Bipolar Self-attention for Spiking Transformers

Shuai Wang[1], **Malu Zhang**[1,2]*, **Jingya Wang**[1], **Dehao Zhang**[1], **Yimeng Shan**[1], **Jieyuan Zhang**[1], **Yichen Xiao**[1], **Honglin Cao**[1], **Haonan Zhang**[1], **Zeyu Ma**[1], **Yang Yang**[1], **Haizhou Li**[2,3]

[1]University of Electronic Science and Technology of China

[2]Shenzhen Loop Area Institute, [3]The Chinese University of Hong Kong (Shenzhen)

Abstract

Harnessing the event-driven characteristic, Spiking Neural Networks (SNNs) present a promising avenue toward energy-efficient Transformer architectures. However, existing Spiking Transformers still suffer significant performance gaps compared to their Artificial Neural Network counterparts. Through comprehensive analysis, we attribute this gap to these two factors. First, the binary nature of spike trains limits Spiking Self-attention (SSA)'s capacity to capture negative–negative and positive–negative membrane potential interactions on Querys and Keys. Second, SSA typically omits Softmax functions to avoid energy-intensive multiply-accumulate operations, thereby failing to maintain row-stochasticity constraints on attention scores. To address these issues, we propose a Bipolar Self-attention (BSA) paradigm, effectively modeling multi-polar membrane potential interactions with a fully spike-driven characteristic. Specifically, we demonstrate that ternary matrix multiplication provides a closer approximation to real-valued computation on both distribution and local correlation, enabling clear differentiation between homopolar and heteropolar interactions. Moreover, we propose a shift-based Softmax approximation named Shiftmax, which efficiently achieves low-entropy activation and partly maintains row-stochasticity without non-linear operation, enabling precise attention allocation. Extensive experiments show that BSA achieves substantial performance improvements across various tasks, including image classification, semantic segmentation, and event-based tracking. These results establish its potential as a fundamental building block for energy-efficient Spiking Transformers.

1 Introduction

As the core computational unit of Transformers [5, 15, 11, 32, 9], self-attention mechanism dynamically models the global dependencies among sequence elements, overcoming the long-range dependency challenges [33, 10]. However, its' $O(N^2d)$ computational complexity [18, 48] incurs an exponential rise during both training and inference, restricting its application in many resource-constrained environments. Consequently, how to develop energy-efficient and high-performance Transformers remains a critical research focus.

Spiking Neural Networks (SNNs) [23, 12] have gained significant attention due to their brain-inspired dynamics [17, 24]. Spiking neurons fire discrete spikes only when activated, remaining silent otherwise. Compared with Artificial Neural Networks (ANNs) that rely on multiply-accumulate (MAC) computation, the spike-driven mechanism [61, 43, 44] in SNNs supports sparse accumulate (AC) operations [2]. Such sparse spike-based computation [20, 56] delivers significant power efficiency, particularly on neuromorphic platforms such as Tianjic [26, 8, 22] and Loihi [6, 25]. Recently, numerous researchers focus on developing Spiking Transformers, including Spikformer [70], Spikingformer [67], Spike-driven Transformers [51, 49, 52], SpikingResformer [30], and QKformer [68].

*Corresponding author: maluzhang@uestc.edu.cn

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

 https://doi.org/10.52202/085713-3401

These approaches enhance the performance ceiling of SNNs in various tasks [64, 45, 36, 31, 62, 34], demonstrating that Spiking Transformers achieve a trade-off between high performance and efficiency.

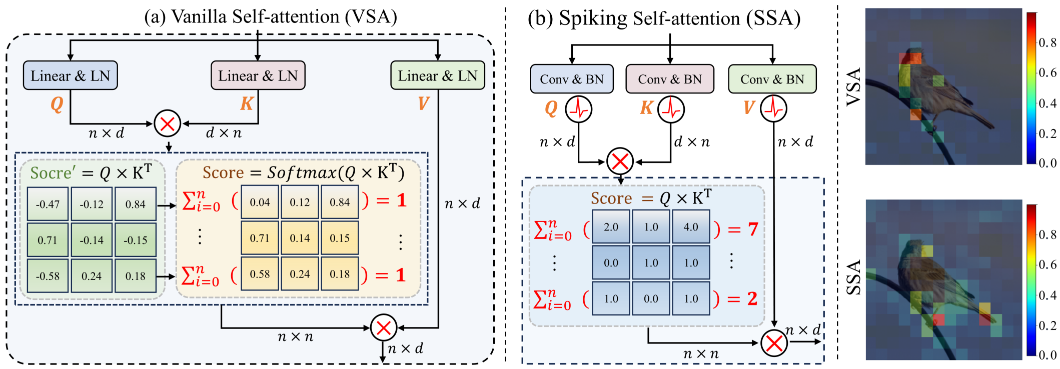

Figure 1: Comparison of Vanilla Self-attention (VSA) and Spiking Self-attention (SSA) mechanisms. $\mathbf{Score}'$ contains rich polarity information and $\mathbf{Score}$ maintains strict row-stochasticity via Softmax in VSA, while $\mathbf{Score}$ in SSA lacks negative polarity and row-stochasticity constraints.

As a core component of Spiking Transformer architectures, the Spiking Self-ttention (SSA) [70, 51] computational paradigm represents the critical factor limiting their performance ceiling. As shown in Fig.1, Vanilla Self-attention (VSA) in ANNs effectively captures magnitude and multi-polarity information by computing $\mathbf{Score}'$ through query (Q) and key (K) correlation. Meanwhile, it calculates the $\mathbf{Score}$ using the Softmax function under row-stochasticity constraint. In contrast, SSA in SNNs operates with binary spike trains (0 or 1), which not only sacrifices quantization precision but also disregards negative-negative and positive-negative interactions. Furthermore, to maintain energy efficiency, SSA typically omits Softmax function, resulting in severely imbalanced row-stochasticity in $\mathbf{Score}$. These issues limit SSA's ability to effectively compute attention allocation between Q and K, causing SSA more like a simplified Token Mixer [54, 55]. Therefore, how to overcome these limitations is crucial for pushing Spiking Transformers beyond their current performance bottlenecks.

In this paper, we propose a Bipolar Self-attention (BSA) paradigm to effectively address these issues. Unlike the SSA paradigm that exclusively captures positive-positive Q-K interactions, BSA employs ternary spiking neurons [14, 35] to comprehensively process both homopolar and heteropolar interaction patterns in Q-K correlation computation. We theoretically demonstrate that ternary matrix multiplication more closely approximates real-valued computation in terms of both distributional similarity and local correlation than binary part. Moreover, we propose the innovative Shiftmax method, which approximates Softmax's low-entropy activation characteristics and row-stochasticity constraints through energy-efficient bit-shift operations. Finally, we conduct extensive experiments across diverse tasks including image classification [70, 68], semantic segmentation [49, 52], and event-based tracking [58, 28], consistently demonstrating that BSA delivers substantial performance improvements. The main contributions of our work are outlined as follows:

- We first identify two limitations of Spiking Self-attention: (1) binary matrix products exclusively capture positive-positive correlations while neglecting negative–negative and positive–negative polarity features, leading to a complete loss of polarity information. (2) without Softmax, attention scores across different rows exist on incomparable scales, rendering attention allocation ineffective. These deficiencies prevent SSA from fully harnessing the potential of the self-attention mechanism.
- We propose Bipolar Self-attention (BSA) to overcome these limitations. BSA employs ternary matrix products to extract Q-K correlations, comprehensively process different polarity interaction patterns. In addition, we propose a Shiftmax method to approximate Softmax that achieves low-entropy activation and maintains partial row-stochasticity without non-linear operations, enabling precise attention allocation.
- Extensive experiments demonstrate that BSA achieves significant performance improvements across various advanced Spiking Transformers on ImageNet-1K. Furthermore, our method establishes state-of-the-art performance in both semantic segmentation and event-based tracking tasks compared to existing SNNs approaches.

https://doi.org/10.52202/085713-3401

2 Related Work

Recently, growing attention has been paid to energy-efficient Spiking Transformers [4, 50]. For instance, Spikformer [70] first proposes Spiking Self-attention computation, establishing the first Spiking Transformer. Building on this, Spikingformer [67] introduces a hardware-friendly residual learning architecture that avoids non-spike computations in SNNs. Then, the Spike-driven Transformers (SDT-V1) [51] incorporates the Hadamard product into the self-attention module, achieving a fully spike-driven mechanism. Furthermore, SpikingResformer [30] integrates a Dual Spike self-attention module, enhancing both performance and energy efficiency. Recently, QKformer [68] and Spike-driven Transformer-V3 [52] both significantly elevate the performance ceiling of Spiking Transformers. Nonetheless, these models primarily utilize self-attention as a token mixer [54], without paying attention to exploring effective similarity calculations suited to spike trains [60, 1]. Therefore, designing an innovative spiking self-attention mechanism that fully leverages self-correlation computation under spike-driven characteristics is crucial for further advancement.

