# TIM: An Efficient Temporal Interaction Module for Spiking Transformer

**Sicheng Shen**[1,2,4] , **Dongcheng Zhao**[1,2] , **Guobin Shen**[1,2,4] and **Yi Zeng**[1,2,3,4,5] *
[1] Brain-inspired Cognitive Intelligence Lab, Institute of Automation, Chinese Academy of Sciences
[2] Center for Long-term Artificial Intelligence
[3] Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology, CAS
[4] School of Future Technology, University of Chinese Academy of Sciences
[5] School of Artificial Intelligence, University of Chinese Academy of Sciences
{shensicheng2024, zhaodongcheng2016, shenguobin2021, yi.zeng}@ia.ac.cn

## Abstract

Spiking Neural Networks (SNNs), as the third generation of neural networks, have gained prominence for their biological plausibility and computational efficiency, especially in processing diverse datasets. The integration of attention mechanisms, inspired by advancements in neural network architectures, has led to the development of Spiking Transformers. These have shown promise in enhancing SNNs' capabilities, particularly in the realms of both static and neuromorphic datasets. Despite their progress, a discernible gap exists in these systems, specifically in the Spiking Self Attention (SSA) mechanism's effectiveness in leveraging the temporal processing potential of SNNs. To address this, we introduce the Temporal Interaction Module (TIM), a novel, convolution-based enhancement designed to augment the temporal data processing abilities within SNN architectures. TIM's integration into existing SNN frameworks is seamless and efficient, requiring minimal additional parameters while significantly boosting their temporal information handling capabilities. Through rigorous experimentation, TIM has demonstrated its effectiveness in exploiting temporal information, leading to state-of-the-art performance across various neuromorphic datasets. The code is available at https://github.com/BrainCog-X/Brain-Cog/tree/main/examples/TIM.

## 1 Introduction

Spiking neural networks (SNNs), representing a novel paradigm in the evolution of artificial neural networks (ANNs), derive their foundational principles from the intricacies of biological neural systems, with a particular focus on dynamic and temporal processing capabilities [Maass, 1997; Zeng *et al.*, 2023]. The event-driven mechanism inherent in SNNs markedly amplifies their energy efficiency, while simultaneously fostering improved interoperability with a broad range of neuromorphic hardware [Roy *et al.*, 2019; Wei *et al.*, 2024]. These networks have emerged as a pivotal tool within the domain of cognitive and intelligent systems research.

In the development of SNNs, researchers initially borrowed structures from biological neural systems but found challenges in scaling these structures for large networks and diverse tasks [Cheng *et al.*, 2020; Zhao *et al.*, 2020; Dong *et al.*, 2023]. To overcome these limitations, well-established designs from the deep learning domain, like ResNet and VGGNet, were introduced to improve flexibility and applicability [Sengupta *et al.*, 2019; Fang *et al.*, 2021a; Shen *et al.*, 2022a]. In addition, the Transformer architecture, known for its excellent parallel processing and ability to handle long-distance dependencies, has become prominent in various domains [Vaswani *et al.*, 2017; Dosovitskiy *et al.*, 2021]. Its self-attention mechanism offers unmatched flexibility and efficiency in handling sequential data. Currently, there's growing interest in combining SNNs with Transformers, using the biological realism and energy efficiency of SNNs alongside the powerful data processing of Transformers, opening up new prospects for simulating complex cognitive functions [Zhou *et al.*, 2022; Li *et al.*, 2022a].

The rapid advancement of event-based cameras has heralded a new era in the realms of image sensing and computer vision [Gallego *et al.*, 2020; Zheng *et al.*, 2023; Zhou *et al.*, 2023b; Rebecq *et al.*, 2019; Gehrig *et al.*, 2019]. Diverging from traditional imaging, event streams store information in the form of events rather than pixels. This approach not only preserves the temporal aspects of images to a certain extent but also reduces the resources required for image processing. Consequently, event streams are regarded as a more biologically congruent, efficient, and promising method for image storage and processing. The integration of SNNs with Transformers offers a highly efficient and adaptable strategy for handling event-driven data. Leveraging the biological principles underlying SNNs, this combination exhibits exceptional performance in dynamic and sequential data processing. Transformers enhance this synergy with their capability for efficient parallel processing and precise handling of long-range dependencies, thereby optimizing the analysis of complex data. This fusion not only emulates the processing mechanisms of biological neural systems but also significantly elevates the efficiency and accuracy of data processing. Particularly in the processing of event-based camera data, this

[*] corresponding author

amalgamation, utilizing spike-based event stream storage, effectively retains temporal information while concurrently reducing resource consumption.

Spikformer [Zhou *et al.*, 2022] represents the first successful integration of the Transformer architecture into the SNN domain. This model innovatively designs Spiking Self Attention to implement Transformer attention. In recent studies of Spiking Transformers, many improvements have been made. Currently, there are two primary approaches to improving Spiking Transformers. The first approach involves enhancing network performance by modifying the attention mechanism (eg. Spikeformer [Li *et al.*, 2022a] and DISTA [Xu *et al.*, 2023]). The second focuses on leveraging the efficiency of SNNs to reduce the computational energy consumption of Transformers (e.g. SPikingformer [Zhou *et al.*, 2023a] and Spike-diriven Transformer [Yao *et al.*, 2023a]). However, these advancements encounter limitations when applied to neuromorphic data. The unique dynamics and temporal complexity of neuromorphic datasets, reflecting biological neural systems, challenge these models. While they enhance network performance in some aspects, they fall short in fully capturing the nuances of neuromorphic data processing, highlighting a need for further specialized adaptations in Spiking Transformers to effectively manage these specific data characteristics.

