Training-free Focus Control for Diffusion Models

Javad Rajabi

Department of Computer Science, University of Toronto

rajabi@cs.toronto.edu

Motivation

- Modern diffusion models (FLUX [1]) generate highly photorealistic images, **but users cannot control focus or depth-of-field (DoF)**.
- In traditional photography, focus and DoF are key tools that guide a viewer's attention. By adjusting the **aperture** or **focal plane**, photographers can selectively emphasize a subject.
- DoF in generative models emerges implicitly from **training data** rather than being an explicit controllable variable.
- Users often **get unwanted shallow DoF**, background blur, or inconsistent focus across samples.

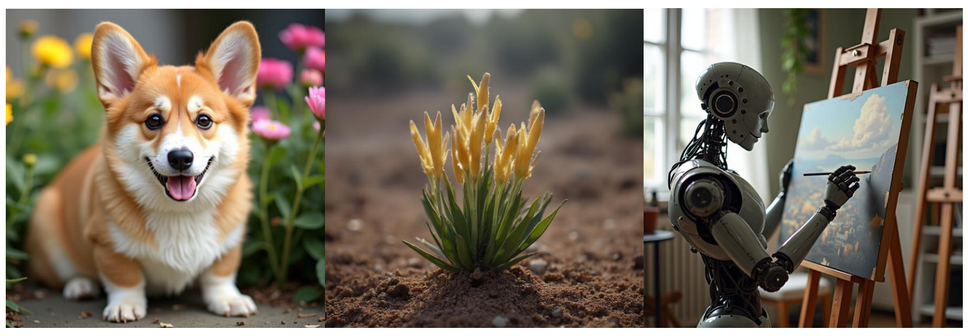

Fig1. Examples of Flux-generated images, each exhibiting unwanted shallow depth of field.

- Existing solutions require large paired datasets , expensive training or fine-tuning and Specialized conditioning modules.
- Our Goal is to Create **a training-free, plug-and-play** method to avoid unwanted focus and bokeh effect during inference using existing diffusion models.

Related Work

- Previous methods for focus control in diffusion models such as **Bokeh Diffusion** [2] and **DiffCamera** [3] achieve controllable blur and refocusing by **training models on large datasets** of sharp/blurred image pairs.
- These approaches rely on **explicit conditioning mechanisms** and **specialized losses** to model depth-of-field, but they require heavy retraining and significant computational resources.
- Recent diffusion and flow matching models such as **Flux** [1] naturally exhibit characteristics of focus, bokeh and depth separation even without being explicitly trained for these effects.
- This suggests that the model has already learned rich internal representations of scene geometry and depth cues from large-scale visual data.

References

[1] B. F. Labs, S. Batifol, A. Blattmann, F. Boesel, S. Consul, C. Diagne,T. Dockhorn, J. English, Z. English, P. Esser et al., "Flux. 1 kontext: Flowmatching for in-context image generation and editing in latent space,"arXiv preprint arXiv:2506.15742, 2025.
[2] A. Fortes, T. Wei, S. Zhou, and X. Pan, "Bokeh diffusion: Defocus blur control in text-to-image diffusion models," arXiv preprintarXiv:2503.08434, 2025.
[3] Y. Wang, X. Chen, X. Xu, Y. Liu, and H. Zhao, "Diffcamera: Arbitraryrefocusing on images," arXiv preprint arXiv:2509.26599, 2025.
[4] Bansal A, Chu HM, Schwarzschild A, Sengupta S, Goldblum M, Geiping J, Goldstein T. Universal guidance for diffusion models. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition 2023

Method Overview

- Diffusion models already contain an **internal representation** of depth and focus.
- Leveraging this, the proposed method introduces a training-free approach that **guides the sampling process** directly using **depth cues and differentiable focus loss.**
- **Pipeline (Repeated at Each Diffusion Step):**

1. **Predict Intermediate $\hat{x}_0$:** Use the model's internal step to estimate the clean image.
2. **Estimate Depth:** Pass $\hat{x}_0$ through a pretrained monocular depth estimator. Provides coarse geometry for deciding focus regions.
3. **Define Target Focus Mask:**

$$M_{\text{target}}(p) = \exp\left(-\frac{\left(D(p)-d_{\text{target}}\right)^2}{2\sigma_{\text{dof}}^2}\right)$$

4. **Compute Current Sharpness Map:**

$$S(p) = \text{sigmoid}(\alpha|\hat{x}_0(p)|)$$

5. **Compute Current Sharpness Map:**

$$L_{\text{focus}} = \frac{1}{N}\sum_p \left(S(p) - M_{\text{target}}(p)\right)^2$$

6. **Gradient-Based Guidance [4]:**

$$z_{t-1} = f_\theta(z_t, t) - \lambda.\nabla_{z_t} L_{\text{focus}}$$

Results

Figure 2: Qualitative evaluation comparing the base model (top row) and our method (bottom row).

Figure 3: Qualitative illustration of failure cases.

	Clip Score
Base	0.3457
Ours	**0.3537**

Table 1: Quantitative comparison of CLIP scores.