BokehCrafter: Taming Video Diffusion Models for Controllable Bokeh Rendering

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

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

Abstract

Bokeh is used in photography to emphasize the selected subject by smoothly blurring the out-of-focus region with appealing highlights. While recent advances have achieved impressive results in rendering realistic blur, existing frameworks typically rely on disparity maps and bokeh-relevant inputs (*e.g.*, focal distance and blur size), and face significant challenges in video bokeh rendering due to limited temporal consistency. In this paper, we propose *BokehCrafter*, the first video diffusion framework that generates temporally coherent and visually pleasing bokeh effects from all-in-focus video inputs under user-friendly input conditions. Specifically, we leverage a dual-stream attention mechanism, integrating a reference image branch and a rendering instruction branch. We propose a Bokeh Image Extraction (BIE) module and a CLIP-based text encoder to extract image and text features, respectively, whose outputs are fused via a Text-Image Fusion (TIF) module to enable fine-grained and controllable bokeh rendering. To support the novel capabilities of our model, we construct Video Bokeh Scenes (VBS), a large-scale dataset containing a wide variety of bokeh videos with corresponding rendering instructions, across various scenes and rendering settings. Extensive experiments demonstrate that our method significantly outperforms state-of-the-art methods in both bokeh rendering quality and temporal consistency.

1 Introduction

Bokeh, a term derived from the Japanese word *boke*, refers to the aesthetically pleasing blur produced in the out-of-focus regions of an image. This effect is highly favored by photography enthusiasts due to its strong visual impact and aesthetic appeal. However, achieving high-quality bokeh effects typically requires expensive professional equipment and expertise. Generating controllable bokeh effects from all-in-focus inputs with simple and user-friendly conditions is highly desirable and promising, yet remains under-explored.

Recent advances have achieved remarkable results in bokeh rendering. However, three fundamental challenges persist: 1) the lack of modeling the temporal consistency; 2) a strong dependency on the input disparity maps; 3) the need for complex input control parameters, such as focal distance and blur size. Classic bokeh rendering methods (Luo et al. 2020; Busam et al. 2019; Zhang et al. 2019; Sheng et al. 2024; Srinivasan et al. 2018; Wadhwa et al. 2018; Yang et al. 2016) simulate the camera physical model through algorithms to generate controllable bokeh effects. Neural rendering methods (Peng et al. 2022a,b; Luo et al. 2024; Wang et al. 2018; Xiao et al. 2018; Zheng et al. 2022; Mandl et al. 2024; Seizinger et al. 2025) learn the underlying mapping from all-in-focus images to corresponding bokeh outputs by training deep neural networks on large-scale datasets. Nevertheless, most methods are designed to perform bokeh rendering on a single all-in-focus image, and lack temporally consistent modeling, which leads to flickering and artifacts between frames during video rendering. Furthermore, the rendering qualities of these methods heavily rely on the accuracy of the input disparity maps. Finally, to enable controllable bokeh rendering, they often require additional input parameters such as focal distance and blur size, further hindering practical applications by users.

To alleviate these issues, we propose *BokehCrafter*, a video diffusion-based framework for controllable bokeh rendering from all-in-focus videos. As shown in Figure 1, unlike existing methods that require disparity maps and bokeh parameters (*e.g.*, focal distance and blur size), BokehCrafter is guided by a textual instruction and a reference image containing the desired bokeh style, offering a more intuitive and user-friendly control approach. To achieve fine-grained and flexible control over the rendering process, we introduce a dual-stream attention mechanism. A Bokeh Image Extraction (BIE) module first derives bokeh-related visual features from the reference image, while a CLIP-based text encoder (Radford et al. 2021) extracts semantic cues from the textual instruction. These two modalities are then fused via a Text-Image Fusion (TIF) module, which jointly conditions the diffusion process and guides the generation of temporally consistent bokeh video. Furthermore, we design dedicated training and inference strategies to improve the model's generalization capability and controllability.

The controllable bokeh rendering capability of our model is largely enabled by our constructed dataset, Video Bokeh Scenes (VBS). This dataset consists of 6,100 diverse scenes, with 5,760 used for training, 300 for testing, and 40 for validation. To generate paired data, we use a rendering engine similar to related works (Yuan et al. 2024; Fortes et al. 2025) to render bokeh effects for each scene. Specifically, each

*Corresponding Author.

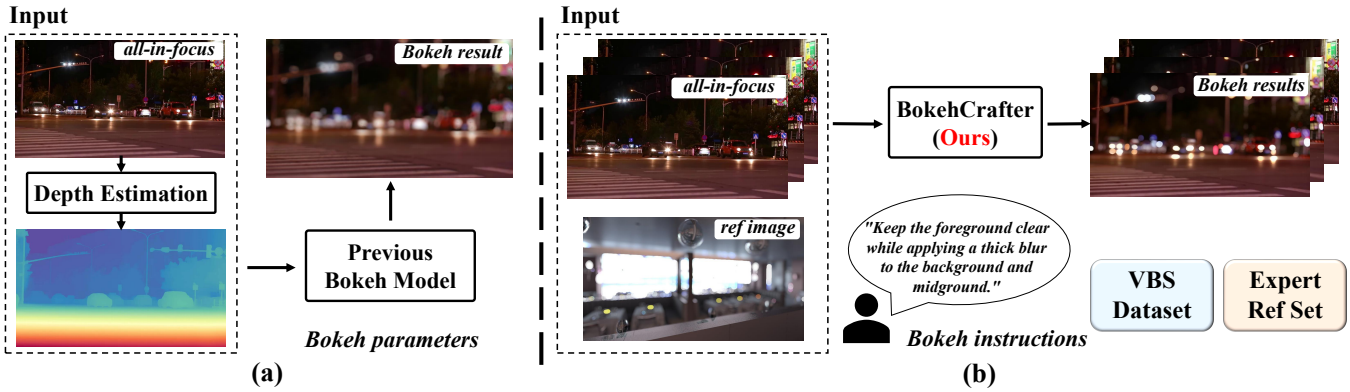


Figure 1: (a) Current methods require disparity maps and complex bokeh-related parameters to control rendering. (b) Our method only needs a reference image and an instruction to guide rendering.

