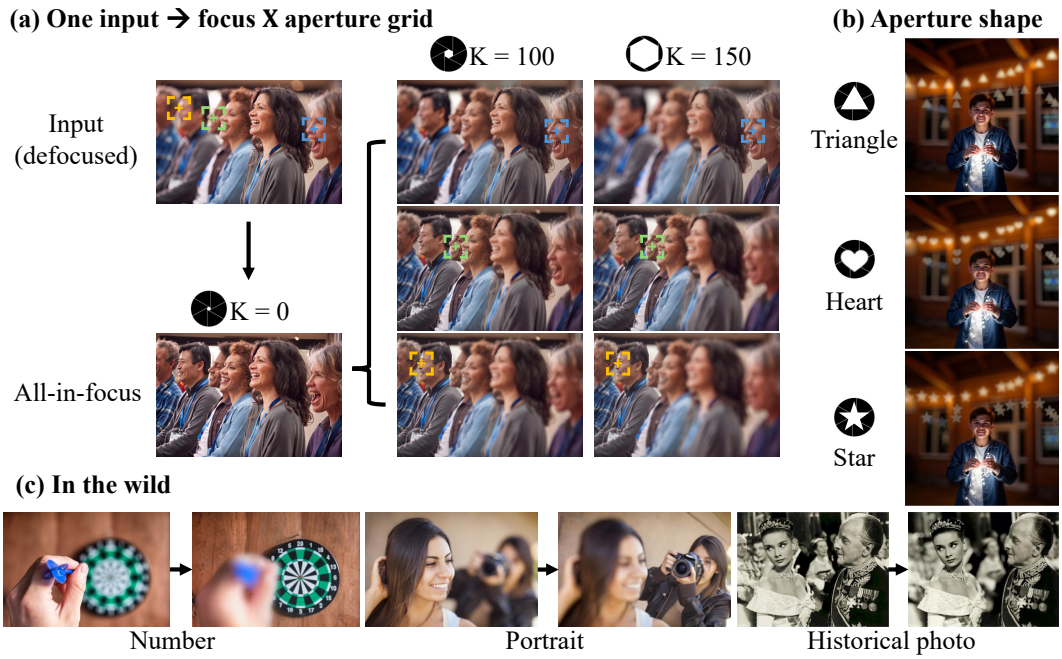


Figure 1: **Generative refocusing with controllable bokeh.** Our model turns a single input image into a virtual camera that users can control. **(a)** Given a single defocused input, our method first recovers an all-in-focus image (K=0) and then generates refocused outputs across increasing bokeh levels K, with each row shifting the focus onto a different subject (colored corner brackets), demonstrating simultaneous focus-plane and aperture-size control from a single input. **(b)** Aperture-shape control synthesizes bokeh shapes—triangle, heart, and star—from point lights in the scene while preserving scene consistency. **(c)** Results on in-the-wild images show that our method generalizes to diverse real-world photos and accurately shifts focus to a user-specified subject.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Controlling depth of field and focus is essential for artistic photography. It helps guide the viewer's eye through selective focus and bokeh effects. Achieving perfect focus often requires multiple shots or specialized equipment, which can be challenging for everyday photographers. Single-image refocusing removes these obstacles by allowing focus and bokeh adjustments after capture. However, it is inherently a two-part problem: one must first recover sharp content from out-of-focus regions, and then render natural-looking bokeh at a user-specified focus plane and blur strength.

Prior work has mostly tackled these two parts separately. Defocus deblurring methods (Ruan et al., 2022; Zamir et al., 2022) aim to recover sharp images, while bokeh synthesis methods (Peng et al., 2022; Seizinger et al., 2025) focus on rendering realistic depth-of-field. While combining both has required dual-camera setups (Alzayer et al., 2023), recent diffusion-based methods (Wang et al., 2025b; Tedla et al., 2025) enable refocusing from a single image.

Despite these advances, three main limitations remain (Tab. 1). First, most methods rely on all-in-focus inputs and accurate depth maps. This precludes their use on images that already contain defocus blur. Second, while synthetic data (Peng et al., 2022; Zhu et al., 2025; Wang et al., 2025b) can be used for training, its realism is limited by simulator fidelity. This approach often fails to capture detailed visual appearances. However, real datasets (Seizinger et al., 2025; Tedla et al., 2025) suffer from a critical limitation: they are restricted to isolated one-dimensional variations (i.e., solely aperture sweeps or focus plane sweeps). Consequently, they do not provide the joint distribution of focus and aperture changes needed to train a model for flexible, dual-parameter refocusing. Third, current methods typically support only aperture size, not its shape. This restriction limits creative bokeh effects.

To tackle these issues, we present **Generative Refocusing**, a flexible single-image refocusing system built on a novel data-completion training scheme. Real-world bokeh data is inherently incomplete: single images lack ground-truth all-in-focus counterparts, while paired captures lack focus-plane and aperture control signals, preventing these heterogeneous sources from being combined at scale. Our key contribution is a training scheme that recovers these missing signals directly from the images themselves and converts every source into a unified format, allowing synthetic renderings, single real bokeh photographs, and paired real captures to be scaled up and trained jointly, rather than in isolation as in prior work. This lets our model inherit the geometric consistency of synthetic data while learning authentic lens characteristics that no single data source alone could provide. Concretely, we decompose refocusing into two modules trained under this scheme: DeblurNet, a diffusion model that recovers a sharp, in-focus image from inputs in any focus state, and BokehNet, which generates customizable bokeh conditioned on user-defined focus planes and bokeh intensity. BokehNet further supports user-specified aperture shapes via a dedicated shape-aware fine-tuning stage (Fig. 1). Our approach achieves state-of-the-art performance in defocus deblurring, bokeh generation, and refocusing while maintaining natural scene consistency.

Our main contributions are:

- A data-completion training scheme that recovers missing focus and bokeh signals, enabling joint training on synthetic and multiple real-world sources previously usable only in isolation.
- A flexible refocusing pipeline that accepts images in any focus state and gives users control over the focus plane, bokeh intensity, and aperture shape.
- Our method is validated on existing benchmarks and our new light-field datasets (LF-Bokeh, LF-Refocus), where it outperforms state-of-the-art baselines across defocus deblurring, bokeh synthesis, and refocusing. It further enables creative aperture-shape control.

2 RELATED WORK

Diffusion Models for Image Restoration. Diffusion models (Ho et al., 2020; Song et al., 2021; Yeh et al., 2024; Tsai et al., 2025) advance image restoration via generative priors, extending to zero-shot restoration and artifact removal. Methods evolved from pixel-space (Saharia et al., 2022) to efficient latent-space approaches (Rombach et al., 2022; Xia et al., 2023), enabling faster inference. ResShift (Yue et al., 2023) cuts steps from 1000 to 4–15 via residual shifting. Deblurring has progressed from merging diffusion priors with regression tasks, such as matting (Wang et al., 2024a) or hierarchical restoration (Chen et al., 2023b), to learning spatially varying kernels in

Table 1: **Comparison across methods.** Each prior method learns from a single data regime, either synthetic renderings or one real sweep, which limits its inputs and controls. Our data completion (Sec. 3.2) trains one model on synthetic data and all three real sources (LFDOF, RealBokeh, ITW), enabling every control from any input. ✓ supported; ✗ not; △ with caveat; – not applicable. *Any focus*: accepts both all-in-focus (AIF) and defocused inputs. DC^2 needs a dual-camera capture; Bokeh Diffusion is text-to-image. ⋆Classical renderer only; not exposed in the public code. †Assumes AIF input; others need fine-tuning. ‡Fixed focal stack, not a chosen focus plane. MagicBokeh is designed for low-quality/defocused zoom inputs and jointly performs super-resolution and bokeh rendering, but its real training data supervises super-resolution rather than bokeh control signals.

	Input		**Training data it can learn from**				**User control**		
			synthetic	*real*					
Method	**Single image**	**Any focus**	**Rendered pairs**	**Focus sweep**	**Aperture sweep**	**Unpaired photos**	**Focus plane**	**Bokeh intensity**	**Aperture shape**
Bokeh rendering									
BokehMe (Peng et al., 2022)	✓	✗	✓	✗	✗	✗	✓	✓	△⋆
Bokehlicious (Seizinger et al., 2025)	✓	△†	✗	✗	✓	✗	✗	✓	✗
BokehDiff (Zhu et al., 2025)	✓	✗	✓	✗	✗	✗	✓	✓	✗
Bokeh Diffusion (Fortes et al., 2025)	–	–	✓	✗	✗	✓	✗	✓	✗
MagicBokeh (Shi et al., 2026)	✓	✓	✓	✗	✗	✗	✓	✓	✗
Refocusing									
DC^2 (Alzayer et al., 2023)	✗	✓	✗	✓	✗	✗	✓	✓	✗
DiffCamera (Wang et al., 2025b)	✓	✓	✓	✗	✗	✗	✓	✓	✗
Learn2Refocus (Tedla et al., 2025)	✓	✗	✗	✓	✗	✗	△‡	✗	✗
GenRefocus (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓

latent space (Kong et al., 2025). Cascaded pipelines (Ho et al., 2022) refine at multiple resolutions, while principled frameworks (Luo et al., 2023; 2025) treat degradations as stochastic processes. Training-free methods (Zhu et al., 2023; Kawar et al., 2022) enable plug-and-play restoration.

Defocus Deblurring. Defocus deblurring recovers sharp images from spatially varying blur. Early deconvolution (Bando & Nishita, 2007; Zhuo & Sim, 2011) introduced artifacts at depth transitions. Dual-pixel sensors (Abuolaim & Brown, 2020), quad-pixel data (Chen et al., 2025), and disparity-aware techniques (Lee et al., 2021; Yang et al., 2023; Wang et al., 2021) improved blur–depth modeling and the recovery of all-in-focus images from focal stacks. Architectures shifted from CNNs (Chen et al., 2022) to transformers (Tsai et al., 2022; Wang et al., 2022), with gains from implicit representations (Quan et al., 2023; Fan et al., 2025) and multi-scale attention (Zhang et al., 2024). Recent work includes vision-language fusion (Yang et al., 2024) for semantic estimation and a shift from supervised (Ren et al., 2025) to diffusion-based learning (Liang et al., 2025; Li et al., 2025), including unpaired data. While these excel as *standalone solutions*, we position deblurring as the *first stage* of refocusing, using FLUX's (Shakker Labs, 2024) generative prior to enable controllable bokeh synthesis.

Bokeh Rendering. Bokeh rendering has progressed from physically based scattering (Potmesil & Chakravarty, 1981; Lee et al., 2008; 2009) and differentiable rendering (Sheng et al., 2024) to neural methods that combine ray tracing with learned enhancements (Peng et al., 2022; 2024). Learning-based approaches evolved from fixed apertures (Ignatov et al., 2020) to variable f-stops (Seizinger et al., 2025; Chen et al., 2024), extending to video (Yang et al., 2025) and 3D scenes (Shen et al., 2025; Wang et al., 2025c; Huang et al., 2025). However, a major limitation is that they fail to synthesize realistic results when the input images already contain defocus blur. Meanwhile, recent diffusion models have enabled advanced bokeh control (Fortes et al., 2025; Yuan et al., 2025; Shrivastava et al., 2026). Yet, these methods are primarily designed for text-to-image synthesis and cannot be readily applied to image editing tasks.

