SpikeDiffusion: A Fully Spiking Structure-Guided Diffusion Framework for Energy-Efficient Image Generation

1st Xue Han
Guangdong Provincial/Zhuhai Key Laboratory of IRADS, and Department of Computer Science
Beijing Normal-Hong Kong Baptist University
Zhuhai, China
u430201701@mail.bnbu.edu.cn

2nd Zhiwen Luo
Concordia University
Montreal, Canada
zhiwen.luo@mail.concordia.ca

3rd Wenchuan Zhang
Guangdong Provincial/Zhuhai Key Laboratory of IRADS, and Department of Computer Science
Beijing Normal-Hong Kong Baptist University
Zhuhai, China
t330201704@mail.bnbu.edu.cn

4th Nizar Bouguila
Concordia University
Montreal, Canada
nizar.bouguila@concordia.ca

5th Weifeng Su
Guangdong Provincial/Zhuhai Key Laboratory of IRADS, and Department of Computer Science
Beijing Normal-Hong Kong Baptist University
Zhuhai, China
wfsu@bnbu.edu.cn

6th Wentao Fan*
Guangdong Provincial/Zhuhai Key Laboratory of IRADS, and Department of Computer Science
Beijing Normal-Hong Kong Baptist University
Zhuhai, China
wentaofan@bnbu.edu.cn

***Abstract*—Spiking Neural Networks (SNNs) offer energy-efficient and biologically plausible computation, yet their discrete spiking dynamics make high-fidelity image generation challenging, particularly in preserving global structure and fine-grained details. To address this challenge, we propose SpikeDiffusion, a fully spiking structure-guided diffusion framework for stable image generation. Unlike attention-based global interaction designs, SpikeDiffusion employs a spiking variational autoencoder (VAE) to produce a structure-preserving prior, which is generated once and reused throughout the denoising process via a lightweight early-fusion strategy, with the greatest benefit in the high-noise early stages. This design enhances structural consistency while avoiding the substantial computational and memory overhead of attention mechanisms. Furthermore, we develop a fully spiking U-Net with multi-step temporal dynamics and membrane-potential integration, enabling end-to-end spike-based diffusion modeling. Extensive experiments on standard benchmarks demonstrate that SpikeDiffusion achieves competitive image quality among SNN-based generative models with significantly reduced energy consumption, highlighting its suitability for neuromorphic and edge deployment. The code is available at https://github.com/hhx0511/spiking-diffusion.**

***Index Terms*—Spiking Neural Networks, Structure-Guided Diffusion, Image Generation, Energy-efficient.**

I. Introduction

Spiking Neural Networks (SNNs) have gained attention as an energy-efficient and biologically inspired computational paradigm, featuring sparse, event-driven communication and temporal dynamics [1], [2], [3], [4]. Unlike conventional artificial neural networks (ANNs), whose performance is typically bounded by peak computational throughput, the deployment of generative models in resource-constrained or neuromorphic settings is often limited by energy consumption and memory access [5], [6]. In this regard, SNNs offer a promising foundation for low-power generative modeling, as spike-based processing naturally reduces redundant computation and memory activity [7]. Nevertheless, achieving high-quality image generation remains challenging, since preserving fine-grained pixel fidelity and global structural coherence is inherently difficult under discrete and sparse spiking representations.

Recent years have seen growing interest in extending Spiking Neural Networks (SNNs) to high-level vision tasks, including image classification [8], [9], [10], object detection [11], [12], and generative modeling [13]. Among these tasks, image generation is particularly challenging, as fine-grained textures, smooth intensity transitions, and global structural coherence are inherently difficult to reconstruct from sparse, spike-based representations [14]. Existing SNN-based generative models, such as SNN-VAEs [15], SNN-GANs [16], and spiking diffusion models [17], have demonstrated the feasibility of spike-driven generation, yet often suffer from structural drift or limited capacity for detail representation. To alleviate these issues, some studies have explored attention-based or Transformer-based mechanisms to more effectively model global contextual information and improve feature fusion [18], [19]. While effective in enhancing global coherence, such designs often incur substantial computational and memory overhead by relying on dense interactions or hybrid non-spiking components, thereby undermining the core motivation for adopting SNNs in energy-efficient and hardware-friendly generative modeling.

