Temporal Dynamics Enhancer for Directly Trained Spiking Object Detectors

Fan Luo[1,3], Zeyu Gao[1,3], Xinhao Luo[1,3], Kai Zhao[2], Yanfeng Lu[1,3*]
[1]State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China
[2] Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China
[3]School of Artificial Intelligence,
University of Chinese Academy of Sciences, Beijing 100049, China
{luofan2024, gaozeyu2023, luoxinhao2023, kai.zhao}@ia.ac.cn,
yanfeng.lv@ia.ac.cn

Abstract

Spiking Neural Networks (SNNs), with their brain-inspired spatiotemporal dynamics and spike-driven computation, have emerged as promising energy-efficient alternatives to Artificial Neural Networks (ANNs). However, existing SNNs typically replicate inputs directly or aggregate them into frames at fixed intervals. Such strategies lead to neurons receiving nearly identical stimuli across time steps, severely limiting the model's expressive power—particularly in complex tasks like object detection. In this work, we propose the Temporal Dynamics Enhancer (TDE) to strengthen SNNs' capacity for temporal information modeling. TDE consists of two modules: a Spiking Encoder (SE) that generates diverse input stimuli across time steps, and an Attention Gating Module (AGM) that guides the SE generation based on inter-temporal dependencies. Moreover, to eliminate the high-energy multiplication operations introduced by the AGM, we propose a Spike-Driven Attention (SDA) to reduce attention-related energy consumption. Extensive experiments demonstrate that TDE can be seamlessly integrated into existing SNN-based detectors and consistently outperforms state-of-the-art methods, achieving mAP@50-95 scores of 57.7% on the static PASCAL VOC dataset and 47.6% on the neuromorphic EvDET200K dataset. In terms of energy consumption, the SDA consumes only 0.240× the energy of conventional attention modules.

Code — https://github.com/Mortal825/TDE.git

Introduction

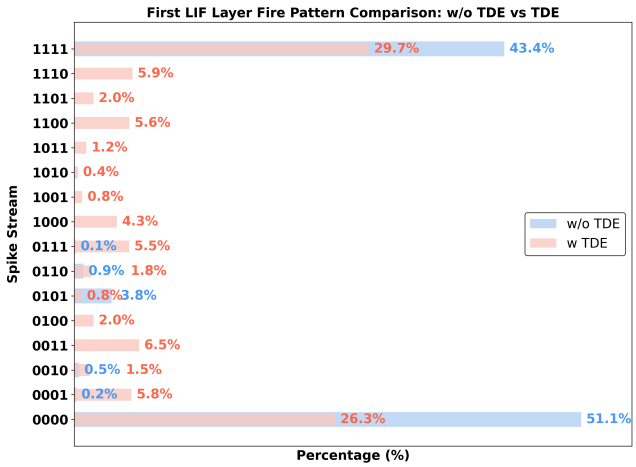
As a representative of third-generation neural networks, Spiking Neural Networks (SNNs) (Maass 1997; Roy, Jaiswal, and Panda 2019; Zhang et al. 2021; Guo et al. 2023; Qu et al. 2024a; Zhang et al. 2025c,b,a) offer a biologically plausible and energy-efficient alternative, gradually attracting increasing attention. SNNs transmit information through discrete spikes rather than continuous values, significantly reducing data transmission and storage costs. Their inherently asynchronous and event-driven nature further eliminates redundant computation and synchronization overhead. As a result, SNNs achieve high energy efficiency on neuromorphic hardware platforms (Poon and Zhou 2011; Merolla et al. 2014).

Nevertheless, SNNs exhibit limited expressiveness, particularly in complex regression tasks such as object detection. This is primarily due to their spike-based communication paradigm, which inherently restricts the precision and continuity of information representation. Although incorporating multiple time steps enhances temporal expressivity, existing SNN-based detection frameworks still fall short of matching the performance of their ANN counterparts. Recent efforts have attempted to bridge this gap by directly or indirectly incorporating integer spikes (Qu et al. 2024b; Luo et al. 2024; Yao et al. 2025). However, direct usage of integer spikes (Qu et al. 2024b) compromises the spike-driven nature of SNNs, while indirect methods (Luo et al. 2024; Yao et al. 2025) often overlook the hardware costs of converting integer spikes into actual spike events. Importantly, these approaches tend to focus on the representation of single spike information but overlook the temporal domain, which is another critical source of expressive power unique to SNNs. In contrast to ANNs, SNNs possess inherent temporal dynamics, which, if effectively leveraged, could offer a richer and more biologically plausible form of computation.

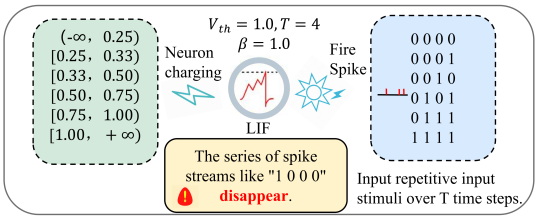
We argue that existing SNN-based object detection frameworks fail to fully utilize the temporal domain. For static datasets, identical stimuli are repeated across time steps, while for neuromorphic datasets, the long integration window causes similar stimuli, both hindering temporal dynamics. As a result, temporal information is primarily captured through the accumulation of membrane potentials, leaving the broader temporal dimension largely underutilized. As shown in the Fig. 1a and Fig. 1b, the spike streams of the first-layer Leaky Integrate-and-Fire (LIF) neurons exhibit significant disappearance, with spike streams such as "1110", "1101". The spike streams tend to be all-silent or all-active, resulting in the temporal dynamics pattern scarcity problem.

