FORMATION-CONSISTENT FLOW MATCHING FOR SINGLE-IMAGE DEFOCUS DEBLURRING

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-image defocus deblurring seeks to recover an all-in-focus image from spatially varying defocus blur. Recent generative restoration methods based on Flow Matching can produce visually plausible results but may introduce hallucinated details that are not well supported by the observed blur. To address this limitation, we propose *Formation-Consistent Flow Matching* (FCFlow), a training framework that incorporates an effective defocus formation model into Flow Matching endpoint optimization. FCFlow learns an effective spatially varying circle-of-confusion (CoC) representation directly from the blurred input without requiring depth or CoC annotations. At a randomly sampled flow time, the predicted velocity is used to obtain a sharp endpoint proposal through single-step extrapolation. The proposal is decoded into image space and reblurred using the estimated CoC map, producing a formation discrepancy that is converted into a bounded latent endpoint correction. This correction refines endpoint optimization, thereby shaping the learned velocity field toward restorations that better explain the observed blur. The formation-aware endpoint optimization branch is omitted at inference, retaining standard Flow Matching sampling without additional formation-related overhead. Extensive experiments on DPDD, LFDOF, and RealDOF demonstrate state-of-the-art performance with 20-step inference, while strong single-step results further highlight the effectiveness of FCFlow in the low number of function evaluations (NFE) regime.

1 INTRODUCTION

Defocus blur is typically spatially varying, with severely blurred and nearly sharp regions often coexisting within the same image, making single-image defocus deblurring particularly challenging. Conventional learning-based methods formulate the task as a direct inverse mapping and improve restoration through adaptive filtering, blur estimation, kernel parameterization, and transformer-based representations (Quan et al., 2023b;a; 2024; Tang et al., 2024; Liang et al., 2024; Zhai et al., 2023). These deterministic approaches are computationally efficient and generally preserve structures supported by the input, but struggle to recover fine details under severe defocus, where high-frequency information is strongly attenuated.

Generative restoration models provide strong image priors for recovering details that are difficult to infer directly from degraded observations. Diffusion-based approaches have demonstrated impressive restoration capabilities (Saharia et al., 2022a;b; Rombach et al., 2022), and recent defocus deblurring methods have further explored blur-aware, depth-guided, and residual conditioning (Feng et al., 2025; Li et al., 2026a;b). However, strong generative priors may also introduce visually plausible but unsupported details, potentially altering reliable structures under severe degradation. Moreover, iterative denoising incurs substantial computational cost, limiting the practicality of diffusion-based methods in resource-constrained or latency-sensitive settings.

Flow Matching (FM) provides an alternative generative formulation by learning a continuous velocity field that transports samples between prescribed endpoints, avoiding the iterative denoising process used by diffusion models (Lipman et al., 2023; Liu et al., 2023; Martin et al., 2025). Flow Matching-based image restoration has also been explored by initializing transport from representations of degraded observations and guiding them toward corresponding clean targets (Kim & Shin, 2026). However, directly applying vanilla FM to defocus deblurring may produce formation-

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

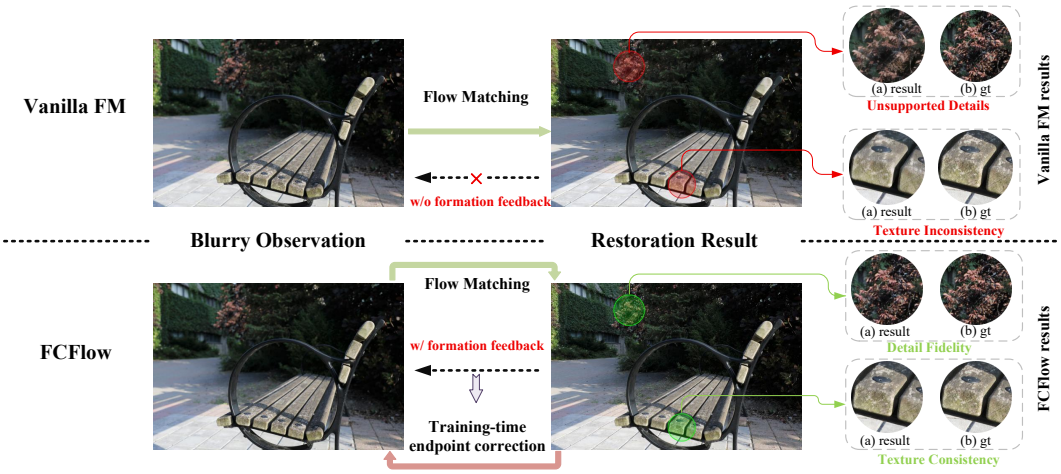


Figure 1: Conceptual comparison between vanilla FM and FCFlow. **Top:** Vanilla FM directly transports the blurred observation toward a sharp endpoint without explicit formation constraints, which may introduce unsupported details or alter reliable textures. **Bottom:** FCFlow introduces a CoC-based defocus formation model to provide training-time endpoint correction, guiding the learned flow toward formation-consistent restoration. The formation pathway is removed at inference.

inconsistent restorations, as illustrated in Fig. 1. Without explicit modeling of the forward defocus formation process, the learned transport may introduce unsupported details in heavily defocused regions or alter reliable textures in nearly sharp regions.

This motivates grounding the learned transport in the underlying image-formation process. Defocus blur is spatially varying and approximately isotropic, with its local extent commonly characterized by the circle of confusion (CoC). Under the thin-lens model, the CoC diameter for a scene point at depth D can be approximated as (Xu et al., 2017):

$$C_d = \frac{|D - D_f|}{D} \frac{F^2}{N(D_f - F)}, \tag{1}$$

where D_f, F, and N denote the focus distance, focal length, and f-number, respectively. Motivated by this spatially varying characterization of defocus, we parameterize the forward defocus process with an effective CoC map C. Given a restored sharp image $\hat{S}$ and the observed defocused image B, we define formation consistency as:

$$\mathcal{B}_C(\hat{S}) \approx B, \tag{2}$$

where $\mathcal{B}_C(\cdot)$ denotes a differentiable defocus renderer parameterized by C. In other words, a formation-consistent restoration should reproduce the observed blur when passed through the estimated forward model. This constraint is particularly important in heavily defocused regions, where weak image evidence makes visually plausible but unsupported details more likely.

We therefore propose *Formation-Consistent Flow Matching* (FCFlow), incorporating formation-aware endpoint optimization into Flow Matching training to shape the learned blur-to-sharp velocity field. At a randomly sampled flow time, the predicted velocity is used to obtain a sharp endpoint proposal through one-step extrapolation. The proposal is decoded into image space and reblurred using the estimated CoC map, producing a formation discrepancy. This discrepancy is converted into a bounded latent endpoint correction for endpoint optimization. The effective CoC representation is learned from paired defocused–sharp images through reblurring supervision, without requiring depth or CoC annotations. Importantly, the formation-aware endpoint optimization branch is used only during training and removed at inference, retaining standard Flow Matching sampling with either single-step or multi-step Euler integration and no additional formation-related overhead.

Our main contributions are:

- We propose *Formation-Consistent Flow Matching* (FCFlow), a framework that integrates the defocus image formation model into Flow Matching training to shape the learned blur-to-sharp velocity field toward formation-consistent restoration.
- We introduce a CoC-based formation feedback mechanism that evaluates random-time endpoint proposals through the forward defocus process and converts the resulting formation discrepancy into a bounded latent endpoint correction, enabling formation-aware endpoint optimization.
- We decouple formation-aware endpoint optimization from inference, allowing FCFlow to leverage formation feedback during training while retaining standard Flow Matching sampling with the same learned velocity field for both one-step and multi-step restoration.

2 Related Work

2.1 Single-Image Defocus Deblurring

Single-image defocus deblurring aims to recover spatially varying defocus blur, whose local extent is commonly characterized by the circle of confusion (CoC). Early methods estimated defocus from local image cues (Liu et al., 2020), while later learning-based approaches employed adaptive filtering, inverse kernels, kernel mixtures, and transformer architectures for spatially varying restoration (Lee et al., 2021; Son et al., 2021; Quan et al., 2023b;a; 2024; Tang et al., 2024; Liang et al., 2024; Zhai et al., 2023). Other methods incorporate depth or defocus cues and mixture-of-experts routing (Feijoo et al., 2026; Zhai et al., 2024; Nazir et al., 2023), while related approaches exploit additional measurements such as dual-pixel or light-field data (Abuolaim & Brown, 2020; Ruan et al., 2021). Most existing methods primarily use degradation cues to improve the inverse restoration mapping. In contrast, FCFlow explicitly models the forward defocus formation process and uses the resulting formation discrepancy during training to shape the learned blur-to-sharp velocity field toward formation-consistent restoration.