Our key insight is that costly global interactions are not strictly required to maintain structural stability in spiking diffusion models. Rather than modeling long-range dependencies through attention mechanisms, we observe that stable low-frequency structural guidance is most beneficial under high-noise early denoising stages [20]. Such global structure can be captured by a generative prior, generated once before reverse diffusion, and reused throughout the denoising process via a lightweight early-fusion strategy. From this perspective, attention-based designs that rely on repeated global interac-

*Corresponding author.

tions can be replaced by a static low-frequency structural prior that better aligns with the event-driven and energy-efficient nature of spiking computation.

In this work, we propose SpikeDiffusion, a fully spiking structure-guided diffusion framework for structure-aware and energy-efficient image generation. Our main contributions are summarized as follows:

- We introduce a fully spiking structure-guided diffusion framework that enables stable and high-fidelity image generation under neuromorphic constraints.
- We design a low-frequency structural prior generator implemented with a spiking VAE, whose output is generated once and reused across all diffusion timesteps, providing structural guidance without iterative recomputation.
- We develop an efficient early-fusion mechanism and a fully spiking U-Net diffusion backbone with multi-step temporal dynamics and membrane potential integration, enabling end-to-end spike-based diffusion modeling during both training and inference.

Extensive experiments on standard benchmarks demonstrate that SpikeDiffusion achieves competitive or superior image quality compared to SNN-based generative models, while significantly reducing parameter count and energy consumption, making it suitable for neuromorphic and edge deployment.

II. Related Work

Recent studies have explored generative modeling with SNNs, including spiking VAEs [15], spiking GANs [16], and autoregressive spike-based generators [21]. These methods demonstrate the feasibility of spike-driven generation, but typically rely on one-shot sampling, which limits their ability to preserve global structure and fine-grained details on complex image distributions. Spiking diffusion models [17] introduce iterative denoising into the spiking domain, improving generation fidelity, yet existing approaches lack explicit mechanisms for providing stable global structural guidance during the diffusion process.

To enhance global context modeling in SNN-based generative models, several recent works introduce attention mechanisms or hybrid ANN-SNN architectures. TCJA-SNN [19] employs temporal-channel joint attention to improve feature integration, while Transformer-based spiking diffusion models [18] leverage attention to capture long-range dependencies. Although effective in certain settings, these designs often rely on dense interactions, repeated global computations, or non-spiking components, which compromise the energy efficiency and hardware friendliness of SNNs. Hybrid frameworks such as SGM-VaGAN [22] further introduce ANN-based discriminators to enhance representation learning. While these approaches improve visual quality, they deviate from a fully spiking generation pipeline and introduce additional training and deployment complexity. In contrast, SpikeDiffusion maintains an end-to-end spiking architecture and demonstrates that stable global structure can be achieved through a self-generated structural prior and diffusion-based refinement, without relying on attention mechanisms or ANN components.

III. Proposed Method

A. Overview of Our Model

As illustrated in Fig. 1, SpikeDiffusion is a two-stage framework for image generation, composed of: (1) a fully spiking VAE [15], which produces a coarse structural prior via prior sampling; and (2) a fully spiking U-Net diffusion model that refines this prior into a high-fidelity image through reverse diffusion. The VAE output $\hat{x}_0$ serves as a static structural prior, which we define as a coarse image representation that preserves global spatial layout and low-frequency content rather than fine-grained textures. Guided by $\hat{x}_0$, the diffusion model generates the final image from Gaussian noise, while reusing the same prior throughout all denoising timesteps.

B. Spiking Neural Networks

Among the various learning algorithms for SNNs, we adopt a state-of-the-art approach [23] that achieves top performance in recognition tasks. Our spiking neurons follow the Leaky Integrate-and-Fire (LIF) model [24], with numerical simulation based on Euler integration. The membrane potential at timestep t is updated as:

$$u_t = \tau_{\text{decay}} \cdot u_{t-1} + x_t, \tag{1}$$

where u_t denotes the membrane potential, x_t is the presynaptic input, and τ_{decay} is a fixed decay factor. A spike is emitted when u_t exceeds a predefined threshold V_{th} [25].

