BokehFlow: Depth-Free Controllable Bokeh Rendering via Flow Matching

Yachuan Huang[1], Xianrui Luo[1], Qiwen Wang[1], Liao Shen[1], Jiaqi Li[1], Huiqiang Sun[1*], Zihao Huang[1], Wei Jiang[1], Zhiguo Cao[1]

[1]School of AIA, Huazhong University of Science and Technology

Abstract

Bokeh rendering simulates the shallow depth-of-field effect in photography, enhancing visual aesthetics and guiding viewer attention to regions of interest. Although recent approaches perform well, rendering controllable bokeh without additional depth inputs remains a significant challenge. Existing classical and neural controllable methods rely on accurate depth maps, while generative approaches often struggle with limited controllability and efficiency. In this paper, we propose BokehFlow, a depth-free framework for controllable bokeh rendering based on flow matching. BokehFlow directly synthesizes photorealistic bokeh effects from all-in-focus images, eliminating the need for depth inputs. It employs a cross-attention mechanism to enable semantic control over both focus regions and blur intensity via text prompts. To support training and evaluation, we collect and synthesize four datasets. Extensive experiments demonstrate that BokehFlow achieves visually compelling bokeh effects and offers precise control, outperforming existing depth-dependent and generative methods in both rendering quality and efficiency.

1 Introduction

Bokeh refers to the shallow depth-of-field effects captured by cameras, serving as an essential component in professional photography. By selectively blurring the input image, bokeh rendering emphasizes regions of interest, improves scene composition, and enhances the overall aesthetic quality of visual content. Classical methods (Wadhwa et al. 2018; Busam et al. 2019; Zhang et al. 2019; Sheng et al. 2024) typically rely on the explicit physical camera model to render bokeh. Neural rendering pipelines (Wang et al. 2018; Xiao et al. 2018; Luo et al. 2020; Peng et al. 2022a,b; Luo et al. 2024; Seizinger et al. 2025) produce visually pleasing results through training on a large-scale dataset.

Existing classical and neural rendering methods achieve controllable bokeh rendering, and recent advances in generative models (Sohl-Dickstein et al. 2015; Song et al. 2020; Ho, Jain, and Abbeel 2020; Rombach et al. 2022) offer a promising alternative to synthesize controllable bokeh from all-in-focus (AiF) images directly. However, three fundamental challenges exist: 1) classical and neural rendering methods are fundamentally limited by the quality and availability of depth maps, which are sometimes inaccurate or unavailable in real-world capture; 2) existing generative methods (Yuan et al. 2025; Fortes et al. 2025) only achieve text-to-image generation and blur intensity control, struggling to achieve image-to-image rendering and are incapable of focus region control, limiting their application in photography; 3) although diffusion-based generative frameworks generate remarkable image fidelity, these models suffer from high computational overhead due to the need for iterative denoising along highly curved generative trajectories.

To address these limitations, we propose **BokehFlow**, a novel generative bokeh rendering framework that is both *depth-free* and *semantically controllable*. We eliminate the need for explicit depth inputs by establishing a connection between generative models and camera optics, where we model the bokeh generation as a direct distribution transport process in latent space. This *depth-free* paradigm not only simplifies the learning process, but also leads to robust and high-quality results, especially in complex real-world scenes where depth estimation tends to be noisy or unreliable, as shown in the top right part of Figure 1.

Beyond depth-free generative rendering, we aim to introduce high-level *semantic controllability* to bokeh rendering. Therefore, we propose an effective **Bokeh Control Adapter (BCA)** that incorporates natural language prompts (e.g., *"focus on the foreground with blur intensity of 30"*) into the vector field dynamics by transformer-style cross-attention. As shown in Figure 1, by leveraging a pre-trained text encoder, the BCA aligns semantic cues with visual features, enabling intuitive user control over the focus region and blur intensity. This approach contrasts with prior depth-dependent methods, providing a flexible control mechanism for bokeh rendering. To improve generative efficiency, we adopt the flow matching framework (Albergo and Vanden-Eijnden 2022; Albergo et al. 2023; Lipman et al. 2022; Liu, Gong, and Liu 2022) to learn vector fields that describe straight, efficient transport paths between data distributions. This enhanced flexibility in trajectory design enables single-step sampling, significantly reducing inference time compared to diffusion models (Ho, Jain, and Abbeel 2020; Song et al. 2020) that rely on curved stochastic paths.

We collect two real-world datasets and two synthetic datasets. Experimental results on these datasets show that

*Coresponding Author.