2.2 Generative Priors for Image Restoration

Generative models provide strong image priors for recovering details that are difficult to infer directly from degraded observations. Conditional and latent diffusion models have been widely adopted for image restoration (Saharia et al., 2022a;b; Rombach et al., 2022), while recent defocus deblurring methods further incorporate blur-aware conditioning, depth guidance, and residual diffusion (Feng et al., 2025; Li et al., 2026a;b). In contrast, FM learns a continuous velocity field along a prescribed probability path and enables efficient few-step inference (Lipman et al., 2023; Liu et al., 2023). Recent work has extended this formulation to image deblurring by initializing transport from degraded observations and guiding it toward clean targets (Kim & Shin, 2026). FCFlow builds on this blur-to-sharp transport formulation and introduces formation-aware endpoint optimization, using differentiable defocus rendering to derive a formation discrepancy that is converted into a bounded latent endpoint correction for endpoint optimization and shaping the learned velocity field.

2.3 Physics-Guided and Forward-Model Restoration

Physics-guided restoration incorporates explicit physical models into learned image priors (Arridge et al., 2019). In generative inverse problems, forward operators are commonly incorporated through posterior sampling, data-consistency updates, or stochastic forward models (Kawar et al., 2022; Chung et al., 2023; Song et al., 2024; Tewari et al., 2023), and are often retained during inference. For defocus restoration, spatially varying image formation has been modeled using disk kernels, Gaussian mixtures, occlusion-aware rendering, and reblurring-based consistency (Srinivasan et al., 2018; Quan et al., 2024; Sheng et al., 2024; Ren et al., 2025; Zhang et al., 2025). FCFlow uses the formation model differently: it predicts an effective CoC map to parameterize a differentiable defocus renderer, uses the formation discrepancy between the reblurred endpoint proposal and the observed blur to provide formation feedback during training, and removes the entire formation pathway at inference, preserving standard Flow Matching sampling.

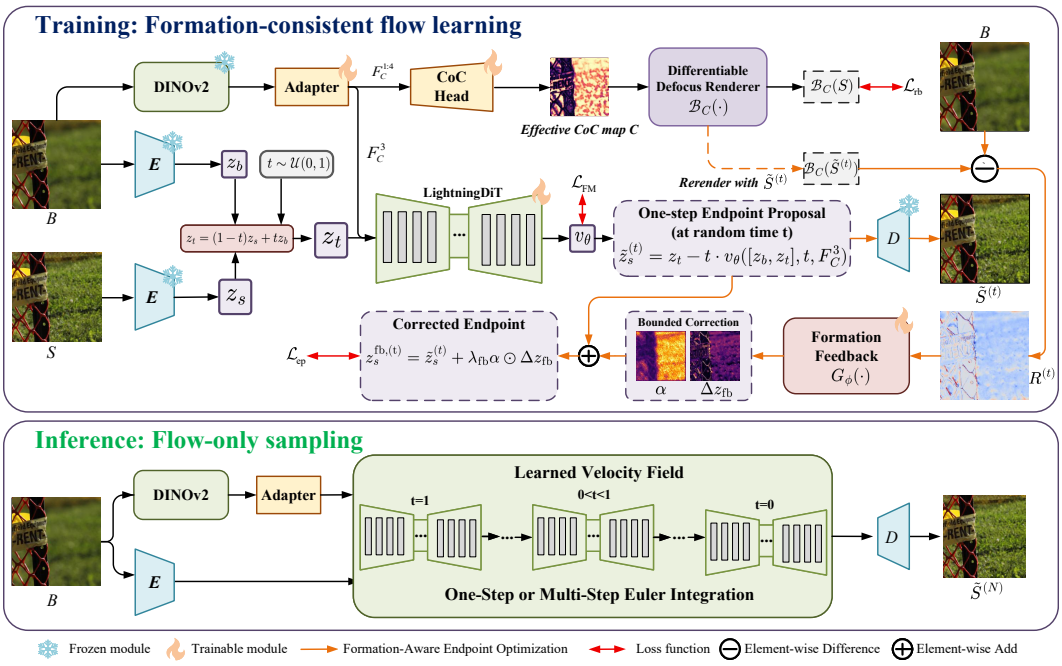


Figure 2: Overview of the proposed FCFlow framework. **Top: Training.** The blurred input is encoded into the source latent and blur-aware features, from which an effective CoC map is predicted to parameterize a differentiable defocus renderer. The renderer reblurs the paired sharp target for CoC supervision and the random-time sharp endpoint proposal to obtain the formation discrepancy, which is converted into a bounded latent endpoint correction for endpoint optimization. **Bottom: Inference.** The formation-aware endpoint optimization branch is removed, retaining standard Flow Matching sampling with the learned velocity field under single-step or multi-step Euler integration.

3 Method

3.1 Framework Overview

Fig. 2 illustrates the overall architecture of FCFlow. The main branch formulates single-image defocus deblurring as latent transport from the blurred observation to its paired sharp target. A frozen visual encoder and a trainable Adapter extract blur-aware features $F_C^{1:4}$ from the blurred input B. The third-level feature F_C^3 conditions the velocity field, while the full set $F_C^{1:4}$ is jointly used to estimate an effective CoC map. During training, the estimated CoC map parameterizes the shared differentiable defocus renderer $\mathcal{B}_C$. Reblurring the paired sharp target provides formation supervision for learning the effective CoC representation. Meanwhile, at a randomly sampled flow time, the predicted velocity is used to obtain a one-step sharp endpoint proposal through extrapolation, which is decoded and reblurred to produce the formation discrepancy $R^{(t)}$. The formation feedback module G_ϕ converts this discrepancy into a bounded latent endpoint correction for endpoint optimization, which shapes the learned velocity field toward formation-consistent restoration. The formation-aware endpoint optimization branch is used exclusively during training and removed at inference. FCFlow therefore retains standard Flow Matching sampling and supports both one-step and multi-step Euler integration without additional formation-related inference overhead.

3.2 Defocus Image Formation

We implement the differentiable defocus renderer $\mathcal{B}_C$ introduced in Eq. 2 using spatially varying disk-shaped point-spread functions (PSFs), whose radii are determined by an effective CoC map. Constructing a distinct disk kernel for every continuous CoC value at every spatial location would be computationally expensive. We therefore approximate the continuous blur range using a bank of

normalized disk kernels $\{K_{\rho_k}\}$ with non-uniformly sampled radii:

$$\rho_k \in \{0, 1, 2, 3, 4, 6, 8, 12, 16, 20\}. \tag{3}$$

The radii are sampled more densely in the small-blur regime, where local image responses are more sensitive to changes in the PSF radius, and more sparsely at larger radii. Related disk-based defocus-kernel parameterizations have been explored in prior restoration methods (Son et al., 2021; Quan et al., 2023a; 2024). For a spatial location p with $C(p) \in [\rho_k, \rho_{k+1}]$, we linearly interpolate the responses of the two neighboring kernels:

$$\mathcal{B}_C(X)(p) = \big(1 - w_k(C(p))\big)(K_{\rho_k} * X)(p) + w_k(C(p))(K_{\rho_{k+1}} * X)(p), \tag{4}$$

where

$$w_k(r) = \frac{r - \rho_k}{\rho_{k+1} - \rho_k}, \qquad r \in [\rho_k, \rho_{k+1}]. \tag{5}$$

Here, $*$ denotes spatial convolution. Although the kernel bank is discrete, linear interpolation makes the rendered response continuous with respect to C, allowing gradients to propagate through the CoC prediction during training. This provides an efficient differentiable approximation of spatially varying defocus without constructing a distinct PSF for every continuous CoC radius.

3.3 Blur-to-Sharp Latent Flow Matching

We formulate defocus deblurring as conditional transport in latent space. A fixed pretrained encoder E maps the blurred input B and its paired sharp target S to latent endpoints $z_b = E(B)$ and $z_s = E(S)$. We define $t = 1$ as the blurred endpoint and $t = 0$ as the sharp endpoint, and construct the linear path $z_t = (1 - t)z_s + tz_b$ with target velocity $u^* = \partial z_t / \partial t = z_b - z_s$, where $t \sim \mathcal{U}(0, 1)$. A frozen DINOv2 encoder together with a trainable Adapter extracts multi-level features $F_C^{1:4}$ from B, of which F_C^3 conditions the velocity estimator as $\widehat{u}_t = v_\theta([z_b, z_t], t, F_C^3)$, where $[\cdot, \cdot]$ denotes channel-wise concatenation.

