Supplementary Material
Spiking Denoising Diffusion Probabilistic Models

Jiahang Cao[1*] Ziqing Wang[1,2*] Hanzhong Guo[3*] Hao Cheng[1] Qiang Zhang[1] Renjing Xu[1†]
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]North Carolina State University, [3]Renmin University of China

Figure 1. More visualization results of CelebA 64×64 and LSUN bedroom 64×64 datasets.

1. Theoretical Energy Consumption Calculation

To calculate the theoretical energy consumption, we begin by determining the synaptic operations (SOPs). The SOPs for each block in the Spiking UNet can be calculated using the following equation:

$$\mathrm{SOPs}(l) = fr \times T \times \mathrm{FLOPs}(l) \tag{1}$$

where l denotes the block number in the Spiking UNet, fr is the firing rate of the input spike train of the block and T is the time step of the spike neuron. $\mathrm{FLOPs}(l)$ refers to floating point operations of l block, which is the number of multiply-and-accumulate (MAC) operations. And SOPs are the number of spike-based accumulate (AC) operations.

To estimate the theoretical energy consumption of Spiking Diffusion, we assume that the MAC and AC operations are implemented on a $45nm$ hardware, with energy costs of $E_{MAC} = 4.6pJ$ and $E_{AC} = 0.9pJ$, respectively. According to [5, 8], the calculation for the theoretical energy consumption of Spiking Diffusion is given by:

$$\begin{aligned} E_{\text{Diffusion}} = {} & E_{MAC} \times \mathrm{FLOP}^{1}_{\mathrm{SNN}_{\mathrm{Conv}}} \\ & + E_{AC} \times \left(\sum_{n=2}^{N} \mathrm{SOP}^{n}_{\mathrm{SNN}_{\mathrm{Conv}}} + \sum_{m=1}^{M} \mathrm{SOP}^{m}_{\mathrm{SNN}_{\mathrm{FC}}} \right) \end{aligned} \tag{2}$$

where N and M represent the total number of layers of Conv and FC, E_{MAC} and E_{AC} represent the energy cost of MAC and AC operation, $\mathrm{FLOP}_{\mathrm{SNN}_{\mathrm{Conv}}}$ denotes the FLOPs of the first Conv layer, $\mathrm{SOP}_{\mathrm{SNN}_{\mathrm{Conv}}}$ and $\mathrm{SOP}_{\mathrm{SNN}_{\mathrm{FC}}}$ are the

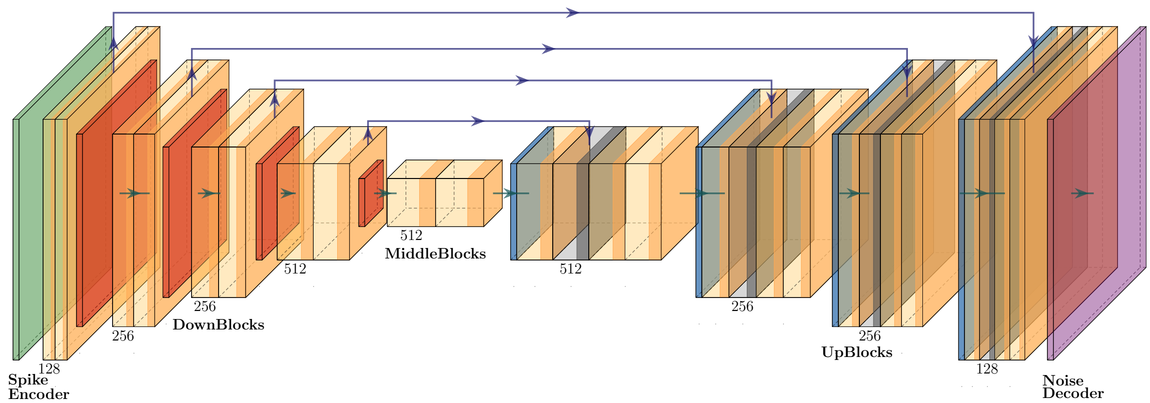

Figure 2. **Detailed architecture of our SNN-UNet.** Our network mainly consists of Pre-spike Resblocks (colored in yellow). The initial noise will first enter the spiking encoder (green) and then be converted into spike series. The forward process is performed by propagating through the DownBlocks, MiddleBlocks and UPBlocks. The orange and the blue blocks indicate the downsampling and upsampling layers, respectively. Eventually, we can get the predicted noise from the last noise decoder (magenta), which in turn reconstructs the image.