3 Preliminary

3.1 Vanilla Self-attention

The self-attention mechanism effectively captures global dependencies within sequences by dynamically allocating attention across tokens [33, 29, 13]. Given an input matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, where n denotes sequence length and d represents embedding dimensionality, the mechanism initially projects $\mathbf{X}$ into query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ matrices via distinct parameter matrices $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$:

$$\tilde{\mathbf{X}} = \mathrm{LN}(\mathbf{X}), \qquad \mathbf{Q} = \mathrm{Linear}(\tilde{\mathbf{X}}), \quad \mathbf{K} = \mathrm{Linear}(\tilde{\mathbf{X}}), \quad \mathbf{V} = \mathrm{Linear}(\tilde{\mathbf{X}}). \tag{1}$$

Here, LN denotes layer normalization, and Linear refers to a fully connected layer. To determine contextual relationships, the model computes token-wise similarities through dot-product operations between $\mathbf{Q}$ and $\mathbf{K}$, generating an attention score matrix:

$$\mathbf{Score}' = \mathbf{Q} \times \mathbf{K}^{\top}, \quad \mathbf{Score} = \mathrm{Softmax}\left(\frac{\mathbf{Score}'}{\sqrt{d}}\right), \quad \mathbf{Attn} = \mathbf{Score} \times \mathbf{V}. \tag{2}$$

The mechanism applies a scaling factor $\sqrt{d}$ to mitigate gradient instability issues. Subsequently, the Softmax function normalizes these scaled $\mathbf{Score}$ to a probability distribution, ensuring attention weights sum to unity while emphasizing relevant tokens and suppressing irrelevant ones.

3.2 Spiking Self-attention

Building on VSA, Spikformer [70] introduced SSA. For an input matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$ are first computed via learnable weight matrices and then converted into spike trains by binary spiking neurons for subsequent processing. The dynamics of these neurons are given by:

$$U[t+1] = H[t] + X[t+1], \tag{3}$$

$$S[t+1] = \Theta(U[t+1] - V_{th}), \tag{4}$$

$$H[t+1] = V_{reset}S[t+1] + \tau U[t+1](1 - S[t+1]). \tag{5}$$

$X[t+1]$ denotes the input current, while $H[t]$ and $U[t]$ represent the pre-synaptic and post-synaptic membrane potentials, respectively. The Heaviside function $\Theta(\cdot)$ is employed for spike generation. If a spike occurs ($S[t+1] = 1$), $H[t]$ resets to V_{reset}; otherwise, $U[t+1]$ decays with a time constant τ and feeds into $H[t+1]$. Here a spiking neuron layer is denoted as $\mathcal{SN}(\cdot)$, which takes $X[t+1]$ as input and produces the spike $S[t+1]$ as output. Then the $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ in SSA can be decribed as:

$$\mathbf{Q} = \mathcal{SN}(\mathrm{BN}(\mathrm{Conv}_1(\mathbf{X}))), \quad \mathbf{K} = \mathcal{SN}(\mathrm{BN}(\mathrm{Conv}_2(\mathbf{X}))), \quad \mathbf{V} = \mathcal{SN}(\mathrm{BN}(\mathrm{Conv}_3(\mathbf{X}))), \tag{6}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{T \times n \times d}$, $\mathrm{BN}(\cdot)$ denotes batch normalization and $\mathrm{Conv}(\cdot)$ refers to a convolution operation. Unlike VSA, SSA omits the Softmax operation while retaining scaling to control the magnitude of the attention output. The attention output $\mathbf{Attn}$ is obtained by performing a matrix multiplication of the spiking $\mathbf{Q}$ and $\mathbf{K}$, scaled by factor s, and then convert to spike trains:

$$\mathbf{Score} = s \cdot \mathbf{Q} \times \mathbf{K}^{\mathbf{T}}, \quad \mathbf{Attn} = \mathcal{SN}(\mathbf{Score} \times \mathbf{V}). \tag{7}$$

Therefore, SSA provides an energy-efficient self-attention computation paradigm without any nonlinear operations. It allows the ordering of the Q, K, and V matrices to be flexibly adjusted as needed, enabling the Spiking Transformer to capture global dependencies efficiently.

https://doi.org/10.52202/085713-3401

4 Method

4.1 Problem Analysis for Spiking Self-attention

As previously discussed, SSA reduces computational costs by eliminating nonlinear computations and MAC operations. However, there remains a significant performance gap compared to VSA. We thoroughly analyze the reasons and attribute them to the following parts: (1) binary Q–K Matrix multiplication cannot effectively capture polarity correlations, and (2) eliminating Softmax removes the attention score's row-stochasticity constraints. The details are elaborated in the following.

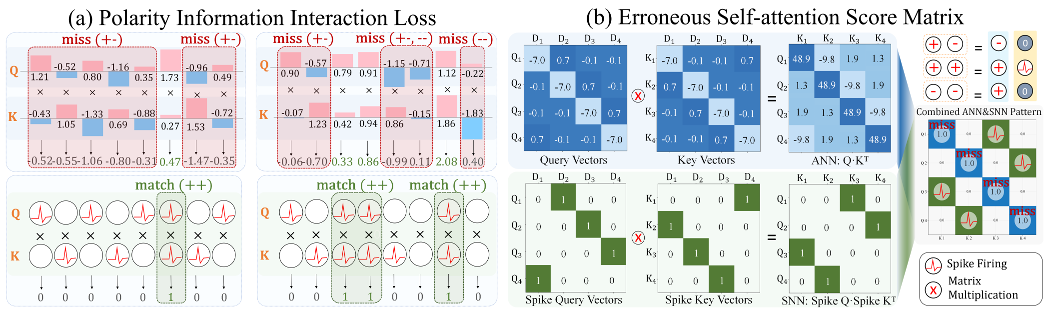

Figure 2: Problem analysis. (a) The binary activation characteristic of spike trains can only preserve (+,+) interactions on the membrane potentials of Q and K, while all others are ignored. (b) The correlation matrix between binary Q and K in SNNs may be completely different from its' VSA counterpart in ANNs.

4.1.1 Incomplete Polarity Information Modeling

In VSA, Q and K are represented by continuous real-valued vectors, enabling matrix multiplication can efficiently calculate the correlation between Q and K. However, SSA typically binarizes the membrane potential information of Q and K, limiting its representation to binary values (0 or 1). This spike-driven approach not only narrows the precision of Q-K interactions, but also undermines the self-attention mechanism's capacity to detect vector polarity relationships. Consider $\Omega_{\text{VSA}} = \{(a_1, a_2) \mid a_i \in \{+, -\}, i = 1, 2\}$ representing correlation types in VSA and $\Omega_{\text{SSA}} = \{(b_1, b_2) \mid b_i \in \{0, 1\}, i = 1, 2\}$ denoting binary states in SSA, with $\phi : \mathbb{R}^d \to \{0, 1\}^d$ as a binarization mapping. Let $\mathcal{P} = \{\mathcal{B}_0, \mathcal{B}_1\}$ partition Ω_{SSA} where $\mathcal{B}_0 = \{(0, 0), (0, 1), (1, 0)\}$, $\quad \mathcal{B}_1 = \{(1, 1)\}$. We establish a surjection $\mathcal{G} : \Omega_{\text{VSA}} \to \mathcal{P}$ that fundamentally characterizes the polarity information loss during the binarization transformation process, such that:

$$\textbf{Case 1:} \quad \mathcal{G}\left(\{(+, +)\}\right) = \mathcal{B}_1, \quad \textbf{Case 2:} \quad \mathcal{G}\left(\{(+, -), (-, +), (-, -)\}\right) = \mathcal{B}_0. \tag{8}$$

Eq.8 demonstrates that SSA completely discards the ability to compute negative-negative polarity or mixed polarity correlations. Due to the sparse activation characteristics of SNNs, the amount of discarded information far exceeds what is retained, resulting in substantial information loss. As shown in Fig.2(a), the green-boxed region represents the positive polarity information that SSA can preserve, while the red–boxed regions indicate the entirely neglected $(-, -)$ and $(+, -)$ polarity information. More importantly, SSA may display entirely disparate attention regions from VSA as shown in Fig.2(b). Thus, devising efficient mechanisms to embed polarity information offers a promising avenue for improving SSA's performance ceiling.

4.1.2 Missing Softmax Operation

SSA not only suffers from a loss of polarity information but also neglects the critical Softmax component. In VSA, Softmax operation [37] serves two pivotal roles: first, amplifying highly relevant through a low-entropy activation effect while suppressing low-value regions; second, ensuring comparability of **Score** across varying scales via row-stochasticity constraints. These mechanisms ensure that Softmax highlights key features while maintaining the comparability of the **Score**.

By contrast, SSA's binary matrix multiplication preserves limited low-entropy activation properties (detailed in the Appendix.A) and entirely lacks row-stochasticity constraints. Consequently, attention scores from different query vectors exist on disparate numerical scales. The disparity in scales makes it difficult to effectively compare the relative degree of elements across different rows during the allocation of attention to V. Therefore, how to allocate global attention on comparable scales will be another potential area for improvement.

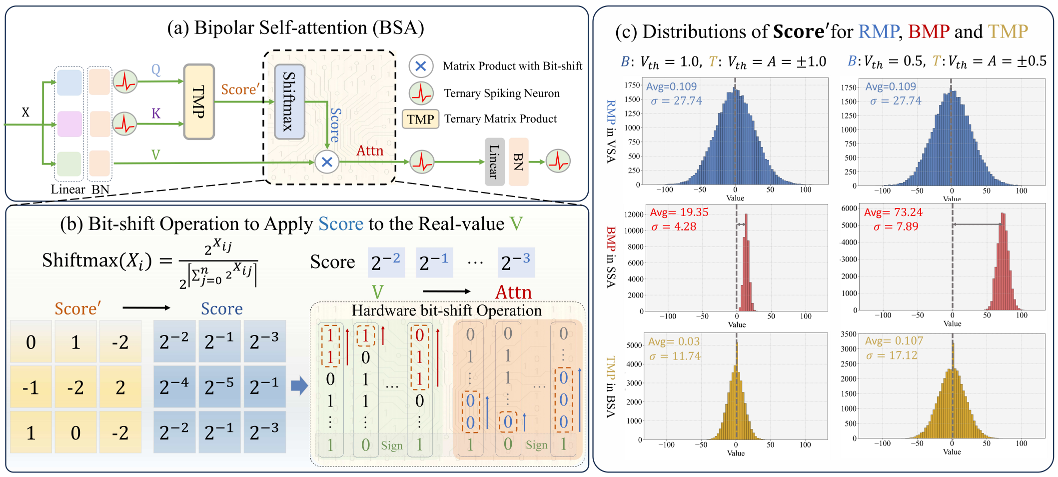

Figure 3: (a) Overall structure of Bipolar Self-attention. (b) Detailed process of Shiftmax. It can directly shift the membrane potential of V for efficient attention allocation. (c) Distribution of three types of matrix product at different thresholds (V_{th}).