In this study, we have developed a Temporal Interaction Module that can be seamlessly integrated into the attention matrix computation of a Spiking Transformer. This module enables the adaptive amalgamation of historical and current information, thereby effectively capturing the intrinsic dynamics of neuromorphic data. Our experiments conducted across various neuromorphic datasets demonstrate that our approach achieves the state-of-the-art performance of Spiking Transformer. In summary, our contributions are as follows:

- Through our analysis, we identified that a primary limitation in current Spiking Transformers is their insufficient handling of temporal information, chiefly because the attention matrix is reliant solely on information from the current moment.
- To address this, we have designed a Temporal Interaction Module to adaptively utilizes information from different time steps and functions as a lightweight addition. It can be integrated with existing attention computation modules without significantly increasing computational load.
- We conducted experiments on neuromorphic datasets, including CIFAR10-DVS, NCALTECH101, NCARS, UCF101-DVS, HMDB51-DVS and SHD. We demonstrated the effectiveness and generalization ability of our method. Our results show that our approach sets a new benchmark in the Spiking Transformer domain, achieving better performance across these datasets.

## 2 Related Work

In this section, we will review and analyze recent research aimed at enhancing the temporal information processing capabilities of SNNs and developments in the Spiking Transformer field.

### 2.1 Advances in SNN Temporal Processing

In terms of enhancing the temporal information processing ability of SNNs, significant progress has been made by researchers. For instance, [Kim *et al.*, 2023] explored the dynamic characteristics of temporal information in SNNs by estimating the Fisher information of weights. PLIF [Fang *et al.*, 2021b] enhanced the integration of information at different time steps in SNNs by introducing a learnable membrane potential constant. IIRSNN [Fang *et al.*, 2020] improved the spatiotemporal information processing capabilities of SNNs through synaptic modeling. [Zhang *et al.*, 2021] discovered that learning based on spike timing in an event-driven manner can yield significant improvements in classification accuracy.TA-SNN [Yao *et al.*, 2021] introduced a temporal-based attention mechanism, enabling the adaptive allocation of importance to different time steps. TCJA [Zhu *et al.*, 2024] incorporates an attention mechanism designed to assess the significance of pulses in both temporal and spatial dimensions. [Shen *et al.*, 2024] significantly bolstered the temporal information processing capability of SNNs by incorporating dendritic nonlinear computations. ETC [Zhao *et al.*, 2023] introduced Temporal Enhancement Consistency constraints to enable SNNs to learn from outputs at different time steps, thereby enhancing performance and reducing latency. BioEvo [Shen *et al.*, 2023] employed a biologically inspired neural circuit search approach, adaptively coordinating different circuits to improve the performance of SNNs in perceptual tasks and reinforcement learning.

### 2.2 Spiking Transformer

In the Transformer domain, Spikformer [Zhou *et al.*, 2022] innovatively implemented an attention mechanism called Spiking Self Attention. This mechanism replaced the traditional activation function of ANNs with spiking neurons and omitted the softmax function typically required before calculating attention [Vaswani *et al.*, 2017], opting instead for direct matrix multiplication. DISTA [Xu *et al.*, 2023] enhanced the capture of spatiotemporal data by designing new neuronal connections. Spikeformer [Li *et al.*, 2022a] introduced separate Temporal and Spatial Multi-head Self Attention in each Transformer block to strengthen data processing capabilities. Spikingformer [Zhou *et al.*, 2023a] achieved greater efficiency by rearranging the positions of neurons and convolutional layers to eliminate float-integer multiplications. In contrast, the Spike-driven Transformer [Yao *et al.*, 2023a] reduced time complexity by altering the order of attention computation and substituting multiplication with addition.

However, despite a series of advancements in the Spiking Transformer domain, these methods still exhibit performance shortcomings when processing information such as neuromorphic data with strong temporal characteristics. This limitation reveals the current technology's inadequacies in capturing and processing data with complex temporal dependencies, particularly in analyzing the deep temporal dynamics and subtle changes within neuromorphic data.

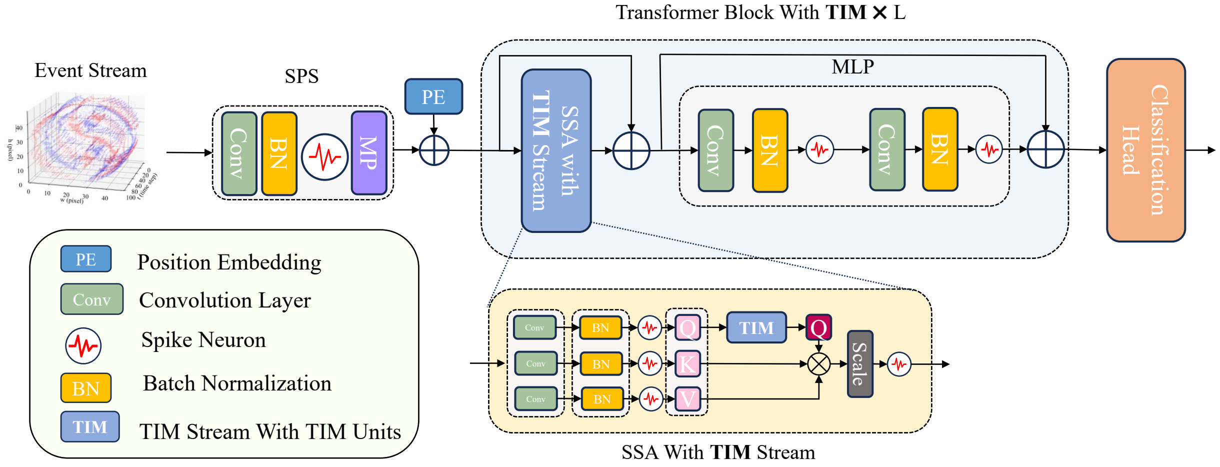

Figure 1: Comprehensive diagram of Spikformer integrated with Temporal Interaction Module (TIM): Demonstrating TIM's Plug-and-Play role within the Spike Self Attention (SSA) Block of the original Spikformer structure.

