Improving Spiking Transformer Via Quantized Spiking Neurons with Shunting Inhibition

Yunhua Chen[1], Chukuo Qu[1,*], Zequan Xie[1,*], Jinsheng Xiao[2,†]
[1]School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China
[2]School of Electronic Information, Wuhan University, Wuhan, China

***Abstract*—Spiking Transformers show promise in bridging the accuracy gap between SNNs and ANNs in complex visual tasks, yet their training efficiency is constrained by the high computational cost of time-step simulations. Quantized spiking neurons, which replace spiking activations with integer activations, improve training efficiency but overlook the critical temporal dynamics and excitatory-inhibitory mechanisms of biological neurons. This omission hinders the model's ability to capture deep spatio-temporal dependencies. To address this, we designed Quantized Inhibition Leaky Integrate and Fire (QI-LIF). By introducing learnable leakage factors and excitatory-inhibitory mechanisms, QI-LIF simulates neural temporal dynamics and refractory periods, preventing overexcitability under continuous stimulation to better capture deep spatio-temporal dependencies. Furthermore, to reduce the computational complexity of self-attention, we propose a Texture Saliency-based dynamic gated Spiking Attention (TSSA). This approach utilizes feature variance across channels as a texture saliency metric to construct dynamic gated masks. These masks block low-variance background noise while allowing high-frequency key features to pass through. Our approach achieves competitive accuracy across multiple datasets. In particular, on the CIFAR-100 dataset, our method achieves an accuracy of 80.92% using only 4 time steps, representing a 3.4% improvement over baseline.**

***Index Terms*—Spiking neural network, Quantized Inhibition neurons, Texture Saliency.**

I. Introduction

Spiking Neural Networks (SNNs) [1]–[3] transmit and process information through biologically plausible spiking signals, achieving exceptional energy efficiency. The Transformer architecture [4], built upon a biologically inspired self-attention module, initially excelled in Natural Language Processing and has since successfully expanded into Computer Vision [5]–[8]. It has established dominance in tasks such as image classification, object detection, and semantic segmentation. The integration of the Transformer architecture into SNNs (i.e., Spiking Transformers) [9] aims to synergize the powerful feature extraction capabilities of Transformers with the inherent low-power nature of SNNs. This approach significantly enhances model performance while preserving high computational energy efficiency, progressively bridging the accuracy gap between SNNs and ANNs in complex visual tasks [10], [31], [33].

However, the training efficiency of Spiking Transformers remains bottlenecked by the high computational cost of multi-step simulations.

Recent studies such as E-SpikeFormer [11] and Spike2Former [12] attempt to alleviate this bottleneck by employing quantized neurons. Their core strategy involves substituting binary spiking activations with integer activations to accelerate training [13], [14]. While this approach yields significant gains in computational efficiency, it neglects critical temporal dynamics and the excitatory-inhibitory mechanisms of biological neurons. Specifically, when activated by strong features, biological neurons temporarily reduce their sensitivity to subsequent weak signals by raising thresholds or activating inhibitory conductances. E-SpikeFormer reduces the dynamic membrane potential accumulation and firing process to a static integer mapping, thereby severing the intrinsic temporal coupling between time steps and disregarding post-spiking inhibitory conductance. This lack of biological fidelity prevents the model from capturing deep spatiotemporal dependencies [15].

To address this, we introduce the Quantized Inhibitory LIF (QI-LIF) neuron. Building upon existing quantized neuron models, QI-LIF integrates a learnable leak factor and an excitatory-inhibitory mechanism. This model simulates the temporal dynamics of biological neuronal membrane potentials and mimics post-spiking inhibitory conductance by suppressing neurons that fire at high frequencies. While retaining the training efficiency of quantization, QI-LIF dynamically adjusts the membrane potential based on its discharge history, preventing neuronal overexcitation under continuous stimulation. To mitigate computational redundancy in self-attention, we propose the S2Iformer, whose core component is the Texture Saliency-based Dynamic Gated Spiking Attention (TSSA). Unlike traditional attention mechanisms that compute global similarity, TSSA constructs an efficient "spatial filter." It utilizes feature variance across the channel dimension as a proxy for texture saliency to construct a dynamic gated mask. Before signals enter neurons, TSSA proactively filters out low-variance background noise, allowing only high-frequency, information-rich key features to pass through [32]. Through this "input purification," TSSA effectively reduces the input

This work was supported by the Natural Science Foundation of Guangdong Province, China (No.2025A1515012243) Guangdong ST programme (No.2026A1010020014).