To model global image structure in the spiking domain, we employ a fully spiking generative module that produces low-frequency structural representations via latent sampling. This module models both the approximate posterior and the prior as autoregressive processes over binary spike events, following prior work on spiking VAE [15]. The factorized distributions are expressed as:

$$q(z_{1:T} \mid x_{1:T}) = \prod_{t=1}^{T} q(z_t \mid x_{\le t}, z_{<t}), \quad p(z_{1:T}) = \prod_{t=1}^{T} p(z_t \mid z_{<t}), \tag{2}$$

where both $q(\cdot)$ and $p(\cdot)$ are implemented by spiking neural networks producing binary outputs.

At each timestep t, the approximate posterior receives both encoder features and previously sampled spikes, whereas the prior depends only on past spike activity:

$$\zeta_{q,t} = \mathcal{F}_q(z_{q,t-1}, x_t^{(\text{enc})}), \quad \zeta_{p,t} = \mathcal{F}_p(z_{p,t-1}), \tag{3}$$

where $\mathcal{F}_q$ and $\mathcal{F}_p$ denote spiking neural networks parameterizing the posterior and prior, respectively.

The binary latent variable $z_{t,c} \in \{0, 1\}$ is sampled from a Bernoulli distribution:

$$z_{t,c} \sim \text{Ber}(\pi_{t,c}), \quad \pi_{t,c} = \text{mean}(\zeta_t[k(c-1) : kc]), \tag{4}$$

where $\pi_{t,c}$ is the average of k binary spikes within ζ_t, providing a probabilistic approximation for Bernoulli sampling.

The structural prior generator is instantiated as a spiking VAE and pretrained independently using a reconstruction loss and a Kullback–Leibler (KL) regularization term before diffusion training. Its parameters are then frozen, and the generated

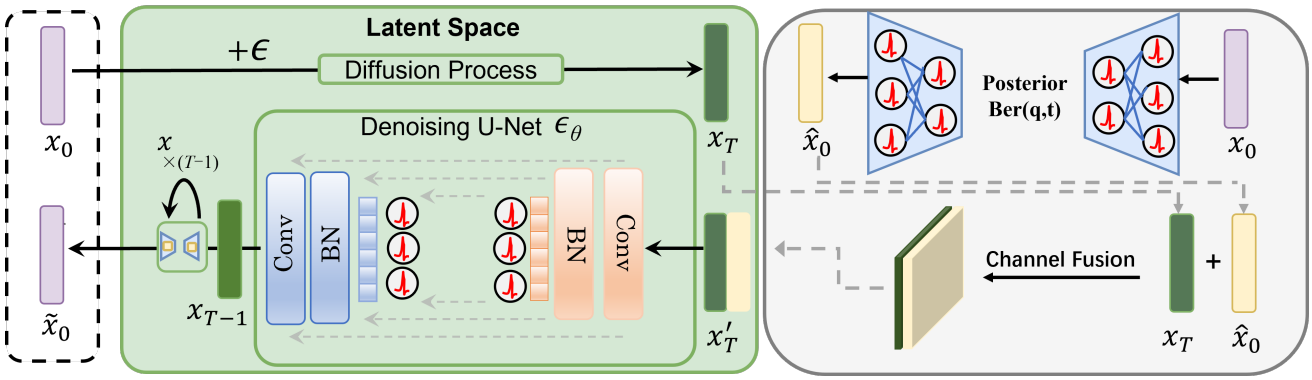

Fig. 1: Overview of the proposed SpikeDiffusion framework. A structure-preserving low-frequency prior $\hat{x}_0$ is generated once and reused throughout the denoising process via channel-wise early fusion to guide denoising. Key variables: x_0, original image; x_T and x_{T-1}, latent variables at adjacent diffusion steps; x'_T, fused latent representation; ϵ, Gaussian noise; ϵ_θ, noise predicted by the spiking U-Net; $\mathrm{Ber}(q,t)$, Bernoulli posterior at timestep t.

prior $\hat{x}_0$ is reused as a fixed structural guidance signal throughout diffusion training and sampling. Non-differentiable spikes are handled using surrogate-gradient-based backpropagation, following standard direct training in SNNs.