## 3 Preliminaries

Currently, the predominant approach in Spiking Transformer methodologies involves substituting the activation functions in conventional Transformer architectures with spiking neurons. In this section, we will provide an in-depth analysis of the limitations encountered during the implementation of this approach.

### 3.1 Spiking Neuron

Spiking neurons, serving as the fundamental computational units of SNNs, exhibit significant differences from conventional ANNs. Their distinctive feature lies in the temporal accumulation of membrane potential; a spike is emitted once this potential surpasses the threshold. Within the realm of SNNs, Leaky Integrate-and-Fire (LIF) neurons are extensively employed due to their optimal balance between computational efficiency and biological plausibility. To facilitate ease of simulation computations, the discrete formulation of LIF neurons is used, and the details are shown in Eq. 1.

$$V[t] = V[t-1] + \frac{1}{\tau}(X[t] - (V[t-1])) \tag{1}$$

where $\tau$ is the membrane time constant, $V[t]$ refers to the membrane potential of the neurons in the step t. $X[t]$ is the input in the step $t$.

Following the emission of a spike, the membrane potential of a spiking neuron is reset to its resting potential $V_{\text{reset}}$. For computational convenience, this resting potential is set to zero. The details of this process are delineated in Eq. 2:

$$V[t] = V[t] \cdot (1 - \Theta(V[t])) \tag{2}$$

$$\Theta(x) = \begin{cases} 0 & \text{if } x < V_{th} \\ 1 & \text{if } x \geq V_{th} \end{cases} \tag{3}$$

$\Theta(x)$ is the Heaviside function. $V_{th}$ is the firing threshold.

To facilitate the application of the backpropagation algorithm for network training, we utilize surrogate gradients as an approximation for the gradients of the spike firing function. This method is implemented as follows:

$$\frac{\partial \Theta}{\partial V} = \begin{cases} 0, & |V - V_{th}| > \frac{1}{a} \\ -a^2|V - V_{th}| + a, & |V - V_{th}| \leq \frac{1}{a} \end{cases} \tag{4}$$

In our study, the variable $a$ serves as a hyperparameter, instrumental in dictating the configuration of the surrogate gradient's shape. We have strategically set the value of $a$ to 4.

### 3.2 Spiking Self Attention Analysis

The primary strength of the Transformer model resides in its innovative self-attention mechanism. This mechanism facilitates a nuanced examination of the relative importance assigned to distinct positions within a sequence, consequently enhancing the model's capacity for sophisticated information processing. Building upon this foundational concept, the Spikformer model introduces an advancement with the Spike Self Attention (SSA) mechanism, a development that draws inspiration from Vaswani et al.'s seminal work [Vaswani *et al.*, 2017]. The intricacies of this mechanism are delineated in Eq. 5.

$$\begin{cases} Q[t], K[t], V[t] = LIFNode(X[t]) \\ A[t] = Q[t]K[t]^T V[t] \end{cases} \tag{5}$$

Upon the amalgamation of SSA and LIF , it is deducible that the conduit of information transmission within the spiking Transformer architecture is characterized by Eq. 6:

$$\begin{aligned} V[t] &= V[t-1] + \frac{1}{\tau}(A[t] - (V[t-1])) \\ &= (1 - \frac{1}{\tau})V[t-1] + \frac{1}{\tau}A[t] \end{aligned} \tag{6}$$

It is apparent that the membrane potential at a given time, $V[t]$, is principally determined by its previous state $V[t-1]$

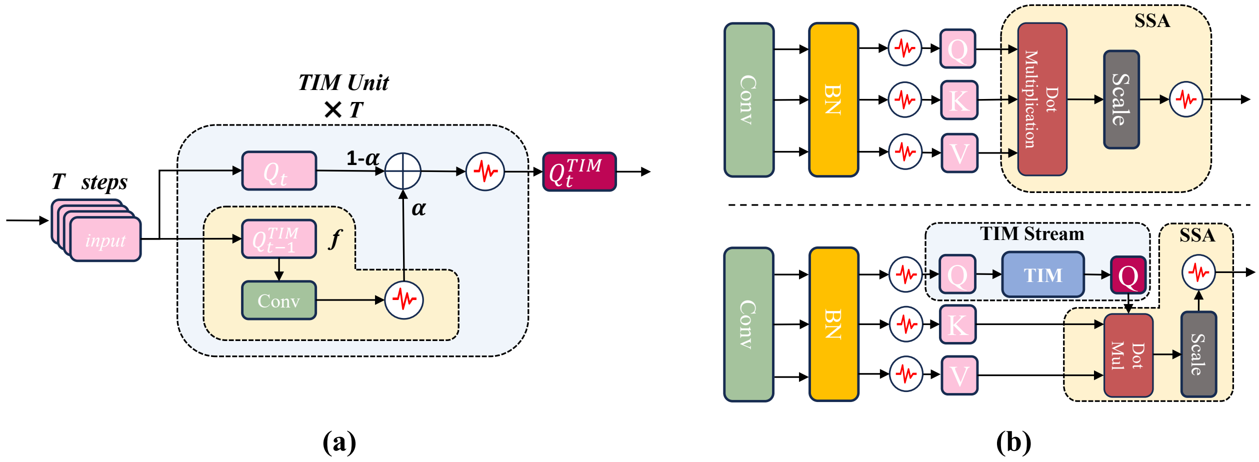

Figure 2: (a): Overview of the Temporal Interaction Module: Adaptive Integration of Historical and Current Temporal Data. (b): The upper part of the figure describes the process of SSA, while the lower part illustrates the process after integrating TIM Stream into SSA.