Single-Image Refocusing. Post-capture refocusing evolved from deconvolution (Bando & Nishita, 2007; Zhang & Cham, 2011) and GANs (Sakurikar et al., 2018) to recent diffusion models (Qin et al., 2025), yet significant limitations remain (see Tab. 1). Constrained by conditioning on explicit focus coordinates, DiffCamera (Wang et al., 2025b) requires fixed-resolution inputs (512×512) and suffers performance drops on arbitrary aspect ratios. Furthermore, it inherits simulator artifacts from its purely synthetic training. Learn2Refocus (Tedla et al., 2025) trains on real data but lacks aperture-size control and generates a focal stack at predefined depth planes that may not align with the user's target focus. Additionally, 3D representations (Shen et al., 2025; Wang et al., 2025c; Lee et al., 2024; Liu et al., 2023) offer controllable DoF but require multi-view capture.

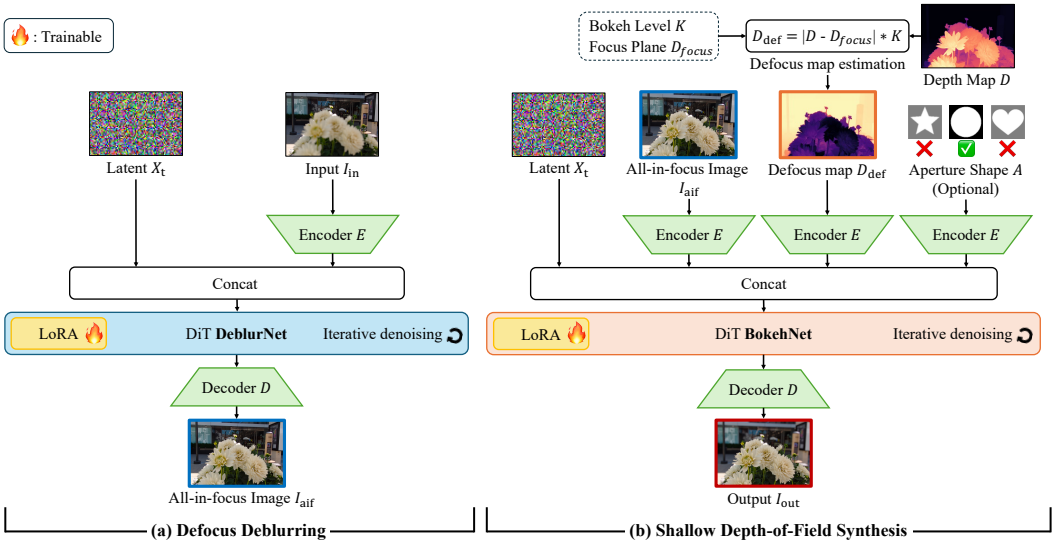


Figure 2: **Pipeline Overview.** Our method decomposes single-image refocusing into two stages: **(a) Defocus Deblurring** and **(b) Shallow Depth-of-Field Synthesis**. **(a)** The noisy latent X_t and the encoded blurry input I_{in} are concatenated into a token sequence. DeblurNet denoises this sequence to reconstruct an all-in-focus image I_{aif}. **(b)** Following the same conditioning design, BokehNet conditions on I_{aif} and the defocus map D_{def} to synthesize the refocused output I_{out}, with an optional aperture-shape condition A. The defocus map D_{def} is computed from the estimated depth map D (Bochkovskii et al., 2025) together with the user-specified focus plane D_{focus} and bokeh level K.

Camera-Conditioned Diffusion. Conditioning diffusion models on camera parameters progressed from extrinsic control for view synthesis (Höllein et al., 2024; Cheng et al., 2024) and video generation (He et al., 2024; Wang et al., 2024b), to intrinsic control of focal length, aperture, and exposure (Fang et al., 2024; Chen et al., 2023a). Direct conditioning often yields inconsistent content. Recent work uses temporal modeling (Yuan et al., 2025), explicit calibration (Bernal-Berdun et al., 2025; Deng et al., 2025), or transformer-based pose handling (Bahmani et al., 2024; Saxena et al., 2023). These focus on *generation*; we operate in *editing*, manipulating images. By *decoupling* deblurring and bokeh stages, we achieve single-image manipulation without multi-frame overhead.

3 Method

3.1 Refocusing Pipeline

We propose Generative Refocusing (GenRefocus), which decomposes the refocusing task into defocus deblurring and bokeh synthesis, realized by two models, DeblurNet and BokehNet, that share a unified conditioning design. At each denoising step t, both networks concatenate the noisy image latent tokens X_t with condition tokens encoded by a pretrained VAE encoder $\mathcal{E}$, and decode the denoised latent with the VAE decoder $\mathcal{D}$ to obtain the target image (see Fig. 2). Such a flexible pipeline allows us to exert greater control over input conditions (see Tab. 1).

Defocus deblurring. Given a blurry input image I_{in}, DeblurNet processes the token sequence $\mathbf{S}_t = [\, X_t \,;\, \mathcal{E}(I_{\text{in}}) \,]$, and after iterative denoising, decodes the image latent part to obtain the all-in-focus output I_{aif} (Fig. 2 (a)). We train DeblurNet in a supervised manner using real paired data (Abuolaim & Brown, 2020; Seizinger et al., 2025).

Shallow depth-of-field synthesis. Given I_{aif} and a user-specified focus plane D_{focus} with bokeh level K, we define the defocus map

$$D_{\text{def}} = K \cdot |D - D_{\text{focus}}| \,, \tag{1}$$

where D is the monocular depth map estimated from I_{aif} using an off-the-shelf depth estimator (Bochkovskii et al., 2025), notation and thin-lens derivation in supplementary material Sec. M. Following the same conditioning design, BokehNet concatenates the condition tokens of I_{aif} and D_{def} with the noisy latent tokens at each denoising step, and denoises X_t to synthesize the refocused output I_{out} (Fig. 2 (b)).

3.2 DATA-COMPLETION TRAINING SCHEME

Training BokehNet requires triplets $(I_{\text{aif}}, I_{\text{out}}, D_{\text{def}})$. Synthetic rendering provides all three, and we use it to pretrain BokehNet, but it inherits simulator bias and lacks real-lens effects such as chromatic aberration and optical vignetting. We therefore fine-tune on real bokeh images, which, however, cannot be used directly: single images have no all-in-focus (AIF) counterpart, and paired captures lack both the focus plane and the bokeh level K. We introduce a training scheme that completes these missing signals rather than discarding such data, turning every real image into a valid training sample (Fig. 3).

(a) Synthetic data. We pretrain on synthetic data first because it provides exact supervision: D_{def} matches the rendered target by construction, and (D_{focus}, K) can be sampled freely, whereas signals completed for real images are only estimates. Starting from real AIF images with estimated depth D, we randomly sample a focus plane D_{focus} and a bokeh level K (Fig. 3 (a)), construct D_{def} via Eq. 1, and render the target bokeh image with a simulator (Peng et al., 2022). This teaches BokehNet to modulate the circle of confusion according to D_{def}, but its realism remains bounded by the renderer, which motivates the real data below.

(b) Single real bokeh images: missing AIF input and focus plane. A single real bokeh image (e.g., ITW (Fortes et al., 2025)) comes with EXIF metadata but no ground-truth AIF counterpart, so it cannot form a training pair directly (Fig. 3 (b)). We instead treat the image as the target I_{out} and estimate I_{aif} with DeblurNet. Since EXIF focus distance is often missing or noisy, we estimate D_{focus} as the median depth inside an in-focus mask M from BiRefNet (Zheng et al., 2024):

$$D_{\text{focus}} = \text{median}\big(D[M]\big), \tag{2}$$

and compute K from EXIF following (Fortes et al., 2025):

$$K \;\approx\; \frac{f^2\, D_{\text{focus}}}{2\,F\,(D_{\text{focus}} - f)} \times \text{pixel_ratio}, \tag{3}$$

where f is the focal length, F the f-number, and pixel_ratio accounts for sensor size and resolution. With D_{focus} and K estimated, D_{def} follows from Eq. 1, the same formulation as the other sources. **Without requiring ground-truth AIF pairs**, BokehNet can thus learn authentic lens characteristics that simulators cannot reproduce, and the noisy estimated inputs match the test-time pipeline, improving robustness.

(c) Paired real captures: missing focus plane and K. Paired datasets such as LFDOF (Ruan et al., 2021) and RealBokeh (Seizinger et al., 2025) provide AIF–bokeh pairs, but no focus plane, and their EXIF is absent or insufficient to estimate K (Fig. 3 (c)). For the focus plane, we obtain an initial in-focus mask M with BiRefNet (Zheng et al., 2024), which is sometimes unreliable on these more diverse scenes. Rather than discarding such samples, we *re-select a small yet reliable in-focus region* to correct M and extract D_{focus}, which retains complex scenes that would otherwise be filtered out and broadens the diversity of real training data. For the bokeh level, we use a *simulator-in-the-loop calibration*: we sweep K and pick the value whose rendering best matches the real bokeh target I_{real},

$$K^\star = \underset{K \in (K_{\min},\, K_{\max})}{\text{argmax}} \; \text{SSIM}\big(R(I_{\text{aif}}, D;\, D_{\text{focus}}, K),\; I_{\text{real}}\big), \tag{4}$$

where R is the bokeh simulator used in Sec. 3.2 (a). We keep $K^\star$ as a pseudo-label only if its SSIM exceeds a threshold, ensuring reliable supervision and allowing real pairs to use the same D_{def} formulation as the other sources.

3.3 BOKEH-SHAPE AWARE SYNTHESIS

State-of-the-art learning-based bokeh synthesis methods deliver compelling results but *do not expose* aperture-shape control. Physics-based renderers, in principle, support arbitrary shapes, yet public implementations typically omit this functionality. We therefore explore explicit *shape-aware* control using only a bokeh shape image within a trainable framework.