scene is rendered under 5 focal distances and 3 levels of blur size, resulting in 15 bokeh variants per scene. This process yields a total of 86,400 training videos, 4,500 test videos, and 600 validation videos. For every video, a corresponding rendering instruction is provided to guide the model in performing bokeh rendering. Additionally, to support real-world use cases where users may not have access to reference bokeh images, we collect a set of real-world bokeh images using a professional DSLR under varying optical conditions. This collection, referred to as the Expert Ref Set, serves as a reference pool during inference.

Experimental results demonstrate that BokehCrafter outperforms state-of-the-art methods by a large margin. Ablation studies further validate the effectiveness of our key designs. We also conduct a user study to evaluate the zero-shot generalization capability of our approach on real-world videos. Our main contributions are summarized as follows:

- We propose the first video diffusion model for controllable bokeh rendering, guided by text instructions and reference images, without relying on disparity maps or bokeh-specific parameters.
- We propose a large-scale dataset Video Bokeh Scenes (VBS), which encompasses a diverse range of scenes and rendering settings with high-quality instructions.
- Extensive experiments demonstrate that our method outperforms existing approaches in terms of generation quality as well as temporal consistency.

2 Related Work

Bokeh Rendering. Bokeh rendering techniques can be broadly categorized into classical rendering methods and neural rendering methods. Classical approaches are typically classified into two subcategories: object space methods (Abadie et al. 2018; Lee, Eisemann, and Seidel 2010; Wu et al. 2013; Yu, Wang, and Yu 2010; Shen et al. 2025) and image space methods (Sheng et al. 2024; Yang et al. 2016; Barron et al. 2015; Bertalmio, Fort, and Sanchez-Crespo 2004; Hach et al. 2015; Soler et al. 2009). The former relies on rich 3D scene data to simulate optical thin-lens models using physical algorithms for photorealistic rendering. The latter only requires an all-in-focus image as input, but usually incorporates auxiliary techniques to recover scene depth. For example, SteReFo (Busam et al. 2019) utilizes stereo image pairs to estimate depth maps, and RVR (Zhang et al. 2019) applies optical flow to smooth input depth maps and renders each frame individually. While classical methods can generate realistic bokeh effects, they often struggle with areas of depth discontinuities, such as object edges, leading to artifacts. To address these issues, neural rendering methods (Luo et al. 2023a; Peng et al. 2022a,b; Luo et al. 2024; Seizinger et al. 2025; Wang et al. 2018; Xiao et al. 2018) leverage deep neural networks to learn the intrinsic mapping from all-in-focus images to bokeh outputs directly from data. BokehMe (Peng et al. 2022a) integrates classical and neural renderers to exploit their complementary strengths. BokehMe++ (Peng et al. 2024) adds highlight and cat-eye effects based on BokehMe to make the bokeh effect more aesthetic. While Generative photography (Yuan et al. 2024) and Bokeh Diffusion (Fortes et al. 2025) utilize image diffusion models for text-to-image generation, our work focuses on the different task of video-to-video generation. VBR (Luo et al. 2024) introduces a temporal fusion block to leverage information from adjacent frames to improve the temporal consistency but it can only process videos in segments. In this paper, we propose a novel model to render video bokeh effects based on video diffusion model. Our method requires only rendering instructions to control the bokeh rendering effects instead of disparity maps and complex input parameters such as focal distance and blur size. Meanwhile, by fine-tuning the video diffusion model, our approach inherently preserves excellent temporal consistency in bokeh rendering.

Video Diffusion Models. Diffusion models (Ho, Jain, and Abbeel 2020; Sohl-Dickstein et al. 2015; Song, Meng, and Ermon 2020) have demonstrated remarkable generative capabilities. In particular, they have shown notable success in text-to-image (T2I) generation (Nichol et al. 2021; Ramesh et al. 2022; Rombach et al. 2022; Saharia et al. 2022; Brooks, Holynski, and Efros 2023), and have also achieved substantial progress in text-to-video (T2V) and image-to-video (I2V) tasks (Blattmann et al. 2023b; Ge et al. 2023;