and the current input $A[t]$. The operation $Q[t]K[t]^TV[t]$ is chiefly responsible for facilitating the interaction of spatial information. However, the retention and extraction of temporal information rely solely on the dynamic changes in the neuronal membrane potential. In the existing spike self-attention mechanisms, temporal information has not been adequately considered, a deficiency that significantly limits the spiking Transformer's capability in processing time-series information.

## 4 Method

To enhance the temporal information processing capability of the Spiking Transformer, we have meticulously developed a plug-in Temporal Interaction Module (TIM). In this section, we will detail the architecture of TIM and explore its integration and application within the SSA computation.

### 4.1 Spikformer Backbone

In our experiments employing the Temporal Interaction Module, we have chosen Spikformer [Zhou *et al.*, 2022] as the primary framework, with its network architecture comprehensively illustrated in Fig. 1. Given TIM's emphasis on temporal enhancement within neuromorphic datasets, our inputs are formatted as Event Streams. We have preserved the key components of Spikformer for consistency across our experimental models. The Event Stream undergoes a transformation into requisite dimensions through the Spiking Patch Splitting (SPS) process, utilizing convolutional techniques. Within the Spiking Self Attention (SSA) module, the TIM Stream is employed for query operations to facilitate temporal enhancement. The Multi-Layer Perceptron (MLP) is structured with a bi-layered Convolution-Batch Normalization-Leaky Integrate-and-Fire (Conv-BN-LIF) framework. Finally, the Classification Head is constituted by a linear layer, aligning with the model's output objectives.

### 4.2 TIM Unit

The effective preservation, processing, and extraction of temporal features are crucial in handling neuromorphic data. As demonstrated in Eq. 5, traditional Spiking Transformers construct an attention matrix at each time step. However, this attention mechanism solely correlates with the current input, leading to a substantial underutilization of information from different time steps.

Here, $f$ signifies an operation for extracting historical information, implemented as a one-dimensional convolution in this study. Consequently, $Q^{TIM}[t]$ comprises two components: the first represents the contribution of historical information to the current attention, while the second signifies the direct impact of the current input. The hyperparameter $\alpha$ facilitates an adaptive balance between the significance of historical and current information. The specific illustration of TIM Unit can be found in Fig 2(a).

### 4.3 TIM Stream

As depicted in Fig. 2(b), this module is integrated into the computational graph of the attention matrix.

$$\begin{aligned} Attention[t] &= A^{TIM}[t]K[t]^TV[t] \\ &= \alpha f(Q^{TIM}[t-1])K[t]^TV[t] + (1-\alpha)Q[t]K[t]^TV[t] \end{aligned} \tag{7}$$

Compared to the traditional SSA, our introduction of the $f$ operation, which employs a one-dimensional convolution, does not significantly increase the number of parameters. By incorporating the TIM-built attention matrix, the model is capable of not only processing information from the current moment but also of utilizing output information from past moments, thereby effectively capturing the intrinsic dynamics of time series. This enhancement significantly bolsters the model's temporal memory capability and allows it to utilize historical information at each time step. This feature enables the model to dynamically adjust its behavior based on the characteristics of different tasks and data, thus improving computational efficiency and the generalization capacity of the model.

# 5 Experiments

To validate the performance of our algorithm, we conducted comprehensive tests on multiple neuromorphic datasets, including DVS-CIFAR10, N-CALTECH101, NCARS, UCF101-DVS, and HMDB51-DVS. Furthermore, to highlight the exceptional performance of our algorithm, we compared it with the current leading SNN models. All experiments were completed on the BrainCog [Zeng *et al.*, 2023] platform. In the experiments, we set the batchsize to 16 and used the AdamW optimizer. The total number of training epochs was set to 500. The initial learning rate was set to 0.005, adjusted with a cosine decay strategy. The time constant ($\tau$ value) of the LIF Node was set to 2, and its firing threshold was set to 1. The simulation step length of the SNN was set to 10. The default $\alpha$ of the TIM Stream was set to 0.5.

## 5.1 CIFAR10-DVS

**CIFAR10-DVS** is an event stream dataset comprising 10,000 images from the CIFAR-10 dataset. These images are converted into event streams using a bicubic interpolation method. As outlined in Tab. 1, our algorithm exhibits a significant improvement over the prevalent Spiking Transformer framework. When benchmarked against Spikformer, our TIM demonstrates enhanced performance, exceeding it by approximately 2.7% under identical configurations. Notably, even when Spikformer is optimized with a longer time step (Time Step=16), our algorithm maintains a performance lead of 0.7%. While Spikeformer attains an accuracy of 81.4% in 4 steps, it requires a substantial 9.28 million parameters. To our knowledge, this achievement marks the most advanced state-of-the-art (SOTA) result for Spiking Transformer applications on the CIFAR10-DVS dataset. In comparison with convolutional structures, the efficacy of TIM is on par with EventMix [Shen *et al.*, 2022b], which records a similar accuracy of 81.45%. However, a striking advantage of our model is its efficiency: while EventMix operates with an extensive 11.69 million parameters, our TIM model achieves comparable results with a significantly reduced parameter count of only 2.59 million.