In addition, attention mechanisms have become a key component in SNN frameworks, providing a clear modulation strategy for sparse, temporally-distributed spike representations, allowing the limited expressive power of SNNs to dynamically focus on informative spatial and temporal

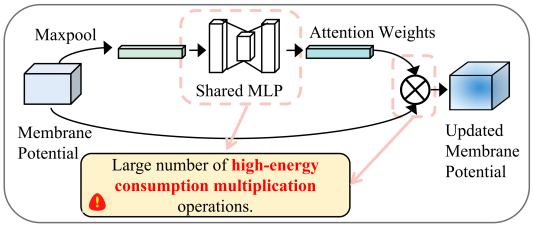
*Corresponding author.



(a) Spike Firing Pattern Before and After the TDE



(b) Membrane Potential Accumulation



(c) Energy Cost of Traditional Attention Mechanisms in SNNs

Figure 1: Research Motivation of the TDE Module. (a) Illustrates the spike firing pattern of the first LIF neuron layer before and after applying TDE. (b) Describes how the LIF neuron in existing SNNs receive repetitive input stimuli, resulting in the disappearance of a series of spike streams. (c) Highlights that traditional attention mechanisms in SNNs introduce a large number of energy-intensive multiplication operations.

regions (Yao et al. 2021, 2023c; Zhu et al. 2024). However, the non-spiking operations introduced by conventional attention mechanisms contradict the energy-efficient nature of SNNs, severely limiting their practical deployment (Qiu et al. 2024).

To address the temporal dynamics pattern scarcity problem, we propose **the Temporal Dynamics Enhancer (TDE)**, a framework that consists of two key components: **a Spiking Encoder (SE)** that generates temporally diverse input stimuli, and **an Attention Gating Module (AGM)** that captures inter-temporal dependencies within the network. By introducing temporal diversity and enriching multi-step spike representations, the TDE effectively unlocks the potential of the temporal dimension in SNNs. Moreover, to mitigate the increased high-energy multiplication operations introduced by the attention mechanism in the AGM, we propose **a Spike-Driven Attention (SDA)** to enable fully accumulation operations reduces attention-related energy consumption.

As shown in Fig. 1a, which illustrates the spike firing pattern of the first LIF layer. After integrating the TDE module, the temporal diversity of fire pattern improves markedly, and all previously missing spike streams in the first-layer LIF neurons are fully recovered.

Our main contributions are summarized as follows:

- We identify that the reliance on repeated inputs across multiple time steps in SNNs restricts the diversity of spike streams, thereby underutilizing the temporal modeling capabilities of the network.
- We propose the Temporal Dynamics Enhancer (TDE), a general framework that enhances spike stream diversity and significantly improves the expressiveness of directly trained spiking Object Detectors.
- We introduce the Spike-Driven Attention (SDA) module, which leverages spatio-temporal spike-based attention to enhance performance while reducing energy consumption to 0.240 times that of non-spiking modules.
- Through extensive experiments on both static (VOC) and neuromorphic (EvDET200K) datasets, we demonstrate that the Temporal Dynamics Enhancer (TDE) consistently improves performance across a variety of state-of-the-art methods, boosting mAP@50-95 by 1.2% on the VOC dataset and 2.2% on the EvDET200K dataset. Our method refreshes the state-of-the-art on both datasets, with the latest mAP@50-95 scores reaching 57.7% and 47.6%, respectively.

Related Works

Direct training of SNN-based object detectors

Direct training of spiking neural network (SNN)-based target detectors has attracted increasing attention, aiming to optimize network architecture and neuronal properties for enhanced performance in object detection tasks. SNNs are typically categorized into CNN-based and Transformer-based models. Owing to their strong inductive bias, CNNs have long dominated the field (d'Ascoli et al. 2021), with representative models such as Spiking ResNet (Zheng et al. 2021), SEW-ResNet (Fang et al. 2021a), and MSResNet (Hu et al. 2024), differing mainly in LIF neuron placement and identity mapping (He et al. 2016). Recently, Transformer-based SNNs (Dosovitskiy et al. 2020; Vaswani et al. 2017) have emerged, achieving state-of-the-art results in classification tasks (Yao et al. 2023a; Zhang et al. 2022) and enabling the development of directly trained target detectors.

Notably, the first such detector was proposed using spike-residue blocks (Su et al. 2023), and SpikeYOLO (Luo et al. 2024) further enhanced detection performance by incorporating the Meta-SpikeFormer design (Yao et al. 2024). However, these frameworks still underexplore the inherent temporal dynamics of SNNs and lack explicit temporal modeling.

Many studies (Fang et al. 2021b; Yao et al. 2022; Guo et al. 2024) show that binary activation of LIF neurons often results in insufficient precision in complex tasks. Current methods, such as I-LIF neurons (Luo et al. 2024), address this by using integer values during training and expanding virtual time steps to maintain spike-driven behavior. Similarly, the SFA method (Yao et al. 2025) uses integer training and spike-driven inference. Ternary spiking neurons (Yuan et al. 2024; Miao et al. 2025) improve deep-layer features, enhancing target detection. However, these methods require additional hardware overhead or disrupt the spike-driven nature of SNNs due to their use of non-traditional LIF neurons. We believe the temporal dimension of SNNs holds significant potential, and this paper focuses on exploring this using the binary activation of original LIF neurons.

Enhancing Temporal Dynamics in SNNs

To enhance the temporal dynamics of SNNs, researchers have improved spike representation encoding schemes. For example, Gated Attention Coding (GAC) (Qiu et al. 2024) transforms static images into robust representations with temporal dynamics. Additionally, Frequency Encoding (FE) (Xu et al. 2024) simulates selective visual attention in the biological brain, removing noise at different frequencies. However, these methods have not clearly elucidated the relationship between temporal dynamics and spike firing pattern, and have only been tested on classification tasks with relatively low precision demands.

