# Training-free Focus Control for Diffusion Models

Final Project Report
Computational Imaging - CSC2529
Fall 2025

Javad Rajabi
*Department of Computer Science*
*University of Toronto*
mj.rajabikhoozani@mail.utoronto.ca

***Abstract*—Recent text-to-image diffusion models can generate highly realistic and semantically rich images, yet they lack explicit control over photographic properties such as focus and depth-of-field (DoF). Although many models implicitly learn shallow-focus aesthetics, users have no reliable way to adjust or correct these effects without retraining or post-processing. In this work, we propose a training-free method for controllable focus manipulation in diffusion models by leveraging intermediate denoised predictions and depth maps estimated during sampling. Our approach constructs a differentiable focus-based loss that measures the discrepancy between the current and desired focus distributions and its gradient is used to guide the sampling trajectory via universal guidance. This enables sharpening of background regions and reduction of unintended bokeh without modifying the underlying model weights. Results show that our method increases background detail while preserving subject clarity across diverse prompts. We further analyze failure cases and discuss limitations of our proposed method. Overall, our findings demonstrate the feasibility of training-free focus control and offer a foundation for more robust controllable photographic effects in future generative models.**

## I. Introduction

Recent advances in text-to-image generative models have shown impressive capabilities in producing realistic, high-quality images. In particular, diffusion models [1] have rapidly become the leading paradigm for high-quality image generation and editing, powering state-of-the-art systems such as Stable Diffusion [2] and FLUX [3]. Despite their remarkable ability to generate photorealistic and semantically coherent images, one crucial aspect of visual control remains underexplored: focus manipulation. In photography, focus and depth-of-field (DoF) are key artistic tools that guide visual attention, emphasize subjects and control background separation. However, in diffusion-based generation, focus is an emergent and implicit property that users cannot easily adjust without retraining the model or performing post-processing.

Photography has offered a rich set of camera controls for shaping visual aesthetics. In traditional photography, DoF offers a wide range of creative possibilities. The intentional blurriness brought by DoF in photographs, known as bokeh effects, emphasizes specific subjects. By varying the lens aperture and focal distance, photographers can decide whether a scene appears deeply in focus or softly blurred in the foreground or background [4].

In contrast, current generative models lack the ability to make such fine-grained adjustments. Diffusion models are not explicitly designed to reproduce photographic effects such as bokeh or depth-of-field. These properties, when they appear, emerge implicitly from the learned data distribution rather than from any explicit modeling of camera optics. The above problems raise a natural question: *Can we have training-free focus control in a diffusion model that has already been trained?*

In the rapidly evolving domain of text-to-image (T2I) generation, reproducing such camera effects is surprisingly difficult. Existing approaches typically rely on prompt engineering (e.g., adding "f/1.8 aperture" to the text prompt), which often yields inaccurate approximations and unintentionally alters the scene content. Recent works have started to explore controlling camera intrinsics in image generation via camera tokens [5]. Others model camera settings along the temporal axis of a text-to-video framework to improve consistency, requiring the generation of multiple frames per prompt at inference for better results [6]. This introduces substantial computational overhead and redundancy.

Previous research [7]–[9] has attempted to address this issue by explicitly training diffusion models for focus control. Works such as Bokeh Diffusion [7] and DiffCamera [8] extend generative frameworks with explicit DoF conditioning or refocusing modules. These methods rely on large datasets of image pairs, consisting of all-in-focus and blurred counterparts, annotated with camera parameters such as aperture and focus distance. While effective, such training-based pipelines are computationally intensive and data-hungry.

Through observation of Flux-generated images, we found that the model often produces outputs with bokeh-style characteristics (a clear subject in the foreground and a softly blurred background) even when the input prompt contains no mention of depth, blur, or focus. This implicit bias toward a shallow depth-of-field suggests that the model has learned to associate such optical cues with aesthetic quality during training.

