
Single-Step Diffusion via Direct Models

Anonymous Author(s)

Affiliation

Address

email

1 Comparison in the Case of Class-Conditioned Generation on ImageNet-256

Table 1: Comparison of different training objectives using the same model architecture (**DiT-B**). FID-50k scores (lower is better) are reported for 128-step, 4-step, and 1-step denoising. Direct Models achieves high-quality samples in a single step and training run, narrowing the gap with two-phase distillation methods. Parentheses indicate evaluations outside the method’s intended scope.

	ImageNet-256 (Class-Conditioned)		
	128-Step	4-Step	1-Step
<i>Two-Phase Training</i>			
Progressive Distillation	(201.9)	(142.5)	35.6
Consistency Distillation	132.8	98.01	136.5
Reflow	16.9	32.8	44.8
<i>End-to-End (Single Training Run)</i>			
Diffusion	39.7	(464.5)	(467.2)
Flow Matching	17.3	(108.2)	(324.8)
Consistency Training	42.8	43.0	69.7
Live Reflow (Ours)	46.3	95.8	58.1
Shortcut Models	15.5	28.3	40.3
Direct Models (Ours)	-	-	36.5

Table 1 presents results for class-conditioned generation on ImageNet-256. Direct Models achieves competitive performance in the single-step generation setting, outperforming several end-to-end baselines and narrowing the gap with multi-phase approaches.

The only difference from the unconditional case is that we now aim to learn $w_\nu(x_0, c, t)$, where c is the class label used for conditioning.

The class-conditioned version of the progressive velocity propagation equation is:

$$w(x, c, t + \delta t) = \frac{t - \delta t}{t} \cdot w(x, c, t) + \frac{\delta t}{t} \cdot v(x_t, c, t) \quad (1)$$

This leads to the same training algorithm as in the unconditional case, with the addition of class conditioning.

Given an initial sample $x_0 \sim \mu_0$, the corresponding transformed sample x_1 is obtained via a single forward pass:

$$x_1 = x_0 + w_\nu(x_0, c, 1). \quad (2)$$

Classifier-Free Guidance [1]. For the ImageNet case, we adopt classifier-free guidance (CFG) during training. A limitation of CFG in Direct Models is that the guidance scale must be chosen before training.



Figure 1: Representative examples generated unconditionally on the CelebA-HQ dataset at 256×256 resolution, using single-step generation with a DiT-B size model trained for 500,000 iterations.

2 Visual Results

Figures 1 and 2 show images generated by our method, trained on the unconditional CelebA-HQ and class-conditioned ImageNet datasets, respectively.

3 Training Details

Table 2 provides detailed training configurations corresponding to the results reported in Table 1 (main paper) and Table 1 (supplementary).

References

- [1] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1

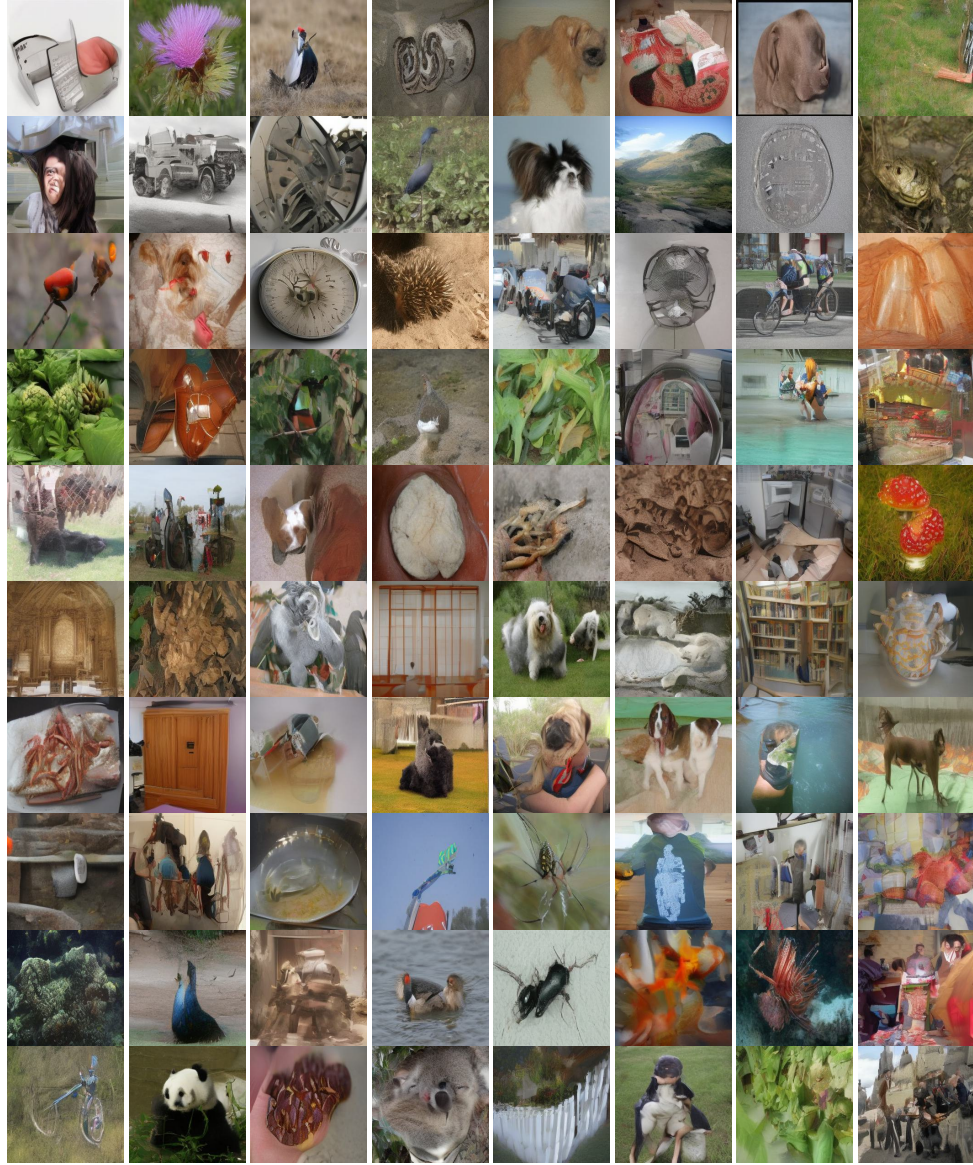


Figure 2: Representative examples generated unconditionally on the ImageNet dataset at 256×256 resolution, using single-step generation with a DiT-B size model trained for 800,000 iterations.

Batch Size	64 (CelebA-HQ), 256 (Imagenet)
Training Steps	500,000 (CelebA-HQ), 800,000 (Imagenet)
Latent Encoder	sd-vae-mse-ft
Latent Downsampling	8 (256x256x3 to 32x32x4)
Classifier Free Guidance	0 (CelebA-HQ), 1.5 (Imagenet)
Class Dropout Probability	0 (CelebA-HQ), 0.1 (Imagenet)
EMA Parameters Used For Evaluation?	Yes
EMA Ratio	0.999
Optimizer	AdamW
Learning Rate	0.00005
Weight Decay	0.0
Hidden Size	768
Patch Size	2
Number of Layers	12
Attention Heads	12
MLP Hidden Size Ratio	4
δt	0.01

Table 2: Hyperparameters used during training.