3.3.1 SIMULATOR AND DATA

Real photographs exhibiting diverse bokeh shapes are rare, and paired AIF–bokeh examples are even scarcer. We thus rely on simulation. A key observation is that when an all-in-focus (AIF) image

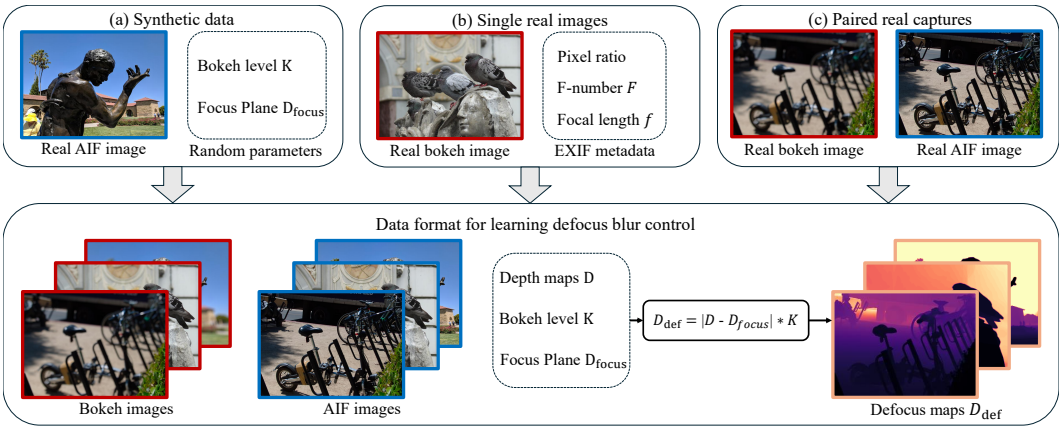


Figure 3: **Training data generation.** Every source is converted into the same training format: a bokeh image, an AIF image, a depth map D, a bokeh level K, and a focus plane D_{focus}. (a) Synthetic: D_{focus} and K are sampled, and the bokeh image is rendered. (b) Single real images miss the AIF input and focus plane: DeblurNet recovers the AIF image, and D_{focus} is estimated from an in-focus mask and depth. (c) Paired real captures miss the focus plane and K: D_{focus} comes from a refined in-focus mask, and K from simulator-in-the-loop calibration (Eq. 4). Rather than discarding such incomplete real data, we recover these missing signals, turning every source into a valid training sample with authentic optical characteristics that simulators cannot reproduce.

lacks point-light stimuli, the simulated bokeh carries weak shape cues, making the model reluctant to learn shape-conditioned responses. To address this, we synthesize a **PointLight-1K** dataset designed to reveal aperture shape. We also extend the classical BokehMe renderer to scatter through a given shape: given an AIF image I_{aif}, depth D, focus-plane proxy D_{focus}, bokeh level K, and a shape kernel s (binary/raster PSF), the simulator R renders

$$I_{\text{syn}} \;=\; \mathcal{R}\big(I_{\text{aif}}, D;\, D_{\text{focus}}, K, s\big), \tag{5}$$

yielding paired supervision for shape-aware training.

3.3.2 Shape-conditioned fine-tuning

To incorporate the aperture shape A, we append its tokens directly to the unified sequence. To ensure that shape edits do not undo the learned bokeh synthesis, we freeze the original LoRA weights and introduce a trainable LoRA module. Only this LoRA is fine-tuned to handle the shape conditioning.

3.4 Evaluation Datasets: LF-Bokeh and LF-Refocus

Existing real-world datasets provide only isolated focus or aperture sweeps and lack ground-truth depth maps and reference bokeh levels K, which makes it difficult to jointly evaluate fidelity and controllability. We therefore construct two benchmarks from light-field captures (Rerabek & Ebrahimi, 2016; Dansereau et al., 2019). **(a) LF-Bokeh.** LF-Bokeh contains about 200 images with diverse focus planes and aperture sizes, synthesized from light-field captures. It is used to evaluate bokeh synthesis in terms of both fidelity and controllability. **(b) LF-Refocus.** LF-Refocus consists of 400 source–target pairs. Each pair comprises a source image focused at a near depth plane and a target focused at the far depth plane of the same scene, all rendered from the central view. It is used to evaluate end-to-end single-image refocusing.

4 Experiments

4.1 Setup

Backbone and training details. Our backbone is FLUX-1-dev (Shakker Labs, 2024), fine-tuned via LoRA (Hu et al., 2021) with a conditioning scheme following (Tan et al., 2025a;b). DeblurNet employs LoRA rank r=128, while BokehNet uses r=64. DeblurNet is trained for 60K steps. BokehNet is trained in two stages: (i) 40K steps on synthetic data, and (ii) 60K steps on real data.

Datasets. We train DeblurNet on 3.5K pairs sourced from DPDD (Abuolaim & Brown, 2020) and a subset of RealBokeh. For BokehNet, to maintain realism and controllability, we use a hybrid dataset comprising ∼70K synthetic pairs from (Yuan et al., 2025; Ignatov et al., 2020), combined with approximately 26K real examples from the ITW dataset (Fortes et al., 2025), RealBokeh (Seizinger et al., 2025), and LFDOF (Ruan et al., 2021).

Table 2: **Defocus deblurring benchmark on RealDOF (Lee et al., 2021) and DPDD (Abuolaim & Brown, 2020) datasets.** We report reference (LPIPS, DISTS) and no-reference (CLIP-IQA, MANIQA, MUSIQ) perceptual metrics. Best, second best, and third best results are highlighted.

	RealDOF (Lee et al., 2021)					DPDD (Abuolaim & Brown, 2020)				
Method	LPIPS ↓	DISTS ↓	CLIP-IQA ↑	MANIQA ↑	MUSIQ ↑	LPIPS ↓	DISTS ↓	CLIP-IQA ↑	MANIQA ↑	MUSIQ ↑
Input	0.5241	0.2865	0.3562	0.2213	28.7087	0.3485	0.1827	0.4337	0.3325	45.5376
DRBNet (Ruan et al., 2022)	0.2550	0.1312	0.3889	0.2609	38.9014	0.1819	0.1063	0.4142	0.3254	46.7229
Restormer (Zamir et al., 2022)	0.2863	0.1573	0.4062	0.2642	40.8355	0.1762	0.1204	0.4332	0.3307	47.9872
INIKNet (Quan et al., 2023)	0.2851	0.1615	0.3984	0.2580	38.9959	0.1860	0.1235	0.4290	0.3265	47.1961
Bokehlicious (Seizinger et al., 2025)	0.2080	0.1070	0.4281	0.2746	42.2314	0.1598	0.0918	0.4322	0.3237	47.6057
DiffCamera (Wang et al., 2025b)	0.4523	0.2140	0.3861	0.2667	37.9951	0.4048	0.2141	0.4438	0.3350	47.1452
GenRefocus (Ours)	0.2408	0.1126	0.4595	0.2884	43.5222	0.1440	0.0772	0.4755	0.3452	49.4122

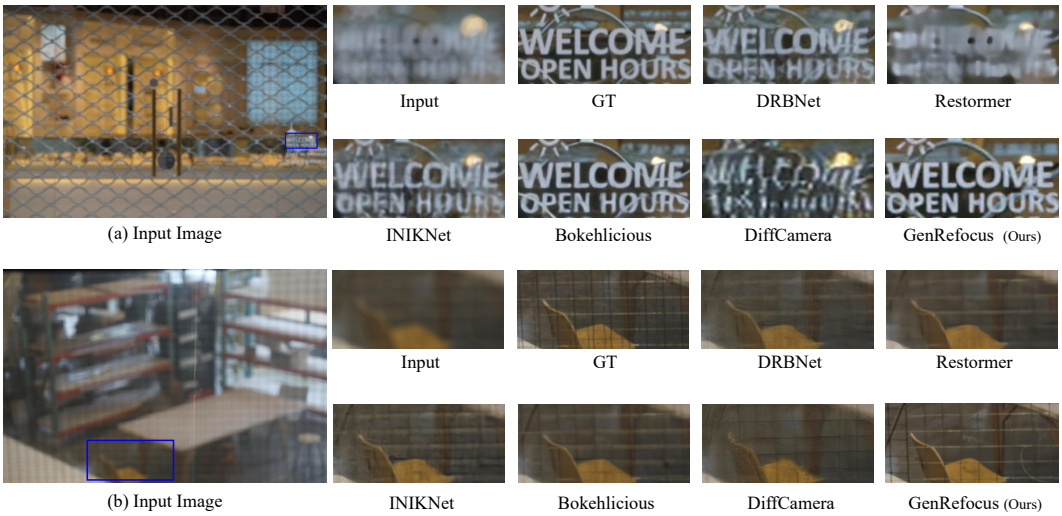


Figure 4: **Qualitative comparison of defocus deblurring.** Visual results on **(a)** RealDOF (Lee et al., 2021) and **(b)** DPDD (Abuolaim & Brown, 2020) datasets. Blue boxes on the left indicate cropped regions shown in detail. In (a), most methods fail to resolve the text, whereas both our method and Bokehlicious (Seizinger et al., 2025) successfully recover the words "WELCOME" and "OPEN HOURS". However, Bokehlicious produces noticeable distortions in the right-side characters (e.g., "ME" and "RS"), whereas our approach faithfully restores the original text structure. For the challenging example (b) with severe defocus blur, all baseline methods remain highly blurry and fail to recover meaningful details. In contrast, our method synthesizes a relatively clear and visually compelling output.

Benchmarks and Metrics. We evaluate defocus deblurring on RealDOF and DPDD, bokeh synthesis on LF-Bokeh, and refocusing on LF-Refocus. Since ground-truth depth maps and reference bokeh levels K are unavailable, for fidelity on LF-Bokeh and LF-Refocus, we follow (Peng et al., 2022; Zhu et al., 2025) and run a per-image binary search over K that maximizes SSIM with the target. To quantify controllability, we adopt LVCorr (Yuan et al., 2025; Fortes et al., 2025; Wang et al., 2025b): for images generated from the same all-in-focus input with a fixed focus plane across varying K, we compute the Pearson correlation between K and the Laplacian variance trend of the outputs. Further details are provided in the supplementary material Sec. D.

4.2 Results

Defocus deblurring. As reported in Tab. 2, DeblurNet outperforms competing methods on most reference and no-reference metrics. Notably, RealDOF is an out-of-distribution benchmark whose blur differs substantially from our DPDD-based training data, yet our method leads on all no-reference metrics, demonstrating strong generalization. Among diffusion-based methods, we outperform DiffCamera (Wang et al., 2025b) by a large margin on both datasets. Other diffusion-based deblurring methods are excluded: Swin-Diff (Liang et al., 2025) and (Li et al., 2025) have no public code, and DeblurDiff (Kong et al., 2025) targets motion rather than defocus deblurring. Although our reference scores on RealDOF (Lee et al., 2021) are below Bokehlicious (Seizinger et al., 2025), our results are more visually appealing and more robust to diverse inputs such as grayscale photographs and low-quality, blurry images (see supplementary material). Qualitatively, Fig. 4(a) shows that our

Table 3: **Bokeh synthesis benchmark.** Evaluation on LF-Bokeh. GenRefocus achieves the best fidelity on all three metrics. In terms of controllability, it is second only to BokehMe, whose classical physically based renderer inherently guarantees near-perfect control; it ties with MagicBokeh and outperforms the other diffusion-based methods (BokehDiff, Bokeh Diffusion, and DiffCamera). "-" indicates not evaluated: Bokehlicious is controlled by f-stop rather than K, so LVCorr, which is defined over K, does not apply.

