Fixed Point Diffusion Models

Supplementary Material

Blocks	*Model*	*FID (DPM-Solver++)*	*FID (DDPM)*	*Params.*
120	UNet	77.9	**89.8**	**46M**
	FP-UNet	**63.1**	95.9	47M
240	UNet	83.0	83.0	**46M**
	FP-UNet	**65.9**	**70.5**	47M
600	UNet	81.8	81.5	**46M**
	FP-UNet	**62.5**	**63.5**	47M

Table 6. **Evaluation of Fixed-Point Variants of Non-ViT Architectures.** We show results with *non-ViT* architectures. The fixed point adaptations (FP-Unet) show significant improvements over standard Unet in both *DPM-Solver++* and *DDPM* samplers. Note that the FP-UNet replaces a single UNet layer with a fixed point layer, so it has marginally more parameters.

A. Diffusion Models

As stated in Sec. 3.3, we provide an overview of diffusion models here in case any readers are not familiar with diffusion models.

Diffusion denoising probabilistic models add noise to a data sample X_0 drawn from a target data distribution $q(X_0)$. This noising process is executed in a series of steps, where each step adds a specific quantity of noise controlled by a variance schedule $\{\beta_t\}_{t=0}^{T}$. At each step, the new data sample X_t is generated from the previous one X_{t-1} according to the distribution $q(X_t|X_{t-1}) = \mathcal{N}(X_t; \sqrt{1-\beta_t}X_{t-1}, \beta_t\mathbf{I})$. The reverse diffusion process, or generative process, starts with a noisy sample from $q(X_T) \sim \mathcal{N}(0, 1)$ and aims to iteratively remove the noise to recover a sample from the original distribution $q(X_0)$. This reverse process is learned by a neural network, approximating the distribution $q(X_{t-1}|X_t)$ as $s_\theta(X_{t-1}|X_t) \approx q(X_{t-1}|X_t)$.

B. Additional Qualitative Examples

We provide examples on CelebA-HQ, LSUN Church, FFHQ, and ImageNet in Figs. 8 to 11. For each dataset, we sample 48 images using DDPM with 560 transformer block forward passes at resolution 256px. Note that the images are not cherry-picked. We also provide additional qualitative comparisons with DiT in Fig. 12.

C. UNet Architectures

Our method is not limited to transformer-based architectures. In **??**, we showcase experiments with non-ViT models: UNet [42] and Hourglass Transformer (HT) [11]. We find that, as with DiT, fixed point variants of UNet (FP-UNet) and HT (FP-HT) demonstrate strong performance. Additionally, we note that HT was released after the submission of our paper, which demonstrates how easily our method can be adapted to new architectures.

Iters. per Step	3	5	6	8	12	26
Constant	48.0	45.8	46.6	47.3	48.5	62.5
Decreasing	48.0	46.3	47.3	48.3	49.1	63.2
Increasing	**46.7**	**44.8**	**45.9**	**45.6**	**48.0**	**61.7**

Table 7. **Performance of Iteration Allocation Heuristics.** *Constant* uses a fixed iteration count per diffusion timestep, while *Increasing* and *Decreasing* vary their iteration counts linearly with respect to the timestep.

D. Additional Results on Reallocating Computation Across Timesteps

As described in Sec. 3.3, we apply very simple heuristics ("constant", "increasing", "decreasing") to vary the number of iterations used at each timestep across the denoising process. Results are shown in Tab. 7. The increasing heuristic outperforms the constant and decreasing ones; this aligns with the intuition that when a little compute is given, allocating resources more toward the later stages of the denoising process improves generation quality and detail. Note that such flexibility in resource allocation is a novel feature of FPDM, not possible in previous explicit diffusion models.

E. Description of an Adaptive Allocation Algorithm

FPDM allows for the adjustment of solution accuracy at different stages of the denoising process. As noted in Sec. 3.3, in addition to implementing straightforward heuristics such as "increasing" and "decreasing", it supports using adaptive algorithms to allocate the forward passes across timesteps. We leave an in-depth investigation of adaptive algorithms to future work, but we give an example below to demonstrate how one such algorithm could work.

We start by considering θ_t, the difference between the last two solving solutions, as a metric of solution quality at each step. This aligns with our observation in Fig. 6b, where θ_t decreases as more fixed point iterations (i.e. forward passes) are applied. Then a simple adaptive algorithm could be to simply set an error threshold Θ and iterate the fixed point iteration process at each timestep t continue until θ_t falls below Θ. Then the global threshold Θ controls the number of forward passes.

The only question left would be how to choose Θ to match a given computational budget (i.e. a number of forward passes). For this, an online binary probing scheme can be employed: use binary search to select a Θ', on which we perform inference for one batch of images. If the number of forward passes used to meet the Θ' threshold exceeds our

budget, we increase Θ' in subsequent iterations; conversely, if the number is below our budget, we decrease Θ'. Note that only constant time of probing is needed to find a sufficiently good threshold at the beginning of the inference. This computational cost would be negligible, especially when sampling many batches of images.