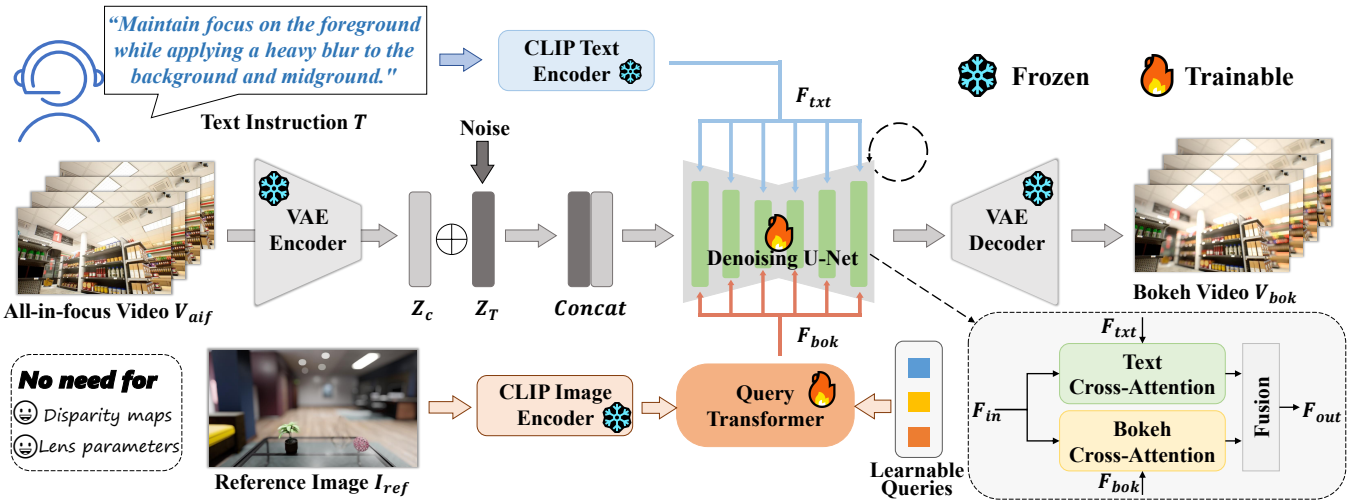


Figure 2: Overview of BokehCrafter. The pipeline takes an all-in-focus video V_{aif}, a text instruction T, and a reference bokeh image I_{ref} as inputs. The text instruction is encoded by a frozen CLIP text encoder to obtain semantic features F_{txt}, while bokeh-related visual cues F_{bok} are extracted via a Bokeh Image Extraction (BIE) module using a CLIP image encoder and a trainable Query Transformer. These features are fused through a Text-Image Fusion (TIF) module to guide the denoising U-Net. The denoised latents are finally decoded into a bokeh video V_{bok}.

He et al. 2022; Ho et al. 2022; Luo et al. 2023b; Singer et al. 2022; Wang et al. 2023, 2024; Zhou et al. 2022; Xing et al. 2024; Liu et al. 2023; Cheng, Xiao, and He 2023; Yang et al. 2024; Brooks et al. 2024; Blattmann et al. 2023a; Chen et al. 2023). Among these methods, VDM (Ho et al. 2022) constitutes the earliest work on video generation with diffusion models, Sora (Brooks et al. 2024) achieves impressive performance in this domain, and SVD (Blattmann et al. 2023a) offers widely-used open-source models for image-to-video generation. Constructing well-curated video datasets for training, video diffusion models can generate high-quality videos and are used as the model prior for various video-related tasks. In this paper, we leverage them to achieve high-quality and temporally consistent video bokeh rendering, thereby maintaining generalization to open-world video scenarios.

3 BokehCrafter

In this section, we introduce **BokehCrafter**, a diffusion-based framework for video bokeh rendering, guided by both text instructions and reference bokeh images. We first formulate the task as a conditional denoising diffusion process (Sec. 3.1). Then, we detail the network architecture, including a Bokeh Image Extraction (BIE) module and a Text-Image Fusion (TIF) module for condition integration (Sec. 3.2). Finally, we present several training and inference strategies designed to improve generalization and controllability (Sec. 3.3).

3.1 Problem Formulation

Different from conventional bokeh rendering methods, we are the first to formulate video bokeh rendering as a conditional denoising diffusion process. Given an all-in-focus input video $V_{\text{aif}} \in \mathbb{R}^{F\times 3\times H\times W}$, a text instruction T, and a reference bokeh image $I_{\text{ref}} \in \mathbb{R}^{3\times H\times W}$, the objective is to render an output bokeh video $V_{\text{bok}} \in \mathbb{R}^{F\times 3\times H\times W}$ by modeling the conditional distribution:

$$p(V_{\text{bok}} \mid V_{\text{aif}}, I_{\text{ref}}, T). \tag{1}$$

To improve efficiency, we perform diffusion in a latent space using a pretrained video variational autoencoder (VAE), consisting of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. The input video is first encoded into a compact latent representation, as $Z_0 = \mathcal{E}(V_{\text{aif}}) \in \mathbb{R}^{F\times C\times h\times w}$, where C, h, and w denote the channel and spatial dimensions of the latent space. To train the denoising network, Gaussian noise is added to Z_0 following a fixed variance schedule:

$$Z_t = \sqrt{\bar{\alpha}_t}\, Z_0 + \sqrt{1-\bar{\alpha}_t}\, \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \tag{2}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$. The denoising network ϵ_θ is trained to recover the noise from the corrupted latent and the condition:

$$\mathcal{L} = \mathbb{E}_{t, Z_0, \boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \epsilon_\theta(Z_t, t, \mathbf{c}_I, \mathbf{c}_T)\|_2^2\right], \tag{3}$$

where $\mathbf{c}_I$ and $\mathbf{c}_T$ denote the reference image and the text instruction embedding, respectively. After denoising, the predicted clean latent Z' is decoded back into the output video $V_{\text{bok}} = \mathcal{D}(Z')$. This formulation enables flexible and controllable rendering of realistic bokeh effects, guided by both visual and semantic cues.

3.2 Network Architecture

As illustrated in Figure 2, BokehCrafter takes an all-in-focus video V_{aif}, a text instruction T, and a reference bokeh image I_{ref} as inputs. The text instruction T is encoded by CLIP (Radford et al. 2021) into semantic features F_{txt}. In parallel, a Bokeh Image Extraction (BIE) module extracts

bokeh-relevant visual representations F_{bok} from the reference image I_{ref}. These two modalities are subsequently integrated by a Text-Image Fusion (TIF) module to guide the bokeh rendering process.