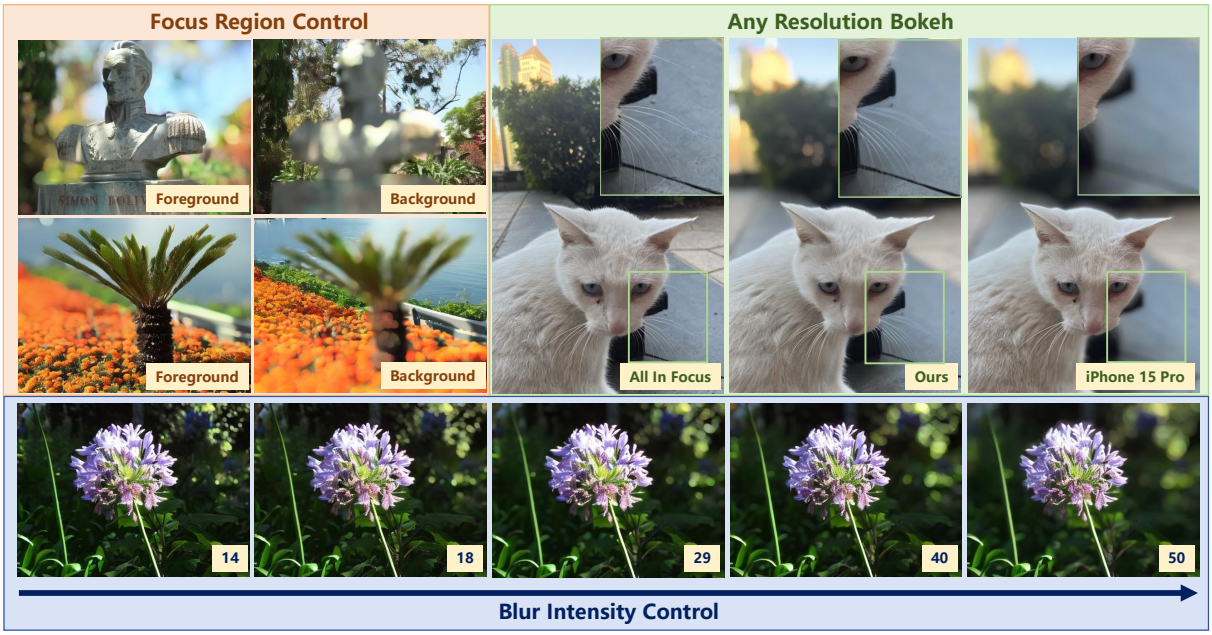

Figure 1: BokehFlow creates photorealistic and controllable bokeh effects from any resolution images without requiring depth maps. Our model achieves focus region control (*top left*), blur intensity control (*bottom*), and renders better edges around the focused object than iPhone (*top right*) in real-world scenes where depth tends to be unreliable. Zoom in for best view.

BokehFlow consistently produces high-fidelity bokeh rendering with text-driven semantic control, outperforming existing approaches in both visual quality and inference speed. Our contributions can be summarized as follows:

- We propose **BokehFlow**, the first depth-free controllable bokeh rendering framework via efficient flow matching.
- We introduce the **Bokeh Control Adapter**, a conditioning mechanism that leverages natural language prompts to enable control over focus regions and blur intensity.
- Extensive experiments demonstrate that our method achieves superior visual quality, controllability, and efficiency over depth-dependent and diffusion baselines.

2 Related Works

2.1 Bokeh Rendering

Existing approaches can be categorized into three classes: classical rendering methods, neural rendering methods, and generative models. Classical methods (Yang et al. 2016; Wadhwa et al. 2018) rely on explicit depth maps to guide the rendering process. DrBokeh (Sheng et al. 2024) introduces an improved compositing formulation that alleviates artifacts under complex occlusion conditions. Neural rendering methods (Xiao et al. 2018; Wang et al. 2018; Peng et al. 2022b,a; Luo et al. 2023; Peng et al. 2024; Seizinger et al. 2025) aim to regress bokeh images from all-in-focus inputs. However, they fundamentally rely on accurate depth inputs. Noisy and imprecise depth tends to introduce artifacts near depth boundaries, undermining the robustness and generalization in real-world scenarios. For controllable rendering without depth map, Bokehlicious (Seizinger et al. 2025) resorts to an aperture-varying DSLR-captured dataset to achieve aperture-controllable bokeh synthesis. However, its controllability remains limited, as it only supports aperture control while lacking the ability to switch the focus region. Recent generative models offer a promising alternative to achieve depth-free bokeh rendering by utilizing scene priors from large-scale datasets. Generative Photography (Yuan et al. 2025) and Bokeh Diffusion (Fortes et al. 2025) apply generative models to produce impressive results. However, these methods are designed for text-to-image generation and do not support user-provided all-in-focus images as input, making them unsuitable for scene-consistent bokeh rendering. Moreover, they struggle to control the focus region. In contrast, our method supports both image-consistent and semantically-controllable bokeh synthesis.

2.2 Flow Matching

Flow matching (Lipman et al. 2022; Liu, Gong, and Liu 2022; Albergo et al. 2023; Neklyudov et al. 2023) formulates generation as learning a deterministic vector field between source and target distributions instead of a stochastic process in diffusion models. It offers substantial improvements in inference efficiency such as single-step sampling. Therefore, flow matching has shown strong performance in a variety of tasks including image generation (Lipman et al. 2022; Esser et al. 2024; Fischer et al. 2023), video understanding (Chen et al. 2024) and 3D perception (Gui et al. 2025; Li et al. 2025). In the context of bokeh rendering, diffusion-based methods typically require Gaussian noise as the starting distribution, which does not intuitively correspond to the natural correlation between an all-in-focus