Figure 8. **Additional qualitative examples on CelebA-HQ.** Examples are sampled using the DDPM sampler with 560 transformer block forward passes at resolution 256px. These images are not cherry-picked.

Figure 9. **Additional qualitative examples on LSUN Church.** Examples are sampled using the DDPM sampler with 560 transformer block forward passes at resolution 256px. These images are not cherry-picked.

Figure 10. **Additional qualitative examples on FFHQ.** Examples are sampled using the DDPM sampler with 560 transformer block forward passes at resolution 256px. These images are not cherry-picked.

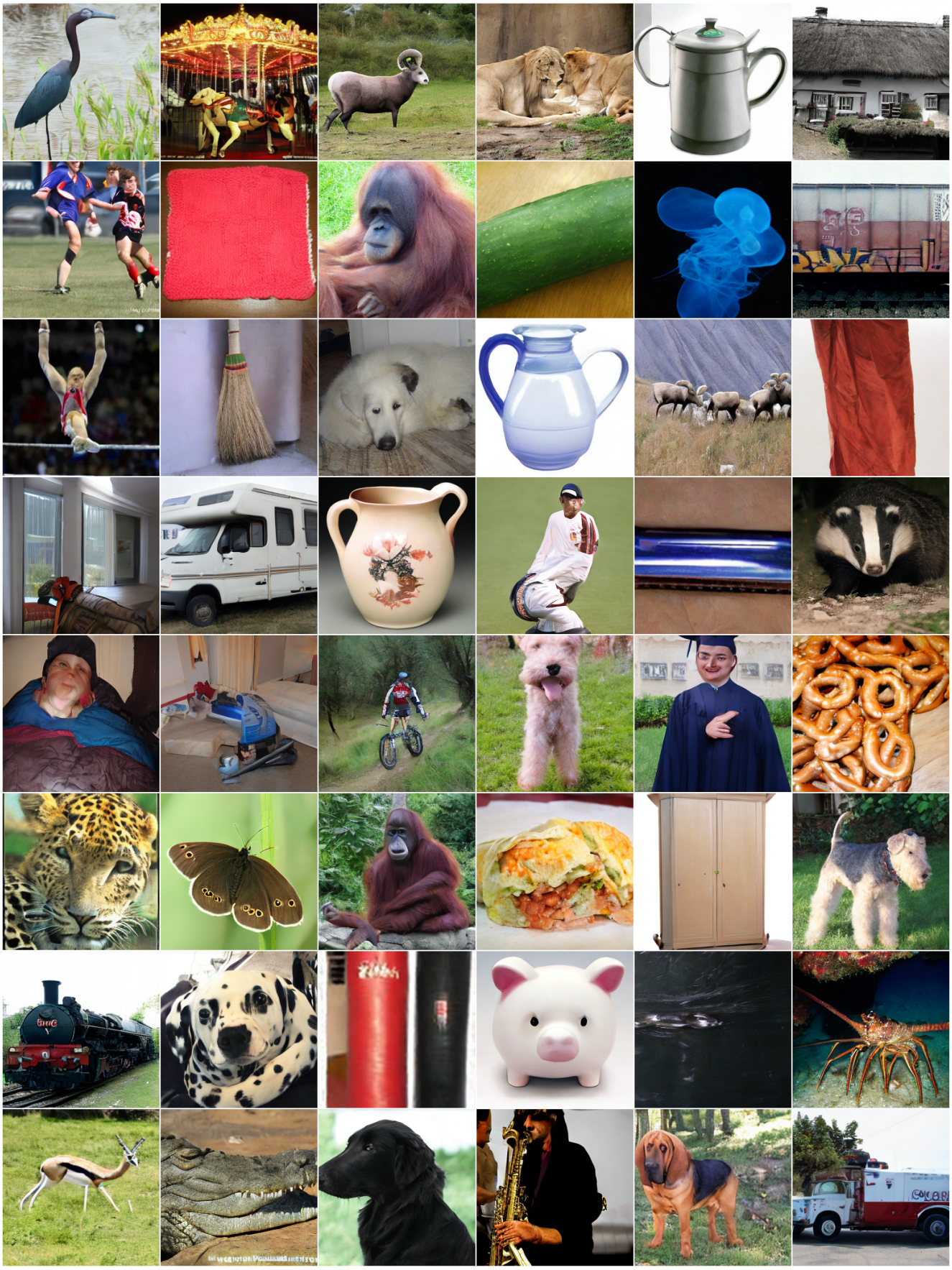

Figure 11. **Additional qualitative examples on ImageNet.** Examples are sampled using the DDPM sampler with 560 transformer block forward passes at resolution 256px. These images are not cherry-picked.

Figure 12. **Additional qualitative comparison with DiT.** We show examples on CelebA-HQ, LSUN Church, FFHQ, and ImageNet. All images are sampled using the DDPM sampler with 560 transformer block forward passes at resolution 256px. The images are not cherry-picked.