BiFM: Bidirectional Flow Matching for Few-Step Image Editing and Generation

Supplementary Material

A. Extensive Backgrounds

Denoising Diffusion Models generate images from noise by learning a reverse process of a predefined forward diffusion process. The forward diffusion process is formulated as a Markov process starting from data space to a prior noise space after multiple time steps. Specifically, given the number of discrete timesteps T, mean schedule $\{\alpha_t\}_{t=1}^T$, and variance schedule $\{\sigma_t^2\}_{t=1}^T$, the forward process is formalized as Eq. (13), where $\mathbf{x}_0 \sim p_{\text{data}}(\mathbf{x})$ and $\mathbf{x}_T \sim \mathcal{N}(0, I)$. The learned reverse process of Eq. (13) is Eq. (14):

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \alpha_t \mathbf{x}_0, \sigma_t^2 I) \tag{13}$$

$$p_\theta(\mathbf{x}_0) = \int p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \tag{14}$$

The log-likelihood of samples from denoising diffusion models can be decomposed as:

$$\mathbb{E}\left[-\log p_\theta(\mathbf{x}_0)\right] \leq \mathbb{E}_q \left[-\sum_{t\geq 1} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \right] - \log p(\mathbf{x}_T) \tag{15}$$

By considering only optimization terms associated with the learned network (expressed as ϵ_θ), the training objective can be expressed by ELBO of Eq. (15) as Eq. (16), a noise-prediction parameterization found effective by Ho et al. [13]:

$$\mathcal{L} = \mathbb{E}_{t,\mathbf{x}_0,\epsilon} \|\epsilon - \epsilon_\theta(\alpha_t \mathbf{x}_0 + \sigma_t \epsilon, t)\|^2 \tag{16}$$

Flow Matching constructs a time dependent path which transports noise distribution $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{x}; 0, I)$ into data distribution $\mathbf{x}_1 \sim p_{\text{data}}(\mathbf{x})$. The transportation is described as the following flow matching ODE:

$$\frac{d}{dt}\phi_t(\mathbf{x}) = v_t(\phi_t(\mathbf{x})), \tag{17}$$

$$\mathbf{x}_t = \phi_t(\mathbf{x}_0), \phi_0(\mathbf{x}) = \mathbf{x} \tag{18}$$

A flow model is uniquely determined by its learned velocity field $v_\theta(\mathbf{x}_t, t)$. Flow matching modifies from the noise prediction in denoising diffusion trajectory to velocity prediction in probability distribution transport flow, which simplifies the overall framework. A practical flow matching training objective, conditional Flow Matching Loss (CFM), can be written as:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1} ||v_\theta(\mathbf{x}_t, t) - v_t(\mathbf{x}_t|\mathbf{x}_0, \mathbf{x}_1)||^2 \tag{19}$$

where the target velocity v_t is the conditional velocity. where $v_t|\mathbf{x}_0, \mathbf{x}_1 := \mathbf{x}_1 - \mathbf{x}_0$ is the per-sample velocity of the flow, $v_\theta(\mathbf{x}_t, t)$ is the velocity prediction from the leaned neural network θ.

B. Additional Implementation Details

Model Configuration. For image editing experiments, we adopt Stable Diffusion 3 Medium (SD3-M) [8], a Multimodal Diffusion Transformer (MMDiT) operating in the latent space of its VAE, conditioned on three pretrained text encoders: CLIP-L/14, CLIP-G/14, and T5-XXL, following the official SD3 design. Following [4], we train LoRA adapters only in the MMDiT blocks. In addition, we introduce a trainable extra time-interval embedding module that augments the SD3 timestep conditioning, which has the same architecture as the time embedding in original SD3, and is zero-initialized. The total trainable parameters are LoRA weights injected into attention/ MLP projections, and extra time-embedding parameters. For SD3 experiments, we use 32 H100 GPUs for fine-tuning; 100 epochs takes $\sim$ 120 hours.

Dataset Configuration. For training BIFM on a pretrained Stable Diffusion 3 model [8], we utilized MagicBrush dataset [41] with 10K manually annotated real image editing triplets. We generate captions for source and target images using BLIP-2. We train our model with batch size of 4 and learning rate $1e^{-5}$ with Adam optimizer.

Training Configuration. We do not train BiFM with CFG guidance (unlike MeanFlow configurations) to preserve sampling flexibility across guidance values. We train without guidance and apply CFG only at inference when appropriate. For T2I results we use CFG scale 4. For ImageNet results we do not apply CFG. ImageNet training takes 80 epochs ($\sim 150k$ steps, batch size 256), and Table A(b) shows BiFM achieves noticeable gains over MeanFlow after 80 epochs.

C. Additional Experimental Details

Sampling and Evaluation Details. In Figure 4, NFE/steps used for each method are: PnP Inv 50, RF-Edit 30, FlowEdit 28, ReNoise 4, SwiftEdit 1, and BiFM 1.