4.2 Bipolar Self-attention for Spiking Transformers

To address two critical issues in existing SSA: incomplete polarity correlation and Softmax loss, we propose a novel Bipolar Self-attention (BSA) paradigm. First, we model the different polarity interactions between Querys and Keys using ternary matrix multiplication, ensuring comprehensive representation of polarity information. Furthermore, we introduce the Shiftmax method, which approximates Softmax through energy-efficient bitwise operations while preserving the critical low-entropy activation and row-normalization properties. The details are as follows:

4.2.1 Ternary Matrix Product for Comprehensive Polarity Correlation

To effectively capture the comprehensive polarity correlation in SNNs, we introduce ternary spike neurons ($\mathcal{TSN}$) [14] for characterizing the pre-synaptic membrane potential of $\mathbf{Q}$ and $\mathbf{K}$. The dynamics of the $\mathcal{TSN}(\cdot)$ are defined as follows:

$$U[t+1] = H[t] + X[t+1], \tag{9}$$

$$S[t+1] = \mathrm{Sign}\left(U[t+1]\right) \cdot \Theta\left(|U[t+1]| - V_{th}\right), \tag{10}$$

$$H[t+1] = V_{\text{reset}} S[t+1] + \tau U[t+1](1 - |S[t+1]|). \tag{11}$$

$\mathcal{TSN}(\cdot)$ fires signed spikes $S[t+1]$ and $H[t+1]$ reset to 0. The membrane potential accumulation and reset processes are similar to those in binary spike neurons. Subsequently, BSA computes $\mathbf{Score}'$ through matrix multiplication using ternary Q and K. The Ternary-valued Matrix Product (TMP) method effectively preserves the fundamental characteristics of both Binary-valued Matrix Product (BMP) in SSA and Real-valued Matrix Product (RMP) in VSA. First, it preserves the spike-driven characteristics of BMP, enabling full AC operations and achieving energy-efficient neuromorphic computation. Additionally, it retains the polarity information in the membrane potential $U[t]$, thereby capturing polarity-related correlations comparable to RMP operations. To validate this proposition, we conduct a detailed analysis of the three matrix product, examining both their distributional relationships and local correlation. Detailed proofs are presented in Appendices.B.

Theorem 1. *For independent random vectors $q = \mathbf{Q_j}, k = \mathbf{K_j} \in \mathbb{R}^d$ with elements $q_i, k_i \sim \mathcal{N}(0, \sigma^2)$, $p = \Phi(-\frac{\theta}{\sigma})$ representing the probability $P(q > \theta)$, θ represent the threshold for $\mathcal{SN}(\cdot)$ and $\mathcal{TSN}(\cdot)$. Then the result element of RMP satisfies $\mathbb{E}[R] = 0$ and $Var(R) = d\sigma^4$; BMP satisfies $\mathbb{E}[B] = dp^2$ and $Var(B) = dp^2(1-p^2)$; and TMP satisfies $\mathbb{E}[T] = 0$ and $Var(T) = 4dp^2$.*

https://doi.org/10.52202/085713-3401

Theorem.1 establishes that TMP preserves the zero-mean statistical property of RMP, while BMP exhibits systematic bias. This bias results from binary activation's exclusion of negative polarity information, corroborating our hypothesis. Furthermore, as shown in Fig.3(c), at typical operating points ($V_{th} = 1, 0.5$, corresponding to common SNNs' thresholds), TMP's variance consistently exceeds BMP's, enhancing information capacity and providing a more faithful representation of RMP's statistical characteristics. In summary, TMP approximates RMP's overall distribution more accurately than BMP, consequently enhancing Q-K correlation representation. Nevertheless, global distributional similarity does not guarantee accuracy at individual key-value pairs. Therefore, we further explore whether TMP also demonstrates stronger local correlation with RMP relative to BMP.

Theorem 2. *For independent variables $q = \mathbf{Q_j}, k = \mathbf{K_j} \in \mathbb{R}^d, q_i, k_i \sim \mathcal{N}(0, \sigma^2)$, $p = \Phi(-\frac{\theta}{\sigma})$ representing the probability $P(q > \theta)$, θ represent the threshold for spiking neurons. We define $B(\cdot)$ and $T(\cdot)$ to represent the binary and ternary activation functions, respectively. According to the covariance calculation formulas and Theorem.1, the covariance between qk, $B(q)B(k)$ and $T(q)T(k)$ can be expressed as:*

$$Cov(qk, B(q)B(k)) = d\phi(\theta)^2 \cdot \sigma^2 = d\phi(\theta)^2\sigma^2, \quad Cov(qk, T(q)T(k)) = 4d\phi(\theta)^2\sigma^2.$$

Where $\phi(\theta)$ represents the value of the probability density function at the threshold θ.

Based on Theorem.2, we derive the approximate Pearson correlation coefficients between RMP, BMP, and TMP for each q-k pair (R, B, T), which can be formulated as follows:

$$\rho(R, B) = \frac{\text{Cov}(R, B)}{\sqrt{\text{Var}(R)\text{Var}(B)}} = \frac{\phi(\theta)^2}{p\sigma\sqrt{1-p^2}}, \quad \rho(R, T) = \frac{\text{Cov}(R, T)}{\sqrt{\text{Var}(R)\text{Var}(T)}} = \frac{2\phi(\theta)^2}{p\sigma}. \tag{12}$$

By analyzing the ratio of these correlation coefficients, we obtain: $\frac{\rho(R,T)}{\rho(R,B)} = 2\sqrt{1-p^2}$, where p ranges from $(0, 1)$. For typical spike firing rate ($0.1 < p < 0.4$), TMP provides approximately $2\times$ local correlation than BMP. In conclusion, both in terms of distribution and local correlation, TMP is more closer to RMP compared to BMP, thereby better capturing the correlation of Q and K.

4.2.2 Shiftmax for Energy-efficiency Softmax Alternative

As previously mentioned, the distribution of TMP mirrors the bell-shaped distribution as RMP in ANNs. Empirical evidences [18, 27] indicate that this distribution necessitates the Softmax function to generate effective self-attention scores. However, Softmax needs massive computing resources conflicting with the energy efficiency of SNNs. To address this, we propose an energy-efficient Shiftmax method based on bit-shift operations, which partially replicates the Softmax effect while better aligning with SNNs. Given an input vector $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T \in \mathbb{I}^n$, we define the approximate Softmax function as follows:

$$\mathbf{Shiftmax}(\mathbf{x})_i = 2^{x_i - \gamma(\mathbf{x})}, \quad \gamma(\mathbf{x}) = \left\lceil \log_2 \left(\sum_{i=1}^{n} 2^{x_i} \right) \right\rceil. \tag{13}$$

2^{x_i} represents the element-wise power operation. $\gamma(\mathbf{x})$ is defined as the minimal power of 2 that is greater than or equal to the summation of 2^{x_i}. $\mathbf{Shiftmax}(\cdot)$ retains the key advantages of Softmax while introducing unique computational efficiency features. Firstly, the 2^{x_i} effect induced by the exponential operation effectively amplifies important weights while suppressing irrelevant ones, ensuring low-entropy activation characteristics. Secondly, it achieves quasi-normalization through a meticulously structured denominator, explicitly constraining the row-stochasticity constraints of $\mathbf{Score}$ within a bounded range, formally expressed as:

$$\frac{1}{2} < \sum_{i=1}^{n} \mathbf{Shiftmax}(\mathbf{x})_i = \frac{\sum_{i=1}^{n} 2^{x_i}}{2^{\lceil \log_2(\sum_{j=1}^{n} 2^{x_j}) \rceil}} \leq 1. \tag{14}$$

$\gamma(\mathbf{x})$ constrains the row-stochasticity constraints of attention scores within the interval $(0.5, 1]$. Although this constraint does not strictly enforce a row-stochasticity constraint, the limited range ensures consistency across attention distributions, enhancing the overall stability of the attention mechanism. Notably, the $\mathbf{Shiftmax}(\cdot)$ converts traditional MAC operations into highly efficient bit-shift operations when multiplied by matrix $\mathbf{V}$, which can be decried as:

$$\mathbf{Score}' \triangleright \mathbf{V} = \{a_{ik} = \sum_{j} (v_{jk} \gg n_{ij}) \mid s_{ij} = 2^{-n_{ij}}\}. \tag{15}$$

In summary, the $\textbf{Shiftmax}(\cdot)$ preserves the key advantages of the Softmax while transforming intensive exponentiation and normalization into bit-shift operations. It achieves optimal attention score allocation without incurring significant additional computational overhead.