The Flow Matching objective is:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{B,S,t}\left[\left\| v_\theta([z_b, z_t], t, F_C^3) - (z_b - z_s)\right\|_2^2\right]. \tag{6}$$

At inference, restoration starts from z_b at $t = 1$ and integrates the learned velocity field toward $t = 0$. For $1 = t_0 > t_1 > \cdots > t_N = 0$, Euler integration is:

$$z^{(i+1)} = z^{(i)} - (t_i - t_{i+1}) v_\theta([z_b, z^{(i)}], t_i, F_C^3), \qquad z^{(0)} = z_b. \tag{7}$$

The final restoration is decoded as $\widehat{S}^{(N)} = D(z^{(N)})$, where D is a frozen pretrained decoder. The same velocity field supports both one-step ($N = 1$) and multi-step ($N > 1$) inference.

3.4 Formation-Aware Endpoint Optimization

Effective CoC estimation. Given the blurred observation B, a frozen DINOv2 encoder and a trainable Adapter extract multi-level features $F_C^{1:4}$, from which the CoC head H_ω predicts an effective CoC map $C = C_{\max}\sigma(H_\omega(F_C^{1:4}))$, where $C_{\max}$ denotes the maximum CoC radius and $\sigma(\cdot)$ is the sigmoid function, bounding C by $C_{\max}$. Without depth or CoC annotations, the effective CoC representation is learned by reblurring the paired sharp target:

$$\widetilde{B}_S = \mathcal{B}_C(S), \qquad \mathcal{L}_{\mathrm{rb}} = \|\widetilde{B}_S - B\|_1. \tag{8}$$

Thus, the effective CoC representation is indirectly supervised by its ability to reproduce the observed blur through the differentiable defocus renderer.

Random-time endpoint proposal. During training, a time t is randomly sampled along the linear transport path. The predicted velocity at z_t is used to extrapolate the latent state over the remaining interval from t to 0 to obtain a one-step sharp-endpoint proposal:

$$\widetilde{z}_s^{(t)} = z_t - t v_\theta([z_b, z_t], t, F_C^3). \tag{9}$$

Here, $\widetilde{z}_s^{(t)}$ denotes the endpoint proposal before formation feedback, with its image-space counterpart given by $\widetilde{S}^{(t)} = D(\widetilde{z}_s^{(t)})$. Random-time sampling enables endpoint optimization along the

transport path, encouraging accurate endpoint prediction from different states and thereby improving restoration under low numbers of function evaluations (NFE). At $t = 1$, Eq. 9 reduces to the direct one-step blur-to-sharp prediction used at inference.

Formation feedback. The endpoint proposal is reblurred using the predicted CoC map,

$$\widetilde{B}^{(t)} = \mathcal{B}_C(\widetilde{S}^{(t)}), \qquad R^{(t)} = B - \widetilde{B}^{(t)}, \tag{10}$$

where $R^{(t)}$ denotes the signed formation discrepancy between the observed blur and the reblurred endpoint proposal. Rather than directly minimizing this discrepancy, we use it as formation feedback to refine endpoint optimization.

The discrepancy $R^{(t)}$ is encoded into a latent-resolution feature h_L by a formation feedback module G_ϕ, i.e., $h_L = G_\phi(R^{(t)})$. Two prediction heads $H_\Delta(\cdot)$ and $H_\alpha(\cdot)$ then estimate a bounded latent endpoint correction and a spatial gate:

$$\Delta z_{\mathrm{fb}} = \Delta_{\max} \tanh\left(H_\Delta(h_L)\right), \qquad \alpha = \sigma\left(H_\alpha(h_L)\right), \tag{11}$$

where $\Delta_{\max}$ bounds the correction magnitude and $\alpha \in (0, 1)$ controls its spatial strength. The corrected endpoint is then obtained as:

$$z_s^{\mathrm{fb},(t)} = \widetilde{z}_s^{(t)} + \lambda_{\mathrm{fb}} \alpha \odot \Delta z_{\mathrm{fb}}, \tag{12}$$

where $\odot$ denotes element-wise multiplication and λ_{fb} controls the overall feedback strength. The corrected endpoint is then supervised by the paired sharp latent:

$$\mathcal{L}_{\mathrm{ep}} = \left\| z_s^{\mathrm{fb},(t)} - z_s \right\|_1. \tag{13}$$

Because the refined endpoint depends on the random-time Flow proposal, this supervision propagates formation feedback to the velocity estimator, shaping the learned transport field toward more accurate sharp-endpoint prediction.

3.5 Joint Training and Flow-Only Inference

The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{FM}} + \lambda_{\mathrm{ep}} \mathcal{L}_{\mathrm{ep}} + \lambda_{\mathrm{rb}} \mathcal{L}_{\mathrm{rb}}, \tag{14}$$

where λ_{ep} and λ_{rb} balance the endpoint and reblurring objectives. The encoder E, decoder D, and DINOv2 encoder are frozen, while the Adapter, velocity field v_θ, CoC head H_ω, and formation feedback module G_ϕ are jointly optimized. Here, $\mathcal{L}_{\mathrm{FM}}$ (Eq. 6) supervises latent transport, $\mathcal{L}_{\mathrm{rb}}$ (Eq. 8) supervises the learning of the effective CoC representation through formation consistency, and $\mathcal{L}_{\mathrm{ep}}$ (Eq. 13) propagates formation feedback to the velocity field. At inference, the CoC head, renderer, and formation feedback module are removed. The learned velocity field performs Euler integration from $t = 1$ to $t = 0$ using z_b and F_C^3, supporting both single-step and multi-step restoration without formation-related inference overhead.

4 Experiments

4.1 Experimental Settings

Datasets and metrics. We train FCFlow on the DPDD training split (Abuolaim & Brown, 2020), using 512×512 crops without depth or CoC annotations. Evaluation is conducted on 76 DPDD test pairs, 725 LFDOF pairs (Ruan et al., 2021), and 50 RealDOF pairs (Lee et al., 2021). Following prior work (Tang et al., 2024), RealDOF images are bilinearly resized to 1680×1120, while LFDOF images are evaluated at their original resolution. We report RGB PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), and DISTS (Ding et al., 2020), together with FID (Heusel et al., 2017) and MUSIQ (Ke et al., 2021) as complementary distributional and no-reference metrics, respectively.

Implementation details. We use the frozen Qwen-Image VAE (Wu et al., 2025) to construct a fixed 16-channel latent space, with frozen DINOv2 ViT-L/14 (Oquab et al., 2024) features. The trainable components include the Adapter, LightningDiT (Yao et al., 2025) velocity field, CoC head, and formation feedback module. The velocity field contains 28 transformer blocks with 1024 hidden channels and 16 attention heads. All modules are jointly optimized with AdamW with a learning rate of 5×10^{-5} using bfloat16 precision. Training runs for 1.4M iterations with a batch size of 1 on a single NVIDIA RTX 4090 GPU. We set $C_{\max} = 20$, $\lambda_{\mathrm{fb}} = 1$, $\Delta_{\max} = 1$, and $\lambda_{\mathrm{ep}} = \lambda_{\mathrm{rb}} = 1$.