While this default behavior often enhances the visual appeal of generated images, it also exposes a key limitation: the user has no direct control over the level or placement of focus. In some cases, users may want to produce all-in-focus images, yet the model inherently introduces artistic blur around the background or edges. Interestingly, this behavior is not

uniform, different generated samples exhibit varying degrees of background blur, indicating that Flux possesses an internal understanding of bokeh, focus, and depth relationships but lacks a controllable interface to manipulate them.

In this project, we argue that retraining is unnecessary for a model like Flux, which already exhibits a strong internal representation of focus and DoF learned from diverse visual data. Instead, we propose a training-free approach to control focus and bokeh levels directly at inference time. By estimating the intermediate denoised image $\hat{x}_0$ at each timestep and computing its corresponding depth map, we can infer the current focus plane of the generated image. Given a target (conditioned) focus plane, we then define a differentiable loss function that measures the deviation between the current and desired focus distributions. The gradient of this loss can be used as a guidance signal to steer the diffusion process toward the specified focus condition, enabling controllable refocusing and depth-of-field adjustment without any fine-tuning or retraining.

Fig. 1. Example images generated by the model, illustrating its tendency to produce bokeh-like compositions with a sharp foreground subject and a softly blurred background, even in the absence of any depth, blur or focus-related cues in the prompt.

## II. Related Works

### A. Simulating bokeh

Bokeh rendering aims at simulat- ing bokeh effects on all-in-focus images based on their depth map. Classic rendering methods simulate the blur effect by estimating the blur level of each pixel based on its location on different depth layers using different algorithms [10]. However, these methods suffer from color leakage and artifacts at the boundaries of depth discontinuities. Neural-network-based rendering methods implicitly simulate the blur calculation to avoid artifacts by learning from image statistics [11].

### B. Text-to-image diffusion

Diffusion models [12] have become the dominant approach in text-to-image (T2I) generation due to their high sample quality and diversity. More recently, Stable Diffusion 3 [13] and FLUX [3] have set a new state-of-the-art by scaling diffusion transformer backbones with rectified flows [14]. Despite their advancements in prompt fidelity and scene composition, these models struggle with fine-grained photographic controls, such as depth-of-field manipulation.

### C. Camera-conditioned diffusion models

Most existing works on camera control in diffusion models focus on extrinsic parameters, such as camera position and viewing angle [15] or trajectory in video generation [16]. Previous works like Bokeh Diffusion [7] and DiffCamera [8] demonstrate training-based approaches for focus control in diffusion models. Bokeh Diffusion fine-tunes a pretrained model to adjust the level of defocus blur by conditioning it on a physical blur parameter, using a hybrid dataset of real and synthetically blurred images to train a ”grounded self-attention” mechanism that ensures scene consistency. Similarly, DiffCamera trains a diffusion transformer to allow for arbitrary refocusing by specifying both a focus point and a blur level. It accomplishes this by training on a large-scale dataset of simulated bokeh image pairs and enforcing a ”stacking constraint”, a regularization term derived from photographic principles, to ensure the model learns physically accurate depth-of-field effects. Both methods achieve high-quality results but are dependent on extensive training with specialized paired data. In contrast to these training-dependent methods, our work is predicated on the observation that state-of-the-art models like FLUX already possess a strong, implicit understanding of focus and DoF. We propose a training-free approach that enables direct control over focus and bokeh without the need for specialized datasets or computationally intensive training.

### D. Guided image generation

The surge in diffusion model research is largely attributed to advancements in sampling guidance techniques. Since the score (predicted noise) represents an explicit gradient field, the sampling trajectory can be guided by incorporating auxiliary gradients, leading to powerful guidance techniques such as classifier-free guidance (CFG) [17]. Works in this category employed a frozen pre-trained diffusion model as a foundation model but modify the sampling method to guide the image generation with feedback from either the model itself or a guidance function. These guidance methods significantly enhance image quality and prompt/condition alignment.