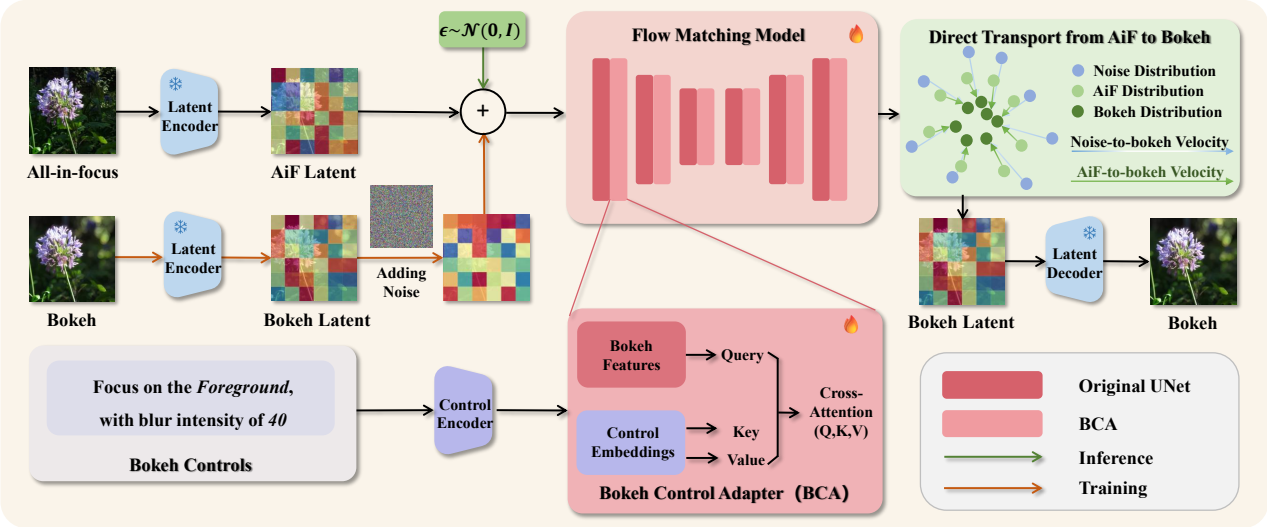

Figure 2: Pipeline of BokehFlow. The all-in-focus and bokeh images are first encoded into latent space using a VAE encoder. Random noise is added to the bokeh latent, and the flow matching model learns to denoise the concatenated all-in-focus and noisy bokeh latents through our direct transport design. Bokeh controls, which include the focus region and blur intensity, are encoded into control embeddings z_C via a control encoder. In our proposed Bokeh Control Adapter (BCA), these features are injected through cross-attention, where bokeh features z_B serve as queries and z_C are used as keys and values. Finally, the denoised bokeh latent is decoded by the VAE decoder to generate the output bokeh image.

image and its corresponding bokeh image. This results in unnecessarily complex transport trajectories and increases sampling latency. In this paper, we leverage the flow matching framework to model the direct transport between the input all-in-focus image and the bokeh image, significantly reducing inference time while enhancing spatial coherence, particularly near occlusion boundaries.

3 Methodology

3.1 Flow Matching for Bokeh Rendering

Latent Flow Matching. To reduce the computational cost of training high-resolution flow matching models for bokeh rendering, we follow prior works (Rombach et al. 2022; Dao et al. 2023) and adopt an autoencoder-based latent space that is perceptually aligned with the image domain. We map both the all-in-focus image $\mathcal{I}_A$ and corresponding bokeh image $\mathcal{I}_B$ into the latent space through the encoder $\mathcal{E}$. We denote $z_A = \mathcal{E}(\mathcal{I}_A)$ and $z_B = \mathcal{E}(\mathcal{I}_B)$ as the latent representations of the input and target images, respectively. Reconstructions can be obtained by a shared decoder $\mathcal{D}$, achieving faithful image synthesis.

Flow Matching (Lipman et al. 2022; Liu, Gong, and Liu 2022; Albergo et al. 2023; Neklyudov et al. 2023) is a class of generative modeling techniques that learns to regress vector fields along fixed conditional probability paths, achieving transformation across data distributions. We exploit its inherent property of straight-line interpolation to define the corruption process using standard Gaussian noise ϵ :

$$\phi_t(z_B) = tz_B + (1-t)\epsilon, \tag{1}$$

where $\phi_t(z_B)$ denotes the corrupted latent representation of bokeh image at time $t \in [0, 1]$, tracing a linear path from noise to data. Given the assumption of uniform linear interpolation, the corresponding time-dependent velocity field at time t is derived as:

$$v_t(\phi_t(z_B)) = \phi_t(z_B) - \epsilon. \tag{2}$$

By the fundamental relation between displacement and velocity, the dynamics follow the ordinary differential equation (ODE):

$$d\phi_t(z_B) = v_t(\phi_t(z_B))\, dt. \tag{3}$$

Direct Transport from All-in-Focus to Bokeh. Prior works (Rombach et al. 2022; Dao et al. 2023) propose training a network $\mathcal{N}_{fm}$ to model a fixed velocity field from Gaussian noise to the target distribution, effectively realizing linear optimal transport during inference. In the context of bokeh rendering, the training objective is formulated as:

$$\mathcal{L}_{FM} = \mathbb{E}_t \left\| \mathcal{N}_{fm}(\phi_t(z_B), z_A, z_C, t) - v_t(\phi_1(z_B)) \right\|, \tag{4}$$

where z_A denotes the latent representation of the all-in-focus image, and z_C is the bokeh control embeddings introduced by our proposed bokeh control adapter in Sec. 3.2. During inference, the model predicts the final-step velocity $v_t(\phi_1(z_B)) = z_B - \epsilon$. By numerically integrating the ODE from $t = 0$ to $t = 1$, we recover the target bokeh latent z_B in a small number of steps.

However, training the network to predict a fixed-scale mapping from Gaussian noise to bokeh distribution can be suboptimal. The reason lies in that during inference, even if partial denoising has been achieved, the network $\mathcal{N}_{fm}$ still needs to predict the global path, leading to a waste of capacity. In contrast, we redefine the transport path directly from

all-in-focus latents z_A to bokeh latents z_B, rather than starting from noise $\varepsilon \sim \mathcal{N}(0, I)$. Therefore, Eq.(4) is rewritten as:

$$\mathcal{L}_{FM}^{\text{Direct}} = \mathbb{E}_t \left\| \mathcal{N}_{fm}(\phi_t(z_B), z_A, z_C, t) - \mathcal{N}_{fm}^{\text{Direct}} \right\|, \quad (5)$$

where $\mathcal{N}_{fm}^{\text{Direct}} = \phi_1(z_B) - \phi_t(z_B) = (1-t)v_t(\phi_1(z_B))$. The new objective $\phi_1(z_B) - \phi_t(z_B)$ enables the flow matching network to directly predict the residual vector for linear transport from the current distribution to the target distribution, at variable scales. This direct transport formulation, contrasting with the fixed global mapping from noise, leads to improved sample efficiency, spatial consistency, and faithful rendering aligned with the input semantics.