We show that BiFM offers benefit in image generation training. Fig. C presents training curves of FID versus epoch for baseline (FM) and FM augmented with BiFM. Across the entire training trajectory, FM+BiFM achieves consistently lower FID than FM alone, indicating faster conver-

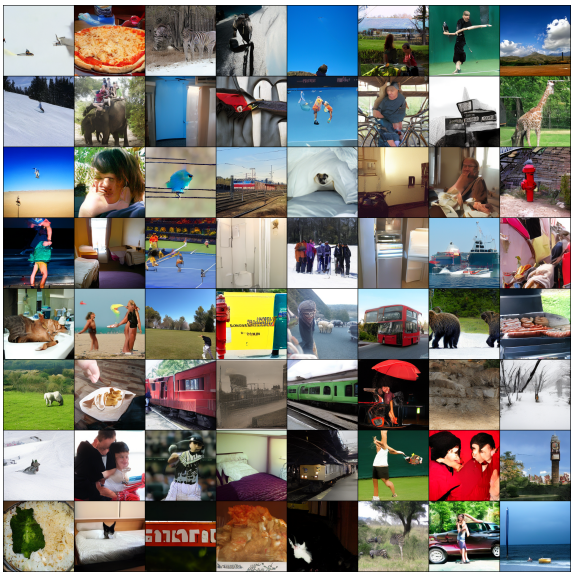

Figure A. **MSCOCO-256 Text-to-Image Generation Visualization Results.** Model trained for 100K iterations.

Figure B. **Image Editing Visualization.**

gence and better generative quality.

Method	Model	FID↓
FM	MMDiT	6.05
REPA	MMDiT	4.73
MeanFlow	MMDiT	5.02
BiFM	MMDiT	**4.57**

(a) Text-to-image generation result on MSCOCO. We re-implement MeanFlow on MMDiT for results.

Model	Method	FID↓
SiT-B/2	MeanFlow	28.2
SiT-B/2	**BiFM**	27.8
SiT-L/2	MeanFlow	17.0
SiT-L/2	**BiFM**	**16.4**

(b) ImageNet-256 generation. We do not distill / apply CFG for this experiment.

Table A. **Image Generation Results.** NFE=50.

We include additional baselines for image generation experiments (see Table A). In Table A, we add MeanFlow results on MSCOCO and MeanFlow training from scratch on

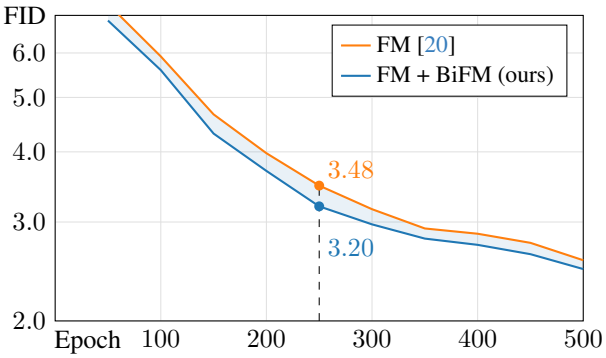

Figure C. **CIFAR-10 Training Epochs vs. FID**

ImageNet, to validate improvements beyond CIFAR-10.

D. More Generation Visualization

In this section we provide more visualization samples from image generation experiments. In Figure B, we show editing examples comparing BiFM and FlowEdit under 1-step, 4-step and 28-step settings. Fig. A shows uncurated samples generated by vanilla MMDiT using random prompts from MSCOCO-256 dataset. For small-resolution datasets, Fig. D and Fig. E display uncurated 32×32 samples from CIFAR-10 and ImageNet-32, respectively, generated by a U-Net model trained with BiFM.