Universal guidance [18] proposed a guidance algorithm that augments the image sampling method of a diffusion model to include guidance from an off-the-shelf auxiliary network. This algorithm is motivated by an empirical observation that the reconstructed clean image $\hat{x}_0$ obtained by Equation (4), while naturally imperfect, is still appropriate for a generic guidance function to provide informative feedback to guide the image generation and only requires loss functions to be differentiable, and it also avoids any retraining. We aim to leverage this type of guidance to apply training-free focus control in image generation.

## III. Background

### A. Diffusion Models

In diffusion models [12], random noise $\epsilon \sim \mathcal{N}(0, I)$ is added during forward path to an image $x_0$ to produce a noisy

image $x_t$ at an arbitrary timestep $t$:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \tag{1}$$

with $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ according to a variance schedule $\beta_1, ..., \beta_t$. A denoising network $\epsilon_\theta$ is learned to predict $\epsilon$ by optimizing an objective

$$\mathcal{L} = \mathbb{E}_{x_0,t,\epsilon\sim\mathcal{N}(0,I)} \left[ \|\epsilon - \epsilon_\theta(x_t, t)\|_2^2 \right], \tag{2}$$

for uniformly sampled $t \in \{1, ..., T\}$.

During sampling, the model produces denoised image $x_{t-1}$ from $x_t$ at each timestep $t$ based on the noise estimation $\epsilon_\theta(x_t, t)$ as follows:

$$x_{t-1} = \frac{1}{\sqrt{\bar{\alpha}_t}} \left( x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) + \sigma_t z, \tag{3}$$

where $z \sim \mathcal{N}(0, I)$ and $\sigma_t^2$ is set to $\beta_t$. Starting with randomly sampled noise $x_T \sim \mathcal{N}(0, I)$, the process is applied iteratively to generate a clean image $x_0$. Note that noise estimation of the diffusion model can be considered as $\epsilon_\theta(x_t) \approx -\sigma_t \nabla_{x_t} \log p(x_t)$, where $p(x_t)$ denotes the distribution of $x_t$.

In addition, using the reparameterization trick, it is possible to obtain the intermediate prediction of $x_0$ [19] at a given timestep $t$ as

$$\hat{x}_0 = (x_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(x_t, t))/\sqrt{\bar{\alpha}_t}. \tag{4}$$

### B. Universal guidance

To guide the generation with information from the external guidance function $f$ and the loss function $\ell$, an immediate thought is to extend classifier guidance [20] to accept any general guidance function. Concretely, given a class prompt $c$, classifier guidance performs classification-guided sampling by replacing $\epsilon_\theta(x_t, t)$ in each sampling step $S(x_t, t)$ with

$$\hat{\epsilon}_\theta(x_t, t) = \epsilon_\theta(x_t, t) - \sqrt{1 - \alpha_t} \nabla_{x_t} \log p(c|x_t). \tag{5}$$

Defining $\ell_{ce}(\cdot, \cdot)$ to be the cross-entropy loss and $f_{cl}$ to be the guidance function that outputs classification probability, Equation (5) can be re-writtern as

$$\hat{\epsilon}_\theta(x_t, t) = \epsilon_\theta(x_t, t) + \sqrt{1 - \alpha_t} \nabla_{x_t} \ell_{ce}(c, f_{cl}(x_t)). \tag{6}$$

However, directly replacing $f_{cl}$ and $\ell_{ce}$ with any off-the-shelf guidance and loss functions does not work in practice, as $f$ is most likely trained on clean images and fails to provide meaningful guidance when the input is noisy.

To address the issue, we leverage the fact that $\epsilon_\theta(x_t, t)$ predicts the noise added to the data point, and we can therefore obtain a *predicted* clean image $\hat{x}_0$ by Equation (3). We propose to instead calculate the guidance based on the predicted clean data point as

$$\hat{\epsilon}_\theta(x_t, t) = \epsilon_\theta(x_t, t) + s(t) \cdot \nabla_{x_t} \ell(c, f(\hat{x}_0)) \tag{7}$$

where $s(t)$ controls the guidance strength for each sampling step and

$$\nabla_{x_t} \ell(c, f(\hat{x}_0)) = \nabla_{x_t} \ell \left( c, f \left( \frac{x_t - \sqrt{1 - \alpha_t}\epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}} \right) \right)$$

as in Equation (3).