	Fidelity			Controllability
Method	LPIPS ↓	DISTS ↓	CLIP-I ↑	LVCorr ↑
BokehMe (Peng et al., 2022)	0.1235	0.0747	0.9509	0.9940
Bokehlicious (Seizinger et al., 2025)	0.2105	0.1038	0.9355	-
BokehDiff (Zhu et al., 2025)	0.1715	0.0923	0.9471	0.8510
Bokeh Diffusion (Fortes et al., 2025)	0.4929	0.2488	0.7487	0.8020
DiffCamera (Wang et al., 2025b)	0.1405	0.0737	0.9450	0.5816
MagicBokeh (Shi et al., 2026)	0.1886	0.1102	0.9160	0.9914
GenRefocus (Ours)	0.0816	0.0490	0.9711	0.9914

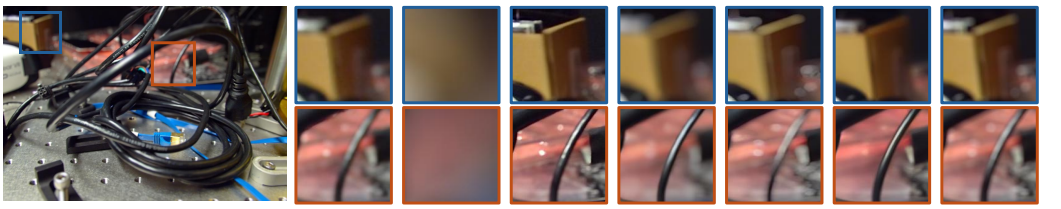


Figure 5: **Qualitative comparison on bokeh synthesis benchmark.** Results on LF-Bokeh with zoomed patches (blue and orange boxes) highlighting detail quality. Our method synthesizes bokeh effects that better match ground truth with realistic blur gradients and natural occlusion handling.

approach faithfully recovers text with well-preserved edges. Fig. 4(b) is a particularly challenging case: competing methods recover only a blurry dark stripe and introduce background distortions, whereas DeblurNet reconstructs more consistent content.

Bokeh synthesis. As presented in Tab. 3, our approach outperforms all other methods across fidelity metrics. In terms of controllability, which measures how well the generated bokeh aligns with the conditioned bokeh level K, BokehMe (Peng et al., 2022) achieves near-perfect control because its hybrid design incorporates a classical, physically motivated renderer. Nevertheless, our method matches MagicBokeh and outperforms the other diffusion-based methods, demonstrating that our learning strategy effectively preserves both realism and accuracy. We additionally conduct a user study with 20 participants evaluating sharp-to-defocused progressions on 4 in-the-wild scenes, following the protocol of (Fortes et al., 2025). Our method is preferred over Bokehlicious, BokehMe, and DiffCamera in 86.2%, 77.5%, and 98.8% of comparisons, respectively, corroborating our quantitative results. Visual results are provided in the supplementary material Sec. E.

Refocusing. As shown in Tab. 4, our approach achieves the best results on both fidelity and perceptual metrics and outperforms other baselines. Qualitatively, Fig. 6 shows that competing methods tend to produce spatially ambiguous blur and fail to place the focal plane correctly, resulting in uniformly soft content. In contrast, our approach restores fine details at the intended focus while synthesizing physically consistent bokeh elsewhere.

Aperture shape control. After fine-tuning BokehNet on PointLight-1K, the model synthesizes bokeh that follows user-specified aperture shapes, as shown by the heart-shaped bokeh in Fig. 1(b).

4.3 Ablation Study

BokehNet training ablation. To quantify the impact of different training data distributions, we evaluate BokehNet training under various dataset combinations on LF-Bokeh. We consider three primary data sources: (a) synthetic data, (b) the ITW dataset (Fortes et al., 2025), and (c) LFDOF (Ruan et al., 2021) and RealBokeh (Seizinger et al., 2025). As shown in Tab. 5, a baseline trained purely on (a) provides a foundational performance but leaves room for improvement. Introducing either (b) or (c) to the synthetic baseline noticeably enhances performance.

Table 4: **Quantitative comparison on refocusing benchmark.** Results on our LF-Refocus dataset. GenRefocus achieves the best results on all fidelity and perceptual metrics. The Bokehlicious→BokehMe baseline chains an existing deblurring method (Seizinger et al., 2025) with an existing bokeh renderer (Peng et al., 2022)

Method	LPIPS ↓	DISTS ↓	CLIP-I ↑	CLIP-IQA ↑	MUSIQ ↑
Bokehlicious (Seizinger et al., 2025)→BokehMe (Peng et al., 2022)	0.1469	0.0913	0.9364	0.4543	44.1177
DiffCamera (Wang et al., 2025b)	0.1426	0.0782	0.9401	0.4669	40.5869
Learn2Refocus (Tedla et al., 2025)	0.1706	0.0967	0.9322	0.4709	38.3440
MagicBokeh (Shi et al., 2026)	0.1996	0.1176	0.9029	0.5537	46.9067
GenRefocus (Ours)	0.1324	0.0759	0.9571	0.5803	52.9706



Figure 6: **Qualitative comparison on the refocusing benchmark.** Results on LF-Refocus (blue and orange boxes show zoomed patches). DiffCamera (Wang et al., 2025b) loses sharp details and misassigns background blur; Learn2Refocus (Tedla et al., 2025) misplaces the focal plane. Ours focuses accurately and renders realistic bokeh closest to the ground truth.

Table 5: **Ablation on BokehNet training datasets.** Comparing combinations of datasets: (a) Synthetic data, (b) ITW dataset (Fortes et al., 2025), and (c) LFDOF (Ruan et al., 2021) and RealBokeh (Seizinger et al., 2025) on LF-Bokeh. The results show that combining all three yields the best optical characteristics.

Datasets			Metrics		
(a) Synthetic	(b) ITW	(c) LFDOF & RealBokeh	LPIPS ↓	DISTS ↓	CLIP-I ↑
✓	-	-	0.1321	0.0755	0.9451
✓	✓	-	0.1161	0.0666	0.9531
✓	-	✓	0.0955	0.0559	0.9667
✓	✓	✓	**0.0816**	**0.0490**	**0.9711**



Figure 7: **Failure case: transparent surfaces.** With a transparent glass, focusing on the foreground may fail to correctly blur the background due to unreliable defocus maps.



Figure 8: **Failure case: hallucinated details.** With severely blurred inputs, our model may hallucinate incorrect results.

5 Conclusion

We present *Generative Refocusing*, a two-stage diffusion framework that turns a single image into a controllable virtual camera, decoupling DeblurNet and BokehNet to enable intuitive control over focus plane, bokeh intensity, and aperture shape, while learning optical behavior beyond what simulators can capture.

Limitations. Our method relies on an estimated defocus map, so depth errors (e.g., transparent surfaces, Fig. 7) can misassign blur, and severe input blur may cause hallucinated details (Fig. 8). Future work should generalize to richer aperture vocabularies, including user-drawn designs.

REFERENCES

Abdullah Abuolaim and Michael S. Brown. Defocus deblurring using dual-pixel data. In *ECCV*, 2020.

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 *CVPR*, 2023.

Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. *arXiv preprint arXiv:2407.12781*, 2024.

Yosuke Bando and Tomoyuki Nishita. Towards digital refocusing from a single photograph. In *PG*, 2007.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In *ICML*, 2023.

Edurne Bernal-Berdun, Ana Serrano, Belen Masia, Matheus Gadelha, Yannick Hold-Geoffroy, Xin Sun, and Diego Gutierrez. Precisecam: Precise camera control for text-to-image generation. In *CVPR*, 2025.

Alexey Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *ICLR*, 2025.

Hang Chen, Yin Xie, Xiaoxiu Peng, Lihu Sun, Wenkai Su, Xiaodong Yang, and Chengming Liu. Quad-pixel image defocus deblurring: A new benchmark and model. In *CVPR*, 2025.

Kang Chen, Shijun Yan, Aiwen Jiang, Han Li, and Zhifeng Wang. Variable aperture bokeh rendering via customized focal plane guidance. *arXiv preprint arXiv:2410.14400*, 2024.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *ECCV*, 2022.

Su-Kai Chen, Hung-Lin Yen, Yu-Lun Liu, Min-Hung Chen, Hou-Ning Hu, Wen-Hsiao Peng, and Yen-Yu Lin. Learning continuous exposure value representations for single-image hdr reconstruction. In *ICCV*, 2023a.

Zheng Chen, Yulun Zhang, Ding Liu, Jinjin Gu, Linghe Kong, Xin Yuan, et al. Hierarchical integration diffusion model for realistic image deblurring. In *NeurIPS*, 2023b.

Ta-Ying Cheng, Matheus Gadelha, Thibault Groueix, Matthew Fisher, Radomir Mech, Andrew Markham, and Niki Trigoni. Learning continuous 3d words for text-to-image generation. In *CVPR*, 2024.

Donald G. Dansereau, Bernd Girod, and Gordon Wetzstein. Liff: Light field features in scale and depth. In *CVPR*, 2019.

Junyuan Deng, Wei Yin, Xiaoyang Guo, Qian Zhang, Xiaotao Hu, Weiqiang Ren, Xiao-Xiao Long, and Ping Tan. Boost 3d reconstruction using diffusion-based monocular camera calibration. In *ICCV*, 2025.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.

Cheng-De Fan, Chen-Wei Chang, Yi-Ruei Liu, Jie-Ying Lee, Jiun-Long Huang, Yu-Chee Tseng, and Yu-Lun Liu. Spectromotion: Dynamic 3d reconstruction of specular scenes. In *CVPR*, 2025.

I-Sheng Fang, Yue-Hua Han, and Jun-Cheng Chen. Camera settings as tokens: Modeling photography on latent diffusion models. In *SIGGRAPH Asia*, 2024.

Armando Fortes, Tianyi Wei, Shangchen Zhou, and Xingang Pan. Bokeh diffusion: Defocus blur control in text-to-image diffusion models. In *SIGGRAPH Asia*, 2025.

Google DeepMind. Gemini models: Product overview. https://deepmind.google/technologies/gemini/, 2025. Accessed: November 21, 2025.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.

Lukas Höllein, Aljaž Božič, Norman Müller, David Novotny, Hung-Yu Tseng, Christian Richardt, Michael Zollhöfer, and Matthias Nießner. Viewdiff: 3d-consistent image generation with text-to-image models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5043–5052. IEEE, 2024.

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.

Yachuan Huang, Xianrui Luo, Qiwen Wang, Liao Shen, Jiaqi Li, Huiqiang Sun, Zihao Huang, Wei Jiang, and Zhiguo Cao. Bokehflow: Depth-free controllable bokeh rendering via flow matching. *arXiv preprint arXiv:2511.15066*, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Andrey Ignatov, Jagruti Patel, and Radu Timofte. Rendering natural camera bokeh effect with deep learning. In *CVPRW*, 2020.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *NeurIPS*, 2022.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, 2021.

Lingshun Kong, Dongqing Zou, Fu Lee Wang, Jimmy Ren, Xiaohe Wu, Jiangxin Dong, Jinshan Pan, et al. Deblurdiff: Real-world image deblurring with generative diffusion models. In *NeurIPS*, 2025.