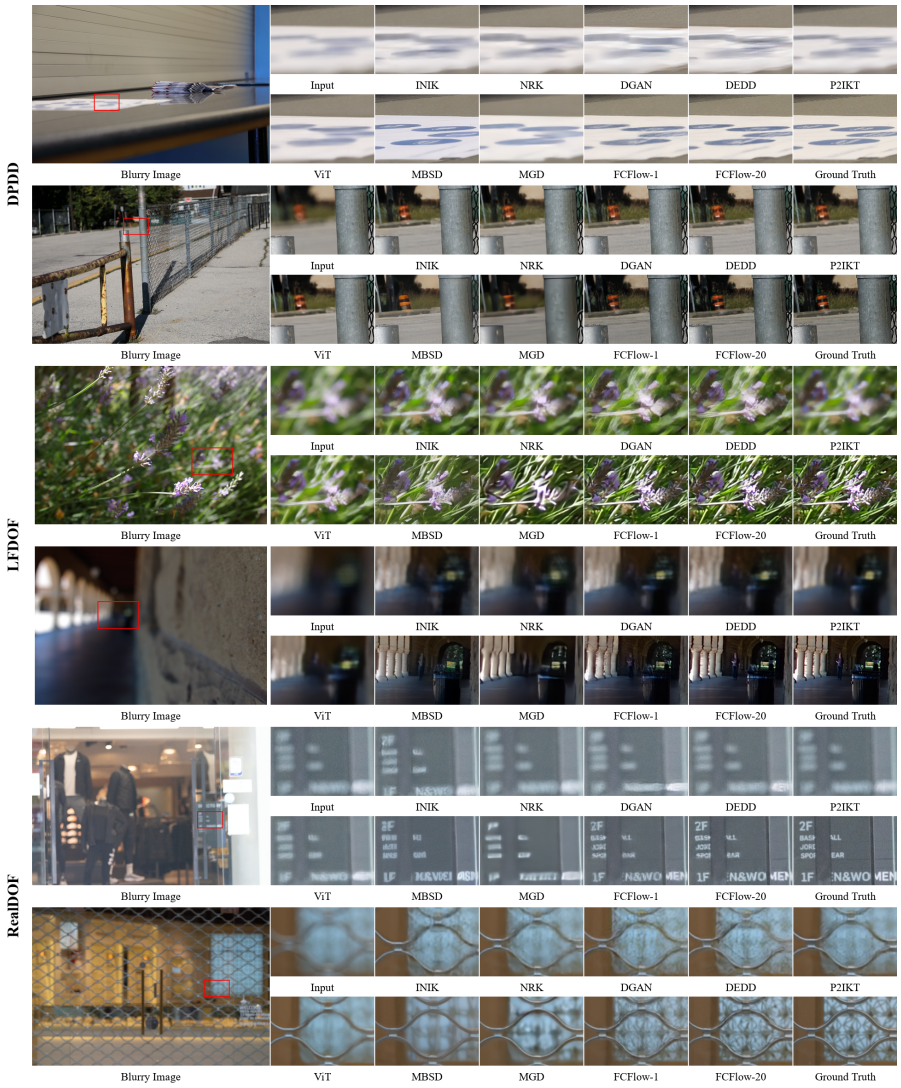


Figure 3: Qualitative comparisons on representative examples from DPDD, LFDOF, and RealDOF. Each dataset includes two examples, with the full blurred input shown on the left and enlarged crops on the right. Red boxes indicate the corresponding crop regions.

4.2 Deblurring Performance Comparisons

We compare FCFlow with representative defocus deblurring methods, including INIKNet (Quan et al., 2023b), NRKNet (Quan et al., 2023a), DefocusGAN (Zhai et al., 2023), DEDDNet (Zhai et al., 2024), P2IKT (Tang et al., 2024), and ViTDeblur (Liang et al., 2024), as well as the degradation-aware generative methods MBSD (Li et al., 2026a) and MGD (Li et al., 2026b).

Quantitative comparison. Table 1 reports results on DPDD, LFDOF, and RealDOF, respectively. With only one Euler step, FCFlow-1 achieves 28.74, 29.88, and 31.29 dB PSNR on the three datasets, surpassing the previous best PSNR results by 2.18, 3.33, and 2.24 dB, respectively. It also achieves competitive performance across the evaluated metrics on RealDOF and LFDOF. Using 20 Euler steps further improves performance, with FCFlow-20 achieving the best results across all six metrics on all three datasets. These results demonstrate strong restoration under low-NFE inference, with further gains from multi-step integration.

Qualitative comparisons. Fig. 3 compares results on DPDD, LFDOF, and RealDOF, respectively. On DPDD, several prior methods fail to fully recover heavily blurred structures such as the elliptical pattern, while MBSD and MGD exhibit noticeable color shifts; FCFlow more closely matches the ground-truth appearance while recovering fine local details, including the roadside post. On LFDOF,

Table 1: Quantitative comparison with various defocus deblurring methods on DPDD, RealDOF, and LFDOF. FCFlow-1 and FCFlow-20 denote inference with 1 and 20 Euler steps, respectively. Best and second-best results are shown in **bold** and underlined, respectively.

Dataset	Metric	INIK	NRK	DGAN	DEDD	P2IKT	ViTDeblur	MBSD	MGD	FCFlow-1	FCFlow-20
DPDD	PSNR↑	26.1473	26.1519	25.5531	25.6061	26.3301	26.0364	26.5608	26.3179	28.7432	**29.3448**
	SSIM↑	0.8116	0.8112	0.7919	0.7951	0.8147	0.8162	0.7990	0.8126	0.8307	**0.8448**
	LPIPS↓	0.1815	0.2228	0.1649	0.1614	0.1920	0.1993	0.1066	0.2372	0.1372	**0.0920**
	DISTS↓	0.1180	0.1339	0.0934	0.0949	0.1260	0.1214	0.0607	0.1477	0.0813	**0.0533**
	FID↓	46.2373	55.1670	50.9651	51.1601	47.8951	49.9173	25.6798	54.5894	37.7003	**23.3330**
	MUSIQ↑	57.1018	57.6717	58.0607	58.1713	57.2999	58.8318	59.5899	56.6249	58.4522	**60.6433**
RealDOF	PSNR↑	21.6153	25.4886	24.4969	24.7997	25.9567	25.5911	26.5468	26.0616	29.8796	**30.2480**
	SSIM↑	0.6645	0.8010	0.7694	0.7837	0.8110	0.8139	0.7839	0.8238	0.8711	**0.8810**
	LPIPS↓	0.2421	0.2420	0.2111	0.1966	0.2180	0.2117	0.1769	0.2182	0.1233	**0.0887**
	DISTS↓	0.1181	0.1355	0.1171	0.1108	0.1253	0.1139	0.0889	0.1247	0.0772	**0.0491**
	FID↓	43.0878	43.3368	43.5964	40.9758	37.4817	39.3147	29.5817	35.7249	21.7713	**15.3918**
	MUSIQ↑	42.2849	40.1355	42.7636	44.4357	41.3828	44.9109	44.3136	42.9811	42.4617	**46.8789**
LFDOF	PSNR↑	26.5138	26.3984	25.5669	25.0939	26.8972	26.8799	27.1026	29.0472	31.2870	**31.7059**
	SSIM↑	0.8046	0.8054	0.7756	0.7760	0.8148	0.8323	0.7624	0.8698	0.8983	**0.9049**
	LPIPS↓	0.2184	0.2327	0.2068	0.1954	0.2262	0.2050	0.1097	0.1536	0.0610	**0.0557**
	DISTS↓	0.1181	0.1184	0.1222	0.1221	0.1250	0.1093	0.0600	0.1003	0.0315	**0.0261**
	FID↓	48.3506	53.2647	51.0763	49.1009	47.9308	42.2473	23.7846	34.0093	11.4860	**9.3508**
	MUSIQ↑	55.7972	56.9806	58.2354	59.4184	55.6946	60.1131	62.7218	59.6981	62.3828	**63.0320**

Table 2: Ablation study of endpoint optimization, feedback design, and test-time correction on the DPDD test set. All variants use the same Flow Matching backbone.

	Training design		Test-time	1-NFE		20-NFE	
Variant	$\mathcal{L}_{rb}$	Random-t endpoint optimization	Correction	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
A0 Vanilla FM	–	–	–	25.6903	0.1804	27.6936	0.1608
A1 Formation-Supervised	✓	–	–	25.9048	0.1732	27.7028	0.1596
A2 Direct Sharp Regression	✓	*Sharp* regression	–	26.8917	0.1621	28.4573	0.1448
A3 Direct Formation Regression	✓	*Formation* regression	–	26.9523	0.1573	28.0725	0.1408
A4 Sharp-Feedback (w/ Inference Correction)	✓	*Sharp* feedback	✓	28.0726	0.1493	28.7516	0.1221
A5 Sharp-Feedback	✓	*Sharp* feedback	–	28.0924	0.1510	28.7445	0.1235
A6 FCFlow (w/ Inference Correction)	✓	*Formation* feedback	✓	**28.7459**	**0.1370**	**29.3459**	**0.0910**
A7 FCFlow	✓	*Formation* feedback	–	28.7432	0.1372	29.3448	0.0920

MBSD and MGD produce less consistent details, including false petals, a darkened head, and missing patterns on the posts and bucket, whereas FCFlow better preserves color, structure, and texture. On RealDOF, FCFlow recovers clearer text and more regular window patterns with fewer visible distortions. Overall, FCFlow improves detail recovery while better preserving formation-consistent appearance and structure. The small visual gap between FCFlow-1 and FCFlow-20 further demonstrates the effectiveness of FCFlow under low-NFE inference.

4.3 Ablation Study

Effect of formation supervision. As shown in Table 2, introducing the reblurring loss $\mathcal{L}_{rb}$ (A1 vs. A0) improves both PSNR and LPIPS under 1-NFE and 20-NFE inference. Because the Adapter is shared between CoC estimation and velocity conditioning, gradients from $\mathcal{L}_{\mathrm{rb}}$ also influence the adapted feature F_C^3 used by the velocity field. The gain therefore reflects the benefit of injecting defocus-formation information into the shared conditional representation.

Effect of random-t endpoint optimization. We further supervise the sharp-endpoint proposal $\widetilde{S}^{(t)}$ constructed at a randomly sampled flow time. We consider two endpoint discrepancies: the *sharp* discrepancy $S-\widetilde{S}^{(t)}$ and the *formation* discrepancy $B-\mathcal{B}_C(\widetilde{S}^{(t)})$, which measure target-space accuracy and formation consistency, respectively. In A2 and A3, these discrepancies are directly minimized with an ℓ_1 loss. Both variants outperform formation supervision alone (A1), demonstrating the benefit of directly optimizing endpoint predictions along the transport path. This indicates that random-t endpoint optimization improves endpoint prediction from different flow states and reduces reliance on iterative integration.