We term Equation (7) forward universal guidance. In practice, applying this guidance effectively brings the generated image closer to the condition while keeping the generation trajectory in the data manifold.

## IV. Method

Diffusion models (e.g. Flux [3]) already contain an internal representation of depth and focus, meaning they implicitly understand how objects are arranged in foreground and background of the image and how sharply each region should appear. Building on this insight, the proposed method introduces a training-free technique that directly influences the model during generation. Instead of requiring additional datasets or retraining, the approach steers the sampling process in inference by leveraging depth cues and a differentiable focus loss. Together, these components allow the model to more accurately control depth-of-field effects.

The core principle is to steer the generative process toward a user-defined focus target by calculating a guidance signal from an intermediate depth estimate at each denoising step. This avoids the need for model retraining or specialized datasets.

The method can be broken down into the following key steps, which are repeated throughout the diffusion sampling process:

### A. Intermediate Image and Depth Estimation

At any given timestep $t$ in the reverse diffusion process, we first obtain an estimate of the final clean image, denoted as the predicted $\hat{x}_0$ in Equation (3). This predicted $\hat{x}_0$ serves as a plausible preview of the final output. We then feed this intermediate image into a pretrained, off-the-shelf monocular depth estimation network, specifically the pretrained Depth Anything [21] model, to compute its corresponding depth map. This map provides a spatial understanding of the scene's geometry as it is being formed.

### B. Focus-based Differentiable Loss

With the depth map from the current step, we can infer the image's emergent focus properties. The user provides a target focus condition, which can be defined as a target depth plane or a desired focus distribution (e.g., a mask specifying sharp and blurred regions). We then formulate a differentiable loss function that quantifies the deviation between the current focus state of the predicted $\hat{x}_0$ and the desired target focus. This loss function effectively measures how far the generation is from the user's intended photographic style, all-in-focus landscape or a portrait with a shallow DoF.

#### 1. Target Focus Mask ($M_{target}$)

First, we use the estimated depth map $D$ of the predicted image $\hat{x}_0$. The user provides two parameters: a target focus depth $d_{\text{target}}$ and a depth-of-field parameter $\sigma_{\text{dof}}$ that controls the "thickness" of the focal plane. The target focus mask $M_{\text{target}}$ is then generated for each pixel $p$ using a Gaussian function. This function assigns a score close to 1 for pixels

at the target depth and a score that smoothly decays to 0 for pixels farther away.

The formula for the target focus mask is:

$$M_{\text{target}}(p) = \exp\left(-\frac{(D(p) - d_{\text{target}})^2}{2\sigma_{\text{dof}}^2}\right) \tag{8}$$

*2. Current Sharpness Map ($S$)*

Next, we need to measure the current sharpness at every pixel of the predicted image $\hat{x}_0$. A common and effective way to measure sharpness is to compute the magnitude of the image Laplacian $|\hat{x}_0|$ which is high in areas with fine details and low in smooth or blurred regions. To create a normalized map that is comparable to $M_{\text{target}}$, we apply a sigmoid function. Here, $\alpha$ is a scaling hyperparameter that controls the sensitivity of the sharpness detection.

The formula for the sharpness map $S$ is:

$$S(p) = \text{sigmoid}(\alpha \cdot |\hat{x}_0(p)|) \tag{9}$$

*3. Loss Function ($\mathcal{L}_{focus}$)*

Finally, the focus loss is calculated as the Mean Squared Error (MSE) between the current sharpness map $S$ and the target focus mask $M_{\text{target}}$. Minimizing this loss encourages the generated image to have sharpness characteristics that align with the desired focus distribution. In the formula, $N$ is the total number of pixels in the image.