C. Spiking Diffusion Model

Diffusion models generate samples by reversing a forward noising process with a parameterized denoising network [26], [27]. Given a noise schedule $\{\beta_t\}$, we define $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. The forward diffusion process corrupts the clean image x_0 as:

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{5}$$

The denoising network ϵ_θ is trained to predict the injected noise ϵ at each diffusion timestep by minimizing:

$$\mathbb{E}_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right]. \tag{6}$$

In SpikeDiffusion, the denoising network ϵ_θ is instantiated as a fully spiking U-Net, enabling diffusion-based generation under neuromorphic constraints. While diffusion operates over discrete timesteps t, the spiking U-Net processes each noisy input x_t over an additional set of spiking timesteps $\tau = 1, \ldots, T_s$. This introduces a two-level temporal structure, where membrane potential dynamics integrate information across spiking steps to enhance denoising stability under discrete spike-based computation.

All nonlinear activations are implemented using spiking neurons, whereas synaptic operations such as convolution and normalization produce real-valued feature maps. The spiking U-Net outputs a sequence of intermediate representations over T_s spiking steps, which are decoded into a single real-valued noise prediction via temporal membrane integration, as detailed in the following paragraph.

During training, the diffusion model is optimized jointly with structural guidance provided by a static low-frequency prior, as summarized in Algorithm 1.

Algorithm 1 SpikeDiffusion Training

repeat
  Sample clean image $x_0 \sim q(x_0)$
  $\hat{x}_0 \leftarrow \mathrm{PriorGen}(x_0)$ ▷ Generate a low-frequency structural prior (instantiated with a spiking VAE)
  Sample timestep $t \sim \mathrm{Uniform}(\{1, \ldots, T\})$
  Sample noise $\epsilon \sim \mathcal{N}(0, I)$
  $x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon$
  $x_t^{\text{input}} = \mathrm{Concat}(x_t, \hat{x}_0)$
  Take gradient step on
    $\|\epsilon - \epsilon_\theta(x_t^{\text{input}}, t)\|_2^2$
until convergence

D. Structural Guidance via Early Fusion

To provide reliable structural guidance under high noise levels, we guide the diffusion process using a coarse, low-frequency structural prior $\hat{x}_0$ generated by a fully spiking structural prior generator. Specifically, $\hat{x}_0$ is generated once before reverse diffusion and remains fixed throughout the denoising process. At each reverse diffusion step, the noisy latent variable x_t is concatenated with the static structural prior $\hat{x}_0$ along the channel dimension:

$$x_t^{\text{input}} = \mathrm{Concat}(x_t, \hat{x}_0), \tag{7}$$

where $\mathrm{Concat}(\cdot)$ denotes channel-wise concatenation. This lightweight early-fusion strategy provides structural guidance throughout the denoising process, with the strongest effect in the high-noise early stages, thereby stabilizing global structure without relying on attention-based global interactions.

At inference time, $\hat{x}_0$ is generated by sampling from the learned structural prior and reused across all denoising timesteps. The complete sampling procedure is described in Algorithm 2, where $c_1(t)$ and $c_2(t)$ follow the standard DDPM posterior [28].

Algorithm 2 SpikeDiffusion Sampling

Sample latent spikes $z \sim p(z)$ from the learned structural prior
$\hat{x}_0 \leftarrow \mathrm{PriorGenDec}(z)$ ▷ Generate coarse structural prior
Sample noise $x_T \sim \mathcal{N}(0, I)$
for $t = T$ to 1 **do**
  $x_t^{\text{input}} = \mathrm{Concat}(x_t, \hat{x}_0)$
  $\hat{\epsilon} \leftarrow \epsilon_\theta(x_t^{\text{input}}, t)$
  $\tilde{x}_0 \leftarrow \frac{1}{\sqrt{\bar{\alpha}_t}} \left(x_t - \sqrt{1-\bar{\alpha}_t}\,\hat{\epsilon}\right)$
  $\mu_\theta \leftarrow c_1(t)\tilde{x}_0 + c_2(t)x_t$ ▷ standard diffusion posterior
  Sample $\xi \sim \mathcal{N}(0, I)$
  $x_{t-1} \leftarrow \mu_\theta + \sigma_t \xi$ ▷ σ_t^2 is the posterior variance
end for
return x_0