†Corresponding author: xiaojs@whu.edu.cn
*Equal Contribution

density to QI-LIF neurons, ensuring that their dynamic inhibition mechanism activates only when processing critical spatio-temporal features.

The main contributions of this paper are summarized as follows:

- We propose Quantized Inhibitory LIF (QI-LIF) neurons to address the limitations of discrete quantization. By incorporating learnable leak factors and excitatory-inhibitory mechanisms, this model simulates biological potential decay to prevent overexcitation, thereby enabling the network to capture deep spatio-temporal dependencies.
- We design a Texture Saliency-based Dynamic Gated Spiking Attention (TSSA) module with linear computational complexity. By leveraging feature channel variance to quantify texture saliency, TSSA acts as a spatial filter that suppresses low-variance background noise before it reaches the QI-LIF neurons.

II. Preliminary

In this paper, we employ the widely used Leaky Integrate-and-Fire (LIF) model [16], which can be described by the following mathematical formula:

$$\mathbf{U}[t] = \beta \mathbf{H}[t-1] + \mathbf{X}[t], \tag{1}$$

$$\mathbf{S}[t] = \text{Hea}(\mathbf{U}[t] - V_{\text{th}}), \tag{2}$$

$$\mathbf{H}[t] = \begin{cases} \mathbf{U}[t](1-\mathbf{S}[t]) + V_{\text{reset}}\mathbf{S}[t], & \text{hard reset} \\ \mathbf{U}[t] - V_{\text{th}}\mathbf{S}[t], & \text{soft reset} \end{cases} \tag{3}$$

Where t is the time step, $\mathbf{U}[t]$ represents the membrane potential, spatial input $\mathbf{X}[t]$ is extracted from the raw spike sequence via Conv or Linear, temporal input $\beta\mathbf{H}[t-1]$ originates from the decay of the membrane potential at the previous time step, β is the leakage factor, $\mathbf{S}[t]$ is the output spike, $\mathbf{H}[t]$ is the post-firing membrane potential, $\text{Hea}(\cdot)$ denotes the Heaviside function where $\text{Hea}(x) = 1$ if $x \geq 0$. When U[t] exceeds the threshold V_{th}, the spiking neuron fires a spike, and the membrane potential is reset to V_{reset}.

III. Method

A. Overall Architecture

The proposed holistic S2Iformer is illustrated in Figure 1(a). It adopts a hierarchical design comprising three stages. The input data is represented as $\mathrm{X}_{in} \in \mathbb{R}^{\mathrm{T}\times\mathrm{C_{in}}\times\mathrm{H}\times\mathrm{W}}$, where T denotes the time step and C_{in} represents the number of input channels (e.g., 3 for static images). In the first stage, the input is processed using the Patch Embedding module. This module employs a 4×4 convolutional kernel with stride $= 4$ to convert the image into a spike sequence, mapping features to the initial channel dimension C. This process achieves initial downsampling, reducing the token count to $\frac{\mathrm{H}}{4} \times \frac{\mathrm{W}}{4}$. Subsequently, the feature sequence is fed into stacked encoder blocks for spatio-temporal feature extraction via TSSA and Spiking MLP [10]. TSSA efficiently filters salient texture features using variance gating and processes high-resolution feature maps with O(N) linear complexity. The second stage introduces a DownSampling module to generate multi-scale feature representations. Before proceeding to the next stage, this module halves the spatial resolution while doubling the channel dimension. The third stage employs SSA(Spiking Self-Attention) to compute global correlations on low-resolution feature maps.

To ensure training stability and temporal dynamics in deep SNNs, we employ QI-LIF neurons as the spiking activation function across all modules (including TSSA, SSA, and MLP). The computational flow of the l_{th} encoder block can be uniformly represented as follows, depending on the current stage:

$$X_l' = \text{Attention}(\text{QILIF}(\text{BN}(X_{l-1}))) + X_{l-1}, \tag{4}$$

$$X_l = \text{MLP}(\text{QILIF}(\text{BN}(X_l'))) + X_l'. \tag{5}$$

where $\text{BN}(\cdot)$ denotes batch normalization. The $\text{Attention}(\cdot)$ function is $\text{TSSA}(\cdot)$ in Stage 1 and Stage 2, and $\text{SSA}(\cdot)$ in Stage 3. Finally, after feature extraction across these three stages, the output features pass through a classification head to produce the final prediction result.