SOPs of n^{th} Conv and m^{th} FC layer, respectively.

2. More visualization on the Celeba and LSUN

We provide more qualitative results on the CelebA and LSUN bedroom datasets at the beginning of this Supplementary Material, hoping to aid the reader in assessing image quality, and artifacts.

3. Implementation Details

The detailed architecture of our Spiking UNet is illustrated in Fig. 2. It is important to note that we adopt the most primitive UNet [6] structure without any transformer blocks since the self-attention mechanism has not been demonstrated to be fully compatible with the spiking transmission process. The encoding (head) layer is composed of 2 Spiking Convolutional (Conv) layers and 1 LIF layer, which converts the floating input into spike sequences. The base latent channel dimension is 128 and the deepest dimension is 1024. Since the predicted noise of the diffusion process must be floating-point numbers, the conversion of discrete features to continuous features is necessary, so we adopt 2 Conv layers and a membrane potential layer [3] as our decoder. As for the spiking neuron (activation function) in the SNN-UNet, we use a special case of LIF: Integrate-and-Fire (IF [1]) model, where the decay rate of the neuron is 1.0. We choose $\mathcal{L}_{mse}$ for the training objective and use a batch size of 128 for the main experiments and our ablation study. The SNN-UNet was trained with a learning rate of 0.0002 using the Adam [4] optimizer. In addition, SDDPM does not use the EMA [7] algorithm in the training process. For fair comparisons, we re-evaluate the results of DDPM [2] using the same UNet architecture and the same training scheme as SDDPM. Our code is available at https://github.com/AndyCao1125/SDDPM.

4. Threshold Guidance on SDDPM

We tested more experimental demonstrations on threshold guidance in Tab. 2, including the results of CIFAR-10 and CelebA. The top-1 and top-2 results are bold and underlined, respectively. However, the order of magnitude regarding threshold adjustment still needs to be further explored.

Method	Threshold	FID↓	IS↑
Baseline	1.000	19.73	7.44
Inhibitory Guidance	0.999	<u>19.25</u>	<u>7.48</u>
	0.998	19.38	**7.55**
	0.997	**19.20**	7.47
	0.995	19.42	7.43
	0.990	19.77	7.45
Excitatory Guidance	1.001	20.00	7.47
	1.002	19.98	<u>7.48</u>
	1.003	20.04	7.46
	1.005	20.42	7.46
	1.010	21.57	7.37

Table 1. **More Results on CIFAR-10 by different threshold guidances.** Experiments are conducted by SDDPM (T=4).

Method	Threshold	FID↓
Baseline	1.000	25.09
Inhibitory Guidance	0.999	**24.69**
	0.998	<u>25.08</u>
	0.997	27.30
Excitatory Guidance	1.001	26.34
	1.002	28.25
	1.003	28.93

Table 2. **More Results on CelebA by different threshold guidances.** Experiments are conducted by SDDPM (T=4).

References

[1] Anthony N Burkitt. A review of the integrate-and-fire neuron model: I. homogeneous synaptic input. *Biological cybernetics*, 95:1–19, 2006. 2

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 2

[3] Hiromichi Kamata, Yusuke Mukuta, and Tatsuya Harada. Fully spiking variational autoencoder. In *AAAI*, volume 36, pages 7059–7067, 2022. 2

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[5] Priyadarshini Panda, Sai Aparna Aketi, and Kaushik Roy. Toward scalable, efficient, and accurate deep spiking neural networks with backward residual connections, stochastic softmax, and hybridization. *Frontiers in Neuroscience*, 14:653, 2020. 1

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*. Springer, 2015. 2

[7] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS*, 30, 2017. 2

[8] Man Yao, Guangshe Zhao, Hengyu Zhang, Yifan Hu, Lei Deng, Yonghong Tian, Bo Xu, and Guoqi Li. Attention spiking neural networks. *IEEE TPAMI*, 2023. 1