Bokeh Image Extraction (BIE). We first employ CLIP (Radford et al. 2021) to extract dense visual features $F_{\text{img}} \in \mathbb{R}^{N\times d}$ from the reference bokeh image I_{ref}, where N is the number of spatial tokens and d is the feature dimension. Since the reference image may not share scene content with the input video, we focus on extracting bokeh-specific attributes such as blur patterns and blur intensity. To distill these attributes, we introduce a set of learnable query tokens $Q \in \mathbb{R}^{N_q\times d}$, where N_q is the number of queries. These queries interact with F_{img} through Q-Former (Li et al. 2023), which applies cross-attention to extract semantically rich and bokeh-relevant cues:

$$F_{\text{bok}} = \text{Q-Former}(Q,\ K = F_{\text{img}},\ V = F_{\text{img}}). \tag{4}$$

The output $F_{\text{bok}} \in \mathbb{R}^{N_q\times d}$ serves as a compact representation of the bokeh style, which is later fused with the text features F_{txt} in the Text-Image Fusion (TIF) module.

Text-Image Fusion (TIF). To effectively integrate the text feature F_{txt} and the reference bokeh feature F_{bok}, we adopt a Text-Image Fusion (TIF) module. Specifically, we employ a dual-stream cross-attention architecture, where the text features F_{txt} and bokeh features F_{bok} independently interact with the input backbone feature F_{in} through dedicated cross-attention modules. The resulting outputs are then fused by addition, as formulated below:

$$F_{\text{out}} = \text{TCA}(F_{\text{in}}, F_{\text{txt}}) + \text{ICA}(F_{\text{in}}, F_{\text{bok}}), \tag{5}$$

where TCA and ICA refer to the text-based and image-based cross-attention modules, respectively.

3.3 Training and Inference Strategies

To improve generalization and controllability during bokeh rendering, we adopt a set of training strategies, including reference content decoupling and condition dropout for training and dual-condition classifier-free guidance for inference.

Reference Content Decoupling (RCD). Our model requires a reference image as input to provide bokeh information, and this choice is intentionally diverse. To mitigate overfitting to scene content, we select the reference frame from a different scene under the same bokeh conditions, rather than from the same scene. This encourages the model to focus on stylistic bokeh cues rather than memorizing specific scene semantics. Additionally, it also reflects the fact that users typically do not have access to the corresponding bokeh video of the input all-in-focus video during inference.

Condition Dropout. To enhance the model's robustness, we apply condition dropout during training. Specifically, 5% of the samples randomly drop the image condition ($c_I = \phi$), another 5% drop the text condition ($c_T = \phi$), and an additional 5% drop both ($c_I = \phi,\ c_T = \phi$). This strategy exposes the model to all combinations of partial or missing conditions and improves generalization capability.

Dual-Condition Classifier-Free Guidance. To effectively leverage both text and image conditions, we adopt a dual-condition classifier-free guidance (CFG) strategy during inference. At each denoising step, we compute three noise predictions: unconditional $\epsilon_\theta(z_t, \phi, \phi)$, image-only $\epsilon_\theta(z_t, c_I, \phi)$, and fully conditioned $\epsilon_\theta(z_t, c_I, c_T)$. The final prediction is computed as:

$$\begin{aligned}\widetilde{\epsilon}_\theta = \epsilon_\theta(z_t,\phi,\phi) &+ \lambda_I\left[\epsilon_\theta(z_t,c_I,\phi) - \epsilon_\theta(z_t,\phi,\phi)\right] \\ &+ \lambda_T\left[\epsilon_\theta(z_t,c_I,c_T) - \epsilon_\theta(z_t,c_I,\phi)\right],\end{aligned} \tag{6}$$

where λ_I and λ_T control the guidance strengths of the image and text conditions, respectively.

4 VBS Dataset

4.1 Data Construction

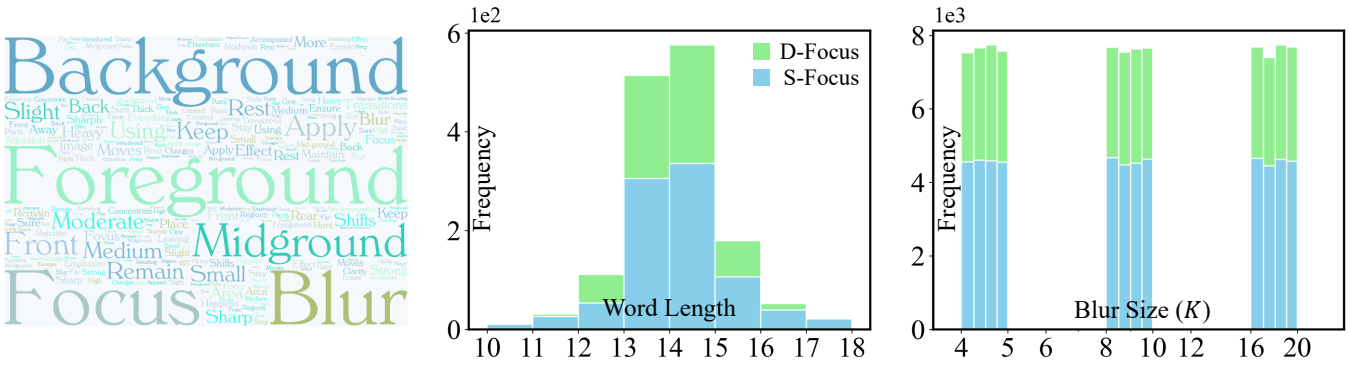
We curate a Video Bokeh Scenes (VBS) dataset with high-quality bokeh video pairs and relevant rendering instructions. Considering that the majority of current bokeh rendering methods require the disparity maps, focal distances, and blur sizes as inputs, we also retain these bokeh-related conditions for fair comparison.