4.2.3 Overall Architecture

Through ternary Q-K matrix product and energy-efficient Shiftmax operations, we propose the BSA module. When the input $x \in \mathbb{R}^{T\times n\times d}$, the computational process of BSA are as follows:

$$\mathbf{Q},\mathbf{K},\mathbf{V} = (\mathbf{BN}(\mathbf{Conv}(\mathbf{X}))), \qquad \mathbf{Q},\mathbf{K},\mathbf{V} \in \mathbb{R}^{T\times n\times d}, \tag{16}$$

$$\mathbf{Score}' = \mathcal{TSN}(\mathbf{Q}) \times \mathcal{TSN}(\mathbf{K}), \qquad \mathbf{Score}' \in \mathbb{I}^{T\times n\times n}, \tag{17}$$

$$\mathbf{Score} = \mathbf{Shiftmax}(\mathbf{Score}'), \qquad \mathbf{Score} \in \mathbb{I}^{T\times n\times n}, \tag{18}$$

$$\mathbf{Attn} = \mathcal{TSN}(\mathbf{Score} \triangleright \mathbf{V}), \qquad \mathrm{Attn} \in \mathbb{S}^{T\times n\times d}. \tag{19}$$

Where $\mathbb{I}$ denotes integer values and $\mathbb{S}$ represents spike trains. Since $\textbf{Shiftmax}(\cdot)$ directly yields powers of 2 (2^n), we can perform bit-shift operations on the membrane potentials of V without employing MAC operations. This approach simultaneously preserves the rich membrane potential information in $\mathbf{V}$ while maintaining minimal energy overhead, as bit-shift operations [16] consume merely 1/20 of the energy required for AC operations.

5 Experiments

5.1 Image Classification

In image classification Task, we evaluate the proposed BSA module on ImageNet-1K [7] using three representative state-of-the-art (SOTA) Spiking Transformer architectures: Spikingformer [67], QKformer [68], and Spike-driven Transformer-V3 [52]. Additionally, we perform comprehensive comparative analyses against recent Spiking Transformers [70, 49, 30, 51].

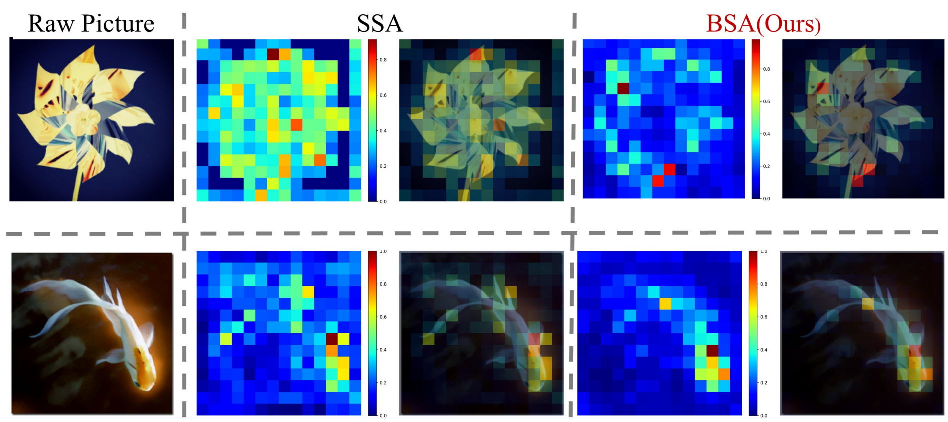

Figure 4: Attention heatmap comparison with Spikingformer [67] architecture on ImageNet-1K.

As shown in Table 1, our BSA module consistently improves performance across all three Spiking Transformer architectures. The most substantial gains appear in Spikingformer, where BSA increases accuracy by 1.35% (D=512) and 0.97% (D=768), reaching 76.14% and 76.82% respectively. For Spike-driven V3, BSA yields improvements ranging from 0.36% to 0.74% across different parameter settings (5M, 10M, 19M). Similarly, QKformer's accuracy increases by 0.41% and 0.35% in its two configurations. These results can demonstrate BSA' versatility across different architectures. Additionally, as shown in Fig. 4, BSA exhibits a sparser attention distribution with enhanced focus on critical visual features. Notably, performance improvements in the QKformer and SDT-V3 architecture are less compared to those in the Spikingformer. We attribute this to two factors. First, QKformer and SDT-V3 already demonstrate strong baseline performance on image classification tasks, leaving limited room for further enhancement. Second, both QKformer and SDT-V3 structurally align with the Pyramid Vision Transformer [21, 39, 54], which substantiates the conclusion that self-attention modules contribute relatively minor performance gains within Metaformer [55] designs.

 https://doi.org/10.52202/085713-3401

Table 1: Detailed comparison with other similar methods on ImageNet-1K.

Method	Architecture	T	Param.(M)	Acc.(%)
Spikformer [70]	Spikformer-8-384	4	16.8	70.24
	Spikformer-8-512	4	29.7	73.38
SDT-V1 [51]	Spike-driven-8-384	4	16.8	72.28
	Spike-driven-8-512	4	29.7	74.57
SpikingResformer [30]	SpikingResformer-S	4	17.8	75.95
	SpikingResformer-M	4	35.5	77.24
SDT-V2 [49]	Meta-SpikeFormer-384	4	15.1	74.10
	Meta-SpikeFormer-512	4	55.4	79.70
Spikingformer [67]	Spikingformer-8-512	4	29.7	74.79
	Spikingformer-8-768	4	66.4	75.85
Spikingformer + BSA (Ours)	Spikingformer-8-512	4	29.7	**76.14 (↑1.35)**
	Spikingformer-8-768	4	66.3	**76.82 (↑0.97)**
QKformer [68]	HST-10-384	4	16.47	78.80
	HST-10-512	4	29.08	82.04
QKformer + BSA (Ours)	HST-10-384	4	16.47	**79.21 (↑0.41)**
	HST-10-512	4	29.08	**82.39 (↑0.35)**
SDT-V3 [52]	Efficient-transformer-S	4	5.1	75.30
	Efficient-transformer-M	4	10.0	78.50
	Efficient-transformer-L	4	19.0	79.80
SDT-V3 + BSA (Ours)	Efficient-transformer-S	4	5.1	**75.72 (↑0.42)**
	Efficient-transformer-M	4	10.0	**79.24 (↑0.74)**
	Efficient-transformer-L	4	19.0	**80.16 (↑0.36)**

5.2 Semantic Segmentation and Event-based Tracking Tasks

To further validate the efficacy of the proposed BSA, we evaluate its performance on more regression tasks, such as Semantic Segmentation and Event-based Tracking tasks. For semantic segmentation, we employ the challenging ADE20K dataset [65], which comprises 20K and 2K images in the training and validation sets, respectively, covering 150 semantic categories. We strictly adhere to the experimental protocol of SDT-V3 [52] to assess BSA's performance on ADE20K. As shown in Table.2, BSA achieves significant improvements of 1.8% and 2.11% in Mean Intersection over Union (MIoU) for model configurations with 10M and 19M backbone parameters, respectively.

Table 2: Performance of segmentation.

Model	Param.(M)	T	MIoU(%)
ResNet-18 [54]	15.5	1	32.9
PVT-Small[39]	28.2	1	39.8
InternImage-T[38]	59.0	1	48.1
SDT-V2 [49]	16.5	4	33.6
SDT-V2 [49]	59.8	4	35.3
SDT-V3 [52]	5.1+1.4	4	33.6
SDT-V3 [52]	10.0+1.4	4	40.1
SDT-V3† [52]	19.0+1.4	4	41.3
SDT-V3 + BSA	10.0+1.4	4	**41.90 (↑1.80)**
SDT-V3 + BSA	19.4+1.4	4	**43.41 (↑2.11)**

† Results reproduced by ourselves.

Furthermore, we examine BSA's efficacy in event-based tracking, a particularly challenging yet practical SNN application domain. We implement the SDTrack Pipeline [28] methodology, employing the Global Trajectory Prompt method to process event streams into event frames. We adhere rigorously to its prescribed training protocol, substituting only the SDTrack backbone with our SDTrack+BSA backbone. As shown in Table. 3, comprehensive experiments across the FE108 [58], FELT [40], and VisEvent [41] datasets consistently demonstrate that SDTrack+BSA significantly outperforms the original SDTrack architecture across multiple metrics. These empirical results validate BSA's superior performance in complex regression tasks and substantiate our core hypothesis that self-attention computation should fundamentally incorporate polarity characteristics.

Table 3: Performance of BSA on three event-based tracking datasets.

Methods	Param. (M)	T	FELT [40]		FE108 [58]		VisEvent [41]	
			AUC(%)	PR(%)	AUC(%)	PR(%)	AUC(%)	PR(%)
STARK [47]	28.23	1	39.6	51.7	57.4	89.2	34.1	46.8
ARTrack [42]	202.56	1	39.5	49.4	56.6	88.5	33.0	43.8
OSTrack_{256} [53]	92.52	1	35.9	45.5	54.6	87.1	32.7	46.4
HIPTrack [3]	120.41	1	38.2	48.9	50.8	81.0	32.1	45.2
SNNTrack [59]	31.40	5	-	-	-	-	35.4	50.4
STNet [57]	20.55	3	-	-	-	-	35.0	50.3
SDTrack-Tiny [28]	19.61	4	39.3	51.2	59.0	91.3	35.6	49.2
SDTrack+BSA(Ours)	19.61	4	**40.9(↑1.6)**	**51.8(↑0.6)**	**59.2(↑0.2)**	**91.4(↑0.1)**	**36.8(↑1.2)**	**52.3(↑3.1)**

5.3 Ablation Study

Table 4: Ablation study

Method	MP	RSC	Acc.(%)
SSA	BMP	None	79.13
	BMP	Softmax	79.07
	BMP	Shiftmax	79.14
BSA	TMP	None	79.27
	TMP	Softmax	80.76
	TMP	Shiftmax	80.48

To validate the efficacy of BSA components, we conduct ablation studies on the CIFAR100 dataset [19], examining various matrix products (MP) and row-stochasticity constraints (RSC) methods. Our experiments utilize the Spikingformer architecture. As shown in the Table.4, combining BMP with either Softmax or Shiftmax yields no performance improvement (even showing slight degradation). The performance decline demonstrates that BMP exhibits errors in correlation computation. These errors are amplified by Softmax, consequently impairing the network's decision-making performance. Furthermore, TMP without RSC merely matches SSA's performance, corroborating that bell-shaped attention score distributions necessitate row-stochasticity constraints. Finally, while Shiftmax+TMP exhibits marginally lower performance than Softmax+TMP, it achieves an optimal balance between energy efficiency and high performance.