The final loss function is:

$$\mathcal{L}_{\text{focus}} = \frac{1}{N}\sum_{p}(S(p) - M_{\text{target}}(p))^2 \tag{10}$$

### C. Gradient-based Guidance

Since the loss function is differentiable, we can compute its gradient with respect to the latent variable $x_t$ at timestep $t$. This gradient indicates the direction in which the latent should be adjusted to minimize the focus error and make the generated image better match the target focus condition. This gradient is then used as a guidance signal, which contributes to steer the sampling direction at the current step. By applying this corrective signal iteratively, we guid the entire diffusion trajectory toward a final image that satisfies the specified focus and bokeh requirements.

$$\hat{\epsilon}_\theta(x_t, t) = \epsilon_\theta(x_t, t) + \lambda \cdot \nabla_{x_t}\ell_{\text{focus}} \tag{11}$$

Here, $\lambda$ is a weighting coefficient that controls the influence of the focus-based gradient guidance during sampling.

## V. Results

In this section, we evaluate the effectiveness of our training-free focus control method through both qualitative and quantitative experiments. We compare our approach against the baseline FLUX [3] model without focus guidance to assess improvements in controllability, visual consistency, and adherence to user-specified focus conditions. Qualitatively, we demonstrate that our method produces images with sharper subject–background separation, addressing the model's inherent bias toward shallow focus observed in earlier analyses. Quantitatively, we report performance using standard metrics such as CLIP score. Together, these results highlight the capability of our approach to introduce focus manipulation without retraining, while maintaining the generative quality of the underlying diffusion model.

In practice, we found that varying the values of $d_{\text{target}}$ and $\sigma_{\text{dof}}$ had only a limited and often negligible effect on the generated results. Despite the formulation of our focus mask, achieving global focus manipulation across the entire image remained challenging, as the diffusion model tends to preserve its inherent depth and sharpness priors. To obtain more consistent behavior, we therefore set $d_{\text{target}}$ to the farthest depth value in the predicted depth map, ensuring that the background region is always selected as the target focus plane. We additionally fixed $\sigma_{\text{dof}} = 0.05$ for all experiments, since smaller values produced unstable gradients while larger values broadened the focus mask to the point that the effect became minimal. We also observed that the guidance weight $\lambda$ must be carefully chosen: values below $0.5$ resulted in nearly imperceptible changes, whereas larger values caused oversharpening, oversaturation or even semantic drift where the content of the generated image changed unintentionally. Finally, to ensure fair comparisons between the baseline and our method, all experiments were conducted using the same random seed and initial noise to guarantee alignment in the sampling trajectory.

### A. Qualitative results

Our qualitative evaluation demonstrates that the proposed training-free focus-control method reduces unintended background blur and produces images with more sharp details. As shown in Figure 2, the baseline FLUX [3] model often exhibits an implicit bias toward shallow focus, generating scenes where the background is noticeably softer which makes our earlier observations. In contrast, our method yields sharper and more detailed background regions while preserving the integrity of the main subject. In the first example, the young boy and surrounding environment retain the original structure, but the background becomes significantly clearer and more in focus. Similarly, in the astronaut example, the lunar surface gains additional texture and detail while maintaining consistency in lighting and geometry. The same trend appears in the robot-painter example, where the workshop environment becomes more defined without altering the semantic content of the scene. Across diverse prompts and visual styles, these results highlight our method's ability to correct for excessive or undesired bokeh while maintaining overall scene coherence.

### B. Quantitative results

To obtain a numerical comparison of our method against the baseline, we use the CLIP score as discussed earlier. Because there is no direct metric for measuring the degree of bokeh reduction or background sharpness, CLIP score serves as an indirect indicator of overall semantic alignment between the generated image and the input prompt. For each method, we generated approximately 1000 samples and report the average