B. Quantized Inhibition Leaky Integrate and Fire

Existing SNN quantization methods [11], [12] primarily focus on accelerating training convergence by reducing quantization error through the introduction of integer activation. However, these approaches often treat membrane potential as a static numerical mapping, overlooking two critical temporal dynamics of biological neurons: Temporal Leakage [17] and Dynamic Inhibition. The leakage mechanism describes the natural decay of membrane potential in the absence of input, preventing information overload. The inhibition mechanism (e.g., refractory period), determined by the neuron's firing history, prevents overexcitability [18]. To address this limitation, we propose a quantization model for inhibitory LIF neurons, integrating LIF's leakage mechanism and neurobiological inhibitory mechanisms into the quantization process. As shown in Figure 1(c), membrane potentials enter the neuron, undergo conditional leakage and inhibitory mechanisms, and are ultimately quantized into integer values.

*1) **Forward**:* First, to simulate the natural decay process of biological neuronal membrane potentials, we introduced a conditional leakage mechanism. The evolution of the input feature (membrane potential) $x[t]$ is described as follows:

$$x_{\text{leak}}[t] = \begin{cases} x[t] \cdot L & \text{if } x[t] \geqslant V_{\text{th}}, \\ x[t] & \text{else}, \end{cases} \tag{6}$$

where, $V_{\text{th}} = V_{\text{max}}/2$ represents the set leakage activation threshold, while L denotes the learnable leakage factor (initially set to 0.9). In Figure 1(c), with Vth set to 2, inputs 2.6 and 3.7 will leak, while 1.3 remains unchanged.

Second, to simulate the refractory period and excitatory inhibition mechanisms of biological neurons following spiking firing, we designed a dynamic inhibition module. When

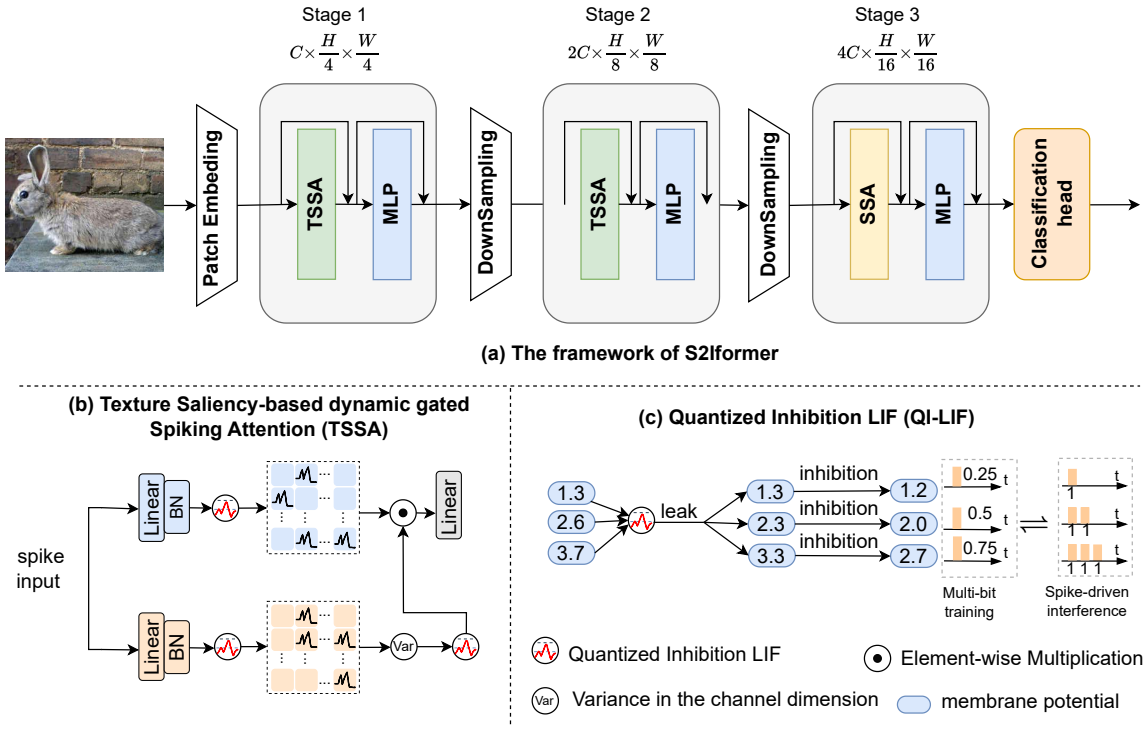

Fig. 1: An overview of S2Iformer. We proposed the TSSA module, which achieves global modeling of $O(N)$ linear complexity by Utilizing the variance of features across channel dimensions as a metric for texture saliency,retaining only key features rich in high-frequency information.