## 5.2 N-CALTECH101

The **N-CALTECH101** dataset, as introduced in [Orchard *et al.*, 2015], is an innovative adaptation of the classic CALTECH101 image library, covering 101 diverse categories. This dataset is notably derived using advanced neuromorphic visual sensors. It uniquely offers a time-based visual event representation of the original static image categories, effectively capturing the dynamic and temporal nuances inherent in each category. In our research, we have compiled and analyzed several benchmarks from the N-CALTECH101 dataset alongside the results of our Temporal Integration Model, as summarized in Tab. 1. These benchmarks primarily stem from ANN architectures. However, our SNN model based TIM exhibits a remarkable performance edge over these benchmarks. Despite being compared with the Event Transformer which is an ANN-based Transformer model, our SNN-based TIM has achieved comparable performance.

## 5.3 NCARS

The **NCARS** dataset, as introduced in [Orchard *et al.*, 2015], represents a cutting-edge binary classification dataset. It consists of a comprehensive collection of dynamic event streams, meticulously capturing cars and non-car objects through the lens of event cameras. These event streams offer a rich, real-time depiction of visual phenomena, distinguishing themselves by their ability to encode temporal information about moving objects. Our TIM continues to set new benchmarks in the analysis of the NCARS dataset. It achieves a remarkable 96.5% accuracy, a testament to its robustness and advanced feature extraction capabilities. This level of performance not only surpasses the Event Transformer by a notable margin of 1.1% but also maintains a weak lead over EventMix, outperforming it by 0.2%. These results underscore the efficacy of TIM in handling the complex dynamics and temporal variations inherent in the NCARS dataset.

## 5.4 UCF101-DVS And HMDB51-DVS

The **UCF101-DVS** and **HMDB51-DVS** datasets represent neuromorphic adaptations of the well-known UCF101 and HMDB51 datasets, respectively. These adaptations transform the original video-based datasets into a neuromorphic format, creating a rich collection of visual event streams. The UCF101-DVS encompasses an array of 101 action categories, while the HMDB51-DVS includes 51 distinct action categories. In these datasets, each action category is meticulously converted into dynamic visual event streams using advanced event cameras, offering a novel perspective on action recognition. Our TIM has showcased exceptional performance on these challenging datasets. On the HMDB51-DVS dataset, TIM achieved a top-1 accuracy of 58.6%, and on the UCF101-DVS dataset, it reached a top-1 accuracy of 63.8%, as detailed in Tab. 2.

## 5.5 SHD

The Spiking Heidelberg Digits (**SHD**) dataset is an audio-based classification dataset containing 1000 spoken digits ranging from zero to nine in both English and German languages, with a total of 20 classes. Human speech contains more explicit information in the temporal sequence, and the interdependence and correlation between different time points in speech are richer than those derived from event datasets converted from static images. TIM demonstrates a 1.2% advantage over the baseline model on this dataset, shown in Tab 2. The results indicates TIM's capability to extract intricate temporal features.

# 6 Discussion

## 6.1 Ablation Study

To thoroughly validate the effectiveness of our algorithm, we embarked on comprehensive ablation studies utilizing the CIFAR10-DVS and N-CALTECH101 datasets. These datasets, known for their complexity and diverse visual event

| **Method** | **Architecture** | **Steps** | **CIFAR10-DVS** | **N-CALTECH101** | **NCARS** |
|---|---|---|---|---|---|
| SALT [Kim and Panda, 2021] | VGG-11 | 20 | 67.1 | 55.0 | - |
| Rollout [Kugele *et al.*, 2020] | VGG-16 | 48 | 66.5 | - | 94.1 |
| SEW-ResNet [Fang *et al.*, 2021a] | Wide-7B-Net | 20 | 74.4 | - | - |
| ResNet-18 [Shen *et al.*, 2022b] | ResNet-18 | 10 | 79.2 | 75.3 | 95.9 |
| EventMix [Shen *et al.*, 2022b] | ResNet-18 | 10 | 81.5 | 79.5 | 96.3 |
| NDA [Li *et al.*, 2022b] | ResNet-19 | 10 | 78.0 | 78.6 | 87.2 |
| Event Transformer [Li *et al.*, 2022c] | Transformer | -* | 71.2 | 78.9 | 95.4 |
| RM SNN [Yao *et al.*, 2023b] | PLIF-SNN | 10 | - | 77.9 | - |
| VMV-GCN [Xie *et al.*, 2022] | VMV-GCN | -* | 69.0 | 77.8 | 93.2 |
| Spikformer [Zhou *et al.*, 2022] | Spikformer | 10 / 16 | 78.9 / 80.9 | - | - |
| DISTA [Xu *et al.*, 2023] | Spikformer | 10 | 79.1 | - | - |
| Spikingformer [Zhou *et al.*, 2023a] | Spikformer | 10 / 16 | 79.9 / 81.3 | - | - |
| Spike-driven Transformer [Yao *et al.*, 2023a] | Spikformer | 16 | 80.0 | - | - |
| 2D-WT-Haar [Wang *et al.*, 2023] | Spikformer | 16 | 81.0 | - | - |
| Spikeformer [Li *et al.*, 2022a] | Spikeformer | 4 | 81.4 | - | - |
| **TIM(Ours)** | Spikformer | **10** | **81.6** | **79.0** | **96.5** |