3.2 Bokeh Control Adapter

In previous bokeh rendering methods (Sheng et al. 2024; Peng et al. 2022a; Wadhwa et al. 2018), the defocus blur effect is simulated by scattering each pixel over its neighborhood, constrained by a blur radius. The blur radius r is computed as:

$$r = \alpha \cdot |d - d_f|, \quad (6)$$

where d denotes the disparity of a pixel, d_f is the disparity of focus region, and α controls overall blur intensity. While interpretable, this approach depends heavily on accurate depth and offers limited semantic flexibility.

To address this, we introduce the Bokeh Control Adapter (BCA), which replaces explicit physical parameters with semantic prompts. User instructions like *"focus on the foreground with blur intensity of 30"* are encoded by a pretrained language model CLIP (Radford et al. 2021) to obtain the control embeddings z_C. These control embeddings z_C are injected by cross-attention layers into the vector field regressor $\mathcal{N}_{fm}^{\text{Direct}}$.

The cross-attention is computed as:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \quad (7)$$

where $Q = W_Q z_B$, $K = W_K z_C$, and $V = W_V z_C$. The features z_C are fused into the latent representation z_B to control the focus region and blur intensity.

By rendering spatially varying blur through attention-driven modulation instead of depth-based kernels, BCA enables flexible and interpretable bokeh generation without relying on depth maps or manual tuning.

3.3 Training Strategy

Although our model is trained in a depth-free manner, we aim to enhance its rendering realism by leveraging prior knowledge from large pretrained models. Since our denoising flow matching model adopts a standard conditional U-Net architecture, similar to those used in diffusion-based generative models, it naturally supports knowledge transfer from various pretrained sources. This enables more efficient training compared to learning from scratch.

Specifically, we explore initializing our model with different types of priors: (1) image-level priors from pretrained generative models such as Stable Diffusion, which help encode scene semantics and appearance structures; and (2) depth-aware priors from generative depth prediction models, including Marigold (Ke et al. 2024) and DepthFM (Gui et al. 2025), which provide implicit cues about foreground and background separation. Among them, we find that depth-oriented initialization leads to better depth-aware rendering, particularly around occlusion boundaries and fine structures, while image-based initialization offers better generalization on diverse appearances. This flexible knowledge transfer strategy empowers our model to produce physically plausible bokeh with natural focus region transitions and layered blur, without requiring an explicit depth map as auxiliary input during training.

Implementation Details. We adopt an $8\times$ downsampling VAE (Kingma, Welling et al. 2013) with 4-channel latent output. The flow matching model is a conditional U-Net (8-in, 4-out channels), following previous works (Ke et al. 2024; Gui et al. 2025), and is built with the Diffusers library. Bokeh control is applied by a pretrained CLIP text encoder (Radford et al. 2021), consistent with Stable Diffusion 2.1 (Rombach et al. 2022). Training is conducted on the CBD dataset (resize to 512×384 resolution) for 10K iterations using Adam (Kingma 2014), with an initial learning rate of $3 \cdot 10^{-5}$ decayed to $3 \cdot 10^{-7}$ after 3K steps. Experiments are run on NVIDIA RTX A6000 GPU for a single time, since our experiments are stable across multiple runs.

4 Experiments

4.1 Experimental Settings

Evaluation Metrics. For the overall image quality evaluation, we adopt the same metrics as DrBokeh (Sheng et al. 2024), namely SSIM, PSNR and LPIPS. In addition, to compare the rendering quality specifically at object boundaries or depth discontinuities, we introduce edge-based evaluation metrics: SSIM_{eg} and PSNR_{eg}. Specifically, we first extract object edges using the Sobel operator, then apply a dilation operation to obtain an edge mask, and compute SSIM and PSNR only within the edge regions. We also report the inference time to assess rendering efficiency.

Datasets. We evaluate BokehFlow on four datasets. There are no available large-scale datasets for depth-free controllable bokeh rendering, so we curated a 35,000-pair Control Bokeh Dataset (**CBD**) for training and evaluation. **CBD** contains 3,500 natural scenes from monocular depth dataset ReDWeb (Xian et al. 2018), each with one all-in-focus image at about 1024×768 resolution and corresponding depth map. Inspired by previous works (Fortes et al. 2025; Yuan et al. 2025), we render bokeh images via the bokeh rendering engine using diverse focal disparities and blur intensities $K \sim \mathcal{U}(10, 50)$, producing a total of 35,000 image pairs. Following BokehMe (Peng et al. 2022a), we also evaluate our method on the DSLR-captured real-world dataset **EBB400**, which includes 400 image pairs randomly selected from EBB! (Ignatov, Patel, and Timofte 2020). Since control parameters are unavailable, we predict a depth map for each sample using Depth Anything V2 (Yang et al. 2024), and manually annotate a bounding box indicating the all-in-focus region. The refocused disparity is then computed as