neurons receive strong stimuli, they generate inhibitory conductance that reduces sensitivity to subsequent inputs. We first estimate the current latent firing rate using a sigmoid function:

$$r_{\text{fire}}[t] = \sigma(x[t]/V_{\max}), \tag{7}$$

Subsequently, the dynamic suppression factor is calculated and applied to the membrane potential following leakage:

$$\gamma_{\text{inh}}[t] = 1.0 - \alpha \cdot r_{\text{fire}}[t], \tag{8}$$

$$x_{\text{inh}}[t] = x_{\text{leak}}[t] \cdot \gamma_{\text{inh}}[t], \tag{9}$$

where α is the inhibition strength parameter. Equation (9) implements a multiplicative gating mechanism: a higher potential firing rate r_{fire} results in a smaller inhibition factor γ_{inh}, which exerts a stronger downward pressure on the membrane potential. In Figure 1(c), the membrane potential of 3.3 exhibits a higher potential firing rate compared to 1.3 and 2.3, thus demonstrating the strongest inhibitory effect. This mechanism effectively prevents neuronal over-excitation when confronted with high-frequency noise inputs.

Finally, the membrane potential, after leakage and suppression adjustments, is quantized into integer values and output as activation values:

$$x_{\text{quant}}[t] = \text{Round}\big(\text{clip}\{x_{\text{inh}}[t], V_{\min}, V_{\max}\}\big) \tag{10}$$

where clip$\{\cdot\}$ operation constrains values within $[V_{\min}, V_{\max}]$, while Round$(\cdot)$ performs rounding. During training, QI-LIF passes integer $x_{\text{quant}}[t]$ to preserve precision and enable backpropagation; During inference, to preserve the SNN's spike-driven nature, we convert the integer $x_{\text{quant}}[t]$ into a binary spike sequence spanning $V_{\max}$ virtual time steps. Specifically, if $x_{\text{quant}}[t] = k$, the neuron fires spike for k virtual time steps and remains inactive during the remaining steps.

Through this approach, QI-LIF ensures that the MAC in the inference process can be fully converted into sparse AC. While maintaining low-power advantages, it significantly enhances the spatio-temporal information representation capability and biological plausibility of the quantized neuron model by introducing LIF leakage mechanisms and bio-inspired inhibition mechanisms.

*2) **Backward Gradient**:* For equation (10), since the gradient of the Round$(\cdot)$ function is nearly zero everywhere (non-differentiable), we employ the Straight-Through Estimator (STE) [19] during backpropagation:

$$\frac{\partial x_{\text{quant}}}{\partial x_{\text{inh}}} \approx \begin{cases} 1 & \text{if } V_{\min} \le x \le V_{\max}, \\ 0 & \text{otherwise} \end{cases} = M_{\text{clamp}} \tag{11}$$

where M_{clamp} is the clamping mask, which equals 1 if $x \in [V_{\min}, V_{\max}]$ and 0 otherwise.

Since γ_{inh} itself also depends on x, but in practice, to stabilize training, γ_{inh} is treated as a constant coefficient. From

equation (9), we have:

$$\frac{\partial x_{\text{inh}}}{\partial x_{\text{leak}}} = \gamma_{\text{inh}} \tag{12}$$

From equation (6), we have:

$$\frac{\partial x_{\text{leak}}}{\partial x} = M_{\text{leak}} = \begin{cases} L & \text{if } x \geqslant V_{\text{th}}, \\ 1 & \text{else} \end{cases} \tag{13}$$

According to the chain rule:

$$\frac{\partial \mathcal{L}}{\partial x} = \frac{\partial \mathcal{L}}{\partial x_{\text{quant}}} \cdot \frac{\partial x_{\text{quant}}}{\partial x_{\text{inh}}} \cdot \frac{\partial x_{\text{inh}}}{\partial x_{\text{leak}}} \cdot \frac{\partial x_{\text{leak}}}{\partial x} \tag{14}$$