Table 1: Comparison with other benchmark results on CIFAR10-DVS, N-CALTECH101 and NCARS. The $*$ indicates unknown time steps from the original studies.

| **Dataset** | **Model** | **Acc@1(%)** |
|---|---|---|
| UCF101-DVS | Event Frames + I3D / 3D-ResNet [Bi *et al.*, 2020] | 53.5 / 57.9 |
| | P3D-63[Qiu *et al.*, 2017] | 53.4 |
| | C3D [Tran *et al.*, 2015] | 47.2 |
| | Res-SNN18 + RM [Yao *et al.*, 2023b] | 63.5 |
| | **TIM(Ours)** | **63.8** |
| HMDB51-DVS | C3D [Tran *et al.*, 2015] | 41.7 |
| | P3D-63 [Qiu *et al.*, 2017] | 40.4 |
| | Res-SNN18 + RM [Yao *et al.*, 2023b] | 44.7 |
| | Spikepoint [Ren *et al.*, 2023] | 55.6 |
| | **TIM(Ours)** | **58.6** |
| SHD | Spikformer | 85.1 |
| | **TIM(Ours)** | **86.3** |

Table 2: Comparison with other benchmark results on UCF101-DVS, HMDB51-DVS and SHD.

| **Dataset** | **Model** | **Acc@1(%)** |
|---|---|---|
| CIFAR10-DVS | Baseline | 78.5 |
| | Baseline + **TIM** | 81.6 |
| N-CALTECH101 | Baseline | 77.7 |
| | Baseline + **TIM** | 79.0 |

Table 3: Ablation study of the temporal interaction module.

representations, provided an ideal testing ground for assessing the capabilities of our model. As detailed in Tab. 3 and Fig. 3, our model demonstrated substantial performance enhancements over the established baseline models in both datasets.

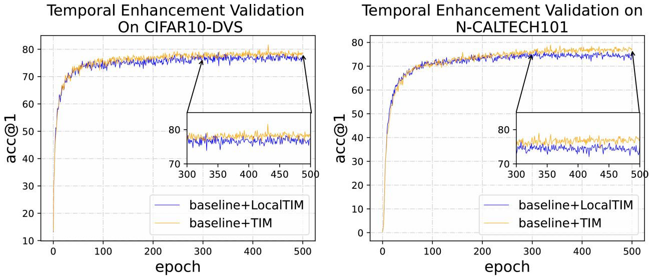

Figure 3: Temporal enhancement validation of TIM on CIFAR10-DVS and N-CALTECH101.

## 6.2 Temporal Enhancement Validation

To rigorously validate that the performance enhancements observed in the Spiking Transformer are indeed due to its improved temporal processing capabilities, and not merely a consequence of an increased parameter count, we undertook a meticulous adjustment of the Temporal Interaction Module's structure. This refinement was crucial in isolating the impact of temporal capability enhancements from the effects of parameter augmentation. As detailed in Eq. (8) and (9), we specifically re-engineered the operational mechanism of the TIM Stream. This modification was aimed at changing the functional dynamics of the TIM Unit. Instead of allowing the TIM Unit to engage in interactions across multiple time steps, which could potentially confound our assessment of its temporal processing proficiency, we constrained its operation to the current time step only (which we call local TIM).

$$Q[t] = LIFNode(TIM(Q[t])) \tag{8}$$

$$Attention[t] = Q[t]K^T[t]V[t] \tag{9}$$

| **Dataset** | **Model** | **Acc@1(%)** |
|---|---|---|
| CIFAR10-DVS | Baseline+local TIM | 78.9 |
| | Baseline + TIM | 81.6 |
| N-CALTECH101 | Baseline+local TIM | 76.6 |
| | Baseline + TIM | 79.0 |

Table 4: Temporal Enhancement Validation on CIFAR10-DVS and N-CALTECH101 datasets.

| **alpha** | **0** | **0.2** | **0.4** | **0.5** | **0.6** | **0.8** |
|---|---|---|---|---|---|---|
| CIFAR10-DVS | 78.5 | 79.8 | 79.7 | 80.7 | **81.6** | 81.2 |
| N-CALTECH101 | 77.7 | 78.5 | **79.0** | 78.5 | 77.6 | 78.7 |

Table 5: The top-1 accuracy of N-CALTECH101 and CIFAR10-DVS with different alpha values.

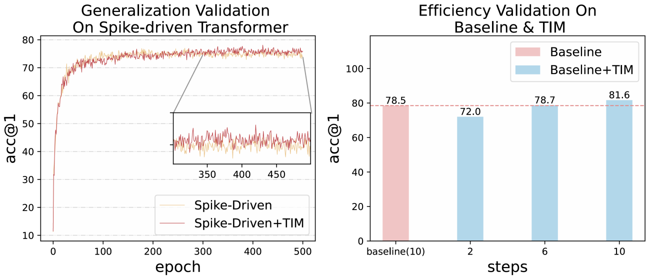

Figure 4: Efficiency and Generalizability Validation of TIM on CIFAR10-DVS.