Method	CBD dataset						EBB400 dataset					
	$PSNR_{eg}\uparrow$	$SSIM_{eg}\uparrow$	PSNR↑	SSIM↑	LPIPS↓	Time(s)↓	$PSNR_{eg}\uparrow$	$SSIM_{eg}\uparrow$	PSNR↑	SSIM↑	LPIPS↓	Time(s)↓
SteReFo	19.22	0.5979	20.89	0.6830	0.4294	0.547	30.77	0.9713	23.09	0.8268	0.2621	2.697
VDSLR	<u>19.80</u>	0.5948	<u>21.52</u>	0.6801	<u>0.4137</u>	0.529	30.79	0.9712	23.28	0.8287	0.2594	2.436
DrBokeh	18.01	0.5373	19.69	0.6369	0.4807	5.344	29.40	0.9692	22.26	0.8170	0.278	8.933
DeepLens	19.58	0.5671	21.19	0.6346	0.4656	0.959	29.94	0.9705	22.29	0.8209	0.2912	2.751
MPIB	19.36	0.5947	21.03	0.6787	0.4592	0.785	30.85	0.9718	23.28	<u>0.8329</u>	<u>0.2570</u>	4.202
BokehMe	19.74	<u>0.6078</u>	21.42	<u>0.6914</u>	0.4636	<u>0.514</u>	<u>30.95</u>	<u>0.9719</u>	<u>23.37</u>	0.8326	0.2574	<u>2.007</u>
Ours	**22.46**	**0.8193**	**21.98**	**0.6948**	**0.1870**	**0.461**	**31.04**	**0.9724**	**23.41**	**0.8336**	**0.2158**	**1.885**

Table 1: Quantitative results compared with depth-dependent methods on the synthetic CBD and real-world EBB400 dataset. The best performance is in boldface, while the second is underlined.

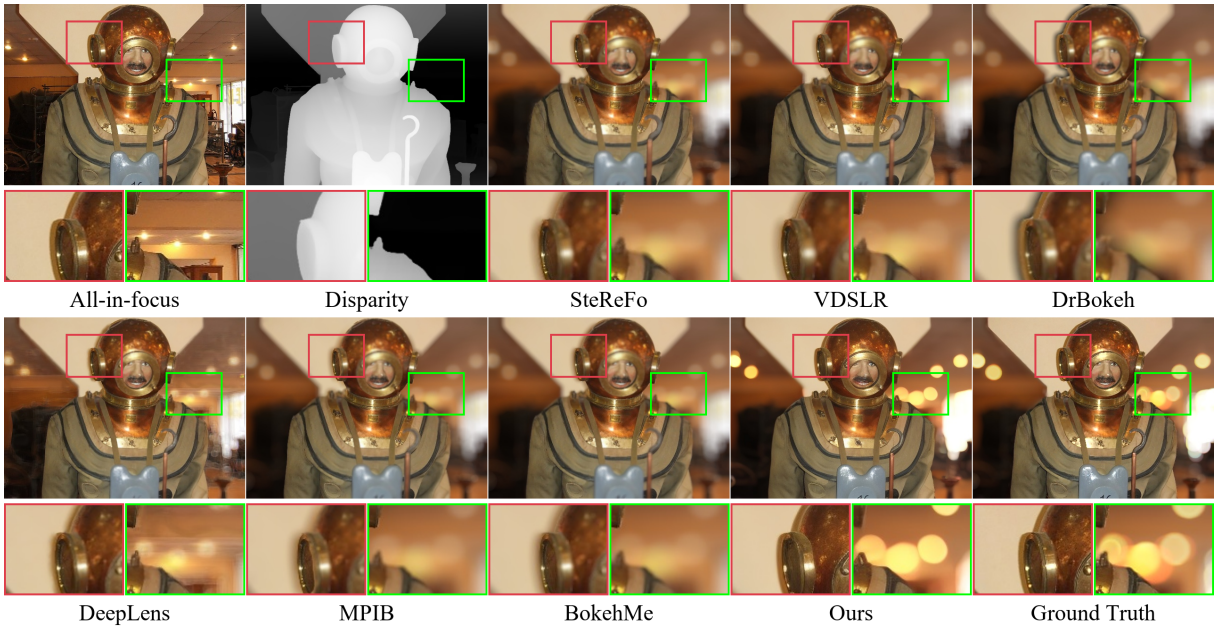

Figure 3: Visual comparison results on CBD dataset. Compared to depth-based methods, our method preserves sharper edges in the focus region and renders the most aesthetically pleasing bokeh effects.

Method	$PSNR_{eg}\uparrow$	$SSIM_{eg}\uparrow$	PSNR↑	SSIM↑	LPIPS↓	Time↓
GenPho	24.35	0.8692	23.60	0.7433	0.2314	1.400s
Ours	**24.53**	**0.8813**	**23.66**	**0.7727**	**0.1389**	**0.217s**

Table 2: Quantitative results on GenPhotoBokeh dataset with depth-free method Generative Photography.

the median depth within the bounding box. All images have a resolution of approximately 1536×1024. In addition, we synthesize **GenPhotoBokeh** dataset to compare our method with the text-to-image diffusion-based method (Yuan et al. 2025), which does not accept an all-in-focus image as input. It includes 1,000 text-driven scenes from Generative Photography (Yuan et al. 2025). We use the original prompts and set multiple blur values {2,10,20,30,40}, extracting the first frame from the generated 5-frame video as the AiF image. Depth is estimated by Depth Anything V2, and bokeh GTs are rendered with the same settings used in CBD for fair comparison, yielding 4,000 pairs at 384×256 resolution. Finally, we collect a real-world dataset **IB30** for the user study. It comprises 30 real-world images at 768×1024 resolution captured by iPhone 15 Pro. We extract the captured AiF and bokeh images, and retrieve the corresponding depth maps via the online photo editor PhotoPea (Kutskir 2016).