Figure 4: Qualitative comparison of ablation variants under 20-NFE inference.

Table 3: Effect of inference NFE on Vanilla FM and FCFlow on the DPDD test set.

Model	Metric	1-NFE	2-NFE	4-NFE	8-NFE	16-NFE	20-NFE
Vanilla FM	PSNR↑	25.6903	26.0835	26.4118	26.6124	27.6861	**27.6936**
	LPIPS↓	0.1804	0.1782	0.1755	0.1702	0.1643	**0.1608**
FCFlow	PSNR↑	28.7432	29.0719	29.2678	29.3473	**29.3513**	29.3448
	LPIPS↓	0.1372	0.1187	0.1041	0.0959	0.0924	**0.0920**

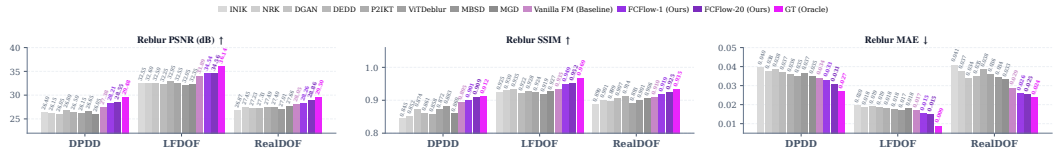


Figure 5: Forward-formation consistency comparison with various methods on DPDD, LFDOF, and RealDOF. Results are evaluated using PSNR, SSIM, and MAE between the reblurred images and the observed blur. GT denotes the paired sharp target reblurred using the estimated formation model.

Effect of endpoint feedback. We then replace direct discrepancy regression with feedback-based endpoint refinement. Rather than directly minimizing the sharp or formation discrepancy, the feedback-based variants first convert the discrepancy into a bounded latent endpoint correction, which refines the predicted endpoint and propagates supervision to the velocity field through the endpoint loss. This design consistently improves over the corresponding direct regression (A5 vs. A2 and A7 vs. A3). Moreover, formation feedback substantially outperforms sharp feedback (A7 vs. A5), indicating that evaluating the predicted endpoint through the forward defocus process provides a more informative refinement signal for shaping the learned velocity field.

Qualitative comparison of ablation variants. Fig. 4 visualizes representative ablation variants. A0 produces an overly smooth restoration with noticeable deviations from the target. Introducing formation supervision in A1 recovers more target-consistent local textures, although with slightly reduced edge sharpness. Sharp-feedback optimization in A5 further improves structural clarity, but discrepancies in fine textures remain. With formation feedback, A7 better preserves both clear structures and detailed leaf patterns that closely match the target.

Test-time correction and inference NFE. Table 2 shows that retaining endpoint correction at inference brings only marginal changes (A4 vs. A5 and A6 vs. A7). We therefore adopt the correction-free A7, allowing the formation branch to be removed during inference. Table 3 further shows that FCFlow already performs strongly at 1-NFE and rapidly approaches saturation as NFE increases, while additional Euler steps provide further improvements.

4.4 Forward-Formation Analysis

Forward-model fidelity. We evaluate the forward model in the observation space by comparing $\mathcal{B}_C(S)$ with B. As shown in Fig. 5, the GT-reblur reference achieves 29.48/0.912/0.027 on DPDD, 36.14/0.969/0.009 on LFDOF, and 29.50/0.935/0.024 on RealDOF in PSNR/SSIM/MAE, respectively. These results indicate that the estimated forward model can reproduce the observed blur with relatively small formation discrepancies, supporting its use for formation feedback.

Formation consistency of restorations. We further evaluate the formation consistency of restored images by comparing $\mathcal{B}_C(\hat{S})$ with the observed blur B using PSNR, SSIM, and MAE. As shown in Fig. 5, FCFlow consistently achieves higher reblur PSNR and SSIM and lower MAE than Vanilla FM and competing methods across all three datasets. This consistent trend indicates that FCFlow produces restorations that better agree with the observed blur under the estimated forward model.

5 Conclusion

We introduced FCFlow for single-image defocus deblurring, incorporating formation-aware endpoint optimization into FM training to promote formation-consistent restoration. The formation-aware branch is removed at inference, retaining standard FM sampling without additional formation-related overhead. Experiments demonstrate state-of-the-art performance across multiple datasets, with strong single-step results and further gains from multi-step integration.

AI USE STATEMENT

Generative AI tools were used solely for language editing.

REFERENCES

Abdullah Abuolaim and Michael S. Brown. Defocus deblurring using dual-pixel data. In *European Conference on Computer Vision*, pp. 111–126, 2020.

Simon Arridge, Peter Maass, Ozan Öktem, and Carola-Bibiane Schönlieb. Solving inverse problems using data-driven models. *Acta Numerica*, 28:1–174, 2019.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2020.

Daniel Feijoo, Paula Garrido-Mellado, Jaesung Rim, Alvaro Garcia, and Marcos V. Conde. Towards unified image deblurring using a mixture-of-experts decoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4946–4956, 2026.

Haoxuan Feng, Haohui Zhou, Tian Ye, Sixiang Chen, and Lei Zhu. Residual diffusion deblurring model for single image defocus deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2960–2968, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23593–23606, 2022.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157, 2021.

Insoo Kim and Jinwoo Shin. Restoration-aligned generative flow models for blind motion deblurring. *arXiv preprint arXiv:2605.08854*, 2026.

Junyong Lee, Hyeongseok Son, Jaesung Rim, Sunghyun Cho, and Seungyong Lee. Iterative filter adaptive network for single image defocus deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2034–2042, 2021.

Xiaopan Li, Yi Jiang, Shiqian Wu, Shoulie Xie, and Sos Agaian. Multi-level blur-aware stable diffusion for region-adaptive defocus deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6504–6512, 2026a.

Xiaopan Li, Shiqian Wu, Qile Zhu, Shoulie Xie, and Sos Agaian. Single image defocus deblurring via multimodal-guided diffusion and depth-aware fusion. *Pattern Recognition*, 171:112133, 2026b.

Pengwei Liang, Junjun Jiang, Xianming Liu, and Jiayi Ma. Decoupling image deblurring into twofold: A hierarchical model for defocus deblurring. *IEEE Transactions on Computational Imaging*, 10:1207–1220, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Shaojun Liu, Qingmin Liao, Jing-Hao Xue, and Fei Zhou. Defocus map estimation from a single image using improved likelihood feature and edge-based basis. *Pattern Recognition*, 107:107485, 2020.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Ségolène Tiffany Martin, Anne Gagneux, Paul Hagemann, and Gabriele Steidl. Pnp-flow: Plug-and-play image restoration with flow matching. In *International Conference on Learning Representations*, 2025.

Saqib Nazir, Lorenzo Vaquero, Manuel Mucientes, Víctor M. Brea, and Daniela Coltuc. Depth estimation and image restoration by deep learning from defocused images. *IEEE Transactions on Computational Imaging*, 9:607–619, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Yuhui Quan, Zicong Wu, and Hui Ji. Neumann network with recursive kernels for single image defocus deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5754–5763, 2023a.

Yuhui Quan, Xin Yao, and Hui Ji. Single image defocus deblurring via implicit neural inverse kernels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12600–12610, 2023b.

Yuhui Quan, Zicong Wu, Ruotao Xu, and Hui Ji. Deep single image defocus deblurring via gaussian kernel mixture learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Dongwei Ren, Xinya Shu, Yu Li, Xiaohe Wu, Jin Li, and Wangmeng Zuo. Reblurring-guided single image defocus deblurring: A learning framework with misaligned training pairs. *International Journal of Computer Vision*, 133(10):6953–6970, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

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

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH Conference Proceedings*, pp. 1–10, 2022a.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–14, 2022b.

Yichen Sheng, Zixun Yu, Lu Ling, Zhiwen Cao, Xuaner Zhang, Xin Lu, Ke Xian, Haiting Lin, and Bedrich Benes. Dr.Bokeh: Differentiable occlusion-aware bokeh rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4515–4525, 2024.

Hyeongseok Son, Junyong Lee, Sunghyun Cho, and Seungyong Lee. Single image defocus deblurring using kernel-sharing parallel atrous convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2642–2650, 2021.

Bowen Song, Soo Min Kwon, Zecheng Zhang, Xinyu Hu, Qing Qu, and Liyue Shen. Solving inverse problems with latent diffusion models via hard data consistency. In *International Conference on Learning Representations*, 2024.

Pratul P. Srinivasan, Rahul Garg, Neal Wadhwa, Ren Ng, and Jonathan T. Barron. Aperture supervision for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6393–6401, 2018.

