Neural Dynamics Self-attention for Spiking Transformers

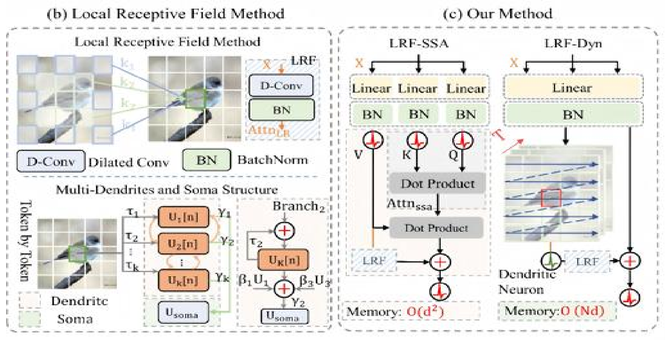
Dehao Zhang[1], Fukai Guo[1], Shuai Wang[1], Jingya Wang[1], Jieyuan Zhang[1], Yimeng Shan[1], Malu Zhang[1,2], ✉Yang Yang[1], Haizhou Li[2,3]

[1]University of Electronic Science and Technology of China, [2]Shenzhen Loop Area Insititue [3]The Chinese University of Hong Kong (Shenzhen)

University of Electronic Science and Technology of China

Motivation & Model Structure

Integrating Spiking Neural Networks (SNNs) with Transformers offers a promising route toward achieving both energy efficiency and strong performance, particularly for edge vision applications. However, although Spiking Self-Attention (SSA) is well suited to neuromorphic computation, it still underperforms Vanilla Self-Attention (VSA). Inspired by local receptive fields and the temporal charge–fire–reset dynamics of neuronal membrane potentials, we propose LRF-Dyn. The overall framework is illustrated as follows.



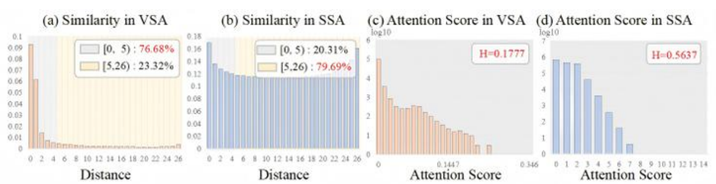
The contribution is as follow:

- **Fundamental Bottlenecks:** Existing Spiking Transformers are constrained by limited local modeling capacity and the memory overhead of storing attention scores.
- **Methodological Innovation:** We propose LRF-Dyn, a locality-aware and memory-efficient spiking attention framework that integrates local receptive fields with neuron-inspired dynamics.
- **Extensive Validation:** Experiments across diverse SNN architectures and vision tasks demonstrate that LRF-Dyn consistently improves performance while reducing inference-time memory requirements.

Method

Limited Local Modeling Capabilty

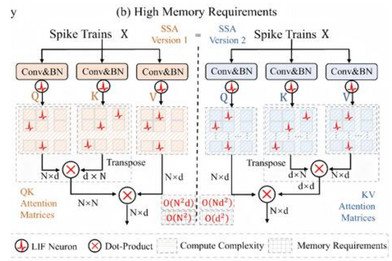
- Due to the stronger correlations among neighboring tokens, ViT inherently exhibit a pronounced locality bias, with approximately **76.8%** of attention concentrated within short-range interactions. In contrast, SSA produces a more **uniform attention distribution**, lacking a clear locality preference.



High Memory Requirements during Inference

$$\mathrm{attn}'_n[t] = \Big(\sum_{j=1}^{N} \mathrm{q}_n[t]\mathrm{k}_j[t]^T e_j^T\Big) \times \mathrm{v}[t] = \mathrm{q}_n[t] \times \sum_{j=1}^{N} k_j[t]^T v_j[t],$$

For the n-th token, it is necessary to store not only the Q, K, and V matrices at each timestep, but also the intermediate results of the KV multiplication, leading to an additional $O(d^2)$ memory overhead. In particular, when d = 512, SSA requires substantial memory, which further slows down model inference. These limitations significantly hinder its deployment on resource-constrained devices, especially neuromorphic chips.