Baseline Comparisons. To comprehensively validate the performance of BokehFlow, we compare it with three types of controllable methods: classical rendering methods[C], neural rendering methods[N] and generative methods[G]. On the CBD dataset, EBB400 dataset and IB30 dataset, we compare with pipelines which support all-in-focus images as input: VDSLR[C] (Yang et al. 2016), SteReFo[C] (Busam et al. 2019), DrBokeh[C] (Sheng et al. 2024), DeepLens[N] (Wang et al. 2018), BokehMe[N] (Peng

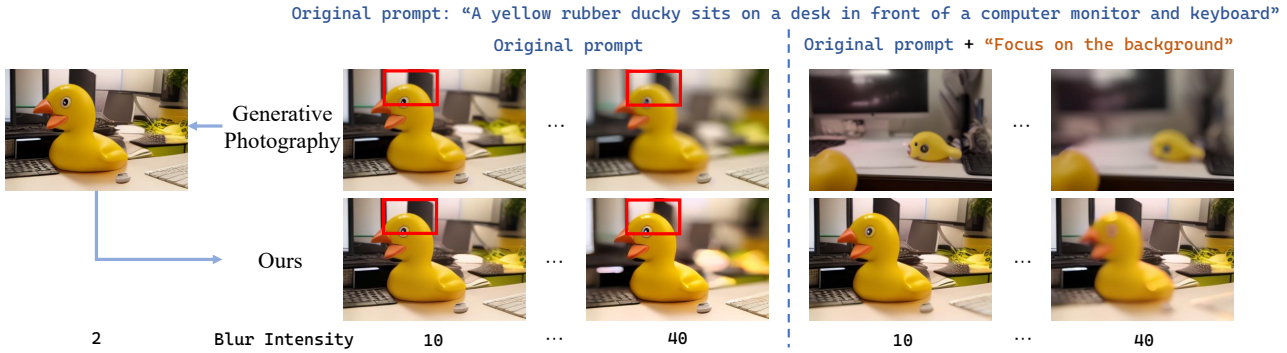

Figure 4: Visual comparison with depth-free text-to-image method. Generative Photography generates a 5-frame video using original prompt and the first frame serves as our all-in-focus input. When adding "focus on the background" to the prompt, it fails to shift focus and alters scene content, while ours preserves the scene and produces appealing foreground bokeh.

et al. 2022a), and MPIB[N] (Peng et al. 2022b). Besides, we compare with text-to-image Generative Photography[G] (Yuan et al. 2025) on the GenPhotoBokeh dataset.

4.2 Comparison with Depth-dependent Methods

Table 1 presents a quantitative comparison between our method and several state-of-the-art classical and neural controllable bokeh rendering approaches on the synthetic dataset CBD and the real-world dataset EBB400. **BokehFlow** outperforms all competing methods in terms of overall rendering quality, achieving the best PSNR, SSIM and LPIPS scores across the board. Leveraging strong generative priors, our method excels particularly at handling challenging regions near depth discontinuities, leading to superior performance on the edge-aware metrics, PSNR_{eg} and SSIM_{eg}. In contrast, methods that rely on explicit depth maps are inherently sensitive to the accuracy of the predicted depth. Imperfect estimated depth, which is common in real-world settings, will lead to a significant performance drop in these approaches, especially in edge fidelity metrics.

We further provide qualitative comparisons on CBD in Figure 3. As shown, our method produces more visually appealing bokeh effects with sharper details and enhanced clarity. When focusing on the foreground, our model renders the most precise boundaries at focused object edges, while bokeh effects in the background exhibit the most natural and aesthetically pleasing defocus patterns, closely resembling the ground truth. These results highlight the effectiveness of our depth-free and controllable framework in achieving high-quality bokeh synthesis. Refer to the appendix for more visual results on EBB400.

4.3 Comparison with Depth-free Method

As shown in Table 2, we compare our method with existing controllable generative approach (Yuan et al. 2025) on the GenPhotoBokeh dataset. We achieve superior performance in PSNR, SSIM and LPIPS metrics, indicating that our method produces higher-fidelity bokeh images with better structural consistency. In addition, our model also obtains the highest scores on the edge-aware metrics, PSNR_{eg} and SSIM_{eg}, which measure the rendering quality near object boundaries and depth discontinuities. These results demonstrate that our approach preserves sharp edges more effectively than existing methods. In terms of efficiency, our method is highly efficient with an inference time that is approximately **1/6** of the compared generative baseline, benefiting from the one-shot sampling.

We present qualitative results in Figure 4, which show that our model better preserves fine structures at the object edges and produces more visually appealing bokeh effects. Furthermore, when adding prompts like "focus on the background" to original prompts, Generative Photography (Yuan et al. 2025) fails to switch the focus region, and the output scene content changes even if the global random seed is fixed. Compared to the baseline, our method achieves more consistent results with all-in-focus images and more precise focus controls, validating the effectiveness of our semantic control and depth-free design.

Strategy	$\text{PSNR}_{\text{eg}}\uparrow$	$\text{SSIM}_{\text{eg}}\uparrow$	PSNR↑	SSIM↑
noise→bokeh	21.96	0.8001	21.54	0.6833
AiF→bokeh (Ours)	**22.46**	**0.8193**	**21.98**	**0.6948**

Table 3: Ablation results for direct transport from all-in-focus to bokeh on CBD. Direct transport is better than starting from noise.

Strategy	$\text{PSNR}_{\text{eg}}\uparrow$	$\text{SSIM}_{\text{eg}}\uparrow$	PSNR↑	SSIM↑
concatenation	20.58	0.7963	20.67	0.6725
cross-attention (Ours)	**22.46**	**0.8193**	**21.98**	**0.6948**

Table 4: Ablation results for Bokeh Control Adapter on CBD. The cross attention mechanism is better than concatenating control parameters to the input all-in-focus image.

4.4 Ablation Study

Transport Strategy. We compare our method with a naïve Flow Matching baseline that also adopts an optimal transport formulation to learn vector fields, but samples from a stan-

Strategy	$PSNR_{eg}$↑	$SSIM_{eg}$↑	PSNR↑	SSIM↑
SD2.1	20.34	0.7978	20.89	0.6763
Marigold	21.66	0.8024	21.86	0.6894
DepthFM (Ours)	**22.46**	**0.8193**	**21.98**	**0.6948**

Table 5: Ablation results for different initialization strategies. Initializing from DepthFM (Gui et al. 2025) achieves best scores.