IV. Experiments

A. Datasets and Training Settings

We evaluate SpikeDiffusion on three widely used image generation benchmarks: MNIST and Fashion-MNIST (28×28 grayscale images with 60k training and 10k test samples), and CIFAR-10 (32×32 RGB images with 50k training and 10k test samples).

SpikeDiffusion consists of a structural prior generator and a spiking diffusion U-Net. The structural prior generator is instantiated using a spiking VAE, with a channel multiplier $k = 20$, a latent dimension of 128, and temporal integration over 16 spiking timesteps to capture low-frequency global structure. The spiking diffusion U-Net adopts a base channel size of 128, channel multipliers $[1, 2, 2, 4]$, and two residual blocks per resolution level. During each denoising stage, the spiking U-Net operates over 4 spiking timesteps to model membrane potential dynamics and temporal accumulation. For training, the structural prior generator is pretrained using the AdamW optimizer with a learning rate of 1×10^{-3}, a batch size of 250, and 150 training epochs. The spiking diffusion model is trained using the Adam optimizer with a learning rate of 1×10^{-5}, a batch size of 32, gradient clipping at 1.0, and an exponential moving average (EMA) decay rate of 0.9999. All reported results are averaged over 10 independent runs with different random seeds.

B. Comparison with Representative SNN Generative Models

To assess the generative capability of SpikeDiffusion, we compare it with a set of baseline models covering both ANN and SNN-based generative paradigms. We report the Inception Score (IS) [29] and Fréchet Inception Distance (FID) [30], which evaluate sample diversity/classifiability and distribution-level fidelity, respectively. The compared methods include: (1) **VAE (ANN)** [15], a conventional non-spiking neural network baseline; (2) **Spike-GAN** [16], a GAN-style SNN generator-discriminator framework based on LIF neurons; (3) **TCJA-SNN** [19], an attention-enhanced SNN design that introduces temporal-channel joint attention for feature integration; (4) **SGM-VaGAN** [22], a hybrid SNN-ANN framework that combines a spiking generator with an ANN discriminator.

Table I reports IS and FID on MNIST, Fashion-MNIST, and CIFAR-10. Overall, SpikeDiffusion achieves consistently superior FID across all datasets, indicating a substantially closer match between generated and real image distributions compared to both ANN and SNN-based baselines. On MNIST and Fashion-MNIST, SpikeDiffusion also attains the highest IS, reflecting improved sample diversity and classifiability. On the more challenging CIFAR-10 dataset, SpikeDiffusion achieves a dramatic reduction in FID compared to prior SNN-based methods (e.g., from 85.34 to 9.51 relative to Spike-GAN), highlighting its ability to model complex image distributions with rich textures and global structure. While attention-enhanced or hybrid frameworks such as TCJA-SNN and SGM-VaGAN improve certain aspects of generation, SpikeDiffusion maintains a clear advantage in distribution-level fidelity, demonstrating that high-quality spiking-based image generation can be achieved without relying on heavy attention mechanisms or ANN discriminators.

TABLE I: Main results under consistent training and evaluation settings. IS (↑) and FID (↓) are reported on MNIST, Fashion-MNIST, and CIFAR-10.

Model	MNIST		Fashion-MNIST		CIFAR-10	
	IS ↑	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓
VAE (ANN) [15]	5.947	112.5	4.252	123.7	2.591	229.6
Spike-GAN [16]	6.512	65.34	5.637	67.87	2.968	85.34
TCJA-SNN [19]	6.45	100.8	5.6	93.41	3.73	170.1
SGM-VaGAN [22]	6.712	58.23	4.849	70.78	**3.008**	74.72
SpikeDiffusion (Ours)	**7.440**	**37.90**	**6.830**	**33.57**	2.970	**9.51**

TABLE II: Comparison of model parameters, energy consumption, and architecture type.