Many studies (Shen et al. 2024; Zhang et al. 2025d; Lee et al. 2025) have explored the advantages of the temporal dimension when using attention mechanisms in SNNs. Inspired by SE-Net (Hu, Shen, and Sun 2018), researchers extended the attention mechanism to the temporal dimension, evaluating the importance of each frame in the final decision during training (Yao et al. 2021), thereby improving the temporal dynamics of SNNs. Modules like TCSA-SNN (Yao et al. 2023c) and TCJA-SNN (Zhu et al. 2024) assess the significance of spike streams across multiple dimensions, further enhancing the temporal dynamics. However, the current temporal-based attention mechanism requires a separate multiplication module to compute attention weights, disrupting the spike-driven nature of SNNs.

Preliminary Knowledge

Leaky Integrate-and-Fire neuron (LIF): In Spiking Neural Networks, one of the most commonly used neuron models is the Leaky Integrate-and-Fire (LIF) model (Tal and Schwartz 1997), as it offers a good balance between biological plausibility and computational efficiency.

Formally, the neuron dynamics at time step t are given by:

$$H_t = V_{t-1} + X_t, \tag{1}$$

$$S_t = \Theta(H_t - V_{th}), \tag{2}$$

$$V_t = \beta(H_t - V_{th}S_t), \tag{3}$$

where $X_t \in \mathbb{R}$ is the input current at time t, V_{t-1} is the membrane potential carried over from the previous step, and H_t is the pre-firing potential after integrating input. The neuron fires a spike $S_t \in \{0, 1\}$ when H_t exceeds the firing threshold V_{th}, determined by the Heaviside step function $\Theta(\cdot)$. After firing, the membrane potential is updated using a soft-reset mechanism, where $\beta \in [0, 1]$ is the leak factor that governs the decay of membrane potential over time.

Neuromorphic data: Neuromorphic data are generated by a Dynamic Vision Sensor (DVS), which triggers events when the logarithmic change in pixel intensity exceeds a threshold. Each event is represented as (x_n, y_n, t_n, p_n), where (x_n, y_n) are spatial coordinates, t_n is the timestamp, and $p_n \in \{-1, 1\}$ indicates polarity (increase or decrease in brightness). A common preprocessing strategy is to accumulate the event streams within a fixed time window into a frame format (Yao et al. 2023b). The resulting event data can be represented as a tensor of shape $\mathbb{R}^{1\times H\times W}$, where the single channel encodes the accumulated event polarity information.

Direct encoding schemes: Current SNN-based object detectors often adopt direct encoding (Wu et al. 2019), where the raw image is injected at each time step and spike streams are implicitly generated by the initial Conv-BN layer.

Method

In this section, we introduce the Temporal Dynamics Enhancer (TDE) to tackle spike firing pattern scarcity in SNNs for object detection. The TDE comprises two modules: a Spiking Encoder (SE) to generate varied temporal stimuli and an Attention Gating Module (AGM) to mitigate irrational exploration within the SE. By adding additional neuron groups, we achieve a fully Spike-Driven Attention (SDA) without multiplication operations. TDE is a general-purpose module. In this section, we integrate it into the SpikeYOLO framework to illustrate its design and implementation.

Spiking Encoder (SE)

Current SNN-based object detection frameworks predominantly employ direct encoding schemes, where the image input is repeatedly fed into LIF neurons. This repetitive operation fails to generate dynamic output, leading to a lack of temporal dynamics in subsequent SNN architectures. The spiking encoder of an SNN is designed to convert an image input $I \in R^{C_{in}\times H\times W}$ into the spike streams $S \in \{0, 1\}^{T\times C_{out}\times H\times W}$. For neuromorphic datasets, we use event-accumulated frames over the entire time window as input.

We first use a convolutional layer to extract initial feature. Considering the membrane potential accumulation property of the LIF neuron model, the feature block at each time step

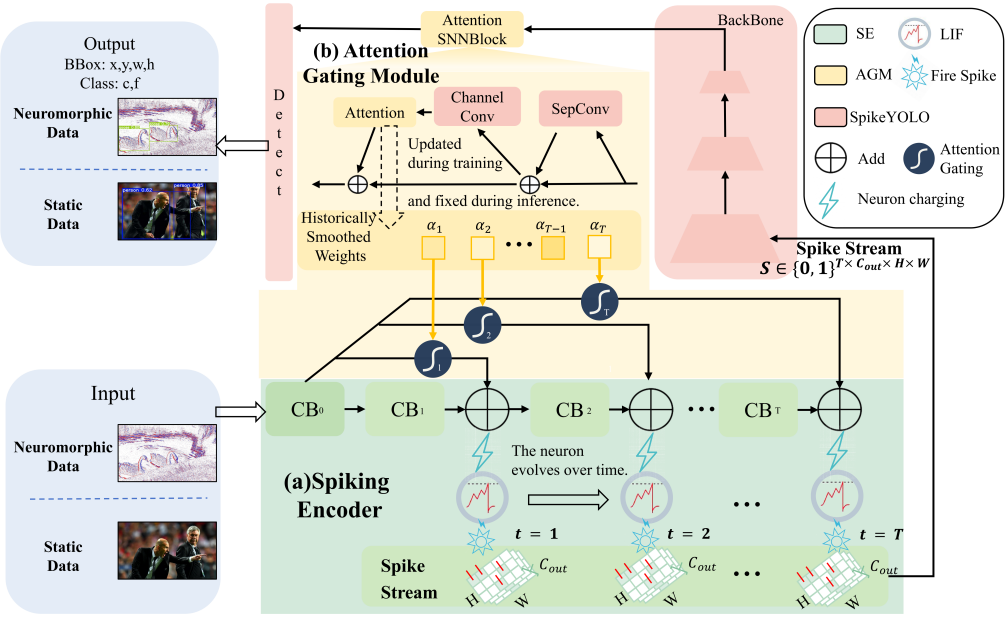


Figure 2: The Temporal Dynamics Enhancer (TDE) consists of two main components: (1) The spiking encoder (SE), using the CB component (Conv-BN), triggers the generation of diverse spikes. Its connection mechanism is based on the charging equation of the LIF neuron and the firing equation. (2) The Attention Gating Module (AGM) enhances the temporal dynamics of neurons within the layer by utilizing a general multi-dimensional attention mechanism. At the same time, it obtains temporal attention weights to regulate the spike stream generation of the SE, suppressing unreasonable exploration.