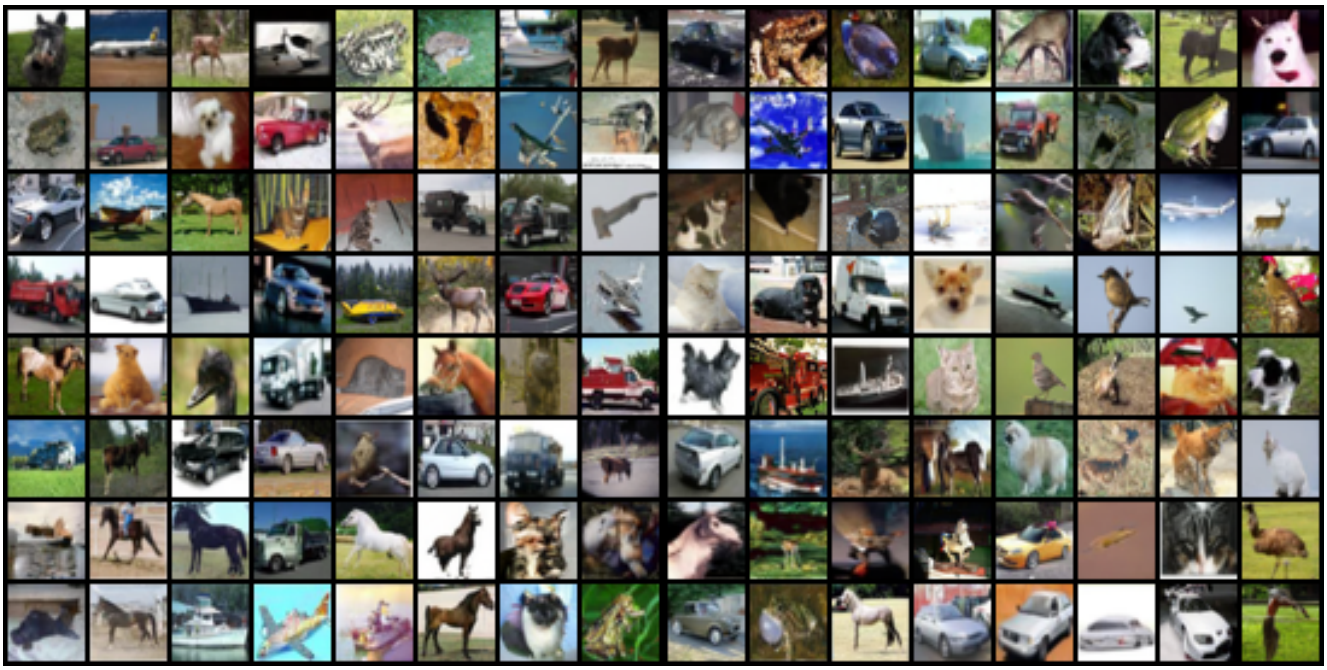

Figure D. **CIFAR-10 Generation Visualization Results.** Model trained for 500 epochs.

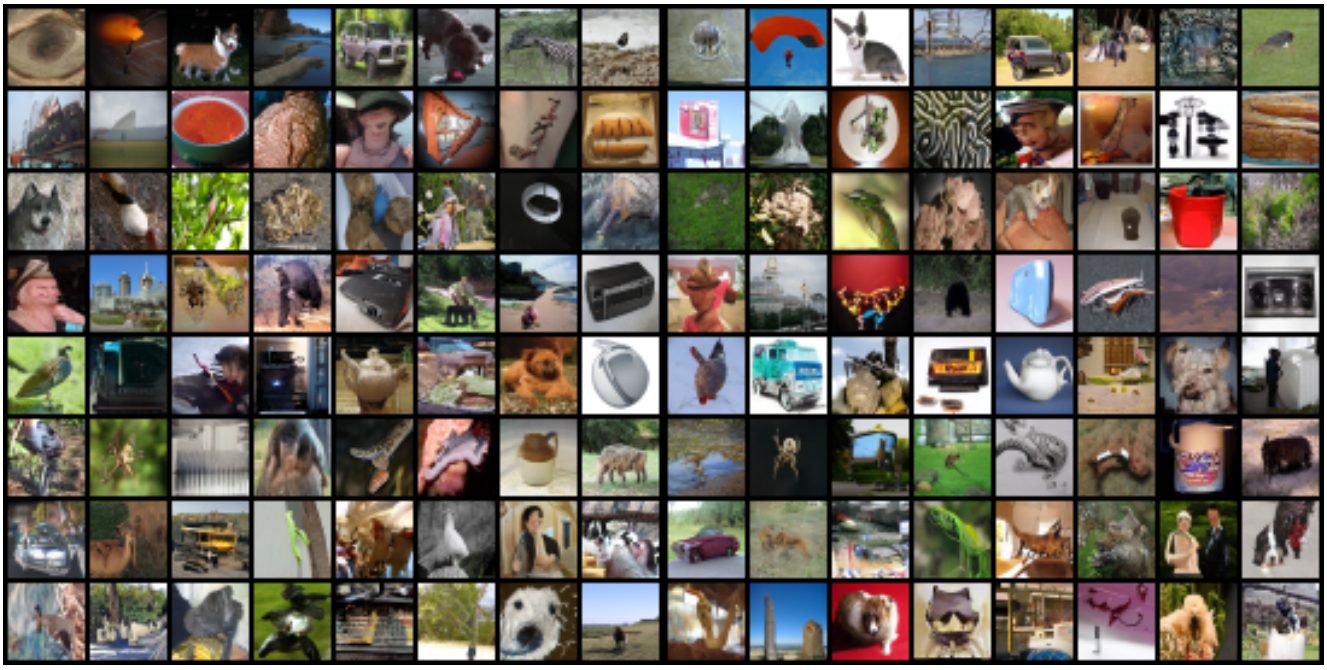

Figure E. **ImageNet-32 Generation Visualization Results.** Model trained for 80 epochs.