Peng Tang, Zhiqiang Xu, Chunlai Zhou, Pengfei Wei, Peng Han, Xin Cao, and Tobias Lasser. Prior and prediction inverse kernel transformer for single image defocus deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5145–5153, 2024.

Ayush Tewari, Tianwei Yin, George Cazenavette, Semon Rezchikov, Josh Tenenbaum, Frédo Durand, Bill Freeman, and Vincent Sitzmann. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. In *Advances in Neural Information Processing Systems*, volume 36, pp. 12349–12362, 2023.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

Guodong Xu, Yuhui Quan, and Hui Ji. Estimating defocus blur via rank of local patches. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5371–5379, 2017.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in neural information processing systems*, 37:21875–21911, 2024.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15703–15712, 2025.

Jucai Zhai, Pengcheng Zeng, Chihao Ma, Jie Chen, and Yong Zhao. Learnable blur kernel for single-image defocus deblurring in the wild. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 3384–3392, 2023.

Jucai Zhai, Yang Liu, Pengcheng Zeng, Chihao Ma, Xinan Wang, and Yong Zhao. Efficient fusion of depth information for defocus deblurring. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2640–2644, 2024.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Zhengbo Zhang, Lin Geng Foo, Hossein Rahmani, Jun Liu, and De Wen Soh. Performing defocus deblurring by modeling its formation process. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5791–5801, 2025.

A Additional Experimental Details

A.1 Dataset Details

DPDD. DPDD (Abuolaim & Brown, 2020) contains real defocused–sharp pairs captured with a Canon EOS 5D Mark IV. For each static scene, the defocused image is captured with a wide aperture ($f/4$), while the all-in-focus reference is captured with a narrow aperture ($f/22$). The two images are therefore acquired sequentially under different aperture settings. Although DPDD also provides dual-pixel sub-aperture views, FCFlow uses only the conventional RGB defocused image as input.

LFDOF. LFDOF (Ruan et al., 2021) is constructed from light-field images captured with a Lytro Illum camera. Defocused and all-in-focus images are generated from the same light-field observation through synthetic-aperture and refocusing operations. This acquisition strategy provides well-aligned image pairs and spatially varying defocus patterns that differ from those produced by conventional DSLR imaging.

RealDOF. RealDOF (Lee et al., 2021) uses a beam-splitter dual-camera system to simultaneously capture defocused and all-in-focus images. One camera operates with a relatively wide aperture to produce optical defocus, while the other uses a narrow aperture to obtain the sharp reference. The synchronized acquisition reduces temporal misalignment between the paired images and provides real optical defocus from a camera system different from DPDD.

Together, the three datasets cover diverse acquisition settings and defocus characteristics. FCFlow is trained only on the DPDD dataset, while the RealDOF and LFDOF test sets are used as held-out datasets for cross-dataset evaluation.

A.2 Evaluation Metrics

We evaluate all methods using the same metric protocols. PSNR and SSIM (Wang et al., 2004) are computed in RGB space and measure pixel-level and structure-level fidelity to the paired sharp reference, respectively, with higher values indicating better reconstruction quality. LPIPS (Zhang et al., 2018) and DISTS (Ding et al., 2020) measure perceptual discrepancy using learned deep features, where lower values indicate better perceptual similarity.

We additionally report MUSIQ (Ke et al., 2021) as a no-reference image-quality metric and FID (Heusel et al., 2017) as a complementary distribution-level metric. For FID, Inception features are extracted from the restored and reference image sets, and a single Fréchet distance is computed for each evaluation dataset rather than averaging per-image scores. Since MUSIQ and FID do not directly measure paired reconstruction accuracy, they are reported as complementary metrics, while paired full-reference metrics remain the primary basis for quantitative comparison.

A.3 Architecture and Module Details

Overall architecture. FCFlow operates in the 16-channel latent space of a frozen pretrained Qwen-Image VAE (Wu et al., 2025). A frozen DINOv2 ViT-L/14 encoder (Oquab et al., 2024) extracts visual features from the blurred input, followed by a lightweight trainable Adapter. The adapted multi-level features $F_C^{1:4}$ are used to estimate the effective CoC representation, while F_C^3 conditions the Flow velocity network. The velocity field is parameterized by LightningDiT (Yao et al., 2025) with a hidden dimension of 1024, 28 transformer blocks, and 16 attention heads. The latent state is tokenized using a 2×2 PatchEmbed.

The Qwen-Image VAE and DINOv2 encoder remain frozen throughout training. The Adapter, LightningDiT, CoC Head, and endpoint-feedback module are jointly optimized using AdamW with a learning rate of 5×10^{-5}. Training is performed with bfloat16 precision on a single NVIDIA RTX 4090 GPU for 1.4 million iterations with a batch size of 1.

At each training iteration, a random time t is sampled along the linear blur-to-sharp path. The predicted velocity is supervised by the Flow Matching loss $\mathcal{L}_{\mathrm{FM}}$ and is also used to construct the endpoint proposal $\tilde{z}_s^{(t)}$. The feedback-refined endpoint $z_s^{\mathrm{fb},(t)}$ is supervised by $\mathcal{L}_{\mathrm{ep}}$, while effective CoC estimation is supervised by the reblurring loss $\mathcal{L}_{\mathrm{rb}}$, as defined in the main text.

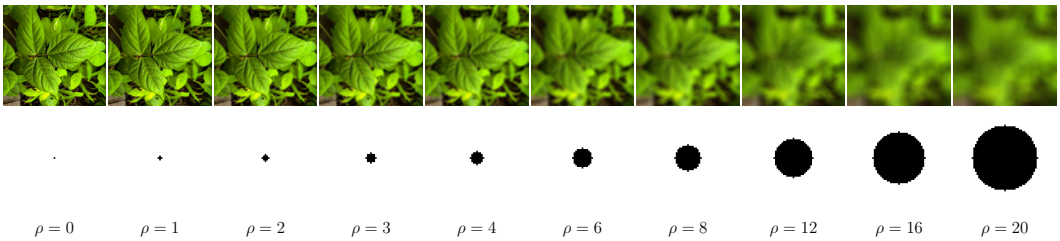


Figure 6: Defocus sequence generated by the discrete kernel bank $\{0, 1, 2, 3, 4, 6, 8, 12, 16, 20\}$. The bank uses dense sampling at small radii and coarser sampling at large radii, providing a better approximation of the nonlinear relationship between radius variation and defocus appearance while maintaining a compact kernel representation.

Adapter. For a 512×512 crop, we pad three pixels on each side to obtain a 518×518 input for the frozen DINOv2 ViT-L/14 encoder. With a 14×14 patch size, this yields a 37×37 token grid, corresponding to 1369 patch tokens with a feature dimension of 1024. Given the blurred input B, we extract intermediate patch-token embeddings from DINOv2 blocks $(l_1, l_2, l_3, l_4) = (4, 11, 17, 23)$:

$$F_D^i = \mathrm{DINO}^{(l_i)}(B), \qquad i \in \{1, 2, 3, 4\}, \tag{15}$$

where $F_D^i \in \mathbb{R}^{1369 \times 1024}$ denotes the feature extracted from the i-th selected block. Each feature is then adapted using a shared residual MLP:

$$F_C^i = F_D^i + (W_2\big(\mathrm{GELU}\big(W_1\,\mathrm{LN}(F_D^i) + b_1\big)\big) + b_2), \qquad i \in \{1, 2, 3, 4\}. \tag{16}$$

Dense CoC Head. The Dense CoC Head follows a four-level DPT-style architecture (Yang et al., 2024) for spatially varying CoC estimation. Each adapted token feature $F_C^i \in \mathbb{R}^{1369 \times 1024}$ is first reshaped to a 37×37 spatial feature map. The four levels are projected by 1×1 convolutions to $\{32, 64, 64, 64\}$ channels and resized to their corresponding spatial scales. They are further mapped to 32 fusion channels and aggregated through four coarse-to-fine feature-fusion blocks. Each block contains residual convolutional units followed by a 1×1 output projection, while the dense branch ends with a $32 \to 16 \to 8 \to 1$ convolutional head.

In parallel, a coarse CoC prediction is obtained from F_C^3 using LayerNorm, a $1024 \to 256$ linear projection, GELU, and a $256 \to 1$ projection. The token-wise prediction is reshaped to 37×37 and bilinearly upsampled to the dense prediction resolution. The coarse and dense predictions are combined before the final sigmoid:

$$C_{\mathrm{norm}} = \sigma\left(C_{\mathrm{coarse}} + C_{\mathrm{dense}}\right), \qquad C = C_{\max} C_{\mathrm{norm}}, \qquad C_{\max} = 20 \text{ pixels}. \tag{17}$$

The final 1×1 convolution of the dense branch is zero-initialized, so the CoC Head initially relies on the coarse prediction and progressively learns spatial refinements. The resulting 518×518 CoC map is cropped by three pixels on each side to recover the original 512×512 resolution.