Bokeh Video Pairs and Parameters. Inspired by related works (Yuan et al. 2024; Fortes et al. 2025), we leverage a rendering engine to get the corresponding bokeh videos. We collect all-in-focus videos and corresponding disparity maps from the IRS (Wang et al. 2021) and TartanAir (Wang et al. 2020) datasets. Specifically, bokeh rendering in an optical system refers to the spreading of points into adjacent regions within a radius defined by the defocus blur. The blur radius r is related to the scene depth according to the following simplified formula:

$$r = K \cdot \left|\frac{1}{z} - \frac{1}{z_f}\right|, \tag{7}$$

where K indicates the blur size, z and z_f are the depth of the pixel and focal plane, respectively. Obviously, the relative position of z and z_f affects the blur radius r. When z is close to z_f, the blur radius r becomes negligible, resulting in a sharp focus region. As z moves away from z_f, the blur radius r increases accordingly, creating a more pronounced bokeh effect.

Bokeh Instructions. For specific rendering settings, such as focusing on the foreground with small blur size, we generate a pool of text prompts with LLMs (*e.g.*, GPT-4o (Achiam et al. 2023)). Since our work focuses on video-to-video generation, the input all-in-focus video already provides sufficient scene information to the model. Unlike other text-to-video (T2V) (Blattmann et al. 2023b; Ge et al. 2023; He et al. 2022; Ho et al. 2022; Luo et al. 2023b; Singer et al. 2022; Wang et al. 2023, 2024; Zhou et al. 2022; Cheng, Xiao, and He 2023; Brooks et al. 2024; Blattmann et al. 2023a; Chen et al. 2023) or image-to-video (I2V) (Xing et al. 2024; Liu et al. 2023; Yang et al. 2024) generation models, we do not require additional captions of the video content. In contrast, we only need an instruction to control the bokeh rendering process. For each video, we randomly sample a sentence from the corresponding prompts pool as the rendering instruction. This ensures that the generated instructions are diverse and can better simulate the situations where different users use the prompts.



(a) Prompts Cloud **(b) Length of Prompts** **(c) Distribution of *K***

Figure 3: Data analysis on (a) word cloud of prompts, (b) length of prompts, and (c) distribution of the sampled blur size K. D/S-Focus denotes Dynamic/Static Focus, respectively.

4.2 Data Statistics

VBS dataset contains $6.1k$ diverse scenes, with a total of $91.5k$ videos and over $1.4M$ frames. We randomly split VBS into training, validation, and test splits with 5760, 40, and 300 scenes, respectively. We show data statistics of our VBS in Figure 3 and describe the details of scenes, instructions, and bokeh-related parameters distribution as follows.

Scenes. VBS dataset features a diverse range of scenes. Indoor/outdoor scenes represent 47%/53%, respectively. Daytime scenes constitute 76% of the dataset, whereas nightfall and night scenes account for 6% and 18%, respectively. This distribution ensures a diverse representation of lighting conditions and environments, which is crucial for comprehensive analysis and robust model training.

Instructions. Figure 3(a) presents a word cloud of the instructions in VBS dataset. As shown in Figure 3(b), over 90% of the instructions fall within the word range of 12 to 16. This consistent length ensures the instructions are concise yet sufficiently informative to guide the bokeh rendering process effectively. D-Focus denotes dynamic focus, where the focus changes over time, while S-Focus denotes static focus, in which the focus remains fixed throughout the video.

Bokeh Parameters. VBS dataset includes a diverse distribution of bokeh-related parameters, as illustrated in Figure 3(c). This diverse distribution is vital for training models that can accurately predict and render bokeh effects under various conditions.

5 Experiments

5.1 Experimental Settings

Implementation Details. To enhance the model's generalization ability under varying camera motions, we apply random zoom and translation augmentations to the input videos. Training is conducted on 8 A100 GPUs with an initial learning rate of 1×10^{-5}, a batch size of 8, and runs for 50K steps. The sampling process employs DDIM (Song, Meng, and Ermon 2020) with dual-conditioning classifier-free guidance (CFG) (Ho and Salimans 2022). During inference, we additionally curate an Expert Ref Set from which users can directly select reference images to guide the model.

Evaluation Metrics. Similar to VBR (Luo et al. 2024), we evaluate bokeh rendering quality and temporal consistency of videos with a series of existing methods. For the bokeh rendering quality, we evaluate the overall rendering quality using PSNR, SSIM and LPIPS. To measure the temporal consistency, we adopt the following formulation:

$$\sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}|(F_{i+1}^{GT}-F_{i}^{GT})-(F_{i+1}^{pred}-F_{i}^{pred})|^2}, \quad (8)$$

where N is the length of the input video, F^{GT}, F^{pred} indicate the GT bokeh video and the predicted output of different methods, respectively.

Datasets and Baselines. Since the SVB dataset introduced in VBR (Luo et al. 2024) is the only existing video bokeh dataset and is not publicly available, we conduct all evaluations exclusively on our proposed VBS dataset. To comprehensively validate the performance of our model, we compare it with a variety of bokeh rendering methods, including classical rendering approaches such as RVR (Zhang et al. 2019) and SteReFo (Busam et al. 2019), as well as neural rendering methods such as DeepLens (Wang et al. 2018), MPIB (Peng et al. 2022b), BokehMe (Peng et al. 2022a), and VBR (Luo et al. 2024). Following (Peng et al. 2022a), we enhance RVR with weight normalization to alleviate its serious artifacts in case of discontinuous depth. Modified method is marked with a superscript †.