Spiking Self-Attention with Local Receptive Fields

To address the limited local receptive field of spiking self-attention, we propose **LRF-SSA**, which integrates a **local convolution** module into the SSA mechanism to increase sensitivity to neighboring positions. Its self-attention output sattn′n[t] can be defined as:

$$\mathrm{sattn}'_n[t] = \underbrace{\mathrm{q}_n[t] \times \sum_{j=1}^{N} k_j[t]^T v_j[t]}_{\text{Global Receptive Fields}} + \underbrace{\sum_{d} \sum_{i,j \in \Omega_d} r_{ij}^d \mathbf{V}^{\rho_k}}_{\text{Local Receptive Fields}},$$

Ω_d = {(i,j)|i,j ∈ {−d,0,d}} represents the positional information of the neighboring region. We introduce multi-scale dilated convolutions to model local receptive fields. Specifically, two 3 × 3 depth-wise convolution kernels with dilation factors d = 3 and d = 5 are employed, where rij denotes the convolutional parameter at position (i,j).

Implementing Self-attention Through Neuronal Dynamics

Inspired by other softmax-free attention, LRF-SSA can be reformulated through causal inference to significantly reduce memory consumption.It can be rewritten as follows:

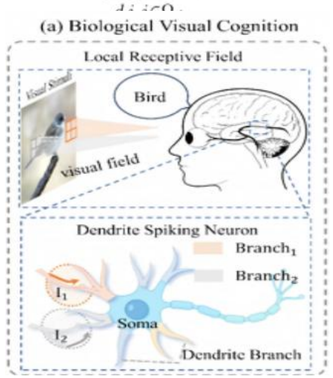
$$\mathrm{sattn}_n[t]' = \mathrm{q}_n[t] \times \underbrace{\sum_{j=1}^{n-1} k_j[t]^T v_j[t]}_{\text{Memory Potential}} + \underbrace{k_n[t]^T v_n[t] + \sum_{d} \sum_{i,j \in \Omega_d} r_{ij}^d \mathrm{v}_{\rho_k}[t]}_{\text{Presynaptic Input}},$$

It closely parallels the **charge–fire-reset dynamics of spiking neurons:** the first term represents membrane potential information and the second represents presynapticinput.

$$\mathrm{X}_n[t] = \mathcal{A} \odot \mathrm{X}_{n-1}[t] + \mathbf{\Gamma} \mathrm{Token}_n[t], \quad \mathrm{sattn}'_n[t] = \mathrm{X}_n[t] + \sum\sum r_{ij}^d \cdot \mathrm{X}_{\rho_k}[t],$$

$$\mathcal{A} = \underbrace{\begin{bmatrix} c_1, \\ c_2, \\ \vdots \\ c_{n-1}, \\ c_n \end{bmatrix}^T}_{\mathcal{C}} \times \begin{bmatrix} -\frac{1}{\tau_1} & \beta_{2,1} & 0 & \cdots & 0 \\ \beta_{1,2} & -\frac{1}{\tau_2} & \beta_{3,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & -\frac{1}{\tau_{n-1}} & \beta_{n,n-1} \\ 0 & 0 & \cdots & \beta_{n-1,n} & -\frac{1}{\tau_n} \end{bmatrix},$$

Here, dn denotes the number of dendrites, and C∈R represents the weights assigned to different dendrites. For the n-th token, different dendritic branches produce distinct responses, which are further integrated by the soma through a specific mechanism to enhance spatial interactions and transform them into spike trains.



Overall Architectures

$$\mathbf{H} = \mathcal{F}^{-1}\{\mathcal{F}(\mathbf{K}) * \mathcal{F}(\mathbf{X})\}, \quad \mathbf{Score} = \mathrm{SN}\{\sum\sum_{i,j \in \Omega_d} r_{ij}^d \cdot \alpha_k \mathbf{H}_{p_k(t)}\},$$

Experiments & Visual Results

- **We examine the performance of ours methods across a variety of visual tasks, including classification and segmentation**

A summary of our method's energy, accuracy and model parameters

Table 1: Comparison with Similar Methods on ImageNet-1K.