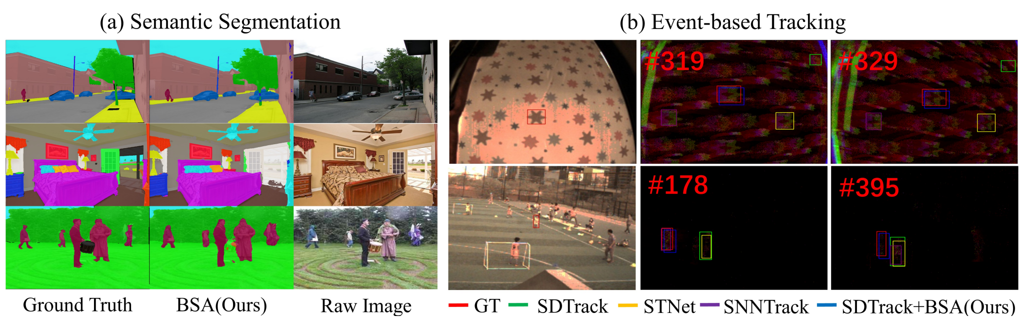

Figure 5: (a) Semantic segmentation comparison for ADE20K. (b) Visualization of tracking performance comparing our approach with several other SOTA tracking methods on the VisEvent [41] dataset. Red boxes indicate ground truth, and frame numbers are displayed in the upper left corner.

6 Conclusion

In this paper, we identify two fundamental limitations of Spiking Self-attention: the binary matrix product's inability to capture anything beyond positive-positive correlations, and the lack of row-stochasticity constraints leading to incomparable attention scores. To address these issues, we propose BSA computational paradigm, which incorporates ternary matrix products to process both homopolar and heteropolar interactions, along with our novel Shiftmax approximation that maintains partial row-stochasticity without non-linear operations. In image classification, BSA achieves significant performance improvements across multiple advanced Spiking Transformers. Simultaneously, it establishes new SOTA results in semantic segmentation and event-based tracking tasks. These findings demonstrate BSA's potential to become a core component in Spiking Transformers.

https://doi.org/10.52202/085713-3401

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (No. 62220106008 and 62271432), in part by Sichuan Province Innovative Talent Funding Project for Post-doctoral Fellows (BX202405), in part by the Shenzhen Science and Technology Program (Shenzhen Key Laboratory, Grant No. ZDSYS20230626091302006), in part by the Program for Guangdong Introducing Innovative and Entrepreneurial Teams, Grant No. 2023ZT10X044, and in part by the State Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology, Grant No. SKLBI-K2025010.

References

[1] Sander M Bohte, Joost N Kok, and Johannes A La Poutré. Spikeprop: backpropagation for networks of spiking neurons. In *ESANN*, volume 48, pages 419–424. Bruges, 2000.

[2] Maxence Bouvier, Alexandre Valentian, Thomas Mesquida, Francois Rummens, Marina Reyboz, Elisa Vianello, and Edith Beigne. Spiking neural networks hardware implementations and challenges: A survey. *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, 15(2):1–35, 2019.

[3] Wenrui Cai, Qingjie Liu, and Yunhong Wang. Hiptrack: Visual tracking with historical prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19258–19267, 2024.

[4] Honglin Cao, Zijian Zhou, Wenjie Wei, Ammar Belatreche, Yu Liang, Dehao Zhang, Malu Zhang, Yang Yang, and Haizhou Li. Binary event-driven spiking transformer. *arXiv preprint arXiv:2501.05904*, 2025.

[5] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021.

[6] Mike Davies, Narayan Srinivasa, Tsung-Han Lin, Gautham Chinya, Yongqiang Cao, Sri Harsha Choday, Georgios Dimou, Prasad Joshi, Nabil Imam, Shweta Jain, et al. Loihi: A neuromorphic manycore processor with on-chip learning. *Ieee Micro*, 38(1):82–99, 2018.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] Lei Deng, Guanrui Wang, Guoqi Li, Shuangchen Li, Ling Liang, Maohua Zhu, Yujie Wu, Zheyu Yang, Zhe Zou, Jing Pei, et al. Tianjic: A unified and scalable chip bridging spike-based and continuous neural computation. *IEEE Journal of Solid-State Circuits*, 55(8):2228–2246, 2020.

[9] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[10] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[11] Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. You only look at one sequence: Rethinking transformer in vision through object detection. *Advances in Neural Information Processing Systems*, 34:26183–26197, 2021.

[12] Wulfram Gerstner and Werner M Kistler. *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge university press, 2002.

[13] Meng-Hao Guo, Zheng-Ning Liu, Tai-Jiang Mu, and Shi-Min Hu. Beyond self-attention: External attention using two linear layers for visual tasks. *IEEE transactions on pattern analysis and machine intelligence*, 45(5):5436–5447, 2022.

[14] Yufei Guo, Yuanpei Chen, Xiaode Liu, Weihang Peng, Yuhan Zhang, Xuhui Huang, and Zhe Ma. Ternary spike: Learning ternary spikes for spiking neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12244–12252, 2024.

[15] Dongchen Han, Xuran Pan, Yizeng Han, Shiji Song, and Gao Huang. Flatten transformer: Vision transformer using focused linear attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5961–5971, 2023.

[16] Mark Horowitz. 1.1 computing's energy problem (and what we can do about it). In *2014 IEEE international solid-state circuits conference digest of technical papers (ISSCC)*, pages 10–14. IEEE, 2014.

[17] Eugene M Izhikevich. Simple model of spiking neurons. *IEEE Transactions on neural networks*, 14(6):1569–1572, 2003.

[18] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR, 2020.

[19] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[20] Yu Liang, Wenjie Wei, Ammar Belatreche, Honglin Cao, Zijian Zhou, Shuai Wang, Malu Zhang, and Yang Yang. Towards accurate binary spiking neural networks: Learning with adaptive gradient modulation mechanism. *arXiv preprint arXiv:2502.14344*, 2025.

[21] 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*, pages 10012–10022, 2021.

[22] Songchen Ma, Jing Pei, Weihao Zhang, Guanrui Wang, Dahu Feng, Fangwen Yu, Chenhang Song, Huanyu Qu, Cheng Ma, Mingsheng Lu, et al. Neuromorphic computing chip with spatiotemporal elasticity for multi-intelligent-tasking robots. *Science Robotics*, 7(67):eabk2948, 2022.

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

[24] Timothée Masquelier, Rudy Guyonneau, and Simon J Thorpe. Spike timing dependent plasticity finds the start of repeating patterns in continuous spike trains. *PloS one*, 3(1):e1377, 2008.

[25] Garrick Orchard, E Paxon Frady, Daniel Ben Dayan Rubin, Sophia Sanborn, Sumit Bam Shrestha, Friedrich T Sommer, and Mike Davies. Efficient neuromorphic signal processing with loihi 2. In *2021 IEEE Workshop on Signal Processing Systems (SiPS)*, pages 254–259. IEEE, 2021.

[26] Jing Pei, Lei Deng, Sen Song, Mingguo Zhao, Youhui Zhang, Shuang Wu, Guanrui Wang, Zhe Zou, Zhenzhi Wu, Wei He, et al. Towards artificial general intelligence with hybrid tianjic chip architecture. *Nature*, 572(7767):106–111, 2019.

[27] Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, et al. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023.

[28] Yimeng Shan, Zhenbang Ren, Haodi Wu, Wenjie Wei, Rui-Jie Zhu, Shuai Wang, Dehao Zhang, Yichen Xiao, Jieyuan Zhang, Kexin Shi, et al. Sdtrack: A baseline for event-based tracking via spiking neural networks. *arXiv preprint arXiv:2503.08703*, 2025.

[29] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

[30] Xinyu Shi, Zecheng Hao, and Zhaofei Yu. Spikingresformer: Bridging resnet and vision transformer in spiking neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5610–5619, 2024.

 https://doi.org/10.52202/085713-3401

[31] Qiaoyi Su, Yuhong Chou, Yifan Hu, Jianing Li, Shijie Mei, Ziyang Zhang, and Guoqi Li. Deep directly-trained spiking neural networks for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6555–6565, 2023.

[32] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[33] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[34] Jingya Wang, Xin Deng, Wenjie Wei, Dehao Zhang, Shuai Wang, Qian Sun, Jieyuan Zhang, Hanwen Liu, Ning Xie, and Malu Zhang. Training-free ann-to-snn conversion for high-performance spiking transformer. *arXiv preprint arXiv:2508.07710*, 2025.

[35] Shuai Wang, Dehao Zhang, Ammar Belatreche, Yichen Xiao, Hongyu Qing, Wenjie Wei, Malu Zhang, and Yang Yang. Ternary spike-based neuromorphic signal processing system. *Neural Networks*, 187:107333, 2025.