5.2 Results on VBS Dataset

Quantitative comparisons. As shown in Table 1, BokehCrafter consistently outperforms state-of-the-art methods, achieving the best PSNR, SSIM and LPIPS scores. Notably, our model does not require input disparity maps or additional controllable bokeh-related parameters, it relies solely on a user-provided instruction and a reference image that encapsulates the desired bokeh characteristics. The metrics demonstrate our model's strong capability in rendering high-quality bokeh effects from all-in-focus videos while remaining highly user-friendly. Our method also achieves the best

Methods	Small blur size				Medium blur size				Large blur size			
	PSNR↑	SSIM↑	LPIPS↓	Consistency↓	PSNR↑	SSIM↑	LPIPS↓	Consistency↓	PSNR↑	SSIM↑	LPIPS↓	Consistency↓
RVR†	25.41	0.794	0.1559	0.0516	26.69	0.870	0.1542	0.0437	24.19	0.805	0.1657	0.0585
SteReFo	26.70	0.838	0.1421	0.0543	26.85	0.865	0.1472	0.0514	24.07	0.789	0.1668	0.0727
DeepLens	25.94	0.871	0.2161	0.0709	25.66	0.867	0.2383	0.0722	25.30	0.864	0.2412	0.0720
MPIB	26.09	0.831	0.1832	0.0521	26.22	0.851	0.1705	0.0517	25.91	0.868	0.1610	0.0559
BokehMe	27.86	0.888	0.1357	0.0385	27.64	0.874	0.1547	0.0401	26.52	0.860	0.1683	0.0542
VBR	26.66	0.838	0.1405	0.0435	28.01	0.890	0.1432	0.0385	26.23	0.848	0.1624	0.0569
Ours	**28.52**	**0.892**	**0.0845**	**0.0352**	**28.35**	**0.895**	**0.1224**	**0.0371**	**26.70**	**0.872**	**0.1587**	**0.0474**

Table 1: Quantitative results on the VBS dataset. We present the evaluation results under different levels of blur. The best performance is in boldface, while the second is underlined.

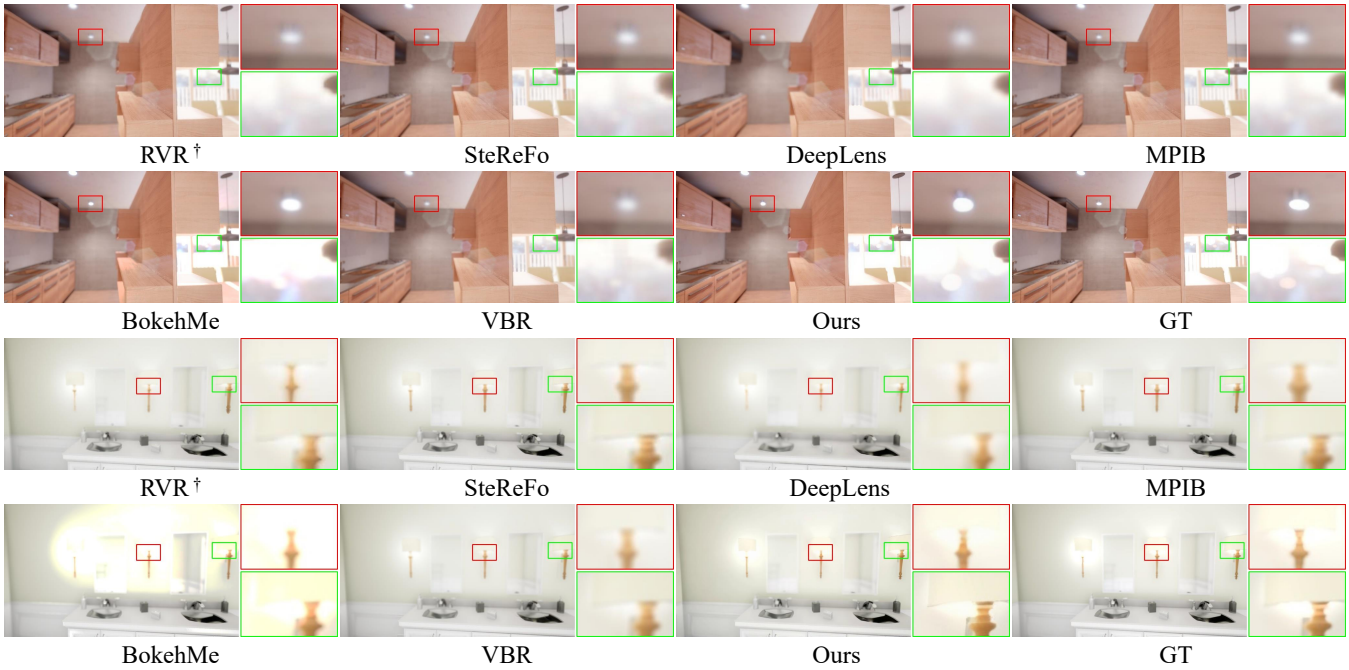


Figure 4: Qualitative comparisons against all baselines. Our method produces more aesthetically pleasing bokeh highlights, and the output is noticeably sharper in the focus position.

temporal consistency, attributed to the temporal attention layers in the diffusion model, which enhance frame-to-frame consistency, highlighting its inherent advantage for video bokeh rendering tasks. Furthermore, most existing methods are highly sensitive to the quality of the input disparity maps. As the accuracy of these maps decreases, the performance of bokeh rendering will deteriorate significantly. We also apply corruption to the disparity maps in the VBS dataset and perform evaluations. Please refer to the supplementary materials for detailed results.

Qualitative comparisons. We present several qualitative examples in Figure 4. The results show that our method produces more aesthetically pleasing bokeh highlights, with outputs that are noticeably sharper and exhibit higher visual clarity compared to existing approaches. Furthermore, our approach demonstrates superior performance in preserving structural details around object boundaries. To valid the effectiveness of our method on real-world videos, as shown in Figure 5, we present a visualization comparison between our method and VBR. Our method outperforms VBR in both bokeh rendering quality and temporal consistency during the focus transition from the foreground to the background.