in every T timestep simulations should satisfy:

$$A_t = \begin{cases} f_0(I), & t = 0 \\ f_t(A_{t-1}, I), & 0 < t <= T \end{cases}. \tag{4}$$

Each time step's feature block is influenced not only by external input but also by the feature block of the previous time step. This is consistent with the LIF neuron update process, enabling the generation of spike inputs with rich temporal dynamics. The function f at timestep t is defined as follows:

$$f_t(A_{t-1}, I) = \alpha_t I + (1 - \alpha_t) f_t^{k\times k}(A_{t-1}), \tag{5}$$

where α_t is the preference coefficient at time step t, and $f_t^{k\times k}$ denotes the 2D convolution operation at time t with a filter size of $k \times k$.

Subsequently, we obtain a feature sequence A. By applying the LIF neuron, we can obtain the spike streams S:

$$S = LIF(A), \quad S \in \{0,1\}^{T\times C_{out}\times H\times W}. \tag{6}$$

Attention Gating Module (AGM)

The Attention Gating Module (AGM) utilizes a general multi-dimensional attention mechanism, which captures temporal attention weights to modulate the output of the SE. This approach, on one hand, leverages the attention mechanism to effectively regulate the membrane potential distribution of the LIF neurons, further enhancing the inter-layer temporal dynamics. On the other hand, by incorporating the attention gating mechanism, we can effectively suppress unreasonable exploration in the SE.

Algorithm 1: Attention Gating

Input: Temporal attention weights $g_{t,\text{temp}}^{\text{float}}$, previous preference coefficient $\bar{\alpha}_t$, batch size B.
Output: Updated preference coefficient α_t.

1: **for** $t = 1$ to T **do**
2: Batch-averaged attention: $\hat{\alpha}_t = \frac{1}{B}\sum_{b=1}^{B} g_{t,b,\text{temp}}^{\text{float}}$
3: Temporal smoothing: $\alpha_t = \frac{1}{2}(\bar{\alpha}_t + \hat{\alpha}_t),\ \bar{\alpha}_t = \alpha_t$
4: **end for**

We first integrate a general multi-dimensional attention mechanism as the basic component of the AGM. This general multi-dimensional attention will be introduced in detail later. The computation process for each dimension can be expressed as:

$$H_{Att}^n = g(H^n) \circ H^n, \tag{7}$$

where $g(H^n)$ denotes the function that generates the attention weights, reflecting the process of focusing on discriminative moments or regions. The attention corresponding to each dimension is denoted with a subscript, such as the temporal attention $g_{\text{temp}}(H^n)$. Here, H^n typically represents the membrane potential of a neuron in the n-th layer prior to reset and H_{Att}^n denotes the attention-modulated membrane potential.

After obtaining temporal attention weights via $g_{\text{temp}}(H^n)$, the AGM module treats them as structural priors to guide the SE's feature construction and avoid unreasonable exploration. The preference coefficient α_t in Eq. (5) is updated as shown in Alg. 1. At each time step, the SE fuses current and

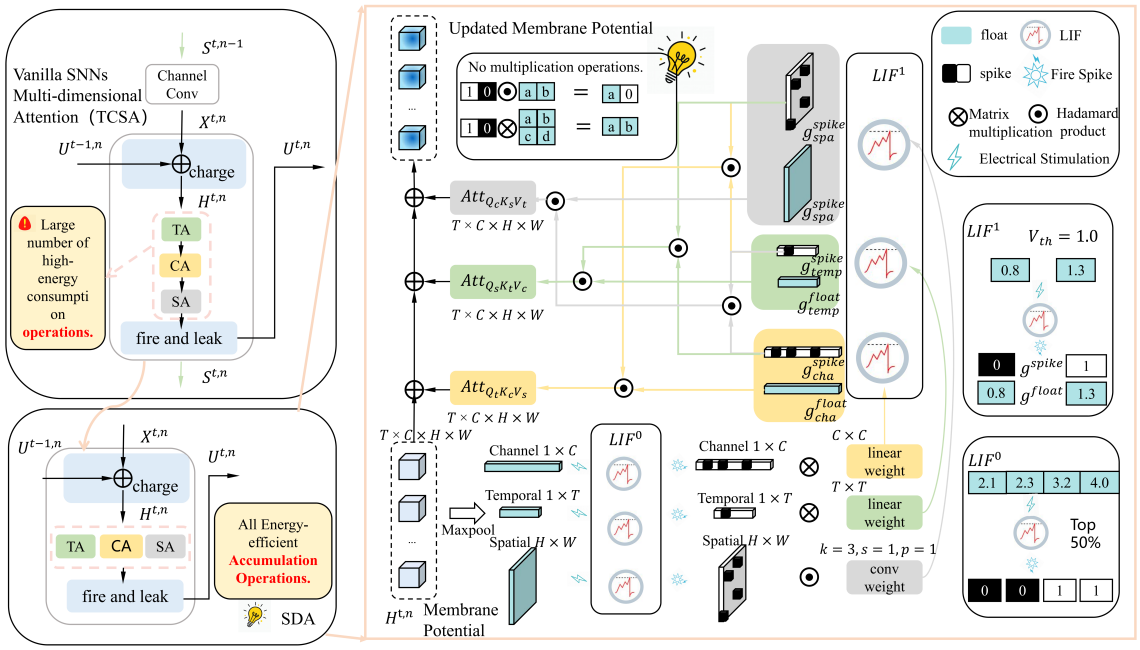


Figure 3: SDA: Schematic of the Spike-Driven Attention. SDA uses two neuron groups to avoid the multiplication operations involved. It fuses the spike and floating-point temporal, spatial, and channel attention weights with cross-attention to obtain the attention-updated membrane potential, eliminating the need for hadamard multiplication and matrix multiplication operations in the membrane potential update.

previous feature blocks based on α_t: higher values emphasize current temporal features, while lower values increase reliance on the original block to suppress unstable or redundant evolution.