Substituting the above formula into Formula 9 yields:

$$\frac{\partial \mathcal{L}}{\partial x} = \frac{\partial \mathcal{L}}{\partial x_{\text{quant}}} \cdot M_{\text{clamp}} \cdot \gamma_{\text{inh}} \cdot M_{\text{leak}} \tag{15}$$

C. *Texture Saliency-based dynamic gated Spiking Attention*

We propose Texture Saliency-based dynamic gated Spiking Attention (TSSA), whose detailed architecture is illustrated in Figure 1(b). The core idea of TSSA lies in its approach: instead of computing pairwise relationships between tokens, it measures the texture saliency of each token through variance [20] and employs a spiking gated mechanism for spatial diversion. For a given input spiking sequence $X \in \mathbb{R}^{N \times C}$ (where N is the number of tokens and C is the number of channels), we first generate the spiking matrices Q and K [21]:

$$Q = \text{QILIF}(\text{BN}(\text{Conv}(XW_Q))), \tag{16}$$

$$K = \text{QILIF}(\text{BN}(\text{Conv}(XW_K))), \tag{17}$$

where$X \in \mathbb{R}^{N \times C}$ represents the input spiking sequence, with N denoting the number of tokens, and C representing the number of channels. The attention mask is then generated using the statistical properties of Q. Specifically, we compute the variance energy map $\mathcal{E}$ of Q across the feature channel dimension:

$$\mathcal{E} = \text{Normalize}\left((Q - \mu_c(Q))^2\right), \tag{18}$$

where $\mu_c(Q)$ denotes the mean of Q in the channel dimension, $(Q - \mu_c(Q))^2$ calculates the centralized variance energy of the feature, and Normalize$(\cdot)$ performs L2 normalization. Subsequently, we feed this energy map into a spike neuron to generate a binary texture saliency mask A:

$$A = \text{SN}(\mathcal{E}), \tag{19}$$

Finally, perform a gated operation based on the Hadamard product between the generated saliency mask A and the content carrier K, and obtain the final output Y through the output projection layer:

$$Y = \text{SN}(\text{BN}(\text{Conv1d}(A \odot K))). \tag{20}$$

IV. EXPERIMENTS

We conducted comprehensive experiments to evaluate the proposed method and compared it with other approaches on multiple widely used benchmark datasets. These experiments covered the static datasets CIFAR-10/100 [22] as well as the neuromorphic datasets CIFAR10-DVS [23] and DVS128 [24] Gesture. Finally, we validated the effectiveness of the proposed modules through a series of ablation experiments.

A. *Results on CIFAR and Neuromorphic Datasets*

*1) **CIFAR Classification**:* On the CIFAR dataset, we trained for 400 epochs with a batch size of 64 and set the virtual time step $D = 4$. We employed the AdamW optimizer with a learning rate of 1×10^{-3} and applied a cosine learning rate decay strategy with a weight decay of 0.06.

The experimental results are shown in Table I. On the CIFAR-10 dataset, S2Iformer achieved an accuracy of 96.79% with 6.70 million parameters, significantly outperforming traditional direct training methods MPBN (96.47%) and GAC-SNN (96.46%). Notably, compared to GAC-SNN, S2Iformer achieves higher accuracy while reducing the number of parameters by nearly 47%. Furthermore, S2Iformer comprehensively outperforms Transformer-based SNN models such as Spikingformer (95.81%) and S-Transformer (96.10%). On the more challenging CIFAR-100 task, S2Iformer demonstrates even greater advantages achieving a competitive accuracy of 80.92%. Compared to the equally high-performing GAC-SNN (80.45%) and Spikingformer-CML (80.37%), S2Iformer not only leads in accuracy but also excels in model lightweighting, fully demonstrating the proposed module's robust modeling capability for complex semantic features.

*2) **Neuromorphic Classification**:* The training process for DVS128 Gesture involves 192 epochs, while CIFAR10-DVS training requires 96 epochs. We set the time step to 16 and employed the AdamW optimizer with an initial learning rate of 5×10^{-3} and a learning rate decay strategy, setting the weight decay to 0.06.

The experimental results are shown in Table II.On the CIFAR10-DVS classification task, S2Iformer achieved 82.5% accuracy with only 1.50M parameters, significantly outperforming Spikformer (80.9%). For the DVS128 Gesture recognition task, S2Iformer attained 98.26% accuracy, demonstrating exceptional feature extraction capabilities at extremely low model capacity.