Method	Architecture	SR.	Param.(M)	Acc.(%)
Spikformer (Zhou et al., 2023b)	Spikformer-8-512	$\mathcal{O}(d^2)$	29.68	73.38
	Spikformer-8-768	$\mathcal{O}(d^2)$	66.34	74.81
Spikformer + LRF-SSA	Spikformer-8-512	$\mathcal{O}(d^2)$	29.71	**74.62** (↑1.24)
	Spikformer-8-768	$\mathcal{O}(d^2)$	66.53	**75.66** (↑0.85)
Spikformer + LRF-Dyn	Spikformer-8-512	$\mathcal{O}(kd)$	29.71	**74.51** (↑1.13)
	Spikformer-8-768	$\mathcal{O}(kd)$	66.53	**75.58** (↑0.77)
QKFormer (Zhou et al., 2024)	HST-10-384	$\mathcal{O}(d^2)$	16.47	78.80
	HST-10-512	$\mathcal{O}(d^2)$	29.08	82.04
QKFormer + LRF-SSA	HST-10-384	$\mathcal{O}(d^2)$	16.55	**79.24** (↑0.44)
	HST-10-512	$\mathcal{O}(d^2)$	29.18	**82.52** (↑0.48)
QKFormer + LRF-Dyn	HST-10-384	$\mathcal{O}(kd)$	16.44	**79.21** (↑0.41)
	HST-10-512	$\mathcal{O}(kd)$	29.18	**82.48** (↑0.44)
SDT-V3 (Yao et al., 2025)	Efficient-Transformer-S	$\mathcal{O}(d^2)$	5.11	75.30
	Efficient-Transformer-L	$\mathcal{O}(d^2)$	18.99	79.80
SDT-V3 + LRF-SSA	Efficient-Transformer-S	$\mathcal{O}(d^2)$	5.24	**76.22** (↑0.92)
	Efficient-Transformer-L	$\mathcal{O}(d^2)$	19.25	**80.31** (↑0.51)
SDT-V3 + LRF-Dyn	Efficient-Transformer-S	$\mathcal{O}(kd)$	5.24	**76.12** (↑0.82)
	Efficient-Transformer-L	$\mathcal{O}(kd)$	19.25	**80.24** (↑0.44)

Table 2: Performance of segmentation.

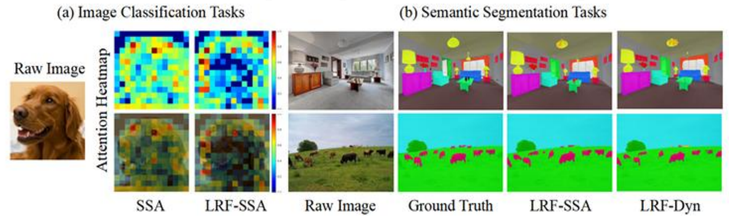
Model	Para. (M)	Attn	T	MIoU(%)
PVT	28.2	✓	1	39.8
(Wang et al., 2021)	48.0	✓	1	41.6
SDT-V3	5.1 + 1.4	✓	4	33.6
(Yao et al., 2025)	18.99 + 1.4†	✓	4	41.3
SDT-V3	5.1 + 1.4	✓	4	**36.2** (↑2.6)
+ LRF-SSA	10.0 + 1.4	✓	4	**43.5** (↑2.2)
SDT-V3	5.24 + 1.4	✗	4	**36.3** (↑2.7)
+ LRF-Dyn	19.25 + 1.4	✗	4	**43.1** (↑1.8)

Table 3: Ablation Experiment.

Method	w/o LRF	Ω ≤ 1	Ω ≤ 3	Ω ≤ 5
LRF-SSA	77.86	78.26	78.52	**78.64**
LRF-Dyn	77.78	78.16	78.50	78.57
Caused SSA†	74.30	75.30	76.20	76.50

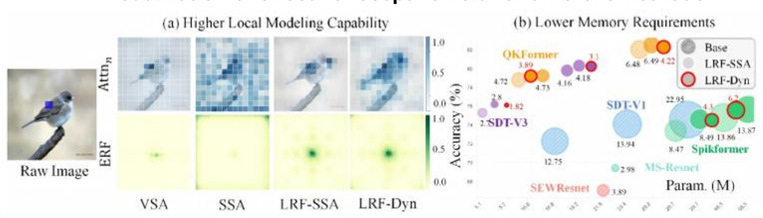
- **We visualize the attention scores of LRF-Dyn on the classification and compare them with the results on the segmentation task.**

Visual results for image recognition and semantic segmentation



- **We provide a visualization of the receptive fields of our method and analyze its memory requirements.**

Visualization of effective receptive field for different methods



Conclusion

We address the challenges of SNNs Transformers: **limited performance gains from mismatched attention distributions and high memory costs.**
Inspired by **local receptive fields and the temporal charge–fire–reset dynamics of neuronal membrane potentials**, we propose LRF-Dyn, which introduces local receptive field modeling and approximates attention via neuronal dynamics.

➢ **Higher Model Performance**

➢ **Lower Memory Requirements**