Spike-Driven Attention (SDA)

This general multi-dimensional attention has various implementations, such as TCSA (Yao et al. 2023c). However, the current integration of attention blocks requires separate multiplication modules in subsequent layers to dynamically compute attention weights, as seen in the attention weight calculation function $g(H^n)$ and the membrane potential update with the Hadamard product in Eq. (7), which disrupts the spike-driven nature of SNNs. To address this, we propose Spike-Driven Attention (SDA), a multidimensional spiking attention block that enhances temporal dynamics while preserving spike-driven characteristics.

To eliminate multiplication operations, we replace independent multiplication units with a group of neurons. The neuron group consists of two sets, LIF^0 and LIF^1. LIF^0 discards the threshold firing characteristic and uses a top-k% strategy for spike generation, where neurons receiving the top k% of stimuli fire, and others remain silent. LIF^1, in contrast, retains the threshold firing mechanism, but its output also includes the membrane potential, providing an additional representation of the spike signal.

In SNNs, Time Attention (TA) utilizes the temporal relationships of membrane potential to enhance the network's temporal features. To compute the 1-D TA weights, the membrane potential of neurons in the n-th layer, $H^n = [\dots, H^n_{\text{temp}}, \dots] \in \mathbb{R}^{T\times c_n\times h_n\times w_n}$, are used as input. The 1-D TA spike weights $g^{\text{spike}}_{\text{temp}} \in \{0,1\}^{T\times 1\times 1\times 1}$ and 1-D TA float weights $g^{\text{float}}_{\text{temp}} \in \mathbb{R}^{T\times 1\times 1\times 1}$ can be represented as:

$$g^{\text{spike}}_{\text{temp}}, g^{\text{float}}_{\text{temp}} = LIF^1_{\text{temp}}(W^n_{\text{temp}}(LIF^0_{\text{temp}}(\text{MaxPool}(H^n)))). \tag{8}$$

Channel Attention (CA) and Spatial Attention (SA) are implemented similarly. Maxpool is used to extract significant regions, LIF^0 is used to excite specific regions, and convolution or fully connected layers map these regions. LIF^1 is then used to obtain the spike attention weights and float attention weights. A detailed implementation can be found in **Supplementary Material (Section I)**.

Finally, we fuse the temporal, spatial, and channel attention weights with cross-attention to obtain the attention-updated membrane potential, H^n_{Att}, in the SDA module as follows:

$$\begin{aligned} H^n_{Att} &= g(H^n) + H^n \\ &= Att_{Q_t-K_c-V_s} + Att_{Q_c-K_s-V_t} + Att_{Q_s-K_t-V_c} + H^n \\ &= g^{\text{spike}}_{\text{temp}} \times g^{\text{float}}_{\text{ch}} \times g^{\text{spike}}_{\text{spa}} + g^{\text{spike}}_{\text{ch}} \times g^{\text{float}}_{\text{spa}} \times g^{\text{spike}}_{\text{temp}} \\ &\quad + g^{\text{spike}}_{\text{spa}} \times g^{\text{float}}_{\text{temp}} \times g^{\text{spike}}_{\text{ch}} + H^n. \end{aligned} \tag{9}$$

Experiments

We integrated the TDE module with the three most advanced directly trained SNN-based object detectors (based on binary spikes) to validate its generalization ability. During the

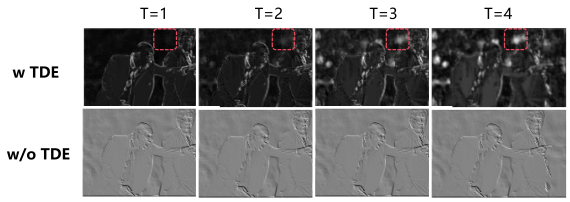


Figure 4: With TDE, the network gradually shifts attention from the object to surrounding regions over time, while the baseline shows mostly static feature maps.

experiments, we used consistent hyperparameters and the same experimental device across all methods, without any additional tricks. We set the time step to $T = 4$ to trade-off efficiency and accuracy. More details are provided in **Supplementary Material (Section II)**.

Dataset and Evaluation Metric

We validated the generalization ability of the TDE module using state-of-the-art object detectors on the static datasets VOC2007 and VOC2012 (Everingham et al. 2015), as well as the neuromorphic dataset EvDET200K (Wang et al. 2025). The VOC2007 and VOC2012 datasets contain 9,963 and 11,530 images, respectively, with annotations for 20 object categories, and are widely used as benchmark datasets for object detection. The EvDET200K dataset consists of 10,054 video streams, covering 10 object categories and providing 202,260 annotations. This dataset focuses on small object detection and includes multiple challenges such as multi-view, multi-illumination, multi-motion, and dynamic backgrounds. For evaluation, we used the mean Average Precision (mAP) at different Intersection over Union (IOU) thresholds, which is the most commonly used metric in object detection. In terms of energy consumption, we calculated the number of floating-point multiplication (MUL) and accumulation (AC) operations for the attention mechanism. Using a 45nm technology node with 32-bit floating-point precision, the energy cost is 3.7 pJ per MUL operation and 0.9 pJ per AC operation (Kim et al. 2020). Additionally, we measured the number of parameters for each detector to provide a more comprehensive and accurate understanding of the models' performance.

Efficiency and Generalizability Validation

The TDE (Temporal Dynamics Enhancer) module demonstrates remarkable consistency improvement when integrated with the current state-of-the-art directly trained SNN object detectors (as shown in Tab. 2).