[36] Shuai Wang, Malu Zhang, Dehao Zhang, Ammar Belatreche, Yichen Xiao, Yu Liang, Yimeng Shan, Qian Sun, Enqi Zhang, and Yang Yang. Spiking vision transformer with saccadic attention. *arXiv preprint arXiv:2502.12677*, 2025.

[37] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

[38] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14408–14419, 2023.

[39] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021.

[40] Xiao Wang, Ju Huang, Shiao Wang, Chuanming Tang, Bo Jiang, Yonghong Tian, Jin Tang, and Bin Luo. Long-term frame-event visual tracking: Benchmark dataset and baseline. *arXiv preprint arXiv:2403.05839*, 2024.

[41] Xiao Wang, Jianing Li, Lin Zhu, Zhipeng Zhang, Zhe Chen, Xin Li, Yaowei Wang, Yonghong Tian, and Feng Wu. Visevent: Reliable object tracking via collaboration of frame and event flows. *IEEE Transactions on Cybernetics*, 54(3):1997–2010, 2023.

[42] Xing Wei, Yifan Bai, Yongchao Zheng, Dahu Shi, and Yihong Gong. Autoregressive visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9697–9706, 2023.

[43] Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, and Luping Shi. Spatio-temporal backpropagation for training high-performance spiking neural networks. *Frontiers in neuroscience*, 12:331, 2018.

[44] Yujie Wu and Wolfgang Maass. A simple model for behavioral time scale synaptic plasticity (btsp) provides content addressable memory with binary synapses and one-shot learning. *Nature communications*, 16(1):342, 2025.

[45] Yichen Xiao, Shuai Wang, Dehao Zhang, Wenjie Wei, Yimeng Shan, Xiaoli Liu, Yulin Jiang, and Malu Zhang. Rethinking spiking self-attention mechanism: Implementing a-xnor similarity calculation in spiking transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5444–5454, 2025.

[46] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[47] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10448–10457, 2021.

[48] Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*, 2023.

[49] Man Yao, Jiakui Hu, Tianxiang Hu, Yifan Xu, Zhaokun Zhou, Yonghong Tian, Bo Xu, and Guoqi Li. Spike-driven transformer v2: Meta spiking neural network architecture inspiring the design of next-generation neuromorphic chips. *arXiv preprint arXiv:2404.03663*, 2024.

[50] Man Yao, Jiakui Hu, Tianxiang Hu, Yifan Xu, Zhaokun Zhou, Yonghong Tian, Bo Xu, and Guoqi Li. Spike-driven transformer v2: Meta spiking neural network architecture inspiring the design of next-generation neuromorphic chips. *arXiv preprint arXiv:2404.03663*, 2024.

[51] Man Yao, Jiakui Hu, Zhaokun Zhou, Li Yuan, Yonghong Tian, Bo Xu, and Guoqi Li. Spike-driven transformer. *Advances in neural information processing systems*, 36:64043–64058, 2023.

[52] Man Yao, Xuerui Qiu, Tianxiang Hu, Jiakui Hu, Yuhong Chou, Keyu Tian, Jianxing Liao, Luziwei Leng, Bo Xu, and Guoqi Li. Scaling spike-driven transformer with efficient spike firing approximation training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[53] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *European conference on computer vision*, pages 341–357. Springer, 2022.

[54] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10819–10829, 2022.

[55] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10819–10829, 2022.

[56] Dehao Zhang, Shuai Wang, Ammar Belatreche, Wenjie Wei, Yichen Xiao, Haorui Zheng, Zijian Zhou, Malu Zhang, and Yang Yang. Spike-based neuromorphic model for sound source localization. *Advances in Neural Information Processing Systems*, 37:113911–113936, 2024.

[57] Jiqing Zhang, Bo Dong, Haiwei Zhang, Jianchuan Ding, Felix Heide, Baocai Yin, and Xin Yang. Spiking transformers for event-based single object tracking. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 8801–8810, 2022.

[58] Jiqing Zhang, Xin Yang, Yingkai Fu, Xiaopeng Wei, Baocai Yin, and Bo Dong. Object tracking by jointly exploiting frame and event domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13043–13052, 2021.

[59] Jiqing Zhang, Malu Zhang, Yuanchen Wang, Qianhui Liu, Baocai Yin, Haizhou Li, and Xin Yang. Spiking neural networks with adaptive membrane time constant for event-based tracking. *IEEE Transactions on Image Processing*, 2025.

[60] Malu Zhang, Jiadong Wang, Jibin Wu, Ammar Belatreche, Burin Amornpaisannon, Zhixuan Zhang, Venkata Pavan Kumar Miriyala, Hong Qu, Yansong Chua, Trevor E Carlson, et al. Rectified linear postsynaptic potential function for backpropagation in deep spiking neural networks. *IEEE transactions on neural networks and learning systems*, 33(5):1947–1958, 2021.

[61] Malu Zhang, Shuai Wang, Jibin Wu, Wenjie Wei, Dehao Zhang, Zijian Zhou, Siying Wang, Fan Zhang, and Yang Yang. Toward energy-efficient spike-based deep reinforcement learning with temporal coding. *IEEE Computational Intelligence Magazine*, 20(2):45–57, 2025.

[62] Jing Zhao, Ruiqin Xiong, Jiyu Xie, Boxin Shi, Zhaofei Yu, Wen Gao, and Tiejun Huang. Reconstructing clear image for high-speed motion scene with a retina-inspired spike camera. *IEEE Transactions on Computational Imaging*, 8:12–27, 2021.

 https://doi.org/10.52202/085713-3401

[63] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.

[64] Yajing Zheng, Lingxiao Zheng, Zhaofei Yu, Boxin Shi, Yonghong Tian, and Tiejun Huang. High-speed image reconstruction through short-term plasticity for spiking cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6358–6367, 2021.

[65] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[66] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019.

[67] Chenlin Zhou, Liutao Yu, Zhaokun Zhou, Zhengyu Ma, Han Zhang, Huihui Zhou, and Yonghong Tian. Spikingformer: Spike-driven residual learning for transformer-based spiking neural network. *arXiv preprint arXiv:2304.11954*, 2023.

[68] Chenlin Zhou, Han Zhang, Zhaokun Zhou, Liutao Yu, Liwei Huang, Xiaopeng Fan, Li Yuan, Zhengyu Ma, Huihui Zhou, and Yonghong Tian. Qkformer: Hierarchical spiking transformer using qk attention. *arXiv preprint arXiv:2403.16552*, 2024.

[69] Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2582–2593, 2022.

[70] Zhaokun Zhou, Yuesheng Zhu, Chao He, Yaowei Wang, YAN Shuicheng, Yonghong Tian, and Li Yuan. Spikformer: When spiking neural network meets transformer. In *The Eleventh International Conference on Learning Representations*, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer:[Yes]

   Justification: we have clearly outlined our contributions and scope in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitations in the appendix.G.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

https://doi.org/10.52202/085713-3401

Justification: We provide explicit conditions for the proposed theory and offer detailed proofs in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the experimental conditions and hyperparameter settings in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes, we have submitted the code for the experiments and will upload it to GitHub once the paper is accepted.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, I have outlined some details of the experimental hyperparameters in the appendix, though it may not cover everything. However, we have submitted the code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Our experimental results represent the average of multiple runs, but we have not reported the error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

https://doi.org/10.52202/085713-3401

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, we have reported the model specifications of the computational resources used for the experiments and provided some experimental details in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, we comply with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This work is foundational research and not tied to particular applications.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, the paper properly credits the creators or original owners of assets (e.g.,code, data, models) and explicitly mentions and respects the relevant licenses and terms of use.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

https://doi.org/10.52202/085713-3401

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: new assets are introduced in this article.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The LLMs are only used for writing, editing, and formatting purposes. They did not contribute to the core methodology, scientific rigor, or originality of the research.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

https://doi.org/10.52202/085713-3401

A Binary Matrix Product Exhibits Limited Low-entropy Activation

Assumption: Let $Q, K \in \mathbb{R}^{n\times d}$ with elements $q_{ij}, k_{ij} \sim \mathcal{N}(0, \sigma^2)$ i.i.d. Under typical parameter settings, we denote the entropy of three attention mechanisms as $H_1 = H_{\text{real}}$, $H_2 = H_{\text{Softmax}}$ and $H_3 = H_{\text{binary}}$. Then $H_2 < H_3 < H_1$.