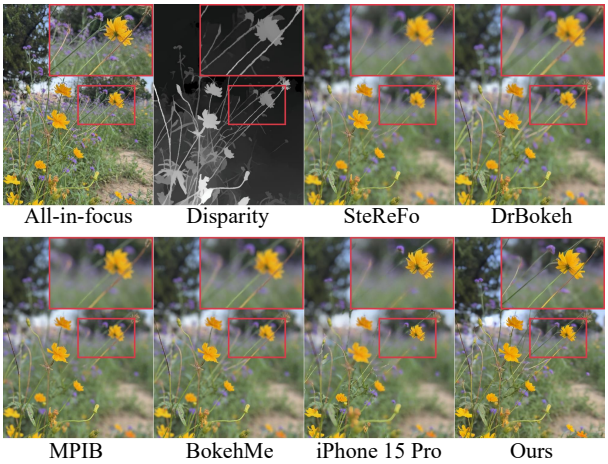

Figure 5: Visual results on IB30 dataset. Our depth-free approach produces more accurate rendering.

dard Gaussian prior $p(z_A) \sim \mathcal{N}(0, I)$, following conventional diffusion-based practices. In contrast, our approach directly starts the transport process from the latent representation of the all-in-focus input image $z_A = \mathcal{E}(\mathcal{I}_A)$. As shown in Table 3, initializing the transport from the latent image representation leads to significantly improved performance in terms of global metrics (PSNR/SSIM) and edge-aware metrics ($PSNR_{eg}$/$SSIM_{eg}$). This highlights the benefit of latent-space alignment between input and target domains for efficient and stable bokeh rendering.

Bokeh Control Adapter . To validate the effectiveness of semantic control mechanism, we compare with a baseline that removes the Bokeh Control Adapter and injects bokeh controls through concatenating bokeh parameters to all-in-focus image. The vector field regressor is conditioned on the concatenated image features and bokeh control parameters without any textual guidance. As shown in Table 4, removing BCA leads to a clear drop in both global metrics (PSNR/SSIM) and edge-aware metrics ($PSNR_{eg}$/$SSIM_{eg}$).

Initialization Strategy. As shown in Table 5, we compare different initialization strategies, including Stable Diffusion 2.1 (Rombach et al. 2022), Marigold (Ke et al. 2024), and our proposed DepthFM-based (Gui et al. 2025) initialization. The results demonstrate that models initialized from depth prediction models consistently outperform those initialized from image generation models like SD2.1, highlighting the effectiveness of depth knowledge transfer. By leveraging geometric priors learned during depth estimation, our model gains a better understanding of scene structure, aligning more closely with physical bokeh rendering models. Furthermore, DepthFM incorporates discriminative supervision during training, which enhances depth awareness, leading to better results than initialization from Marigold.

Comparison	Human Preference
Ours vs. DrBokeh (Sheng et al. 2024)	**81.13**% / 18.87%
Ours vs. SteReFo (Busam et al. 2019)	**75.47**% / 24.53%
Ours vs. MPIB (Peng et al. 2022b)	**83.55**% / 16.45%
Ours vs. BokehMe (Peng et al. 2022a)	**83.17**% / 16.83%
Ours vs. iPhone 15 Pro	**69.45**% / 30.55%

Table 6: User study results indicate that users prefer our method regarding visual quality.

4.5 User Study

Since bokeh is an inherently aesthetic effect with strong subjectivity, we conducted a user study on the IB30 dataset to better evaluate the perceptual quality. We compare BokehFlow with SOTA methods and the iPhone 15 Pro from a human-centric perspective. Notably, as Generative Photography (Yuan et al. 2025) does not support images as input, it is excluded. For each baseline, bokeh images are rendered from AiF images and depth maps captured by iPhone, while our method only uses AiF images. During the user study, participants are asked to choose the more realistic and aesthetically pleasing result, with random image order and methods to reduce bias. A total of 52 volunteers participated in the study. As reported in Table 6 and Figure 5, our method was consistently preferred across most scenes, demonstrating superior perceptual quality, particularly in terms of foreground edge sharpness and natural bokeh appearance.

5 Conclusion

In this work, we propose BokehFlow, a depth-free controllable bokeh rendering framework with efficient flow matching. Current methods either rely on accurate depth or suffer from slow sampling and limited control. BokehFlow formulates bokeh rendering as direct distribution transport in latent space. By leveraging the flow matching paradigm, our model enables one-shot generation of high-quality bokeh images with fast inference and enhanced edge fidelity. Extensive experiments on our synthetic and real-world datasets show the superiority of our method over existing baselines. BokehFlow achieves state-of-the-art performance not only in standard perceptual metrics, but also in edge-aware evaluations, while significantly improving inference efficiency. Qualitative results and user study further confirm the aesthetic quality and controllability of our rendered results.

Limitations and Future Work. BokehFlow currently covers most practical control scenarios, but its control granularity is limited to discrete focus regions. In future work, we plan to extend it to continuous focal region control.

References

Albergo, M. S.; Goldstein, M.; Boffi, N. M.; Ranganath, R.; and Vanden-Eijnden, E. 2023. Stochastic interpolants with data-dependent couplings. *arXiv preprint arXiv:2310.03725*.

Albergo, M. S.; and Vanden-Eijnden, E. 2022. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*.

Busam, B.; Hog, M.; McDonagh, S.; and Slabaugh, G. 2019. Sterefo: Efficient image refocusing with stereo vision. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 0–0.

Chen, H.; Zhang, Y.; Cun, X.; Xia, M.; Wang, X.; Weng, C.; and Shan, Y. 2024. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7310–7320.