From a global perspective, whether on static datasets like VOC2007 and VOC2012, or on neuromorphic datasets like EvDET200K, integrating the TDE (TCSA) module consistently improves the mAP50 by over 1.1%. Notably, when combined with SpikeYOLO, the performance on VOC2007 improves by 3.0%. From a more localized perspective, the integration of TDE enables underperforming methods to surpass others. For example, the original EMSYOLO without TDE achieves a mAP50 of only 44.9% on EvDET200K, significantly lower than SpikeYOLO's 46.5%. However, with the addition of the TDE module, EMSYOLO's performance on EvDET200K improves to 47.1%, surpassing other frameworks. A similar trend is observed on VOC2007, where the original EMSYOLO without TDE has a mAP50 of 31.9%, well below SpikeYOLO's 33.2%. After integrating TDE, EMSYOLO's performance on VOC2007 increases to 34.9%, achieving a similar leap. When TDE is integrated into Meta-SpikeFormer, consistent performance improvements are observed. However, its performance on the VOC dataset remains low, likely due to the lack of CNN's prior inductive bias in Meta-SpikeFormer. It is also worth mentioning that integrating the TDE module into the three directly trained SNN object detectors increases the parameter count by no more than 0.26M, further highlighting the efficiency of the TDE module. Fig. 4 qualitatively illustrates how TDE enhances the temporal dynamics of SNNs.

Method	MUL	AC	Energy (μJ)	Energy Ratio
TDE(TCSA)	5.75 E6	5.76 E5	2.18 E1	1.0
TDE(SDA)	0	5.82 E6	5.24 E0	**0.240**

Note: Energy computed for attention over membrane potentials with T=4, C=128, H=80, W=40.

Table 1: Energy analysis of TDE (TCSA vs. SDA) attention mechanisms in SpikeYOLO on VOC2007.

To reduce energy consumption, we replaced the non-spiking attention TCSA module with the SDA module, eliminating all multiplication operations. As a result, SDA's energy consumption is only 0.240 times that of TCSA (see Tab. 1). While reducing energy consumption, TDE(SDA) still demonstrates continuous performance improvements. Specifically, when tested with the SpikeYOLO method on the VOC2007 dataset, mAP@50 reached 35.6%, surpassing the 34.9% achieved by the non-spiking version TDE(TCSA). However, on the EvDET200K dataset, the performance improvement with TDE(SDA) was lower than that of TDE(TCSA). We believe this is due to the limited learning capacity of the SDA Spike-Driven Attention, as it struggles to capture the intrinsic relationships in data flows (such as time, space, and channels) when working with larger datasets. Nonetheless, its sparse attention mechanism helps prevent overfitting on smaller datasets.

Component Analysis and Ablation Study

To validate the effectiveness of the proposed components, we conducted an ablation study on the SpikeYOLO and EMSYOLO methods using the VOC2007 dataset (see Tab. 3). It is important to note that when AGM is tested in isolation, the gating mechanism is absent, reducing it to a simple multi-dimensional attention mechanism.

Both TCSA and SDA consistently outperform the baseline, supporting prior research. The attention mechanism improves performance by focusing on salient regions and ignoring redundant information, while the SE module generates high-temporal dynamic pulses, addressing the issue of disappearing pulse patterns in traditional object detection, enhancing the expressiveness of binary SNNs. In terms of

Dataset	Methods	Models	Params (M)	Time Steps	mAP@50 (%)	mAP@50:95 (%)
VOC	ANN2SNN	SUHD(Qu et al. 2024b)	-	4	75.3	-
		SpikeYOLO(Luo et al. 2024)	23.156	4	78.0	56.6
		+TDE (TCSA)	23.416	4	79.1 (+1.1)	57.7 (+1.1)
		+TDE (SDA)	23.645	4	78.3 (+0.3)	57.1 (+0.5)
	Direct training	EMSYOLO(Su et al. 2023)	33.889	4	76.8	49.6
		+TDE (TCSA)	34.126	4	77.1 (+0.3)	50.8 (+1.2)
		+TDE (SDA)	34.358	4	77.3 (+0.5)	50.2 (+0.6)
		Meta-SpikeFormer(Yao et al. 2024)	16.652	4	49.8	24.3
		+TDE (TCSA)	16.823	4	51.7 (+1.9)	25.7 (+1.4)
EvDET200K	Direct training	SpikeYOLO(Luo et al. 2024)	23.156	4	75.2 (*74.8*)	46.5 (*41.2*)
		+TDE (TCSA)	23.416	4	76.0 (+0.8)	47.6 (+1.1)
		+TDE (SDA)	23.645	4	75.8 (+0.6)	47.2 (+0.7)
		EMSYOLO(Su et al. 2023)	33.889	4	77.2 (*66.6*)	44.9 (*32.1*)
		+TDE (TCSA)	34.126	4	78.2 (+1.0)	47.1 (+2.2)
		+TDE (SDA)	34.358	4	77.7 (+0.5)	45.9 (+1.0)
		Meta-SpikeFormer(Yao et al. 2024)	16.652	4	75.2	45.9
		+TDE (TCSA)	16.823	4	76.3 (+1.1)	47.2 (+1.3)
VOC2007	Direct training	SpikeYOLO(Luo et al. 2024)	23.156	4	51.7	31.9
		+TDE (TCSA)	23.416	4	55.9 (+4.2)	34.9 (+3.0)
		+TDE (SDA)	23.645	4	56.2 (+4.5)	35.6 (+3.7)
		EMSYOLO(Su et al. 2023)	33.889	4	59.8	33.2
		+TDE (TCSA)	34.126	4	61.3 (+1.5)	35.1 (+1.9)
		+TDE (SDA)	34.358	4	61.3 (+1.5)	34.3 (+1.1)

Note: Numbers in parentheses (in italic) indicate baseline results directly quoted from (Wang et al. 2025) as reference values.The VOC in the table refers to the VOC2007 and VOC2012 datasets.

Table 2: Performance improvement of spiking object detection models with TDE on different datasets.