B. *Ablation Study*

*1) **QI-LIF and TSSA Modules**:* We validated the effectiveness of the QI-LIF and TSSA modules on CIFAR-100, with results shown in Table III. Using Spikformer as the baseline (Accuracy: 78.21%), introducing the QI-LIF module alone improved accuracy by 2.03%, while introducing the TSSA module alone yielded a significant gain of 2.58%. When both modules worked together, the model achieved optimal performance (80.92%), representing a total improvement of 2.71% over the baseline. This demonstrates that the improvements in QI-LIF's neural firing mechanism and TSSA's spatial attention

TABLE I: Comparisons on CIFAR10 and CIFAR100 datasets.

Dataset	Methods	Architecture	Param (M)	Time Step	Top-1 Acc (%)
CIFAR10	MPBN [25]	ResNet19	-	2	96.47
	GAC-SNN [26]	MS-ResNet-18	12.63	6	96.46
	Spikformer [9]	Spikformer-4-384	9.32	4	95.51
	SD-Transformer [27]	SD-Transformer-2-512	10.28	4	95.60
	Spikingformer [28]	Spikingformer-4-384	9.32	4	95.81
	Spikingformer-CML [29]	Spikformer-4-384	9.32	4	95.95
	S-Transformer [30]	SpikingTransformer-2-512	10.23	4	96.42
	S2Iformer(Ours)	**S2Iformer**	**6.70**	4	**96.79**
CIFAR100	MPBN [25]	ResNet19	-	2	79.51
	GAC-SNN [26]	MS-ResNet-18	12.67	6	80.45
	Spikformer [9]	Spikformer-4-384	9.32	4	78.21
	SD-Transformer [27]	SD-Transformer-2-512	10.28	4	78.40
	Spikingformer [28]	Spikingformer-4-384	9.32	4	79.21
	Spikingformer-CML [29]	Spikformer-4-384	9.32	4	80.37
	S-Transformer [30]	SpikingTransformer-2-512	10.23	4	79.90
	S2Iformer(Ours)	**S2Iformer**	**6.74**	4	**80.92**

TABLE II: Comparisons on DVS128 Gesture and CIFAR10-DVS datasets.

Dataset	Methods	Architecture	Param (M)	Time Step	Top-1 Acc (%)
DVS128	Spikformer [9]	Spikformer-2-256	2.57	16	98.3
	SD-Transformer [27]	SD-Transformer-2-256	2.57	16	**99.3**
	Spikingformer [28]	Spikingformer-2-256	2.57	16	98.3
	Spikingformer-CML [29]	Spikformer-2-256	2.57	16	98.6
	S2Iformer(Ours)	**S2Iformer**	**2.00**	16	98.3
CIFAR10-DVS	MPBN [25]	-	-	4	76.7
	GAC-SNN [26]	-	-	10	78.7
	Spikformer [9]	Spikformer-2-256	2.57	16	80.9
	SD-Transformer [27]	SD-Transformer-2-256	2.57	16	80.0
	Spikingformer [28]	SD-Transformer-2-256	2.57	16	81.3
	Spikingformer-CML [29]	Spikformer-2-256	2.57	16	81.4
	S2Iformer(Ours)	**S2Iformer**	**1.50**	16	**82.5**

TABLE III: Ablation Study of QI-LIF and TSSA Modules

QI-LIF	TSSA	Param(M)	Top-1 Acc(%)
baseline		6.78	78.21
✓		6.74	80.24(+2.03)
	✓	6.78	80.79(+2.58)
✓	✓	6.74	80.92(+2.71)

optimization are complementary. Notably, the introduction of these modules slightly reduced the model's parameter count (from 6.78M to 6.74M), achieving a better performance-efficiency tradeoff.

*2) **QI-LIF Modules**:* For the internal components of the QI-LIF module, we conducted a deeper ablation analysis on CIFAR-10, as shown in Figure 2. Experimental data indicate that QI-LIF (96.79%) outperforms the traditional SFA(w/o leak and inhibition) (96.73%), reflecting its superiority in the quantization integration process. Notably, the configuration of the inhibition mechanism significantly impacts performance: removing learnable properties and adopting a fixed inhibition threshold (w/o learnable inhibition) causes accuracy to drop sharply to 96.64%, even lower than when the inhibition mechanism is completely removed (96.77%). This reveals that unreasonable fixed inhibition disrupts normal neuronal firing. By introducing learnable inhibition factors, QI-LIF adaptively regulates the firing frequency of neurons in each layer, thereby enhancing the model's overall performance.