Byeonghyeon Lee, Howoong Lee, Xiangyu Sun, Usman Ali, and Eunbyung Park. Deblurring 3d gaussian splatting. In *ECCV*, 2024.

Junyong Lee, Hyeongseok Son, Jaesung Rim, Sunghyun Cho, and Seungyong Lee. Iterative filter adaptive network for single image defocus deblurring. In *CVPR*, 2021.

Sungkil Lee, Gerard Jounghyun Kim, and Seungmoon Choi. Real-time depth-of-field rendering using point splatting on per-pixel layers. In *Computer Graphics Forum*, volume 27, pp. 1955–1962, 2008.

Sungkil Lee, Elmar Eisemann, and Hans-Peter Seidel. Depth-of-field rendering with multiview synthesis. *ACM TOG*, 28(5):1–6, 2009.

Yuhao Li, Haoran Fang, Xiang Lei, Qi Wang, Gang Hu, Jiaqing Dong, Zilong Li, Jiabin Lin, Qiegen Liu, and Xianlin Song. Real-world defocus deblurring via score-based diffusion models. *Scientific Reports*, 15(1):22942, 2025.

Hanyan Liang, Shuyao Chai, Xixuan Zhao, and Jiangming Kan. Swin-diff: a single defocus image deblurring network based on diffusion model. *Complex & Intelligent Systems*, 11(3):170, 2025.

Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. Robust dynamic radiance fields. In *CVPR*, 2023.

Wenyang Luo, Haina Qin, Zewen Chen, Libin Wang, Dandan Zheng, Yuming Li, Yufan Liu, Bing Li, and Weiming Hu. Visual-instructed degradation diffusion for all-in-one image restoration. In *CVPR*, 2025.

Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Image restoration with mean-reverting stochastic differential equations. In *ICML*, 2023.

Juewen Peng, Zhiguo Cao, Xianrui Luo, Hao Lu, Ke Xian, and Jianming Zhang. Bokehme: When neural rendering meets classical rendering. In *CVPR*, 2022.

Juewen Peng, Zhiguo Cao, Xianrui Luo, Ke Xian, Wenfeng Tang, Jianming Zhang, and Guosheng Lin. Bokehme++: Harmonious fusion of classical and neural rendering for versatile bokeh creation. *IEEE TPAMI*, 2024.

Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Michael Potmesil and Indranil Chakravarty. A lens and aperture camera model for synthetic image generation. *SIGGRAPH*, 1981.

Xinran Qin, Zhixin Wang, Fan Li, Haoyu Chen, RenJing Pei, WenBo Li, and XiaoChun Cao. Camedit: Continuous camera parameter control for photorealistic image editing. In *NeurIPS*, 2025.

Yuhui Quan, Xin Yao, and Hui Ji. Single image defocus deblurring via implicit neural inverse kernels. In *ICCV*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Dongwei Ren, Xinya Shu, Yu Li, Xiaohe Wu, Jin Li, and Wangmeng Zuo. Reblurring-guided single image defocus deblurring: A learning framework with misaligned training pairs. *IJCV*, pp. 1–18, 2025.

Martin Rerabek and Touradj Ebrahimi. New light field image dataset. In *8th international conference on quality of multimedia experience (QoMEX)*, 2016.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Lingyan Ruan, Bin Chen, Jizhou Li, and Miu-Ling Lam. Aifnet: All-in-focus image restoration network using a light field-based dataset. *IEEE TCI*, 7:675–688, 2021.

Lingyan Ruan, Bin Chen, Jizhou Li, and Miu-Ling Lam. Learning to deblur using light field generated and real defocus images. In *CVPR*, 2022.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE TPAMI*, 45(4):4713–4726, 2022.

Parikshit Sakurikar, Ishit Mehta, Vineeth N Balasubramanian, and PJ Narayanan. Refocusgan: Scene refocusing using a single image. In *ECCV*, 2018.

Saurabh Saxena, Junhwa Hur, Charles Herrmann, Deqing Sun, and David J Fleet. Zero-shot metric depth with a field-of-view conditioned diffusion model. *arXiv preprint arXiv:2312.13252*, 2023.

Tim Seizinger, Florin-Alexandru Vasluianu, Marcos V. Conde, and Radu Timofte. Bokehlicious: Photorealistic bokeh rendering with controllable apertures. In *ICCV*, 2025.

Shakker Labs. Flux.1-dev-controlnet-union-pro. `https://huggingface.co/Shakker-Labs/FLUX.1-dev-ControlNet-Union-Pro`, 2024. Accessed: 2025-11-13.

Liao Shen, Tianqi Liu, Huiqiang Sun, Jiaqi Li, Zhiguo Cao, Wei Li, and Chen Change Loy. Dof-gaussian: Controllable depth-of-field for 3d gaussian splatting. In *CVPR*, 2025.

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 *CVPR*, 2024.

Linxiao Shi, Siming Zheng, Zerong Wang, Hao Zhang, Jinwei Chen, Bo Li, Shifeng Chen, and Peng-Tao Jiang. Towards photorealistic and efficient bokeh rendering via diffusion framework. *arXiv preprint arXiv:2605.07429*, 2026.

Ayush Shrivastava, Connelly Barnes, Xuaner Zhang, Lingzhi Zhang, Andrew Owens, Sohrab Amirghodsi, and Eli Shechtman. Fine-grained defocus blur control for generative image models. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4335–4344. IEEE, 2026.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *ICCV*, 2025a.

Zhenxiong Tan, Qiaochu Xue, Xingyi Yang, Songhua Liu, and Xinchao Wang. Ominicontrol2: Efficient conditioning for diffusion transformers. *arXiv preprint arXiv:2503.08280*, 2025b.

SaiKiran Tedla, Zhoutong Zhang, Xuaner Zhang, and Shumian Xin. Learning to refocus with video diffusion models. In *SIGGRAPH Asia*, 2025.

Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *ECCV*, 2022.

Shr-Ruei Tsai, Wei-Cheng Chang, Jie-Ying Lee, Chih-Hai Su, and Yu-Lun Liu. Lightsout: Diffusion-based outpainting for enhanced lens flare removal. In *ICCV*, 2025.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023.

Ning-Hsu Wang, Ren Wang, Yu-Lun Liu, Yu-Hao Huang, Yu-Lin Chang, Chia-Ping Chen, and Kevin Jou. Bridging unsupervised and supervised depth from focus via all-in-focus supervision. In *ICCV*, 2021.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details, 2025a. URL https://arxiv.org/abs/2507.02546.

Yiyang Wang, Xi Chen, Xiaogang Xu, Yu Liu, and Hengshuang Zhao. Diffcamera: Arbitrary refocusing on images. In *SIGGRAPH Asia*, 2025b.

Yujie Wang, Praneeth Chakravarthula, and Baoquan Chen. Dof-gs: Adjustable depth-of-field 3d gaussian splatting for post-capture refocusing, defocus rendering and blur removal. In *CVPR*, 2025c.

Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, 2022.

Zhixiang Wang, Baiang Li, Jian Wang, Yu-Lun Liu, Jinwei Gu, Yung-Yu Chuang, and Shin'Ichi Satoh. Matting by generation. In *SIGGRAPH*, 2024a.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *SIGGRAPH*, 2024b.

Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. In *ICCV*, 2023.

Hao Yang, Liyuan Pan, Yan Yang, Richard Hartley, and Miaomiao Liu. Ldp: Language-driven dual-pixel image defocus deblurring network. In *CVPR*, 2024.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *CVPR*, 2022.

Yan Yang, Liyuan Pan, Liu Liu, and Miaomiao Liu. K3dn: Disparity-aware kernel estimation for dual-pixel defocus deblurring. In *CVPR*, 2023.

Yang Yang, Siming Zheng, Jinwei Chen, Boxi Wu, Xiaofei He, Deng Cai, Bo Li, and Peng-Tao Jiang. Any-to-bokeh: One-step video bokeh via multi-plane image guided diffusion. *arXiv preprint arXiv:2505.21593*, 2025.

Chang-Han Yeh, Chin-Yang Lin, Zhixiang Wang, Chi-Wei Hsiao, Ting-Hsuan Chen, Hau-Shiang Shiu, and Yu-Lun Liu. Diffir2vr-zero: Zero-shot video restoration with diffusion-based image restoration models. *arXiv preprint arXiv:2407.01519*, 2024.

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. In *CVPR*, 2025.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. In *NeurIPS*, 2023.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Wei Zhang and Wai-Kuen Cham. Single-image refocusing and defocusing. *IEEE TIP*, 21(2):873–882, 2011.

Yuelin Zhang, Pengyu Zheng, Wanquan Yan, Chengyu Fang, and Shing Shin Cheng. A unified framework for microscopy defocus deblur with multi-pyramid transformer and contrastive learning. In *CVPR*, 2024.

Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *CAAI Artificial Intelligence Research*, 3(9150038):1–12, 2024.

Chengxuan Zhu, Qingnan Fan, Qi Zhang, Jinwei Chen, Huaqi Zhang, Chao Xu, and Boxin Shi. Bokehdiff: Neural lens blur with one-step diffusion. In *ICCV*, 2025.

Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *CVPR*, 2023.

Shaojie Zhuo and Terence Sim. Defocus map estimation from a single image. *PR*, 44(9):1852–1858, 2011.

A Supplementary Materials Overview

This supplementary material provides implementation and training details, benchmarks, metrics, and user-study protocols. It also presents additional comparisons with DiffCamera, Bokehlicious, a vision-language model, and $\text{RealBokeh}_{\text{bin}}$, along with qualitative controllability results and an ITW dataset analysis. Further analyses cover robustness to depth estimators, the defocus-map derivation, direct refocusing versus our two-stage pipeline, and reduced-step inference.

B Implementation Details

During inference, each stage employs 28 denoising steps. While our backbone (Shakker Labs, 2024) intrinsically supports arbitrary resolutions and aspect ratios without resizing, and our conditioning setup is completely coordinate-independent, ultra-high-resolution inputs (e.g., 4K) still incur prohibitive VRAM consumption and quality degradation. We implement a tiling strategy inspired by (Bar-Tal et al., 2023) during inference. Specifically, we process the image as overlapping patches and blend the tile-wise denoised results. This strategy enables us to process images exactly at their original resolutions, fully preserving image fidelity. All models are trained with the latent flow-matching objective only. Let $z_0 = \mathcal{E}(I_{\text{GT}})$ be the VAE latent of the target image, $\epsilon \sim \mathcal{N}(0, I)$, and $z_t = (1-t)z_0 + t\epsilon$. The loss is

$$\mathcal{L} = \mathbb{E}\Big[\big\|v_\theta(z_t, t, c) - (\epsilon - z_0)\big\|_2^2\Big], \tag{6}$$

where c denotes the condition tokens, and I_{GT} is the sharp image for DeblurNet and the bokeh image for BokehNet. The loss weight is constant, with no timestep-dependent weighting and no auxiliary pixel-space, perceptual, or adversarial loss. We use the same objective for all data sources, since Sec. 3.2 converts every source into the same training format. The timestep is sampled as $t = \text{clip}(u, 0, 1)$ with $u \sim \mathcal{U}(-0.05, 1.05)$, and this distribution is fixed throughout training. We optimize only the LoRA weights, with the VAE, text encoders, and transformer base weights frozen. We use the Prodigy optimizer with weight decay 0.01 and gradient clipping at 0.5, without an additional learning-rate scheduler. Training is performed in bf16 on 512×512 crops with a per-device batch size of 1 and gradient accumulation of 8.