Non-uniform radius bank. Fig. 6 illustrates the motivation for the non-uniform radius bank used in our renderer. Starting from the same sharp image, increasing the disk radius progressively suppresses high-frequency details and broadens local edges. In the small-radius regime ($\rho = 0 \to 4$), even a one-pixel change produces noticeable differences in leaf veins, thin boundaries, and local textures, indicating that the rendered appearance is more sensitive to small radius variations. In contrast, at larger radii ($\rho = 8 \to 20$), most fine structures have already been strongly attenuated, and further increases mainly produce gradual large-scale smoothing. We therefore sample small radii more densely and large radii more sparsely using $\{0, 1, 2, 3, 4, 6, 8, 12, 16, 20\}$, with intermediate blur strengths obtained by linear interpolation between neighboring kernel responses.

Formation feedback module. Given the random-time sharp-endpoint proposal $\widetilde{S}^{(t)} = D(\widetilde{z}_s^{(t)})$ decoded from the latent space, we compute the formation discrepancy $R^{(t)}$ as defined in Eq. 10. The feedback module encodes this image-space discrepancy to the latent resolution using three 3×3 convolutions with stride 2 and padding 1, with channel widths of 32, 64, and 64. Each convolution

is followed by GroupNorm and SiLU. For a 512×512 discrepancy, this yields a 64×64 feature map, matching the spatial resolution of the latent endpoint.

The feature is then projected to a 64-channel hidden representation by a 3×3 convolution and processed by two residual blocks, each containing two GroupNorm–SiLU–depthwise-separable convolution stages. Two 3×3 heads predict a 16-channel latent endpoint correction Δz_{fb} and a one-channel spatial gate map α, which are combined with the random-time endpoint according to Eq. 12. The gate is broadcast across latent channels, yielding a bounded, spatially varying latent correction from the image-space formation discrepancy.

The latent-update head is zero-initialized, while the spatial-gate head uses zero-initialized weights with a bias of -2. Thus, the feedback branch initially leaves $\widetilde{z}_s^{(t)}$ unchanged, while the gate starts at $\sigma(-2) \approx 0.12$. All GroupNorm layers use 32 groups under the default channel configuration.

B Additional Comparison Experiments

Qualitative comparison. Fig. 7 presents six additional qualitative comparisons across the three evaluation datasets. For DPDD, the first example focuses on flowers and foliage in the defocused background. Most methods recover the coarse scene appearance, but fine leaf textures and flower structures remain partially smoothed. MBSD recovers more visible local details than several competing methods, while FCFlow further preserves finer foliage and flower structures that more closely match the sharp target. The second example focuses on distant people and a bench. Most methods recover the overall silhouettes of the people, but local color and texture variations remain inconsistent with the target. For the bench, most competing methods produce blurred structures or lose substantial fine details, whereas FCFlow preserve richer structural information.

For LFDOF, the third example focuses on a transparent object with fine internal structures. Several methods recover its overall shape but leave the internal details partially blurred or distorted. FCFlow recovers more of these structures while maintaining well-defined object boundaries, yielding a closer match to the target. The fourth example highlights fine flower textures, leaves, and thin stems under spatially varying defocus. Most methods recover the dominant structures, but fine textures and narrow boundaries are weakened by residual smoothing. FCFlow better preserves detailed leaf textures and thin stem structures across regions with different defocus strengths.

For RealDOF, the fifth example focuses on text displayed on an LED sign. Several methods recover the overall appearance of the sign but fail to restore fine text patterns, resulting in blurred or distorted character structures. FCFlow reconstructs the text more clearly and more closely matches the target. The sixth example focuses on a railing with repetitive geometric structures. While most methods recover the overall railing, individual bars remain blurred or locally irregular. FCFlow better preserves the repetitive geometry and local boundaries.

Notably, FCFlow-1 already achieves competitive visual restorations across all six examples, while FCFlow-20 further improves local details and overall fidelity. This demonstrates that FCFlow supports effective single-step inference while still benefiting from additional integration steps when higher restoration quality is desired.

Overall, these examples show that FCFlow consistently improves the recovery of fine textures, weak structures, internal details, and repetitive geometric patterns across different datasets and defocus conditions, while maintaining strong visual quality under both one-step and 20-step inference.

C Additional Ablation Studies

Sensitivity to Formation-Model Approximation To evaluate the robustness of FCFlow to approximations of the assumed forward formation model, we replace the default disk-based defocus renderer with a spatially varying Gaussian renderer. The disk PSF explicitly models the circle of confusion induced by defocus under our formation assumption, while the Gaussian PSF serves as an approximate formation model that deviates from this defocus process. For a fair comparison, all other components, including the flow-matching backbone, visual conditioning, CoC predictor, endpoint-feedback module, training schedule, and loss configuration, are kept unchanged.

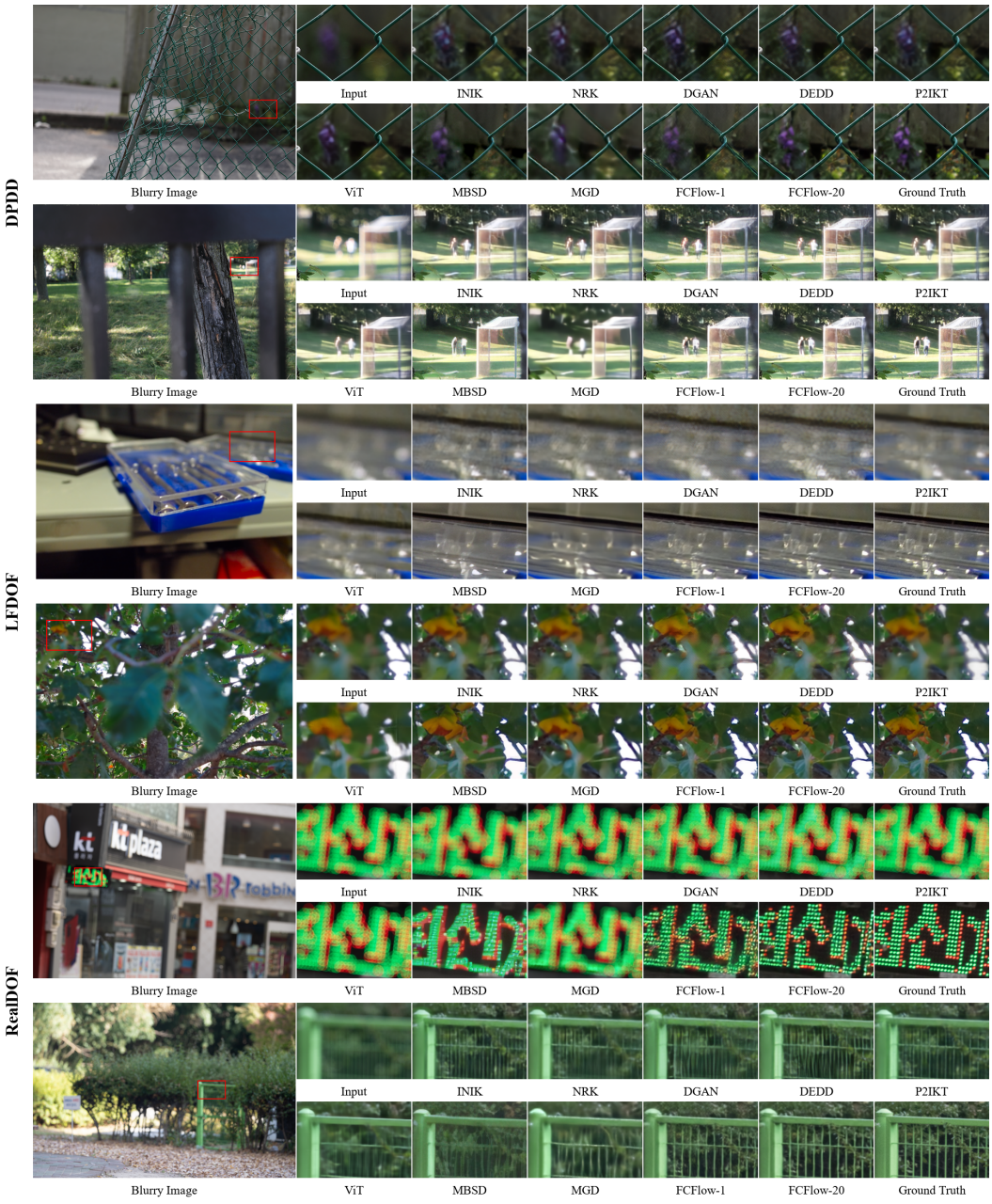


Figure 7: Additional qualitative comparisons on representative examples from DPDD, LFDOF, and RealDOF. Each dataset includes two examples, with the full blurred input shown on the left and enlarged crops on the right. Red boxes indicate the corresponding crop regions.