Dao, Q.; Phung, H.; Nguyen, B.; and Tran, A. 2023. Flow matching in latent space. *arXiv preprint arXiv:2307.08698*.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Fischer, J. S.; Gui, M.; Ma, P.; Stracke, N.; Baumann, S. A.; and Ommer, B. 2023. Boosting latent diffusion with flow matching. *arXiv preprint arXiv:2312.07360*.

Fortes, A.; Wei, T.; Zhou, S.; and Pan, X. 2025. Bokeh Diffusion: Defocus Blur Control in Text-to-Image Diffusion Models. *arXiv preprint arXiv:2503.08434*.

Gui, M.; Schusterbauer, J.; Prestel, U.; Ma, P.; Kotovenko, D.; Grebenkova, O.; Baumann, S. A.; Hu, V. T.; and Ommer, B. 2025. DepthFM: Fast Generative Monocular Depth Estimation with Flow Matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3203–3211.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ignatov, A.; Patel, J.; and Timofte, R. 2020. Rendering natural camera bokeh effect with deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 418–419.

Ke, B.; Obukhov, A.; Huang, S.; Metzger, N.; Daudt, R. C.; and Schindler, K. 2024. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9492–9502.

Kingma, D. P. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P.; Welling, M.; et al. 2013. Auto-encoding variational bayes.

Kutskir, I. 2016. Photopea Online Photo Editor. https://www.photopea.com.

Li, J.; Wang, Y.; Zheng, J.; Zhang, J.; Shen, L.; Liu, T.; and Cao, Z. 2025. CH3Depth: Efficient and Flexible Depth Foundation Model with Flow Matching. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 7222–7232.

Lipman, Y.; Chen, R. T.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Liu, X.; Gong, C.; and Liu, Q. 2022. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*.

Luo, X.; Peng, J.; Xian, K.; Wu, Z.; and Cao, Z. 2020. Bokeh rendering from defocus estimation. In *European Conference on Computer Vision*, 245–261. Springer.

Luo, X.; Peng, J.; Xian, K.; Wu, Z.; and Cao, Z. 2023. Defocus to focus: Photo-realistic bokeh rendering by fusing defocus and radiance priors. *Information Fusion*, 89: 320–335.

Luo, Y.; Shi, M.; Shen, L.; Huang, Y.; Ye, Z.; Peng, J.; and Cao, Z. 2024. Video Bokeh Rendering: Make Casual Videography Cinematic. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7677–7685.

Neklyudov, K.; Brekelmans, R.; Severo, D.; and Makhzani, A. 2023. Action matching: Learning stochastic dynamics from samples. In *International conference on machine learning*, 25858–25889. PMLR.

Peng, J.; Cao, Z.; Luo, X.; Lu, H.; Xian, K.; and Zhang, J. 2022a. Bokehme: When neural rendering meets classical rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16283–16292.

Peng, J.; Cao, Z.; Luo, X.; Xian, K.; Tang, W.; Zhang, J.; and Lin, G. 2024. BokehMe++: Harmonious Fusion of Classical and Neural Rendering for Versatile Bokeh Creation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Peng, J.; Zhang, J.; Luo, X.; Lu, H.; Xian, K.; and Cao, Z. 2022b. Mpib: An mpi-based bokeh rendering framework for realistic partial occlusion effects. In *European Conference on Computer Vision*, 590–607. Springer.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Seizinger, T.; Vasluianu, F.-A.; Conde, M. V.; Wu, Z.; and Timofte, R. 2025. Bokehlicious: Photorealistic bokeh rendering with controllable apertures. *arXiv preprint arXiv:2503.16067*.

Sheng, Y.; Yu, Z.; Ling, L.; Cao, Z.; Zhang, X.; Lu, X.; Xian, K.; Lin, H.; and Benes, B. 2024. Dr. bokeh: differentiable occlusion-aware bokeh rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4515–4525.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. pmlr.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Wadhwa, N.; Garg, R.; Jacobs, D. E.; Feldman, B. E.; Kanazawa, N.; Carroll, R.; Movshovitz-Attias, Y.; Barron, J. T.; Pritch, Y.; and Levoy, M. 2018. Synthetic depth-of-field with a single-camera mobile phone. *ACM Transactions on Graphics (ToG)*, 37(4): 1–13.

Wang, L.; Shen, X.; Zhang, J.; Wang, O.; Lin, Z.; Hsieh, C.-Y.; Kong, S.; and Lu, H. 2018. Deeplens: Shallow depth of field from a single image. *arXiv preprint arXiv:1810.08100*.

Xian, K.; Shen, C.; Cao, Z.; Lu, H.; Xiao, Y.; Li, R.; and Luo, Z. 2018. Monocular relative depth perception with web stereo data supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 311–320.

Xiao, L.; Kaplanyan, A.; Fix, A.; Chapman, M.; and Lanman, D. 2018. Deepfocus: Learned image synthesis for computational display. In *ACM SIGGRAPH 2018 Talks*, 1–2.

Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024. Depth anything v2. *Advances in Neural Information Processing Systems*, 37: 21875–21911.

Yang, Y.; Lin, H.; Yu, Z.; Paris, S.; and Yu, J. 2016. Virtual dslr: High quality dynamic depth-of-field synthesis on mobile platforms. *Electronic Imaging*, 28: 1–9.

Yuan, Y.; Wang, X.; Sheng, Y.; Chennuri, P.; Zhang, X.; and Chan, S. 2025. Generative photography: Scene-consistent camera control for realistic text-to-image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 7920–7930.

Zhang, X.; Matzen, K.; Nguyen, V.; Yao, D.; Zhang, Y.; and Ng, R. 2019. Synthetic defocus and look-ahead autofocus for casual videography. *arXiv preprint arXiv:1905.06326*.