C Additional Training Details

In this section, we detail how we process datasets for training *DeblurNet* and *BokehNet*.

C.1 DeblurNet Training Data

DeblurNet is trained using a combination of paired defocus-blur data and high-quality sharp images. Specifically, we adopt all images from the official training split of the DPDD (Abuolaim & Brown, 2020) dataset, which provides supervised pairs of defocused and all-in-focus images. To supplement this, we incorporate data from RealBokeh_3MP (Seizinger et al., 2025). Since this dataset contains images with varying degrees of focus, we apply a quality filter by computing the Laplacian variance of each image as a focus measure. By ranking the dataset based on this metric, we retain the top 3,000 sharpest images to serve as additional supervision for the deblurring task.

C.2 BokehNet Training Data

Training *BokehNet* requires diverse, high-resolution, all-in-focus inputs to simulate realistic bokeh effects. We draw candidate images from (Yuan et al., 2025) and the EBB (Ignatov et al., 2020) collections, and, as in the *DeblurNet* stage, we use Laplacian variance to filter out blurry examples. This yields a refined pool of approximately 1.7K sharp images. We also use a comprehensive dataset of 26K real bokeh images. This collection comprises 13K previously filtered and verified images from the ITW dataset (Fortes et al., 2025), alongside 13K images newly curated for this work. The newly collected data consists of focus-consistent series captured with varying apertures, containing 2 to 4 images per set. To establish accurate ground truth for these series, we first manually verified and refined the in-focus masks. This annotation step required 4 to 8 seconds per image, totaling approximately 8 hours of manual effort. Utilizing these masks, we then optimized the parameter K using a simulator (Peng et al., 2022). After obtaining the optimized K values, we applied a Structural Similarity (SSIM) threshold to filter out suboptimal results, ensuring that only reliable pairs are used for supervision.

D Additional Details on Benchmarks and Metrics

We evaluate our method on three tasks using reference-based fidelity and no-reference image quality assessment metrics.

- **Defocus Deblurring.** We conduct deblurring experiments on RealDOF and DPDD datasets. To assess reconstruction fidelity and perceptual naturalness, we report LPIPS Zhang et al. (2018), DISTS Ding et al. (2020), and no-reference metrics including CLIP-IQA Wang et al. (2023), MANIQA Yang et al. (2022), and MUSIQ Ke et al. (2021).
- **Bokeh Synthesis.** We introduce LF-Bokeh, featuring 200 images with diverse focus planes and aperture sizes, synthesized from light-field captures Rerabek & Ebrahimi (2016); Dansereau et al. (2019), comprising 170 images from the LIFF and 30 images from the EPFL Light Field Dataset. Notably, images derived from the EPFL Light Field Dataset are entirely excluded from BokehNet training, forming a held-out subset that provides an uncontaminated evaluation of generalization. All images are rendered from the central view of each light-field capture. For each scene, bokeh images are synthesized at two focus planes, one sampled from the near depth range and one from the far depth range, selected randomly within each range to ensure diversity. To assess fidelity, following prior works Zhu et al. (2025); Peng et al. (2022), we conduct a per-image *binary search* over K and select the value that *maximizes SSIM* with the target, reporting LPIPS, DISTS, and CLIP-I Radford et al. (2021). To quantify controllability under these constraints, we adopt the LVCorr metric following Yuan et al. (2025); Fortes et al. (2025); Wang et al. (2025b). Specifically, for images generated from the same all-in-focus input with a fixed focus plane across varying bokeh levels K, we compute the Pearson correlation coefficient (CorrCoef) between these K values and the Laplacian variance trend, thereby reflecting the model's accuracy in rendering the physical blur progression.
- **Refocusing.** The LF-Refocus dataset consists of 400 source–target pairs from LF-Bokeh, of which 50 pairs come exclusively from the EPFL Light Field Dataset and are unseen during training. Each pair comprises a source image focused at a near depth plane and a target focused at the far depth plane of the same scene (or vice versa), all rendered from the central view. We employ an optimal K search for each pair. We report LPIPS, DISTS, and CLIP-I for fidelity, alongside CLIP-IQA and MUSIQ to evaluate perceptual quality across varying depths of field.

D.1 Evaluation Protocols

D.1.1 Bokeh Synthesis (LF-Bokeh)

Since ground-truth bokeh levels are unavailable for real captures, we follow (Peng et al., 2022; Zhu et al., 2025) and run a per-image binary search over K that maximizes SSIM with the target, for every method that takes K as its control input. The baselines require the following adaptations.

Bokehlicious. Bokehlicious (Seizinger et al., 2025) is controlled by f-stop rather than K, so we perform the binary search over f-stop values instead. Because of this fundamentally different control mechanism, we exclude it from the LVCorr evaluation.

Bokeh Diffusion. Bokeh Diffusion (Fortes et al., 2025) targets text-to-image generation. To adapt it to image editing, we invert the input image using (Song et al., 2021). This inversion prolongs inference and strips the image of its original high-frequency details, reducing overall fidelity.

D.1.2 Refocusing (LF-Refocus)

Bokeh level. As in Sec. D.1.1, for every method that takes K as its control input, we determine the optimal bokeh level by a per-image binary search following (Zhu et al., 2025; Peng et al., 2022).

Learn2Refocus. Learn2Refocus (Tedla et al., 2025) takes a specific focal position as the input condition and synthesizes the remaining images in the focal stack. To cover the full range of focal planes, we run inference twice: once conditioned on the minimum focal position (1) and once on the maximum (9). We evaluate all outputs from both generated stacks and report the one yielding the highest SSIM.

E User Study Details

We recruit 20 participants to evaluate sharp-to-defocused progressions on 4 in-the-wild scenes. For each scene, we generate a sharp-to-defocused progression using our method and each baseline

Table 6: **User study win rates (%) of GenRefocus vs. baselines, evaluated with 20 participants on 4 in-the-wild scenes.**

Comparison	GenRefocus win rate
Ours vs. Bokehlicious	86.2%
Ours vs. BokehMe	77.5%
Ours vs. DiffCamera	98.8%

(Bokehlicious, BokehMe, DiffCamera). Participants are shown pairs of progressions side by side, generated from the same input under matched conditions, and asked to select the one that appears more realistic and consistent across the sequence. As summarized in Tab. 6, our method is consistently preferred over all baselines. Fig. 9 shows a representative progression from one scene, comparing our method against the baselines.



Figure 9: **Sharp-to-defocused progression on a representative in-the-wild scene, comparing our method against Bokehlicious, BokehMe, and DiffCamera across increasing bokeh strength.**

We further evaluate the perceptual realism of the recovered all-in-focus (AIF) images. We recruit about 30 participants and select 10 cases from RealDOF. For each case, participants are shown outputs from our method and a baseline side by side and asked to select the one that appears clearer and more realistic. Our method achieves win rates of 92.8% against Bokehlicious, 95.2% against DiffCamera, and 95.5% against DRBNet, as summarized in Tab. 7.

F PointLight-1K Collection Pipeline

We construct the PointLight-1K dataset to better model scenes with strong point-light bokeh. The pipeline consists of the following steps:

Table 7: **User study win rates (%) of GenRefocus vs. baselines on AIF realism, evaluated with about 30 participants on 10 RealDOF cases.**

Comparison	GenRefocus win rate
Ours vs. Bokehlicious	92.8%
Ours vs. DiffCamera	95.2%
Ours vs. DRBNet	95.5%

Table 8: **Quantitative comparison on the DPDD dataset using 512×512 central crops for the defocus deblurring task.** Both methods are evaluated without any tiling strategy to ensure a strictly fair comparison of core architectural capacity. The best results are highlighted in **bold**.

Method	LPIPS ↓	DISTS ↓	CLIP-IQA ↑	MANIQA ↑	MUSIQ ↑
DiffCamera (Wang et al., 2025b)	0.2833	0.1793	0.4602	0.3502	43.2997
Ours	**0.1285**	**0.1045**	**0.4637**	**0.3765**	**47.2070**

1. **Keyword mining.** We first mine keywords from photographs on Flickr[1] that are tagged with terms such as "night" or "bokeh". These keywords capture typical compositions and scene elements with prominent point lights.
2. **Prompt expansion.** We then expand the collected keywords into diverse, natural-language prompts using GPT-4o (Hurst et al., 2024). This step enriches the textual descriptions while preserving point-light characteristics.
3. **Image generation with fine-tuned FLUX.** We use a fine-tuned version of FLUX.1-Dev (Shakker Labs, 2024) (FLUX.1-dev-LoRA-AntiBlur LoRA) to generate images from the prompts. The fine-tuning encourages sharper scenes.
4. **Deblurring with DeblurNet.** Despite fine-tuning, we observe that images involving point lights may still exhibit mild residual blur. To further reduce this blur, we pass all generated images through our *DeblurNet* module, optimizing each image toward an all-in-focus appearance.
5. **Final dataset.** After this process, we obtain a curated set of 1k night-time images with prominent point lights and minimal residual blur. We refer to this dataset as POINTLIGHT-1K and use it as a specialized training set to pretrain and refine our learning strategy for controllable bokeh shape.

G ADDITIONAL COMPARISON WITH DIFFCAMERA

To ensure a strictly fair comparison of core model performance, we evaluate both methods under identical input conditions. Because DiffCamera (Wang et al., 2025b) is constrained to its native 512×512 training resolution and a fixed aspect ratio, we extract 512×512 central crops from the DPDD benchmark. This guarantees that both models are evaluated at DiffCamera's optimal setting.

As shown in Tab. 8, our method consistently outperforms DiffCamera across all full-reference and no-reference metrics. Notably, our approach achieves a significant reduction in LPIPS (0.1285 vs. 0.2833) and DISTS (0.1045 vs. 0.1793). This comprehensive superiority under a strictly controlled baseline confirms our robust generative capacity and highlights the critical advantage of leveraging real-world data during training.

H QUALITATIVE COMPARISON WITH BOKEHLICIOUS

In the main paper, we state that our model is more robust to diverse input conditions. To further substantiate this claim, we provide additional qualitative comparisons with Bokehlicious (Seizinger et al., 2025) in Fig. 10.