Model	#Parameters (M)	Energy / Image (mJ)	Fully SNN
VAE (ANN) [15]	1.13	40.7	✗
Spike-GAN [16]	0.67	1.63	✗
TCJA-SNN [19]	1.83	1.9	✗
SGM-VaGAN [22]	4.1	0.72	✗
SpikeDiffusion (Ours)	**0.26**	**0.27**	✓

Table II compares model size, energy consumption per generated image, and architectural properties of different generative models on MNIST under a consistent evaluation protocol. The reported energy values are operation-level analytical estimates of inference-time compute cost rather than full hardware-level measurements. ANN computation is modeled using 32-bit MAC operations, while SNN computation is modeled using spike-driven accumulations. Following commonly used measurements under 45 nm CMOS technology, we use 4.6 pJ for a 32-bit ANN MAC operation and 0.9 pJ for a spike-driven SNN accumulation [31]. Under this setting, SpikeDiffusion achieves both the lowest parameter count and the lowest estimated energy consumption per image while maintaining a fully spiking generation pipeline. We note that this estimate does not include memory traffic, data movement, hardware control overhead, or training cost, and that the total inference cost also depends on the number of diffusion sampling steps and spiking timesteps.

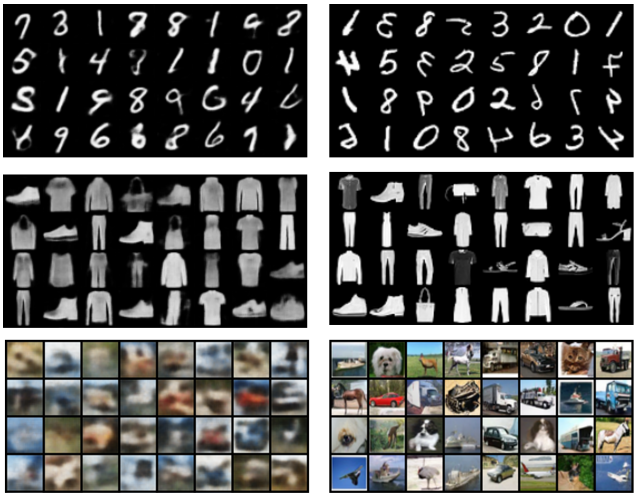

Fig. 2: Qualitative comparison between **SpikeDiffusion** and **Spiking VAE** on MNIST, Fashion-MNIST, and CIFAR-10.

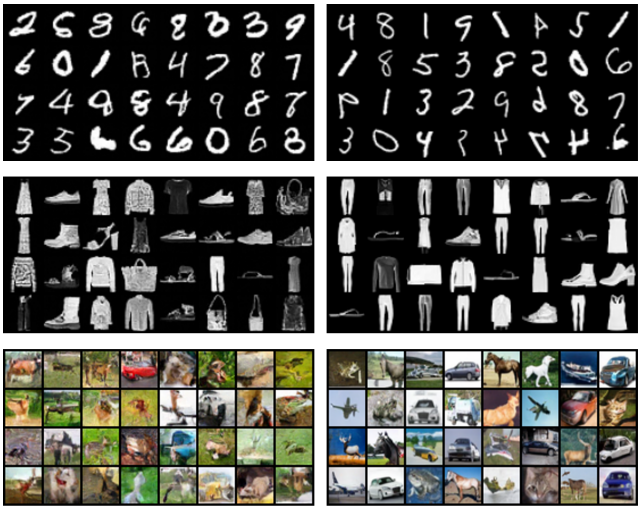

Fig. 3: Qualitative comparison between **SpikeDiffusion** and a diffusion-based SNN baseline without structural guidance.

C. Ablation Analysis

We conduct an ablation study to examine the roles of structural modeling and diffusion-based refinement in SpikeDiffusion, with results summarized in Table III. The spiking VAE alone performs poorly on complex data, as reflected by its high CIFAR-10 FID. Adding diffusion-based refinement reduces the CIFAR-10 FID from 175.5 to 66.27. Further incorporating structural priors lowers it to 9.51, yielding the strongest performance on the more complex datasets while remaining competitive on MNIST.