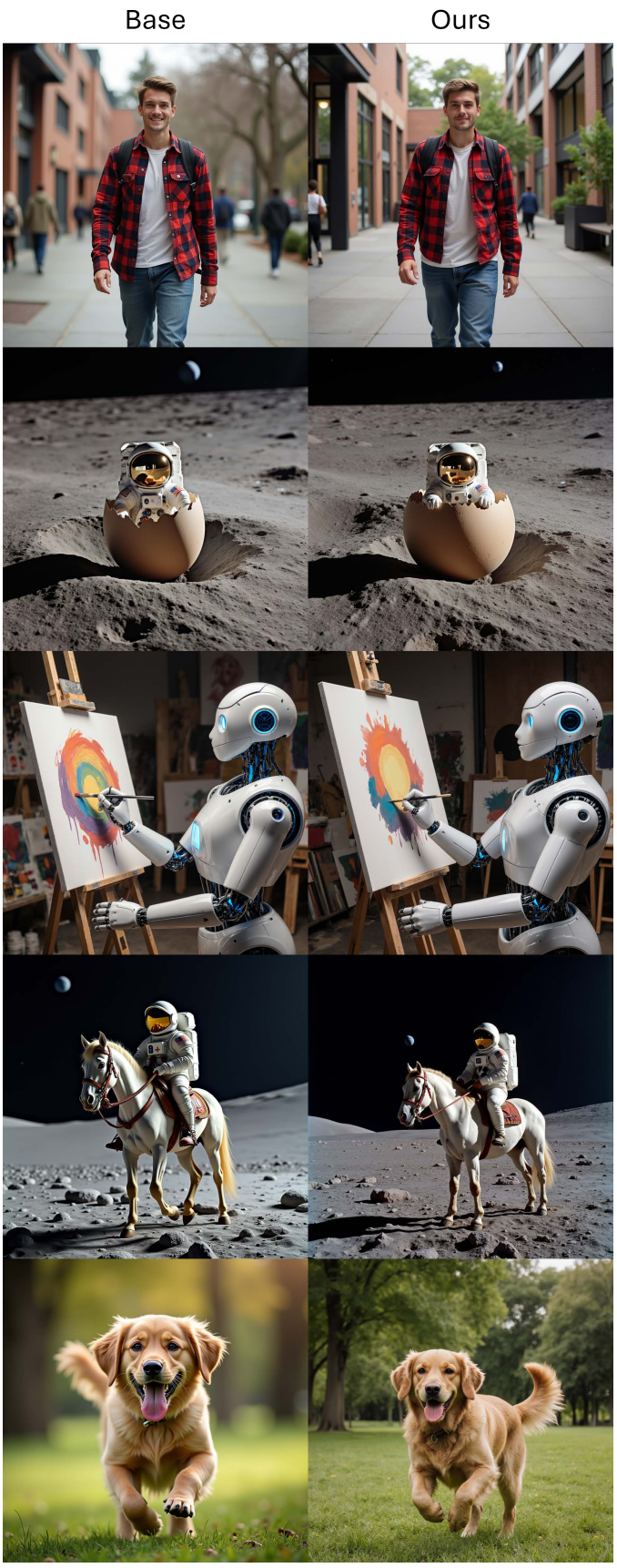


Fig. 2. Qualitative comparison between the baseline FLUX model and our training-free focus-control method. Our approach produces noticeably sharper background regions while preserving subject clarity, demonstrating improved depth-of-field control across diverse scenes.

| | Clip Score |
|---|---|
| Base | 0.3457 |
| Ours | **0.3537** |

TABLE I
QUANTITATIVE COMPARISON OF CLIP SCORES.

scores in Table I. As shown, our approach achieves a higher CLIP score than the baseline, suggesting that images produced with our focus-guidance technique generally exhibit improved semantic consistency and better preservation of scene content. This improvement can be interpreted as evidence that reducing excessive background blur leads to images with sharper details and more coherent global structure, which in turn enhances the model's alignment with the prompt.