[1] https://www.flickr.com/

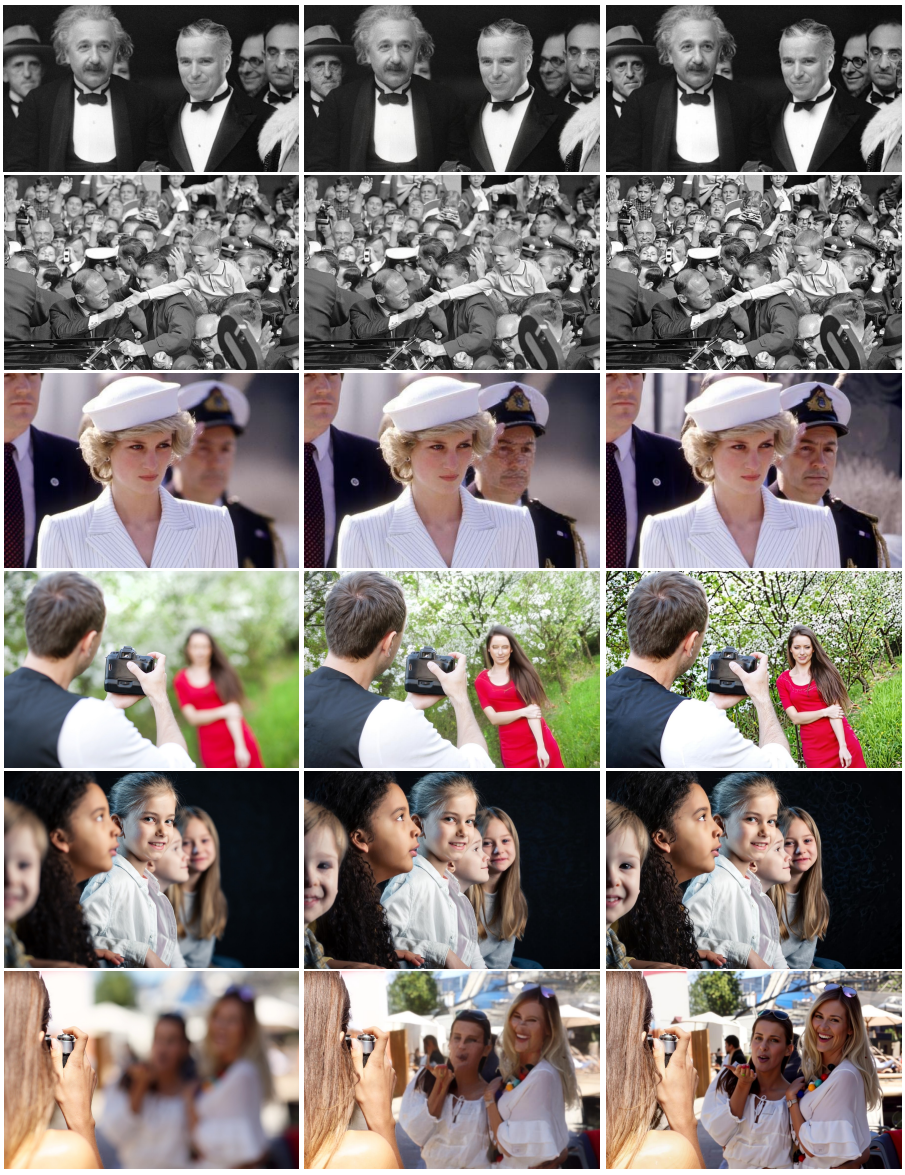


Figure 10: **Comparison with Bokehlicious (Seizinger et al., 2025) on defocus deblurring.** Results are shown for historical images (top three rows) and in-the-wild images (bottom three rows). Bokehlicious (Seizinger et al., 2025) either fails to remove the blur entirely or introduces severe structural artifacts. Conversely, our proposed method generalizes well, robustly handling diverse blur conditions to synthesize crisp, visually pleasing outputs.

A primary real-world application of defocus deblurring lies in restoring invaluable legacy photographs that can no longer be recaptured, as well as rescuing modern in-the-wild images suffering from accidental misfocus. As shown in the first three rows of Fig. 10, we evaluate the models on historical images, including monochrome photographs. Under these conditions, Bokehlicious either yields negligible deblurring effects or introduces severe structural artifacts. Similarly, in in-the-wild scenarios (the last three rows), Bokehlicious fails to recover sharp latent images and suffers from visual degradation.

In contrast, our proposed method successfully restores details across all these highly diverse conditions. From aging black-and-white photos to everyday digital captures, our model consistently



Input Nano Banana Pro GenRefocus (Ours)

Figure 11: **Qualitative comparison with VLM model.** Qualitative refocusing results comparing our Generative Refocusing framework with *Gemini 3 Nano Banana Pro* (Google DeepMind, 2025) given the prompt "focus on the man on the right".



K=0(Input) K=5 K=10 K=15

Figure 12: **Qualitative evaluation of controllability.** Refocusing results across different blur levels ($K \in \{0, 5, 10, 15\}$). The blue boxes highlight specific text regions to illustrate the defocus effect. As K increases, the scene blur increases monotonically, demonstrating our model's precise control over the bokeh effect.

produces visually appealing results, thereby validating its strong generalization capabilities and practical applicability.

I Qualitative comparison with VLM model

We also compare our method with the vision–language model *Nano Banana Pro* (Google DeepMind, 2025), which explicitly supports text-driven refocusing. For this model, we use the prompt "focus on the man on the right" to specify the desired focus point. As shown in Fig. 11, Nano Banana Pro can change the focus to some extent, but it also alters the facial expressions and appearances of people in both the foreground and the background. In contrast, our method preserves the original content while producing a realistic bokeh effect and accurate refocusing.

J Qualitative Results for Controllability

In the main paper, we quantitatively demonstrated that our framework outperforms all other diffusion-based baselines in terms of controllability. In this section, we qualitatively evaluate this controllability by examining the influence of the blur parameter K and the defocus map. As shown in Fig. 12, scene blur increases monotonically as K increases. This monotonic transition effectively proves that our model achieves precise controllability.

K ITW Dataset Analysis

This dataset is utilized to estimate the image bokeh level K via EXIF metadata. We provide further analysis regarding the dataset's diversity. Crucially, all data consists of real optical bokeh images captured directly by digital cameras, rather than artificial defocus synthesized through computational aperture techniques. In addition to the hardware diversity represented by the various camera models

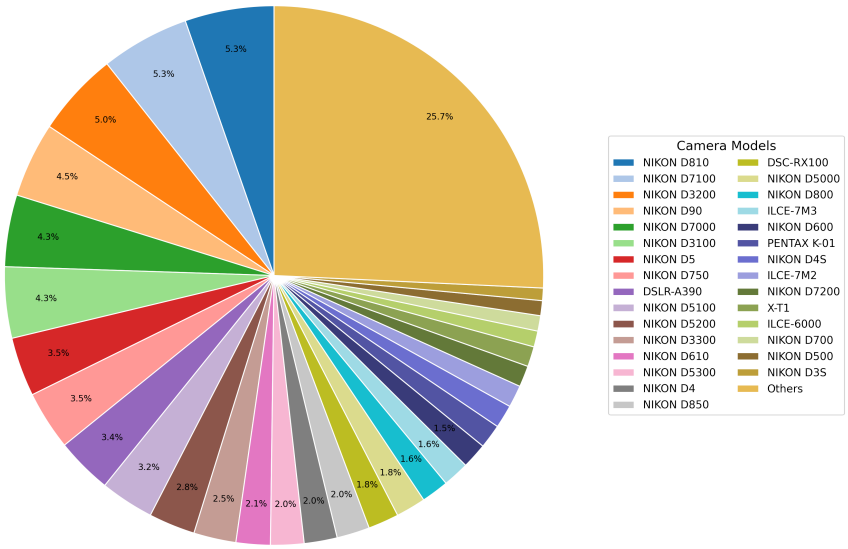


Figure 13: **Camera model distribution.** Percentage of images captured by the top 30 camera models in our dataset. Less frequent models are aggregated into the "Others" category.

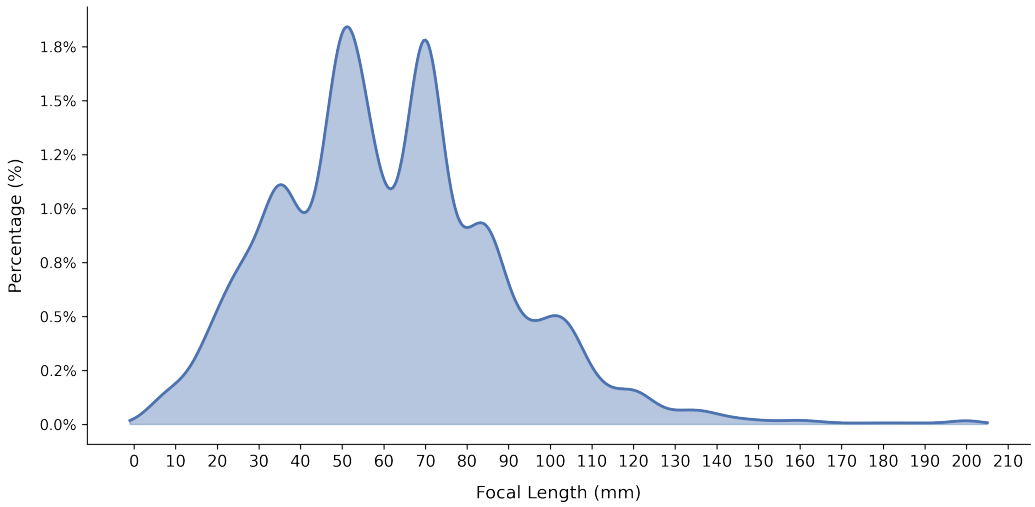


Figure 14: **Focal length distribution.** The percentage distribution of focal lengths across ITW dataset.

(Fig. 13), we analyze the distributions of the key parameters used in our K estimation. Specifically, we provide the distributions for focal length (Fig. 14), F-number (Fig. 15), and pixel ratio (Fig. 16).

L Robustness to Different Depth Estimators

To verify that our method is not overly sensitive to this choice, we evaluate three off-the-shelf depth estimators on the bokeh synthesis task: MoGe-2 (Wang et al., 2025a), UniDepthV2 (Piccinelli et al., 2025), and Depth Pro (Bochkovskii et al., 2025). As shown in Tab. 9, all three estimators yield competitive results, confirming that our pipeline is robust to the depth estimator used. Depth Pro provides the best performance across all metrics and is therefore adopted as our default.