Qualitative comparisons in Fig. 2 and Fig. 3 show that SpikeDiffusion produces clearer structures and suffers less structural drift than the baselines, especially on complex scenes. These results suggest that structural guidance is most beneficial for more complex data distributions.

TABLE III: IS (↑) and FID (↓) on MNIST, Fashion-MNIST, and CIFAR-10. Higher IS and lower FID are better.

Model	**MNIST**		**Fashion-MNIST**		**CIFAR-10**	
	IS ↑	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓
Spiking VAE	6.209	97.06	4.551	90.12	2.945	175.5
Diffusion baseline	6.510	**35.10**	5.930	65.18	2.840	66.27
SpikeDiffusion (Ours)	**7.440**	37.90	**6.830**	**33.57**	**2.970**	**9.510**

We further study the effect of the number of spiking timesteps T_s in the spiking U-Net. Increasing T_s enhances temporal integration and improves generation quality, but also leads to higher computational and energy cost. When T_s is too small, limited temporal dynamics degrade denoising stability, while larger T_s yields only marginal fidelity gains with substantially increased energy consumption. As shown in Table IV, $T_s = 4$ provides the best trade-off between generation fidelity and energy efficiency in our setting.

TABLE IV: Effect of spiking timesteps T_s on FID and energy consumption on CIFAR-10.

Model	**FID ↓**	**Energy (mJ) ↓**
SpikeDiffusion ($T_s = 4$)	9.51	**0.27**
SpikeDiffusion ($T_s = 8$)	8.92	0.81
SpikeDiffusion ($T_s = 16$)	**8.74**	1.98

V. Conclusion

We proposed SpikeDiffusion, a fully spiking structure-guided diffusion framework for energy-efficient image generation. By generating a low-frequency structural prior once and reusing it throughout denoising via early fusion, SpikeDiffusion stabilizes global structure, especially in high-noise early stages, while progressively recovering fine-grained details. Experiments on multiple benchmarks show that SpikeDiffusion achieves competitive or superior image quality among SNN-based generative models with lower energy consumption and parameter count. Ablation studies further show that neither structural modeling nor diffusion-based refinement alone is sufficient, and that their combination is critical for high-fidelity spiking image generation. This work highlights the potential of diffusion-based frameworks for neuromorphic computing. Future work will explore more efficient sampling, improved temporal modeling, and evaluation on higher-resolution and more complex datasets.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (62276106, 62576143), the Guangdong Basic and Applied Basic Research Foundation (2024A1515011767), and the Guangdong Provincial Key Laboratory of IRADS (2022B1212010006).

REFERENCES

[1] G. Liu, W. Deng, X. Xie, L. Huang, and H. Tang, "Human-level control through directly trained deep spiking q-networks," *IEEE Transactions on Cybernetics*, vol. 53, no. 11, pp. 7187–7198, 2022.

[2] W. Maass, "Networks of spiking neurons: the third generation of neural network models," *Neural Networks*, vol. 10, no. 9, pp. 1659–1671, 1997.

[3] K. Roy, A. Jaiswal, and P. Panda, "Towards spike-based machine intelligence with neuromorphic computing," *Nature*, vol. 575, no. 7784, pp. 607–617, 2019.

[4] G. Li, L. Deng, and W. R. Maass, "Brain-inspired computing: a systematic survey and future trends," *Proceedings of the IEEE*, vol. 112, no. 6, pp. 544–584, 2024.

[5] Z. Luo, W. Fan, M. Amayri, and N. Bouguila, "Dynamic deep clustering of high-dimensional directional data via hyperspherical embeddings with bayesian nonparametric mixtures," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.1, KDD 2025, Toronto, ON, Canada, August 3-7, 2025*. ACM, 2025, pp. 938–949.

[6] X. Han, Z. Li, W. Zhang, H. Huang, and W. Fan, "Robust generalized zero-shot learning via dual-stream variational autoencoders and out-of-distribution detection," in *IEEE International Conference on Multimedia and Expo, ICME 2025, Nantes, France, June 30 - July 4, 2025*. IEEE, 2025, pp. 1–6.