As illustrated in Tab. 4, a notable experiment was conducted applying our model to the CIFAR10-DVS and N-CALTECH101 datasets. The results were quite revealing: the model's accuracy decreased by 2.7% and 2.4% on these datasets, respectively. This phenomenon strongly indicates that the improvement in accuracy previously attributed to the Temporal Interaction Module is primarily a result of its enhanced ability to process temporal information, rather than the increased parameter count. The stability of the parameter count during the experiment reinforces this conclusion, highlighting the intrinsic value of TIM in specifically enhancing temporal data processing capabilities. The comprehensive training details are presented in Fig. 3.

We further concentrated on investigating the impact of the hyperparameter $\alpha$ on the performance of the Temporal Integration Module. This exploration was conducted through a series of experiments using the CIFAR10-DVS and N-CALTECH101 datasets. The top-1 accuracy of TIM on both datasets with different $\alpha$ values are illustrated in Tab. 5. The findings from these experiments were quite revealing. We observed that regardless of the specific value, setting $\alpha$ to any non-zero number consistently resulted in better performance compared to the baseline scenario where $\alpha$ was set to zero. This pattern indicates that the introduction of temporal interaction in TIM significantly enhances its performance. More specifically, for the CIFAR10-DVS dataset, the model's peak performance, with an accuracy of 81.6%, was achieved when $\alpha$ was set to 0.6. Similarly, for the N-CALTECH101 dataset, the optimal performance occurred at an $\alpha$ value of 0.4, reaching a top accuracy of 79.0%. These results highlight an important aspect of TIM's functionality. While the model exhibits robustness to a range of $\alpha$ values, fine-tuning this hyperparameter allows for more precise optimization relative to specific datasets. The ability to adjust $\alpha$ effectively tailors the model to different data distributions, enhancing TIM's versatility and efficacy in diverse neuromorphic data processing applications.

### 6.3 Efficiency and Generalizability Validation

TIM introduces additional operations into the model, resulting in extra computational overhead. To alleviate concerns about TIM's reduced efficiency, we reduced the training steps. We found that with 6 steps, TIM achieves the same performance as the baseline does with 10 steps, shown in Fig 4. This implies that TIM can save approximately 40% of the time, suggesting that TIM retains the efficiency of SNN.

Furthermore, to validate the generalizability of TIM's capabilities to other Spiking Transformers, we conducted experiments on the Spike-driven Transformer as well. Before computing the Spike Driven Self Attention (SDSA, shown in Eq. 10) [Yao *et al.*, 2023a], where $\odot$ refers to element-wise multiplication while $\otimes$ denotes Hadamard product. We applied the same procedure to let Q pass through TIM, resulting in the curve shown in Fig 4.

$$\begin{cases} Q[t], K[t], V[t] = LIFNode(X[t]) \\ A[t] = LIFNode(K[t] \odot V[t]) \\ Attention[t] = A[t] \otimes Q[t] \end{cases} \tag{10}$$

SDSA achieved a top-1 accuracy of 77% on CIFAR10-DVS with 10 steps, while SDSA+TIM achieved a top-1 accuracy of 78.5%. We consider the 1.5% difference to be statistically significant, indicating that TIM gains traction in SDSA, demonstrating its ability to generalize across a broader range of Spiking Transformer architectures.

## 7 Conclusion

In our research, we conduct a detailed examination of the current Spiking Transformer models, focusing particularly on their suboptimal use of temporal information. To address this issue, we introduce the Temporal Interaction Module (TIM), a groundbreaking, plug-and-play component designed to enhance the Spiking Transformer's ability to process temporal data more effectively. This module, skillfully constructed using one-dimensional convolutions, is notable for its minimal parameter count. Our methodology ensures simplicity in implementation while adhering to the Spiking Transformer's principles of efficiency and compactness. Rigorous experimental validations underscore the efficacy of TIM, revealing its superior performance, especially prominent in neuromorphic dataset applications. TIM not only excels in these contexts but also establishes new state-of-the-art benchmarks. Furthermore, Our discussion and ablation study demonstrate that TIM effectively extracts temporal information on complex datasets, leading to improvements in model performance. Meanwhile, TIM retains the efficiency characteristic of SNNs and the ability to generalize across various Spiking Transformer architectures.

## Acknowledgments

This research is financially supported by a funding from Institute of Automation, Chinese Academy of Sciences (Grant No. E411230101).

## Contribution Statement

S.S. and D.Z. are equal contribution and serve as co-first authors. S.S., D.Z. and Y.Z. designed the study. S.S. D.Z. and G.S. performed the experiments. S.S., D.Z., G.S. and Y.Z. wrote the paper.

## References

[Bi *et al.*, 2020] Yin Bi, Aaron Chadha, Alhabib Abbas, Eirina Bourtsoulatze, and Yiannis Andreopoulos. Graph-based spatio-temporal feature learning for neuromorphic vision sensing. *IEEE Transactions on Image Processing*, 29:9084–9098, 2020.

[Cheng *et al.*, 2020] Xiang Cheng, Yunzhe Hao, Jiaming Xu, and Bo Xu. Lisnn: Improving spiking neural networks with lateral interactions for robust object recognition. In *IJCAI*, pages 1519–1525. Yokohama, 2020.

[Dong *et al.*, 2023] Yiting Dong, Dongcheng Zhao, Yang Li, and Yi Zeng. An unsupervised stdp-based spiking neural network inspired by biologically plausible learning rules and connections. *Neural Networks*, 2023.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2021.

[Fang *et al.*, 2020] Haowen Fang, Amar Shrestha, Ziyi Zhao, and Qinru Qiu. Exploiting neuron and synapse filter dynamics in spatial temporal learning of deep spiking neural network. *arXiv preprint arXiv:2003.02944*, 2020.