Architecture	AGM	SE	mAP@50 (%)	mAP@50:95 (%)
SpikeYOLO	✗	✗	51.7	31.9
	TCSA	✗	54.3 (**+2.6**)	33.9 (**+2.0**)
	SDA	✗	52.6 (**+0.9**)	32.6 (**+0.7**)
	✗	✓	55.2 (**+3.5**)	34.6 (**+2.7**)
	TCSA	✓	55.9 (**+4.2**)	35.1 (**+3.2**)
	SDA	✓	56.2 (**+4.5**)	35.6 (**+3.7**)
EMSYOLO	✗	✗	59.8	33.2
	TCSA	✗	60.2 (**+0.4**)	33.9 (**+0.7**)
	SDA	✗	60.6 (**+0.8**)	33.8 (**+0.6**)
	✗	✓	60.9 (**+1.1**)	34.1 (**+0.9**)
	TCSA	✓	61.3 (**+1.5**)	35.1 (**+1.9**)
	SDA	✓	61.3 (**+1.5**)	34.3 (**+1.1**)

Note: When AGM is tested alone, the gating mechanism is absent, reducing it to a simple multi-dimensional attention.

Table 3: Ablation study of the proposed modules on the VOC2007 dataset.

mAP@50-95, SpikeYOLO improved by 2.7%, and EMSYOLO by 0.9%, outperforming TCSA by 2.0% and 0.7%, respectively. By using Attention Gating to integrate SE and multi-dimensional attention, performance improves consistently, with SpikeYOLO showing a 3.7% increase and EMSYOLO a 1.9% increase. This highlights the effective coupling of the two components through the attention gating mechanism. To further assess its impact, an experiment was conducted on the SpikeYOLO framework (see Tab. 4), resulting in a 0.7% boost in mAP@50-95. More results can be found in **Supplementary Material (Section III)**.

Method	mAP@50 (%)	mAP@50:95 (%)
TCSA + SE	78.5	57.0
AGM(TCSA) + SE	79.1 (+0.6)	57.7 (+0.7)

Table 4: Effect of the Attention Gating (Alg. 1) on the SpikeYOLO Framework for VOC Datasets.

Conclusion

We propose the Temporal Dynamics Enhancer (TDE) to address the limitations of Spiking Neural Networks (SNNs) in temporal modeling for object detection. Through extensive experiments on both static (VOC) and neuromorphic (EvDET200K) datasets, we demonstrate that TDE consistently enhances performance across a variety of state-of-the-art methods. We hope our research provides new insights into the potential of temporal dynamics in SNNs and inspires future research on more efficient and biologically inspired spike-driven learning paradigms.

Acknowledgements

This work is supported by the Strategic Priority Research Program of the Chinese Academy of Sciences under (Grants XDA0450000, XDA0450202), and Central Government Guidance Funds for Local Science and Technology Development (YDZX2025124)

References

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

d'Ascoli, S.; Touvron, H.; Leavitt, M. L.; Morcos, A. S.; Biroli, G.; and Sagun, L. 2021. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, 2286–2296.

Everingham, M.; Eslami, S. A.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2015. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111: 98–136.

Fang, W.; Yu, Z.; Chen, Y.; Huang, T.; Masquelier, T.; and Tian, Y. 2021a. Deep residual learning in spiking neural networks. *Advances in Neural Information Processing Systems*, 34: 21056–21069.

Fang, W.; Yu, Z.; Chen, Y.; Masquelier, T.; Huang, T.; and Tian, Y. 2021b. Incorporating learnable membrane time constant to enhance learning of spiking neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2661–2671.

Guo, L.; Gao, Z.; Qu, J.; Zheng, S.; Jiang, R.; Lu, Y.; and Qiao, H. 2023. Transformer-based spiking neural networks for multimodal audiovisual classification. *IEEE Transactions on Cognitive and Developmental Systems*, 16(3): 1077–1086.

Guo, Y.; Chen, Y.; Liu, X.; Peng, W.; Zhang, Y.; Huang, X.; and Ma, Z. 2024. Ternary spike: Learning ternary spikes for spiking neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 12244–12252.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, 630–645.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Hu, Y.; Deng, L.; Wu, Y.; Yao, M.; and Li, G. 2024. Advancing spiking neural networks toward deep residual learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2): 2353–2367.

Kim, S.; Park, S.; Na, B.; and Yoon, S. 2020. Spiking-yolo: spiking neural network for energy-efficient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11270–11277.

Lee, D.; Li, Y.; Kim, Y.; Xiao, S.; and Panda, P. 2025. Spiking transformer with spatial-temporal attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13948–13958.

Luo, X.; Yao, M.; Chou, Y.; Xu, B.; and Li, G. 2024. Integer-valued training and spike-driven inference spiking neural network for high-performance and energy-efficient object detection. In *European Conference on Computer Vision*, 253–272.

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

Merolla, P. A.; Arthur, J. V.; Alvarez-Icaza, R.; Cassidy, A. S.; Sawada, J.; Akopyan, F.; Jackson, B. L.; Imam, N.; Guo, C.; Nakamura, Y.; et al. 2014. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science*, 345(6197): 668–673.

Miao, W.; Shen, J.; Xu, Q.; Hamalainen, T.; Xu, Y.; and Cong, F. 2025. SpikingYOLOX: Improved YOLOX Object Detection with Fast Fourier Convolution and Spiking Neural Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1465–1473.

Poon, C.-S.; and Zhou, K. 2011. Neuromorphic silicon neurons and large-scale neural networks: challenges and opportunities. *Frontiers in neuroscience*, 5: 108.

Qiu, X.; Zhu, R.-J.; Chou, Y.; Wang, Z.; Deng, L.-j.; and Li, G. 2024. Gated attention coding for training high-performance and efficient spiking neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 601–610.

Qu, J.; Gao, Z.; Li, Y.; Lu, Y.; and Qiao, H. 2024a. Spike-based high energy efficiency and accuracy tracker for Robot. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1428–1434. IEEE.