[7] J. Zhang, W. Fan, and X. Liu, "Spiking generative networks in lifelong learning environment," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, 2023, pp. 353–364.

[8] S. Deng, Y. Li, S. Zhang, and S. Gu, "Temporal efficient training of spiking neural network via gradient re-weighting," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

[9] Z. Wang, Y. Fang, J. Cao, Q. Zhang, Z. Wang, and R. Xu, "Masked spiking transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 1761–1771.

[10] Z. Zhou, Y. Zhu, C. He, Y. Wang, S. Yan, Y. Tian, and L. Yuan, "Spikformer: When spiking neural network meets transformer," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[11] P. Kirkland, G. D. Di Caterina, J. Soraghan, and G. Matich, "Spikeseg: Spiking segmentation via stdp saliency mapping," in *Proceedings of the IJCNN*, 2020, pp. 1–8.

[12] D. Hareb and J. Martinet, "Evsegsnn: Neuromorphic semantic segmentation for event data," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2024, pp. 1–8.

[13] J. Zhang, W. Fan, and X. Liu, "Spiking generative networks empowered by multiple dynamic experts for lifelong learning," *Expert Systems with Applications*, vol. 238, p. 121845, 2024.

[14] Q. Zhan, R. Tao, X. Xie, G. Liu, M. Zhang, H. Tang, and Y. Yang, "Es-vae: An efficient spiking variational autoencoder with reparameterizable poisson spiking sampling," *arXiv preprint arXiv:2310.14839*, 2023.

[15] H. Kamata, Y. Mukuta, and T. Harada, "Fully spiking variational autoencoder," in *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, 2022, pp. 7059–7067.

[16] V. Kotariya and U. Ganguly, "Spiking-gan: A spiking generative adversarial network using time-to-first-spike coding," in *Proceedings of the IJCNN*, 2022, pp. 1–7.

[17] J. Cao, Z. Wang, H. Guo, H. Cheng, Q. Zhang, and R. Xu, "Spiking denoising diffusion probabilistic models," in *Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 4912–4921.

[18] S. Yang, H. Ma, C. Yu, A. Wang, and E. P. Li, "Spiking diffusion transformer (sdit): A transformer-based architecture for spiking diffusion models," *arXiv preprint arXiv:2402.11588*, 2024.

[19] R. J. Zhu, M. Zhang, Q. Zhao, H. Deng, Y. Duan, and L. J. Deng, "Tcja-snn: Temporal-channel joint attention for spiking neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[20] Y. Qian, Q. Cai, Y. Pan, Y. Li, T. Yao, Q. Sun, and T. Mei, "Boosting diffusion models with moving average sampling in frequency domain," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 8911–8920.

[21] D. M. Arribas, Y. Zhao, and I. M. Park, "Rescuing neural spike train models from bad MLE," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1–12.

[22] W. Zhang, J. Zhang, R. Y. T. Hou, W. Su, and W. Fan, "Spiking generative models based on variational autoencoder and adversarial training," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2025.

[23] H. Zheng, Y. Wu, L. Deng, Y. Hu, and G. Li, "Going deeper with directly-trained larger spiking neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 11 062–11 070.

[24] Y. Wu, L. Deng, G. Li, J. Zhu, Y. Xie, and L. Shi, "Direct training for spiking neural networks: Faster, larger, better," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 1311–1318.

[25] J. Zhang, W. Fan, and X. Liu, "An ann-guided approach to task-free continual learning with spiking neural networks," in *Chinese Conference on Pattern Recognition and Computer Vision*, 2023, pp. 217–228.

[26] P. Dhariwal and A. Q. Nichol, "Diffusion models beat gans on image synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 8780–8794.

[27] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *Advances in Neural Information Processing Systems*, vol. 35, 2022.

[28] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[29] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Advances in Neural Information Processing Systems*, vol. 29, 2016, pp. 2226–2234.

[30] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 6626–6637.

[31] M. Horowitz, "Computing's energy problem (and what we can do about it)," in *Proceedings of the 2014 IEEE International Solid-State Circuits Conference (ISSCC)*. IEEE, 2014, pp. 10–14.