5.3 Ablation Studies

To validate the effectiveness of our training strategy and model architecture, we conduct ablation studies on each component. In the "w/o RCD" experiment, we evaluate the impact of the Reference Content Decoupling (RCD) strategy. Specifically, we compare our method, which selects the first frame from a different scene video, with selecting the first frame from the same scene. As shown in Table 2, our strategy that selects the reference image from differ-

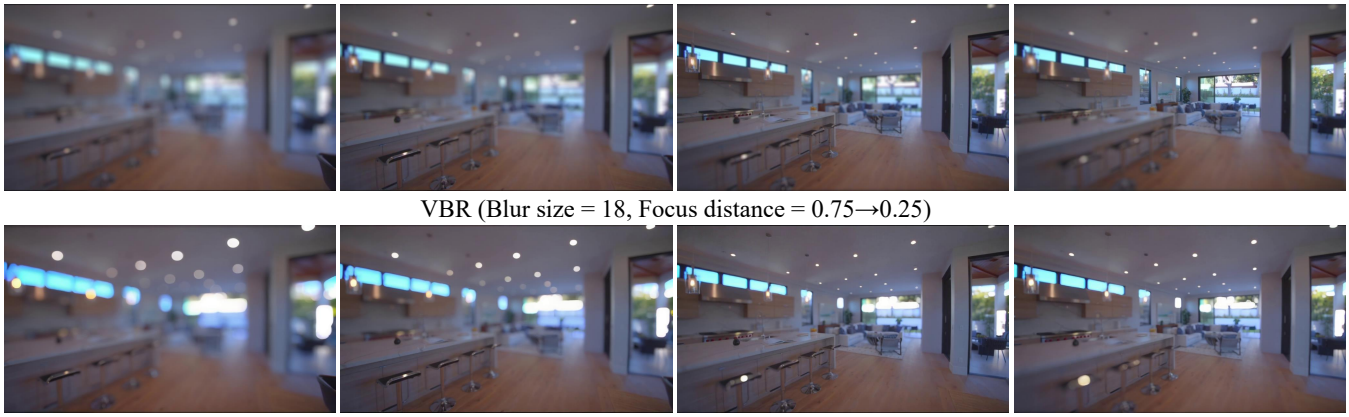


Figure 5: Qualitative results on a real-world video. For this refocusing scene, which transitions from foreground-focused to background-focused, our method produces more aesthetically pleasing light spot effects compared to VBR (Luo et al. 2024).

Methods	PSNR↑	SSIM↑	LPIPS↓	Consistency↓
w/o RCD	26.70	0.871	0.1590	0.0492
w/o BIE	23.31	0.779	0.3527	0.0787
w/o TIF	26.93	0.869	0.1499	0.0470
Ours	**27.86**	**0.886**	**0.1219**	**0.0399**

Table 2: Ablation studies on each component of our method demonstrate their contributions to the overall performance.

Comparison	Human preference
Ours vs. RVR† (Zhang et al. 2019)	**68.44**% / 31.56%
Ours vs. SteReFo (Busam et al. 2019)	**69.03**% / 30.97%
Ours vs. DeepLens (Wang et al. 2018)	**84.76**% / 15.24%
Ours vs. MPIB (Peng et al. 2022b)	**70.11**% / 29.89%
Ours vs. BokehMe (Peng et al. 2022a)	**65.45**% / 34.55%
Ours vs. VBR (Luo et al. 2024)	**62.35**% / 37.65%

Table 3: User study results indicate that users prefer our method as of better quality.

ent scenes yields better results. This demonstrates that using reference images from the same scene during training may lead to overfitting on the training set, thereby reducing the model's generalization ability. In the "w/o BIE" experiment, where we directly fuse the output of the CLIP image encoder with the text stream in TIF, the generation quality significantly degrades. This is because the absence of a dedicated parsing mechanism for these tokens prevents the model from effectively interpreting visual features, resulting in poor outputs. In the "w/o TIF" experiment, where only the text stream is retained, the absence of bokeh reference image information limits the model's understanding of the desired bokeh effect, thereby reducing rendering quality.

5.4 User Study on Real-World Videos

Bokeh is an aesthetic effect with strong subjectivity. PSNR, SSIM, LPIPS and temporal consistency metrics are insufficient to reflect how users perceive the bokeh results. To address this, we further conducted a user study to investigate the performance of our method compared to all baselines from a human perspective. Specifically, we collected 30 real-world videos at a resolution of 1024×576 and generated bokeh videos with identical settings. During the user study, participants were shown two videos at a time, where one is generated by our method and the other is randomly chosen from results produced by competing methods. The order of method selection and the placement of videos were randomized. A total of 91 volunteers were invited to choose the method with better perceptual quality and realism. As shown in Table 3, our method consistently outperforms other approaches by a large margin.

6 Conclusion

In this paper, we introduce the first video diffusion models for the task of bokeh rendering. Leveraging the rich priors inherent in diffusion models and carefully designed modules, our method produces high-quality bokeh effects for input all-in-focus videos without relying on disparity maps or rendering-specific parameters. The user is only required to provide a textual instruction and a reference bokeh image, making the approach both intuitive and user-friendly. To support this novel model, we construct a well-curated paired video bokeh dataset annotated with bokeh instructions for training the video diffusion model. We hope this work will inspire further research in the area of bokeh rendering.

Limitations and future work. As a video diffusion model, our method necessitates multi-step denoising during the inference process, which requires a relatively higher computing cost. We will explore more efficient sampling strategies, such as adaptive step reduction or distillation-based acceleration techniques, in the future to reduce inference time while preserving generation quality.

References

Abadie, G.; McAuley, S.; Golubev, E.; Hill, S.; and Lagarde, S. 2018. Advances in real-time rendering in games. In *ACM SIGGRAPH 2018 Courses*, 1–1.