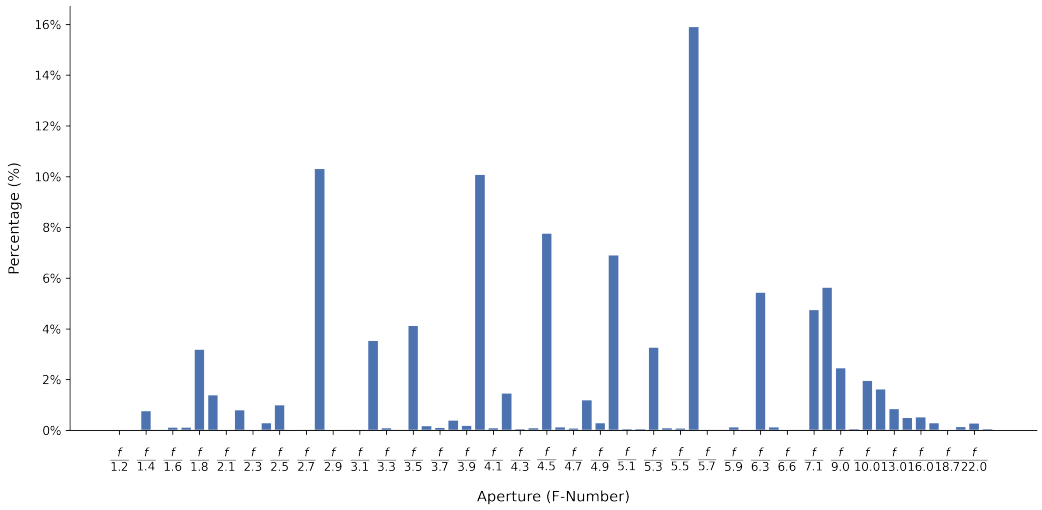


Figure 15: **Aperture (F-number) distribution.** The percentage distribution of aperture settings across the ITW dataset. Labels are presented as vertical fractions ($\frac{f}{N}$) for clarity.

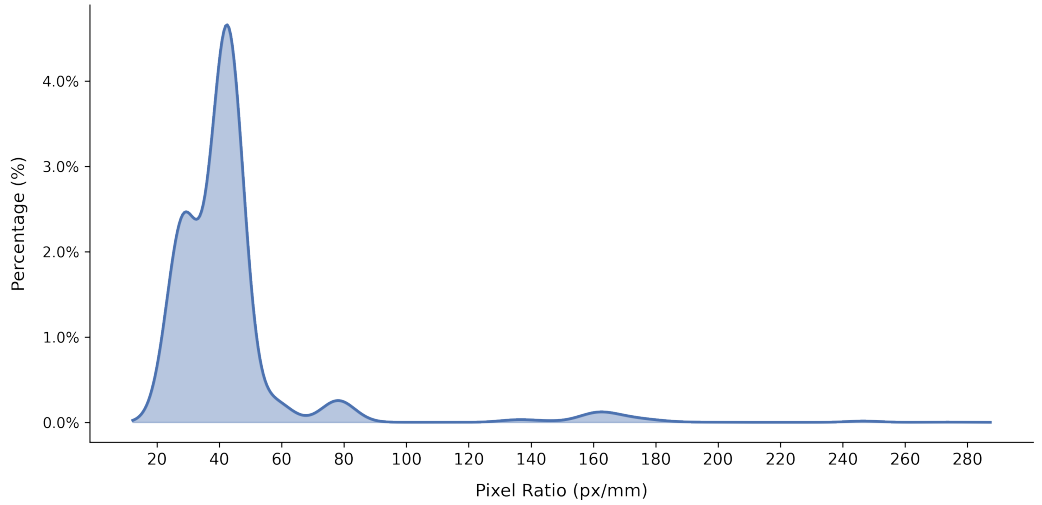


Figure 16: **Pixel ratio distribution.** The percentage distribution of the pixel-to-sensor ratio across the ITW dataset. This ratio, defined as the image's largest edge length divided by the physical sensor width (px/mm).

M Defocus Map Parameterization

Notation clarification. In the main text, D in Eq. 1 and Eq. 2 denotes the inverse-depth map, i.e., $D = 1/z$, where z is the metric depth predicted by the depth estimator (Bochkovskii et al., 2025), and D_{focus} is the focus plane expressed in the same inverse-depth space. The thin-lens formula in Eq. 3, however, requires the physical focus distance. For brevity, Eq. 3 writes this quantity as D_{focus}; to avoid ambiguity, we denote it by $z_{\text{focus}} = 1/D_{\text{focus}}$ throughout this appendix, so that Eq. 3 reads

$$K \approx \frac{f^2\, z_{\text{focus}}}{2F\,(z_{\text{focus}} - f)} \times \text{pixel ratio}. \tag{7}$$

Thin-lens derivation. For a thin lens with focal length f, f-number F (aperture diameter f/F), and focus distance z_{focus}, the circle-of-confusion (CoC) diameter of a point at depth z is

$$c(z) = \frac{f^2}{F\,(z_{\text{focus}} - f)} \cdot \frac{|z - z_{\text{focus}}|}{z}. \tag{8}$$

Table 9: **Ablation on different depth estimators.** We evaluate BokehNet on the bokeh synthesis benchmark (LF-Bokeh) using three different monocular depth estimators. Our method is robust to the choice of depth estimator, and Depth Pro (Bochkovskii et al., 2025) yields the best performance.

Depth Estimator	LPIPS ↓	DISTS ↓	CLIP-I ↑
MoGe-2	0.1312	0.0762	0.9584
UniDepthV2	0.1168	0.0669	0.9653
Depth Pro (Ours)	**0.0816**	**0.0490**	**0.9711**

Table 10: **Comparison of pipeline designs.** Direct refocusing vs. our two-stage pipeline (deblurring, then bokeh synthesis), evaluated on LF-Refocus.

Method	LPIPS ↓	DISTS ↓	CLIP-I ↑
Direct refocusing	0.1723	0.0976	0.9345
Deblurring then bokeh synthesis (Ours)	**0.1324**	**0.0759**	**0.9571**

Since

$$\frac{|z - z_{\text{focus}}|}{z} = z_{\text{focus}} \left| \frac{1}{z} - \frac{1}{z_{\text{focus}}} \right|, \tag{9}$$

this becomes

$$c(z) = \frac{f^2 \, z_{\text{focus}}}{F \, (z_{\text{focus}} - f)} \, \big| D - D_{\text{focus}} \big|. \tag{10}$$

Taking the radius (half the diameter) and converting to pixels via the pixel ratio gives $r(z) = K \, |D - D_{\text{focus}}|$, with

$$K \approx \frac{f^2 \, z_{\text{focus}}}{2F \, (z_{\text{focus}} - f)} \times \text{pixel ratio}, \tag{11}$$

which is Eq. 3 with the focus distance written as z_{focus}. Hence Eq. 1 is the pixel-space CoC radius, and K collects all lens- and sensor-dependent terms.

Units. We express z, z_{focus}, and f in millimeters (the metric depth from the estimator is converted from meters to millimeters) and the pixel ratio in px/mm. Then K has units of px·mm and D has units of mm^{-1}, so D_{def} is expressed in pixels. For synthetic data, K is sampled directly rather than derived from EXIF.

N Direct Refocusing vs. Deblurring then Bokeh Synthesis

We evaluate a baseline that *directly* predicts a refocused image from a blurry input and a defocus map. Trained on the same synthetic data with the BokehNet backbone, this approach underperforms our two-stage pipeline (Tab. 10). We attribute this gap to two factors: (1) *control fidelity*: estimating depth from blurry inputs degrades defocus map accuracy, causing misaligned blur fields; and (2) *data leverage*: a direct formulation precludes injecting tailored real-data supervision into separate deblurring and synthesis subtasks.

O Reduced-Step Inference Ablation

We evaluate GenRefocus on LF-Refocus under fewer denoising steps (20, 10, 4) to study the efficiency–quality trade-off. As shown in Tab. 11, inference time drops from 143.371s (28 steps) to 20.556s (4 steps), while fidelity and perceptual metrics remain largely stable, indicating that GenRefocus tolerates aggressive step reduction.

Table 11: **Reduced-step inference ablation.** We evaluate GenRefocus on LF-Refocus under varying numbers of denoising steps. Per-image inference time drops substantially with fewer steps, while fidelity and perceptual quality remain largely stable, demonstrating a favorable efficiency–quality trade-off. Best results are highlighted in **bold**. All timings are measured on a single NVIDIA RTX A6000.

Setting	Time / image (s) ↓	LPIPS ↓	DISTS ↓	CLIP-I ↑	CLIP-IQA ↑	MUSIQ ↑
28 steps	143.371	**0.1324**	**0.0759**	**0.9571**	0.5803	52.9706
20 steps	101.289	0.1360	0.0778	0.9566	0.5927	**53.5401**
10 steps	50.782	0.1359	0.0780	0.9558	**0.6019**	53.2023
4 steps	20.556	0.1366	0.0788	0.9544	0.5880	51.8423

Table 12: Comparison on $\text{RealBokeh}_{\text{bin}}$ (Seizinger et al., 2025). Bokehlicious is trained end-to-end on this single, pixel-aligned, real-world benchmark and only supports aperture-size control at a fixed focus point, while our method jointly supports flexible focus-plane control and is trained on a broader mixture of synthetic and real sources, trading some in-domain fidelity here for generalization across more diverse control settings (see Tab. 3 and Fig. 5).

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Bokehlicious (Seizinger et al., 2025)	30.16	0.9236	0.1102
GenRefocus (Ours)	24.30	0.8503	0.2414

P Comparison on $\text{RealBokeh}_{\text{BIN}}$

To further contextualize our comparison with Bokehlicious (Seizinger et al., 2025), we evaluate GenRefocus on $\text{RealBokeh}_{\text{bin}}$ (Seizinger et al., 2025), the exact single-aperture benchmark on which Bokehlicious is trained and tuned.

As shown in Tab. 12, our method trails Bokehlicious on this in-domain benchmark. We attribute this gap primarily to a difference in control scope. Bokehlicious only supports aperture size control at a fixed, single focus point aligned with the input image's optical focus, and is trained end-to-end on a single, pixel-precisely aligned real-world dataset (RealBokeh) tailored to exactly this setting. Consequently, it is naturally well-suited to excel on its own specialized benchmark.

In contrast, GenRefocus targets a substantially broader control space, jointly supporting flexible focus-plane selection in addition to aperture size, and is trained on a hybrid mixture of synthetic and multiple real-world sources to support this added flexibility. This broader scope trades some in-domain fidelity on a single-condition benchmark like $\text{RealBokeh}_{\text{bin}}$ for generalization across more diverse control settings. This is reflected in our results on LF-Bokeh (Tab. 3, Fig. 5), which features varying focus planes in addition to aperture size and is unseen by both methods during training: since Bokehlicious cannot adjust the focus point away from the input image's native focus, it is structurally unable to correctly render scenes requiring a shifted focus plane, while GenRefocus consistently outperforms it under this more general setting.

This quantitative trade-off is further corroborated by qualitative comparisons of bokeh synthesis on in-the-wild images. As shown in Fig. 17.

Q AI Use Statement

LLMs were used in the following ways: (1) to polish the manuscript's language (e.g., grammar, wording, and clarity); (2) to help find related work, with all cited papers read and verified by the authors; (3) as coding assistants during implementation, with all code reviewed and tested by the authors; and (4) for synthetic data generation, where GPT-4o expanded mined keywords into natural-language prompts and a fine-tuned FLUX.1-dev model generated the images that make up PointLight-1K. All technical content, claims, experimental designs, and analyses were verified by the authors, who take full responsibility for the content of this paper.



Figure 17: **In-the-wild comparison with Bokehlicious (Seizinger et al., 2025) on bokeh synthesis.** Qualitative results on in-the-wild scenes comparing GenRefocus with Bokehlicious, which is trained and tuned specifically on $\text{RealBokeh}_{\text{bin}}$.