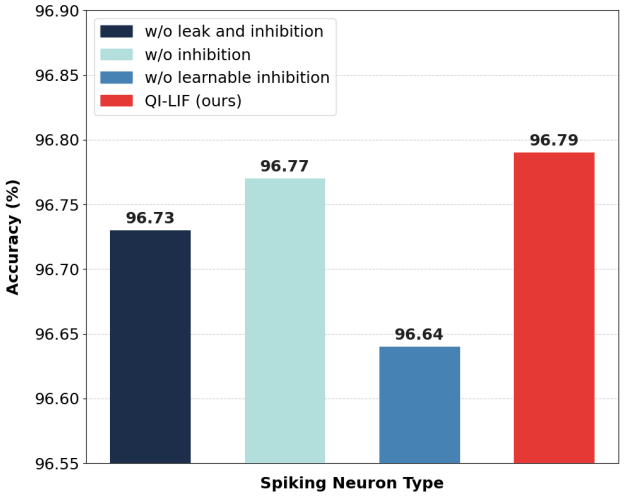

Fig. 2: Ablation Study on QI-LIF Components.

V. Conclusion

In this paper, we addressed the critical limitations of existing quantized Spiking Transformers, specifically their inability to capture deep spatiotemporal dependencies due to the omission of biological temporal dynamics. We proposed a novel architecture that integrates Quantized Inhibitory Leaky Integrate-and-Fire (QI-LIF) neurons with Texture Saliency-based dynamic gated Spiking Attention (TSSA). The QI-LIF neuron successfully reintroduces learnable leakage factors and excitatory-inhibitory mechanisms into the quantization framework. By simulating the refractory period and preventing neuronal overexcitation under continuous stimulation, QI-LIF significantly enhances the model's temporal representation capabilities. Furthermore, the TSSA module serves as an efficient spatial filter with linear complexity. By leveraging channel variance to identify texture saliency, TSSA performs "input purification," physically blocking low-variance background noise and ensuring that the inhibitory mechanisms of QI-LIF are focused on critical high-frequency features. Extensive experiments on both static (CIFAR-10/100) and neuromorphic (CIFAR10-DVS, DVS128 Gesture) datasets demonstrate that our method achieves state-of-the-art performance with low latency. Notably, on CIFAR-100, we achieved an accuracy of 80.92% using only 4 time steps, surpassing the baseline by 3.4%. This work provides a compelling strategy for bridging the performance gap between SNNs and ANNs while preserving the energy-efficient nature of spike-based computing.

References

[1] Wolfgang Maass. Networks of spiking neurons: the third generation of neural network models. Neural networks, 10(9):1659–1671, 1997.

[2] Kaushik Roy, Akhilesh Jaiswal, and Priyadarshini Panda. Towards spike-based machine intelligence with neuromorphic computing. Nature, 575(7784):607–617, 2019.

[3] Schuman, C.D., Kulkarni, S.R., Parsa, M., Mitchell, J.P., Kay, B., et al.: Opportu nities for neuromorphic computing algorithms and applications. Nature Computational Science 2(1):10–19 2022

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS), volume 30, 2017.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representa- tions (ICLR), 2020.

[6] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159, 2020.

[7] Li Q, Gkoumas D, Sordoni A, et al. Quantum-inspired neural network for conversational emotion recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(15): 13270-13278.

[8] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012–10022, 2021.

[9] Z. Zhou, Y. Zhu, C. He, Y. Wang, S. Yan, Y. Tian, and L. Yuan, "Spikformer: When spiking neural network meets transformer," in ICLR,2023.

[10] C. Zhou, H. Zhang, Z. Zhou, L. Yu, L. Huang, X. Fan, L. Yuan,Z. Ma, H. Zhou, and Y. Tian, "QKFormer: Hierarchical spiking transformer using qk attention," arXiv preprint arXiv:2403.16552, 2024.

[11] Yao M, Qiu X, Hu T, et al. Scaling spike-driven transformer with efficient spike firing approximation training[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025.

[12] Lei Z, Yao M, Hu J, et al. Spike2former: Efficient spiking transformer for high-performance image segmentation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39(2): 1364-1372.