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Barron, J. T.; Adams, A.; Shih, Y.; and Hernández, C. 2015. Fast bilateral-space stereo for synthetic defocus. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4466–4474.

Bertalmio, M.; Fort, P.; and Sanchez-Crespo, D. 2004. Real-time, accurate depth of field using anisotropic diffusion and programmable graphics cards. In *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004.*, 767–773. IEEE.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023a. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023b. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22563–22575.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18392–18402.

Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.; et al. 2024. Video generation models as world simulators. *OpenAI Blog*, 1: 8.

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.; Xia, M.; He, Y.; Zhang, Y.; Cun, X.; Yang, S.; Xing, J.; Liu, Y.; Chen, Q.; Wang, X.; et al. 2023. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*.

Cheng, J.; Xiao, T.; and He, T. 2023. Consistent video-to-video transfer using synthetic dataset. *arXiv preprint arXiv:2311.00213*.

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*.

Ge, S.; Nah, S.; Liu, G.; Poon, T.; Tao, A.; Catanzaro, B.; Jacobs, D.; Huang, J.-B.; Liu, M.-Y.; and Balaji, Y. 2023. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22930–22941.

Hach, T.; Steurer, J.; Amruth, A.; and Pappenheim, A. 2015. Cinematic bokeh rendering for real scenes. In *Proceedings of the 12th European Conference on Visual Media Production*, 1–10.

He, Y.; Yang, T.; Zhang, Y.; Shan, Y.; and Chen, Q. 2022. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*.

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

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022. Video diffusion models. *Advances in Neural Information Processing Systems*, 35: 8633–8646.

Lee, S.; Eisemann, E.; and Seidel, H.-P. 2010. Real-time lens blur effects and focus control. *ACM Transactions on Graphics (TOG)*, 29(4): 1–7.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Liu, G.; Xia, M.; Zhang, Y.; Chen, H.; Xing, J.; Wang, Y.; Wang, X.; Yang, Y.; and Shan, Y. 2023. Stylecrafter: Enhancing stylized text-to-video generation with style adapter. *arXiv preprint arXiv:2312.00330*.

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. 2023a. 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.

Luo, Z.; Chen, D.; Zhang, Y.; Huang, Y.; Wang, L.; Shen, Y.; Zhao, D.; Zhou, J.; and Tan, T. 2023b. Videofusion: Decomposed diffusion models for high-quality video generation. *arXiv preprint arXiv:2303.08320*.

Mandl, D.; Mori, S.; Mohr, P.; Peng, Y.; Langlotz, T.; Schmalstieg, D.; and Kalkofen, D. 2024. Neural Bokeh: Learning Lens Blur for Computational Videography and Out-of-Focus Mixed Reality. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, 870–880. IEEE.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

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.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

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.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Seizinger, T.; Vasluianu, F.-A.; Conde, M. V.; and Timofte, R. 2025. Bokehlicious: Photorealistic Bokeh Rendering with Controllable Apertures. *arXiv preprint arXiv:2503.16067*.

Shen, L.; Liu, T.; Sun, H.; Li, J.; Cao, Z.; Li, W.; and Loy, C. C. 2025. DoF-Gaussian: Controllable Depth-of-Field for 3D Gaussian Splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 26462–26471.

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.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

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.

Soler, C.; Subr, K.; Durand, F.; Holzschuch, N.; and Sillion, F. 2009. Fourier depth of field. *ACM Transactions on Graphics (TOG)*, 28(2): 1–12.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Srinivasan, P. P.; Garg, R.; Wadhwa, N.; Ng, R.; and Barron, J. T. 2018. Aperture supervision for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6393–6401.

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, J.; Yuan, H.; Chen, D.; Zhang, Y.; Wang, X.; and Zhang, S. 2023. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*.

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*.

Wang, Q.; Zheng, S.; Yan, Q.; Deng, F.; Zhao, K.; and Chu, X. 2021. Irs: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Wang, W.; Zhu, D.; Wang, X.; Hu, Y.; Qiu, Y.; Wang, C.; Hu, Y.; Kapoor, A.; and Scherer, S. 2020. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4909–4916. IEEE.

Wang, Y.; Chen, X.; Ma, X.; Zhou, S.; Huang, Z.; Wang, Y.; Yang, C.; He, Y.; Yu, J.; Yang, P.; et al. 2024. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, 1–20.

Wu, J.; Zheng, C.; Hu, X.; and Xu, F. 2013. Rendering realistic spectral bokeh due to lens stops and aberrations. *The Visual Computer*, 29: 41–52.

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.

Xing, J.; Xia, M.; Zhang, Y.; Chen, H.; Yu, W.; Liu, H.; Liu, G.; Wang, X.; Shan, Y.; and Wong, T.-T. 2024. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, 399–417. Springer.

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.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*.

Yu, X.; Wang, R.; and Yu, J. 2010. Real-time depth of field rendering via dynamic light field generation and filtering. In *Computer Graphics Forum*, volume 29, 2099–2107. Wiley Online Library.

Yuan, Y.; Wang, X.; Sheng, Y.; Chennuri, P.; Zhang, X.; and Chan, S. 2024. Generative Photography: Scene-Consistent Camera Control for Realistic Text-to-Image Synthesis. *arXiv preprint arXiv:2412.02168*.

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*.

Zheng, B.; Chen, Q.; Yuan, S.; Zhou, X.; Zhang, H.; Zhang, J.; Yan, C.; and Slabaugh, G. 2022. Constrained predictive filters for single image bokeh rendering. *IEEE Transactions on Computational Imaging*, 8: 346–357.

Zhou, D.; Wang, W.; Yan, H.; Lv, W.; Zhu, Y.; and Feng, J. 2022. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*.