### C. Discussion

Even though our method is often effective, the additional guidance can alter the sampling trajectory of the diffusion model more than what we usually expect, which sometimes may lead to failure cases. Some representative examples of these failures are shown in Figure 3. We briefly analyze each example to illustrate the types of issues that may arise: Row 1: although the method is designed to reduce unintended bokeh and reveal sharper background structure, it fails to meaningfully alter the background in this example, the level of blur remains nearly identical to the baseline. Row 2: while attempting to sharpen the background, the method unintentionally changes the stylistic attributes of the generated scene, shifting the image away from the original cartoon-like aesthetic. Such style drift is undesirable, as the method is expected to modify focus properties without altering the artistic domain. Row 3: the method produces even more severe artifacts by substantially altering scene content, demonstrating that excessive guidance can overpower the semantic structure of the underlying diffusion process. Row 4: although the background becomes sharper, the resulting background content becomes inconsistent with the main subjects of the prompt, revealing a mismatch between improved sharpness and semantic alignment. Row 5: in some cases the method produces overall lower-quality images, with reduced detail and degraded textures. Together, these examples highlight key limitations—insufficient effect, style distortion, semantic drift, misalignmen, and quality degradation that motivate future refinements of the guidance strategy.

### D. Conclusion

In this work, we introduced a training-free method for controlling focus and depth-of-field in diffusion models by leveraging intermediate depth estimates and a differentiable focus-based guidance loss. Our approach demonstrates that state-of-the-art models such as FLUX [3] already contain rich internal representations of depth and focus, enabling meaningful refocusing without any retraining or additional datasets. Through qualitative evaluations, we showed that the method effectively reduces unintended background blur and enhances


Fig. 3. Failure cases of our proposed training-free focus-control method across a variety of samples.

scene detail across diverse prompts. Quantitatively, our approach achieves higher CLIP scores compared to the baseline, indicating improved semantic alignment with the input prompt. At the same time, our analysis of failure cases highlights key challenges, such as style drift, semantic changes, and limited global focus manipulation. Overall, this work provides a first step toward practical, training-free focus control in generative models and suggests several promising directions for future research, including more robust guidance formulations, adaptive weighting strategies and improved integration of depth priors to achieve finer control over photographic attributes in synthetic imagery.

## References

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[3] 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: Flow matching for in-context image generation and editing in latent space," *arXiv preprint arXiv:2506.15742*, 2025.

[4] M. Langford, A. Bruce, A. Fox, and R. S. Smith, *Langford's basic photography: the guide for serious photographers*. Routledge, 2024.

[5] I.-S. Fang, Y.-H. Han, and J.-C. Chen, "Camera settings as tokens: Modeling photography on latent diffusion models," in *SIGGRAPH Asia 2024 Conference Papers*, 2024, pp. 1–11.

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

[7] A. Fortes, T. Wei, S. Zhou, and X. Pan, "Bokeh diffusion: Defocus blur control in text-to-image diffusion models," *arXiv preprint arXiv:2503.08434*, 2025.

[8] Y. Wang, X. Chen, X. Xu, Y. Liu, and H. Zhao, "Diffcamera: Arbitrary refocusing on images," *arXiv preprint arXiv:2509.26599*, 2025.

[9] Y. Yang, S. Zheng, J. Chen, B. Wu, X. He, D. Cai, B. Li, and P.-T. Jiang, "Any-to-bokeh: One-step video bokeh via multi-plane image guided diffusion," *arXiv preprint arXiv:2505.21593*, 2025.

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

[11] S. Dutta, S. D. Das, N. A. Shah, and A. K. Tiwari, "Stacked deep multi-scale hierarchical network for fast bokeh effect rendering from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2398–2407.

[12] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

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

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

[15] T.-Y. Cheng, M. Gadelha, T. Groueix, M. Fisher, R. Mech, A. Markham, and N. Trigoni, "Learning continuous 3d words for text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6753–6762.

[16] Z. Xiao, W. Ouyang, Y. Zhou, S. Yang, L. Yang, J. Si, and X. Pan, "Trajectory attention for fine-grained video motion control," *arXiv preprint arXiv:2411.19324*, 2024.

[17] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[18] A. Bansal, H.-M. Chu, A. Schwarzschild, S. Sengupta, M. Goldblum, J. Geiping, and T. Goldstein, "Universal guidance for diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 843–852.

[19] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[20] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[21] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 10 371–10 381.