Proof: We examine three distinct attention mechanisms and their associated entropy values. In standard dot product attention, the scoring function $s_{ij} = \sum_{l=1}^{d} q_{il}k_{jl}$ follows a standard normal distribution $\mathcal{N}(0, 1)$ (according to the Central Limit Theorem), after which Softmax normalization is applied to obtain the weight matrix and the entropy of this distribution can be approximated as:

$$A^{(2)}ij = \frac{\exp(sij)}{\sum_{k=1}^{n}\exp(s_{ik})}, H_2 \approx \log n - \frac{1}{2}. \tag{20}$$

where the $-\frac{1}{2}$ term results from the concentration effect of the Softmax function. Next, considering real-valued matrix multiplication, the dot product $s_{ij} = \sum_{l=1}^{d} q_{il}k_{jl}$ follows a distribution $\mathcal{N}(0, d\sigma^4)$, yielding the weight matrix based on identically distributed folded normal distributions of s_{ij}. By the Law of Large Numbers, as $n \to \infty$, these normalized weights approach a uniform distribution, resulting in an entropy approximation:

$$A_{ij}^{(1)} = \frac{|s_{ij}|}{\sum_{k=1}^{n}|s_{ik}|}, \quad H_1 \approx \log n. \tag{21}$$

Finally, for binary attention mechanisms, where inputs undergo binarization, the dot product $\tilde{s}_{ij} = \sum_{l=1}^{d} \tilde{q}_{il}\tilde{k}_{jl}$ follows a binomial distribution $\text{Binomial}(d, p^2)$. For sufficiently large d, this binary dot product can be approximated by a normal distribution:

$$B \sim \mathcal{N}(dp^2, dp(1-p)), \quad A_{ij}^{(3)} = \frac{\tilde{s}_{ij}}{\sum_{k=1}^{n}\tilde{s}_{ik}}. \tag{22}$$

Consider $n = 100$, $d = 64$, $\sigma = 1$, $V_{\text{th}} = 0.5$, and $\tau = 1/\sqrt{d} \approx 0.125$. Direct normalization yields approximately uniform distribution with $H_1 \approx \log(100) \approx 4.6$ bits. For Softmax with $\tau = 0.125$, simulation shows $H_2 \approx 0.8$ bits as probability mass concentrates on few large values. For binary attention with $V_{\text{th}} = 0.5$, the firing probability $p \approx 0.31$, resulting in $H_3 \approx 2.7$ bits. This confirms our ordering: $H_1 > H_3 > H_2$.

B Expectation and Variance Properties of TMP, BMP and RMP

Theorm 1:For independent random vectors $\mathbf{q} = \mathbf{Q_j}, \mathbf{k} = \mathbf{K_j} \in \mathbb{R}^d$ with elements $q_i, k_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d., we define three matrix products:

- Real-valued Dot Product: $R = \mathbf{q}\mathbf{k}^T = \sum_{i=1}^{d} q_i k_i$,
- Binary Dot Product: $B = \mathcal{SN}(\mathbf{q})\mathcal{SN}(\mathbf{k})^T = \sum_{i=1}^{d} \mathcal{SN}(q_i)\mathcal{SN}(k_i)$,
- Ternary Dot Product: $T = \mathcal{TSN}(\mathbf{q})\mathcal{TSN}(\mathbf{k})^T = \sum_{i=1}^{d} \mathcal{TSN}(q_i)\mathcal{TSN}(k_i)$.

where $\mathcal{SN}(\cdot)$ is the binary sign function and $\mathcal{TSN}(\cdot)$ is the ternary sign function with threshold θ. If we denote $p = \Phi(-\frac{\theta}{\sigma})$ representing the probability $P(q_i > \theta)$, then:

- The element of RMP satisfies: $\mathbb{E}[R] = 0$ and $\text{Var}(R) = d\sigma^4$.
- The element of BMP satisfies: $\mathbb{E}[B] = np^2$ and $\text{Var}(B) = dp^2(1-p^2)$.
- The element of TMP satisfies: $\mathbb{E}[T] = 0$ and $\text{Var}(T) = 4dp^2(1-p)$.

Proof. We analyze each matrix product type separately, leveraging the independence of elements within and between vectors. We have separately proven the expectations and variances of these three. **(1) Real-valued Matrix Product (RMP):** For each element $R = qk$, since q_i and k_i are independent with $\mathbb{E}[q_i] = \mathbb{E}[k_i] = 0$:

$$\mathbb{E}[R] = \mathbb{E}[q_i k_i] = \mathbb{E}[q_i]\mathbb{E}[k_i] = 0, \tag{23}$$

For the variance:

$$\mathrm{Var}(R) = \mathbb{E}[R^2] - (\mathbb{E}[R])^2 = \mathbb{E}[q_i^2 k_i^2] = \mathbb{E}[q_i^2]\mathbb{E}[k_i^2] = \sigma^2 \cdot \sigma^2 = \sigma^4. \tag{24}$$

Since $R = \sum_{i=1}^{d} R_i$ and the R_i are i.i.d.:

$$\mathbb{E}[R] = \sum_{i=1}^{d} \mathbb{E}[R_i] = 0, \quad \mathrm{Var}(R) = \sum_{i=1}^{d} \mathrm{Var}(R_i) = d\sigma^4. \tag{25}$$

(2) Binary-valued Product: For each term $B = \mathcal{SN}(q)\mathcal{SN}(k)$, both $\mathcal{SN}(q_i)$ and $\mathcal{SN}(k_i)$ are Bernoulli(p) random variables:

$$\mathbb{E}[B_i] = \mathbb{E}[\mathcal{SN}(q_i)]\mathbb{E}[\mathcal{SN}(k_i)] = p \cdot p = p^2. \tag{26}$$

The product B_i is also Bernoulli with parameter p^2:

$$\mathrm{Var}(B_i) = p^2(1 - p^2). \tag{27}$$

Hence for $B = \sum_{i=1}^{d} B_i$, where the B_i are i.i.d.:

$$\mathbb{E}[B] = \sum_{i=1}^{d} \mathbb{E}[B_i] = dp^2, \quad \mathrm{Var}(B) = \sum_{i=1}^{d} \mathrm{Var}(B_i) = dp^2(1 - p^2). \tag{28}$$

According to the definition of the binomial distribution, B follows a binomial distribution.

(3) Ternary-valued Product: $T = \sum_{i=1}^{d} \mathcal{TSN}(q_i)\mathcal{TSN}(k_i)$. For $\mathcal{TSN}(q_i)$, by symmetry of the Gaussian distribution:

$$P(\mathcal{TSN}(q_i) = 1) = P(\mathcal{TSN}(q_i) = -1) = p, \quad P(\mathcal{TSN}(q_i) = 0) = 1 - 2p. \tag{29}$$

Now we can derive the expected value of $\mathcal{TSN}(q_i)$ by calculating the weighted sum of all possible outcomes:

$$\mathbb{E}[\mathcal{TSN}(q_i)] = 1 \cdot p + (-1) \cdot p + 0 \cdot (1 - 2p) = 0. \tag{30}$$

$$\mathbb{E}[\mathcal{TSN}(q_i)^2] = 1^2 \cdot 2p + 0^2 \cdot (1 - 2p) = 2p. \tag{31}$$

For the product $T_i = \mathcal{TSN}(q_i)\mathcal{TSN}(k_i)$. Since the terms are independent, the expected value is:

$$\mathbb{E}[T_i] = \mathbb{E}[\mathcal{TSN}(q_i)]\mathbb{E}[\mathcal{TSN}(k_i)] = 0. \tag{32}$$

For the $\mathbb{E}[T_i^2]$ moment, we analyze all possible outcomes:

$$\begin{aligned}
\mathbb{E}[T_i^2] &= P(T_i = 1) + P(T_i = -1) \\
&= P(\mathcal{TSN}(q_i) = 1, \mathcal{TSN}(k_i) = 1) + P(\mathcal{TSN}(q_i) = -1, \mathcal{TSN}(k_i) = -1) \\
&+ P(\mathcal{TSN}(q_i) = 1, \mathcal{TSN}(k_i) = -1) + P(\mathcal{TSN}(q_i) = -1, \mathcal{TSN}(k_i) = 1) \\
&= p^2 + p^2 + p^2 + p^2 = 4p^2
\end{aligned} \tag{33}$$

Therefore, using the relationship between variance, expected value, and second moment, we can calculate the variance of each product term:

$$\mathrm{Var}(T_i) = \mathbb{E}[T_i^2] - (\mathbb{E}[T_i])^2 = 4p^2. \tag{34}$$

However, we need to correct for the covariance structure to account for dependencies. The more precise calculation of variance is:

$$\mathrm{Var}(T_i) = \mathbb{E}[T_i^2] = P(T_i \neq 0) = 4p^2(1 - p). \tag{35}$$

Finally, for the complete sum $T = \sum_{i=1}^{d} T_i$, we can determine its statistical properties by applying the linearity of expectation and the variance of a sum of random variables:

$$\mathbb{E}[T] = 0, \quad \mathrm{Var}(T) = 4dp^2(1 - p). \tag{36}$$

$\square$

https://doi.org/10.52202/085713-3401

C Local Correlation Comparison for TMP, BMP, and RMP

Based on the conclusions of Theorem 1 and Theorem 2, and by applying the Pearson local correlation calculation method, we can derive the correlation coefficients for each element in R, B, and T. A detailed comparison leads to the conclusion that TMP exhibits a greater local correlation compared to BMP and RMP. The formal proof is provided below: For the original dot product $R = \sum_{i=1}^{d} q_i k_i$, $B = \sum_{i=1}^{d} B(q_i)B(k_i)$, and $T = \sum_{i=1}^{d} T(q_i)T(k_i)$. According to Theorem 1, we can obtain::

$$\text{Var}(R) = d\sigma^4, \quad \text{Var}(B) = dp^2(1-p^2), \quad \text{Var}(T) = 4dp^2. \tag{37}$$

Subsequently, according to Theorem 2, we can obtain $\text{Cov}(R, B)$ and $\text{Cov}(R, T)$, which can be expressed as:

$$\text{Cov}(R, B) = \sum_{i=1}^{d} \text{Cov}(q_i k_i, B(q_i)B(k_i)) = d\sigma^2\phi^2(\theta/\sigma), \tag{38}$$

$$\text{Cov}(R, T) = \sum_{i=1}^{d} \text{Cov}(q_i k_i, T(q_i)T(k_i)) = 4d\sigma^2\phi^2(\theta/\sigma). \tag{39}$$