Qu, J.; Gao, Z.; Zhang, T.; Lu, Y.; Tang, H.; and Qiao, H. 2024b. Spiking neural network for ultralow-latency and high-accurate object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3): 4934–4946.

Roy, K.; Jaiswal, A.; and Panda, P. 2019. Towards spike-based machine intelligence with neuromorphic computing. *Nature*, 575(7784): 607–617.

Shen, S.; Zhao, D.; Shen, G.; and Zeng, Y. 2024. TIM: An Efficient Temporal Interaction Module for Spiking Transformer. In *IJCAI*, 1519–1525.

Su, Q.; Chou, Y.; Hu, Y.; Li, J.; Mei, S.; Zhang, Z.; and Li, G. 2023. Deep directly-trained spiking neural networks for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6555–6565.

Tal, D.; and Schwartz, E. L. 1997. Computing with the leaky integrate-and-fire neuron: logarithmic computation and multiplication. *Neural Computation*, 9(2): 305–318.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30: 5998–6008.

Wang, X.; Jin, Y.; Wu, W.; Zhang, W.; Zhu, L.; Jiang, B.; and Tian, Y. 2025. Object detection using event camera: A moe heat conduction based detector and a new benchmark dataset. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 29321–29330.

Wu, Y.; Deng, L.; Li, G.; Zhu, J.; Xie, Y.; and Shi, L. 2019. Direct training for spiking neural networks: Faster, larger, better. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 1311–1318.

Xu, M.; Ma, D.; Tang, H.; Zheng, Q.; and Pan, G. 2024. FEEL-SNN: Robust spiking neural networks with frequency encoding and evolutionary leak factor. *Advances in Neural Information Processing Systems*, 37: 91930–91950.

Yao, M.; Gao, H.; Zhao, G.; Wang, D.; Lin, Y.; Yang, Z.; and Li, G. 2021. Temporal-wise attention spiking neural networks for event streams classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10221–10230.

Yao, M.; Hu, J.; Hu, T.; Xu, Y.; Zhou, Z.; Tian, Y.; Xu, B.; and Li, G. 2024. Spike-driven transformer v2: Meta spiking neural network architecture inspiring the design of next-generation neuromorphic chips. *arXiv preprint arXiv:2404.03663*.

Yao, M.; Hu, J.; Zhou, Z.; Yuan, L.; Tian, Y.; Xu, B.; and Li, G. 2023a. Spike-driven transformer. *Advances in Neural Information Processing Systems*, 36: 64043–64058.

Yao, M.; Qiu, X.; Hu, T.; Hu, J.; Chou, Y.; Tian, K.; Liao, J.; Leng, L.; Xu, B.; and Li, G. 2025. Scaling spike-driven transformer with efficient spike firing approximation training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47: 2973 – 2990.

Yao, M.; Zhang, H.; Zhao, G.; Zhang, X.; Wang, D.; Cao, G.; and Li, G. 2023b. Sparser spiking activity can be better: Feature refine-and-mask spiking neural network for event-based visual recognition. *Neural Networks*, 166: 410–423.

Yao, M.; Zhao, G.; Zhang, H.; Hu, Y.; Deng, L.; Tian, Y.; Xu, B.; and Li, G. 2023c. Attention spiking neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 9393–9410.

Yao, X.; Li, F.; Mo, Z.; and Cheng, J. 2022. Glif: A unified gated leaky integrate-and-fire neuron for spiking neural networks. *Advances in Neural Information Processing Systems*, 35: 32160–32171.

Yuan, M.; Zhang, C.; Wang, Z.; Liu, H.; Pan, G.; and Tang, H. 2024. Trainable spiking-yolo for low-latency and high-performance object detection. *Neural Networks*, 172: 106092.

Zhang, J.; Tang, L.; Yu, Z.; Lu, J.; and Huang, T. 2022. Spike transformer: Monocular depth estimation for spiking camera. In *European Conference on Computer Vision*, 34–52.

Zhang, M.; Luo, X.; Wu, J.; Belatreche, A.; Cai, S.; Yang, Y.; and Li, H. 2025a. Toward Building Human-Like Sequential Memory Using Brain-Inspired Spiking Neural Models. *IEEE transactions on neural networks and learning systems*, 10143–10155.

Zhang, M.; Wang, J.; Wu, J.; Belatreche, A.; Amornpaisannon, B.; Zhang, Z.; Miriyala, V. P. K.; Qu, H.; Chua, Y.; Carlson, T. E.; et al. 2021. Rectified linear postsynaptic potential function for backpropagation in deep spiking neural networks. *IEEE transactions on neural networks and learning systems*, 33(5): 1947–1958.

Zhang, M.; Wang, S.; Wu, J.; Wei, W.; Zhang, D.; Zhou, Z.; Wang, S.; Zhang, F.; and Yang, Y. 2025b. Toward Energy-Efficient Spike-Based Deep Reinforcement Learning With Temporal Coding. *IEEE Computational Intelligence Magazine*, 20(2): 45–57.

Zhang, M.; Wei, W.; Zhou, Z.; Liu, W.; Zhang, J.; Belatreche, A.; and Yang, Y. 2025c. Spike-Driven Lightweight Large Language Model With Evolutionary Computation. *IEEE Transactions on Evolutionary Computation*, 1–1.

Zhang, T.; Yu, K.; Zhong, X.; Wang, H.; Xu, Q.; and Zhang, Q. 2025d. STAA-SNN: Spatial-Temporal Attention Aggregator for Spiking Neural Networks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13959–13969.

Zheng, H.; Wu, Y.; Deng, L.; Hu, Y.; and Li, G. 2021. Going deeper with directly-trained larger spiking neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 11062–11070.

Zhu, R.-J.; Zhang, M.; Zhao, Q.; Deng, H.; Duan, Y.; and Deng, L.-J. 2024. Tcja-snn: Temporal-channel joint attention for spiking neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3): 5112–5125.