1. Appendix

1.1. Nonuniform Distribution Selection

Notably, we have carefully chosen the nonuniform distribution for sampling timestep t (Eq. 15 in the main paper). Specifically, except for the normal distribution in the paper, we also consider the Poisson and exponential distributions. Also, a series of hyperparameter selection experiments are conducted. More details are presented in Tab. 1.

Table 1. Hyperparameter selection for non-uniform distribution in Algorithm 1 in the main paper.

Task	Method	IS ↑	FID ↓	sFID ↓
Other Nonuniform Distributions	FP	14.88	21.63	17.66
	Poisson	13.29	34.54	25.84
	Exponential	12.87	39.91	30.04
Normal Distribution with Different Mean μ, and Variance σ	$\mu = \frac{T}{2},\ \sigma = 0.5\sqrt{\frac{T}{2}}$	15.45	25.11	17.35
	$\mu = \frac{T}{2},\ \sigma = 1.0\sqrt{\frac{T}{2}}$	15.65	24.83	18.90
	$\mu = \frac{T}{2},\ \sigma = 2.0\sqrt{\frac{T}{2}}$	15.85	24.27	17.92
	$\mu = \frac{1.5T}{2},\ \sigma = \sqrt{\frac{T}{2}}$	12.63	39.09	35.81
	$\mu = \frac{1.0T}{2},\ \sigma = \sqrt{\frac{T}{2}}$	15.65	24.83	18.90
	$\mu = \frac{0.5T}{2},\ \sigma = \sqrt{\frac{T}{2}}$	15.88	23.96	17.67

1.2. Actual Acceleration

We test the latency(ms) of the original network (provided checkpoint) and the quantized network on Nvidia RTX A6000 GPU. The results in Table 2 show that the 8-bit quantization achieves about 2x speedup. The speedup can be more significant on NPU.

Table 2. Inference speed test with Nvidia RTX A6000.

Task	Batch Size	FP32	INT8
ImageNet 64x64	1	9.80	4.99
	16	64.42	28.16
CIFAR 32x32	1	5.92	2.98
	16	23.15	14.13