[13] Luo X, Yao M, Chou Y, et al. Integer-valued training and spike-driven inference spiking neural network for high-performance and energy-efficient object detection[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 253-272.

[14] Qiu X, Zhang M, Zhang J, et al. Quantized spike-driven transformer[J]. arXiv preprint arXiv:2501.13492, 2025.

[15] Xiao J, Ma H, Chen R, et al. STKPS-Net: Spatio-Temporal Key Patch Selection Network for Few Shot Anomalous Action Recognition[J]. IEEE Transactions on Information Forensics and Security, 2026, 21: 827-838.

[16] Y. Guo, Y. Chen, X. Liu, W. Peng, Y. Zhang, X. Huang, and Z. Ma,"Ternary spike: Learning ternary spikes for spiking neural networks," in Proceedings of the AAAI conference on artificial intelligence, vol. 38, no. 11, 2024, pp. 12244–12252.

[17] Rast A D, Galluppi F, Jin X, et al. The leaky integrate-and-fire neuron: A platform for synaptic model exploration on the spinnaker chip[C]//The 2010 International Joint Conference on Neural Networks (IJCNN). IEEE, 2010: 1-8.

[18] Sun K, Duan P, Kuhlmann L, et al. ILIF: Temporal Inhibitory Leaky Integrate-and-Fire Neuron for Overactivation in Spiking Neural Networks[J]. arXiv preprint arXiv:2505.10371, 2025.

[19] Yin P, Lyu J, Zhang S, et al. Understanding straight-through estimator in training activation quantized neural nets[J]. arXiv preprint arXiv:1903.05662, 2019.

[20] Wang P, Li T, Li P. Saliency-enhanced infrared and visible image fusion via sub-window variance filter and weighted least squares optimization[J]. PLoS One, 2025, 20(7): e0323285.

[21] Chen Y, Xie Z, Zhong J, et al. SQKformer: Spiking sparse QKformer with adaptive batch normalization for membrane potential[J]. Neurocomputing, 2026: 132666.

[22] A. Krizhevsky, V. Nair, and G. Hinton, "Cifar-10 (canadian institute for advanced research)," URL http://www. cs. toronto. edu/kriz/cifar. html, vol. 5, no. 4, p. 1, 2010

[23] H. Li, H. Liu, X. Ji, G. Li, and L. Shi, "Cifar10-dvs: an event-stream dataset for object classification," Frontiers in neuroscience, vol. 11, p.309, 2017.

[24] A. Amir, B. Taba, D. Berg, T. Melano, J. McKinstry, C. Di Nolfo, T. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza et al., "A low power, fully event-based gesture recognition system," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7243–7252.

[25] Guo Y, Zhang Y, Chen Y, et al. Membrane potential batch normalization for spiking neural networks[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 19420-19430.

[26] Qiu X, Zhu R J, Chou Y, et al. Gated attention coding for training high-performance and efficient spiking neural networks[C]//Proceedings of the AAAI conference on artificial intelligence. 2024, 38(1): 601-610.

[27] Yao M, Hu J, Zhou Z, et al. Spike-driven transformer[J]. Advances in neural information processing systems, 2023, 36: 64043-64058.

[28] Zhou C, Yu L, Zhou Z, et al. Spikingformer: Spike-driven residual learning for transformer-based spiking neural network[J]. arXiv preprint arXiv:2304.11954, 2023.

[29] Zhou C, Zhang H, Zhou Z, et al. Enhancing the performance of transformer-based spiking neural networks by SNN-optimized downsampling with precise gradient backpropagation[J]. arXiv preprint arXiv:2305.05954, 2023.

[30] Y. Guo, X. Liu, Y. Chen, W. Peng, Y. Zhang, and Z. Ma, "Spiking transformer: Introducing accurate addition-only spiking self-attention for transformer," in Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 24398–24408.

[31] Lee D, Li Y, Kim Y, et al. Spiking transformer with spatial-temporal attention[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 13948-13958.

[32] Y. Fang, D. Zhou, Z. Wang, H. Ren, Z. Zeng, L. Li, S. Zhou, and R. Xu,"Spiking transformers need high frequency information," arXiv preprint arXiv:2505.18608, 2025.

[33] Chen Y, Zhong J, Guo Y, et al. C3Net: A cross-modal collaborative calibration of features for object detection using frames and events[J]. Neural Networks, 2026: 108651.