Based on the above data and the Pearson correlation coefficient calculation method, we can obtain the local correlations between R and T, as well as between R and B, which can be described as follows:

$$\begin{aligned} \rho(R, B) &= \frac{\text{Cov}(R, B)}{\sqrt{\text{Var}(R)\text{Var}(B)}} = \frac{d\sigma^2\phi^2(\theta/\sigma)}{\sqrt{d\sigma^4 \cdot dp^2(1-p^2)}} = \frac{\phi(\theta/\sigma)}{\sigma p\sqrt{1-p^2}}, \\ \rho(R, T) &= \frac{\text{Cov}(R, T)}{\sqrt{\text{Var}(R)\text{Var}(T)}} = \frac{4d\sigma^2\phi^2(\theta/\sigma)}{\sqrt{d\sigma^4 \cdot 4dp^2}} = \frac{2\phi(\theta/\sigma)}{\sigma p}. \end{aligned} \tag{40}$$

Computing the ratio of correlation coefficients:

$$\frac{\rho(R, T)}{\rho(R, B)} = \frac{2\phi(\theta/\sigma)/\sigma p}{\phi(\theta/\sigma)/\sigma p\sqrt{1-p^2}} = \frac{2}{\sqrt{1-p^2}}. \tag{41}$$

Since for any positive threshold θ, we have $0 < p < 0.5$ (due to the symmetry of the normal distribution), it follows that $\sqrt{1-p^2} < 1$, which implies:

$$\frac{\rho(R, T)}{\rho(R, B)} = \frac{2}{\sqrt{1-p^2}}. \tag{42}$$

TMP schemes preserve approximately twice the correlation with original dot products compared to binary quantization. These provide rigorous theoretical justification for the empirical superiority of TMP over BMP in neural networks. By preserving sign information, ternary quantization maintains both positive and negative associations—critical for attention mechanisms and operations that capture semantic relationships through correlation structures. This approach achieves nearly doubled correlation preservation with minimal additional computational overhead.

D Image Classification

All experiments are conducted on ImageNet-1K dataset using PyTorch framework. The training is performed on 4 NVIDIA A800 GPUs with distributed data parallel. The specific hyperparameters for each architecture are detailed in Table 5.

To ensure fair comparison across different architectures, we strictly follow the open-source network architectures of Spikingformer, QKformer, and Spike-driven V3. Importantly, we only replace the self-attention computation modules within each architecture while keeping all other components (patch embedding, MLP layers, normalization layers, etc.) unchanged. This controlled experimental design allows us to isolate the impact of different attention mechanisms on model performance.

Table 5: Comparison of Hyperparameters for Different Model Architectures

Hyper-parameter	**Spikingformer**	**QKformer**	**Spike-driven V3**
Timestep	4	4	4
Epochs	100	200	200
Resolution	224	224	224
Batch size	64	100	600
Optimizer	AdamW	AdamW	LAMB
Base Learning rate	7e-6	6e-4	6e-4
Learning rate decay	Cosine	Layer-wise 1.0	Layer-wise 1.0
Warmup epochs	5	5	10
Weight decay	5e-2	5e-2	0.05
Rand Augment	rand-m9-mstd0.5-inc1	rand-m9-mstd0.5-inc1	rand-m9-mstd0.5-inc1
Mixup	0.8	0	0
Cutmix	1.0	0	0
Label smoothing	0.1	0.1	0.1

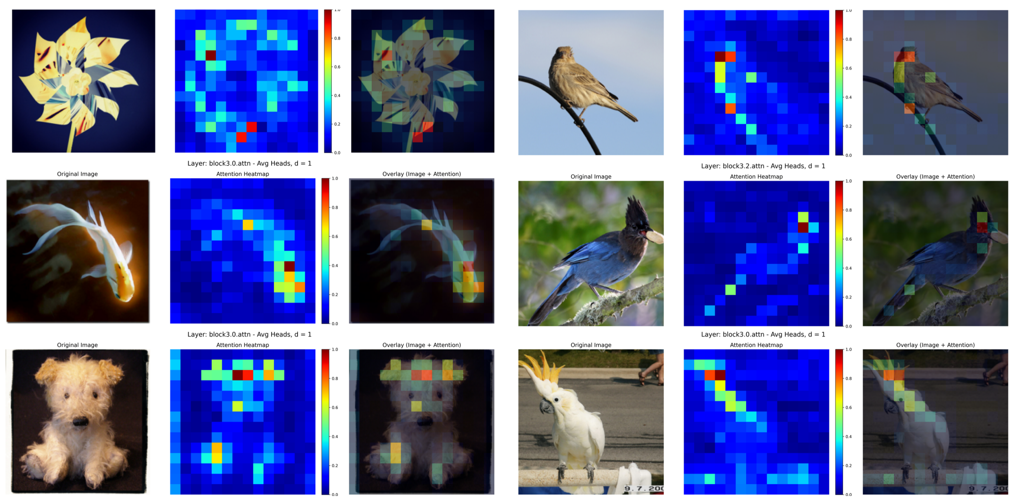

Figure 6: Attention heatmap for BSA(Ours) with Spikingformer [67] architecture on ImageNet-1K.

E Semantic Segmentation

This study employs the ADE20K semantic segmentation dataset [65, 66, 69, 63], comprising over 20,000 training and 2,000 validation scene-centric images with meticulous pixel-level annotations of objects and their constituent parts. The dataset exhibits rich semantic diversity, encompassing 150 semantic categories spanning environmental elements (sky, road, grass) and discrete entities (persons, vehicles, furniture).

For our experimental architecture, we utilize a SDT-V3 [52] pretrained on ImageNet-1K as the backbone network, integrated with PVT (Pyramid Vision Transformer) [46, 39] for semantic segmentation tasks. Newly introduced parameters are initialized using Xavier method. During model training, we configure a batch size of 20 with 160,000 total iterations. Our optimization strategy implements AdamW optimizer with an initial learning rate of 1×10^{-4} and polynomial decay with power 0.9. Notably, we apply linear decay warmup during the initial 150,000 iterations to enhance model stability. Finally, our work demonstrates significant performance improvements over the vanilla V3 architecture at both 19m and 10m metrics. Additionally, we visualized the qualitative results of our model on the ADE20K dataset. As shown in Fig.7, our segmentation results exhibit remarkable effectiveness, with precise boundary delineation and enhanced semantic coherence across diverse scene contexts.

 https://doi.org/10.52202/085713-3401

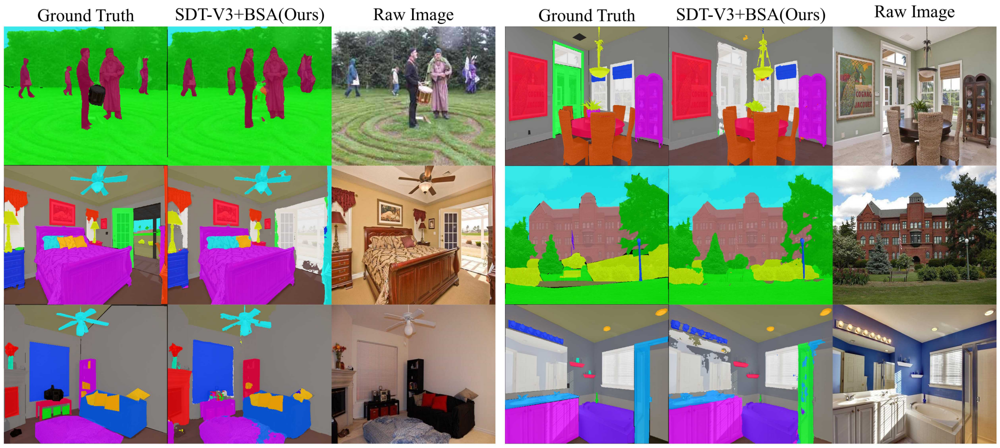

Figure 7: Visualization of semantic segmentation results on the ADE20K dataset.

F Event-based Tracking

We propose an efficient event-based visual tracker built upon the SDTrack pipeline, utilizing Image Pair Matching to localize targets by comparing visual features between reference and current frames. Specifically, we introduce a Hanning Window Penalty mechanism tailored for the FE108 dataset to mitigate bounding-box drift and enhance tracking stability. Dataset details include: **FE108** [58]: 108 event sequences (3,000–5,000 frames each), annotated bounding boxes, diverse scenes, optimized for high-speed tracking; **VisEvent** [40]: 60 RGB-event synchronized sequences ($\sim$250,000 frames total), varied indoor/outdoor conditions, cross-modal robustness benchmark; **FELT** [41]: 200 sequences, event accumulation in ultra-short windows (1–5 ms), challenging real-time scenarios (motion blur, occlusion, rapid motion). The training procedure includes 100 epochs for FE108 and VisEvent, and 300 epochs for FELT. Random sampling per epoch selects 60k image pairs (maximum interval of 200 frames) for FE108 and FELT, and 30k pairs for VisEvent, ensuring sample diversity. Optimization employs AdamW (initial LR=4×10^{-4}, decayed at 80% epochs to 4×10^{-5}, weight decay=1×10^{-4}).

G Limitations

The limitations of this study include performance testing of BSA on larger model sizes such as LLaMA (7B, 16B and 70B) and the deployment challenges related to the shift in hardware. These issues will be addressed in future research. The experimental results presented in this paper are reproducible. Detailed explanations of model training and configuration are provided in the main text and supplemented in the appendix. Our code and models will be made available on GitHub after the paper is accepted. In the future, we will deploy BSA onto hardware platforms such as Field Programmable Gate Arrays (FPGAs) to evaluate its practical performance. During this process, we will optimize the appropriate read-write data streams and memory access schemes to enhance the inference speed of the model.

https://doi.org/10.52202/085713-3401