For the Gaussian variant, the spatially varying standard deviation is derived from the predicted CoC radius C by matching the second-order spatial moment of the disk kernel, which gives $\sigma = C/2$. We compare the Gaussian-based formation feedback with vanilla flow matching without formation feedback and the default disk-based FCFlow. Results are reported on DPDD with both one-step and 20-step inference.

Table 4 evaluates the sensitivity of FCFlow to the choice of formation model. The Gaussian-based variant consistently improves reconstruction metrics over vanilla flow matching at both inference budgets, demonstrating that the effectiveness of formation feedback does not rely exclusively on the exact disk-PSF parameterization. This indicates that even an approximate forward formation model can provide useful observation-space guidance for shaping the restoration field.

Table 4: Sensitivity to formation-model approximation on DPDD. The Gaussian variant replaces only the disk-based formation renderer, while all other model components and training configurations remain unchanged. Results are reported for both 1-step and 20-step inference.

NFE	Formation Model	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	MUSIQ↑
1	Vanilla FM	25.6903	0.7928	0.1804	0.1232	57.7459
	Gaussian PSF	<u>28.1039</u>	<u>0.8290</u>	<u>0.1503</u>	<u>0.1083</u>	<u>58.1124</u>
	Disk PSF (FCFlow)	**28.7432**	**0.8307**	**0.1372**	**0.0813**	**58.4522**
20	Vanilla FM	27.6936	0.8297	0.1608	0.1017	58.2571
	Gaussian PSF	<u>29.1038</u>	<u>0.8397</u>	<u>0.1146</u>	<u>0.0847</u>	<u>59.0410</u>
	Disk PSF (FCFlow)	**29.3448**	**0.8448**	**0.0920**	**0.0533**	**60.6433**

Table 5: Conditioning-level ablation on the DPDD test split. The velocity field is conditioned on different feature levels or their fusion. F_C^3 denotes the default configuration.

Velocity condition	PSNR@1↑	PSNR@20↑	LPIPS@1↓	LPIPS@20↓
F_C^1	27.8197	28.5105	0.1450	0.1180
F_C^2	28.0165	28.7240	0.1427	0.1128
F_C^3 (default)	<u>28.7432</u>	**29.3448**	<u>0.1372</u>	<u>0.0920</u>
F_C^4	28.1132	28.9351	0.1401	0.1098
$F_C^{1:4}$ (fused control)	**28.7513**	<u>29.3412</u>	**0.1328**	**0.0901**

The disk-based FCFlow achieves the best overall performance under both one-step and 20-step inference. Compared with the Gaussian formulation, the disk renderer improves PSNR by 0.64/0.24 dB and SSIM by 0.0017/0.0051 for 1-step/20-step inference, while providing consistently better perceptual quality with lower LPIPS and DISTS scores. The larger perceptual gap at 20-step inference suggests that a more accurate formation model provides more reliable guidance when the learned velocity field is integrated with finer numerical steps.

Overall, these results show that FCFlow is robust to moderate formation-model approximation, while a formation model that better matches the underlying defocus process provides more effective endpoint feedback and shapes a more accurate restoration field.

Conditioning-Level Ablation. Table 5 evaluates the feature level used to condition the velocity field. Among the single-level variants, F_C^3 achieves the best overall performance, suggesting that the representation from the corresponding intermediate block provides a favorable balance between local visual information and higher-level context.

Fusing $F_C^{1:4}$ further improves most reported metrics, indicating that the selected feature levels contain complementary information. However, multi-level conditioning requires additional feature projection and fusion. We therefore use F_C^3 as the default velocity condition, providing a simpler conditioning pathway with only a modest performance gap to the fused variant.

Formation-Model Fidelity. Fig. 8 further visualizes representative examples from the three datasets. Most image regions exhibit small residuals, while larger discrepancies are mainly concentrated around high-frequency structures, thin edges, object boundaries, and depth transitions. This spatially localized error pattern indicates that the predicted effective CoC field, together with the differentiable renderer, captures the dominant defocus formation over most regions, with larger mismatches primarily occurring in areas that are more challenging to reproduce using the adopted spatially varying disk-kernel model. Importantly, the remaining discrepancy is neither spatially diffuse nor unstructured; instead, it highlights locations where the forward model fails to fully reproduce the observed blur. Such localized discrepancies provide an informative formation-consistency signal for guiding the learned flow field toward restorations that better agree with the observed blur.



Figure 8: Qualitative evaluation of formation-model fidelity on DPDD, LFDOF, and RealDOF. For each dataset, the five columns show the paired sharp target S, predicted effective CoC map C, input blurred image B, reblurred image $\mathcal{B}_C(S)$, and the formation discrepancy map $R = B - \mathcal{B}_C(S)$. The colorbars beside the CoC and discrepancy maps indicate the corresponding quantitative values, representing the effective CoC radius in pixels and the pixel-intensity difference between the observed and reblurred images, respectively.

D Computational Analysis and Additional Discussion

D.1 Computational accounting

We report the test-stage parameter count and forward computation of FCFlow for a single 512×512 sample with 1-NFE. The CoC head, defocus renderer, and formation feedback module are excluded from the final A7 flow-only inference path. Component-wise FLOPs are rounded independently, while the reported total corresponds to the directly measured end-to-end cost of the Flow-only inference pipeline.

Table 6: Parameters loaded during FCFlow inference for a single 512×512 input.

Module	Parameters
DINOv2-L	304.369 M
Qwen-Image VAE	44.600 M
LightningDiT	486.558 M
Adapter	4.199 M
Total loaded parameters	839.726 M

Table 7: Forward FLOPs for a single 512×512 input with 1-NFE.

Test-stage component	Forward FLOPs
DINOv2-L	1.014 T
Qwen-Image VAE encode	0.636 T
Adapter	0.046 T
LightningDiT	2.602 T
Qwen-Image VAE decode	1.100 T
Total	5.403 T

Table 8: Single-image inference latency for a 512×512 input with 1-NFE on an RTX 4090 using FP32/TF32.

Inference mode	Mean	Median	Throughput
A6 FCFlow (w/ Inference Correction)	193.44 ms	193.57 ms	∼5.17 images/s
A7 FCFlow (w/o Inference Correction)	156.84 ms	157.02 ms	∼6.38 images/s

Runtime is measured on a single 512×512 input using two warm-up runs followed by eight timed runs with CUDA events. Model loading, image I/O, and PNG saving are excluded. FCFlow without test-time feedback uses only the Flow inference path and takes 156.84 ms on average, corresponding to approximately 6.38 images/s. The optional test-time feedback incurs an additional 36.60 ms on average, corresponding to a 23.3% runtime overhead relative to Flow-only inference.

Model complexity and inference cost. Tables 6 and 7 report the complexity of the final A7 Flow-only inference path. For a 512×512 input with 1-NFE, the CoC and feedback branches used during training are excluded at inference, resulting in 5.403 T FLOPs and 839.726M loaded parameters. Table 8 additionally reports the runtime of the optional feedback-enabled configuration for reference; this branch is not used in the final deployment setting.

D.2 Discussion

Continuous Flow Matching supervises the blur-to-sharp velocity field over sampled intermediate states, while the proposed formation-aware training additionally evaluates the realized sharp-endpoint proposal through the forward formation process. On the DPDD test split, FCFlow achieves 28.7432 dB PSNR with a single Euler step, compared with 25.6903 dB for Vanilla FM. This substantial improvement demonstrates the effectiveness of the complete formation-aware training framework, in which CoC-based formation modeling and discrepancy-driven endpoint feedback jointly shape the learned restoration field toward more accurate endpoint transport.

The NFE analysis further shows that FCFlow substantially reduces the dependence on iterative integration. FCFlow already achieves strong restoration quality at 1-NFE and rapidly approaches saturation as the number of integration steps increases, whereas Vanilla FM benefits much more strongly from additional steps. Multi-step integration nevertheless provides further improvements, showing that the learned field supports both effective single-step inference and higher-quality iterative refinement.

Evaluation on the held-out LFDOF and RealDOF datasets further demonstrates strong cross-dataset generalization. We emphasize that the forward formation process is parameterized by a learned effective CoC map, which serves as a practical representation of spatially varying defocus rather than a fully calibrated optical model.

D.3 Limitations

Despite its strong performance, FCFlow has certain limitations. The differentiable renderer is based on normalized disk-shaped kernels and adopts an idealized defocus formation assumption, without explicitly accounting for complex optical effects such as lens aberrations, sensor noise, exposure variations, ISP processing, and potential spatial misregistrations in training pairs. These modeling assumptions may introduce bias into the estimated CoC map and formation discrepancy, reducing the reliability of the endpoint feedback and potentially affecting the learned velocity field and restoration performance.