[Fang *et al.*, 2021a] Wei Fang, Zhaofei Yu, Yanqi Chen, Tiejun Huang, Timothée Masquelier, and Yonghong Tian. Deep residual learning in spiking neural networks. *Advances in Neural Information Processing Systems*, 34:21056–21069, 2021.

[Fang *et al.*, 2021b] Wei Fang, Zhaofei Yu, Yanqi Chen, Timothée Masquelier, Tiejun Huang, and Yonghong Tian. Incorporating learnable membrane time constant to enhance learning of spiking neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2661–2671, 2021.

[Gallego *et al.*, 2020] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020.

[Gehrig *et al.*, 2019] Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5633–5643, 2019.

[Kim and Panda, 2021] Youngeun Kim and Priyadarshini Panda. Optimizing deeper spiking neural networks for dynamic vision sensing. *Neural Networks*, 144:686–698, 2021.

[Kim *et al.*, 2023] Youngeun Kim, Yuhang Li, Hyoungseob Park, Yeshwanth Venkatesha, Anna Hambitzer, and Priyadarshini Panda. Exploring temporal information dynamics in spiking neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8308–8316, 2023.

[Kugele *et al.*, 2020] Alexander Kugele, Thomas Pfeil, Michael Pfeiffer, and Elisabetta Chicca. Efficient processing of spatio-temporal data streams with spiking neural networks. *Frontiers in Neuroscience*, 14:439, 2020.

[Li *et al.*, 2022a] Yudong Li, Yunlin Lei, and Xu Yang. Spikeformer: A novel architecture for training high-performance low-latency spiking neural network. *arXiv preprint arXiv:2211.10686*, 2022.

[Li *et al.*, 2022b] Yuhang Li, Youngeun Kim, Hyoungseob Park, Tamar Geller, and Priyadarshini Panda. Neuromorphic data augmentation for training spiking neural networks. In *European Conference on Computer Vision*, pages 631–649. Springer, 2022.

[Li *et al.*, 2022c] Zhihao Li, M Salman Asif, and Zhan Ma. Event transformer. *arXiv preprint arXiv:2204.05172*, 2022.

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

[Orchard *et al.*, 2015] Garrick Orchard, Ajinkya Jayawant, Gregory K Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9:437, 2015.

[Qiu *et al.*, 2017] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017.

[Rebecq *et al.*, 2019] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):1964–1980, 2019.

[Ren *et al.*, 2023] Hongwei Ren, Yue Zhou, Yulong Huang, Haotian Fu, Xiaopeng Lin, Jie Song, and Bojun Cheng. Spikepoint: An efficient point-based spiking neural network for event cameras action recognition. *arXiv preprint arXiv:2310.07189*, 2023.

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

[Sengupta *et al.*, 2019] Abhronil Sengupta, Yuting Ye, Robert Wang, Chiao Liu, and Kaushik Roy. Going deeper in spiking neural networks: Vgg and residual architectures. *Frontiers in neuroscience*, 13:95, 2019.

[Shen *et al.*, 2022a] Guobin Shen, Dongcheng Zhao, and Yi Zeng. Backpropagation with biologically plausible spatiotemporal adjustment for training deep spiking neural networks. *Patterns*, 3(6), 2022.

[Shen *et al.*, 2022b] Guobin Shen, Dongcheng Zhao, and Yi Zeng. Eventmix: An efficient augmentation strategy for event-based data, 2022.

[Shen *et al.*, 2023] Guobin Shen, Dongcheng Zhao, Yiting Dong, and Yi Zeng. Brain-inspired neural circuit evolution for spiking neural networks. *Proceedings of the National Academy of Sciences*, 120(39):e2218173120, 2023.

[Shen *et al.*, 2024] Guobin Shen, Dongcheng Zhao, and Yi Zeng. Exploiting nonlinear dendritic adaptive computation in training deep spiking neural networks. *Neural Networks*, 170:190–201, 2024.

[Tran *et al.*, 2015] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2023] Qingyu Wang, Duzhen Zhang, Tielin Zhang, and Bo Xu. Attention-free spikformer: Mixing spike sequences with simple linear transforms. *arXiv preprint arXiv:2308.02557*, 2023.

[Wei *et al.*, 2024] Wenjie Wei, Malu Zhang, Jilin Zhang, Ammar Belatreche, Jibin Wu, Zijing Xu, Xuerui Qiu, Hong Chen, Yang Yang, and Haizhou Li. Event-driven learning for spiking neural networks. *arXiv preprint arXiv:2403.00270*, 2024.

[Xie *et al.*, 2022] Bochen Xie, Yongjian Deng, Zhanpeng Shao, Hai Liu, and Youfu Li. Vmv-gcn: Volumetric multi-view based graph cnn for event stream classification. *IEEE Robotics and Automation Letters*, 7(2):1976–1983, 2022.

[Xu *et al.*, 2023] Boxun Xu, Hejia Geng, Yuxuan Yin, and Peng Li. Dista: Denoising spiking transformer with intrinsic plasticity and spatiotemporal attention, 2023.

[Yao *et al.*, 2021] Man Yao, Huanhuan Gao, Guangshe Zhao, Dingheng Wang, Yihan Lin, Zhaoxu Yang, and Guoqi Li. Temporal-wise attention spiking neural networks for event streams classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10221–10230, 2021.

[Yao *et al.*, 2023a] Man Yao, Jiakui Hu, Zhaokun Zhou, Li Yuan, Yonghong Tian, Bo Xu, and Guoqi Li. Spike-driven transformer, 2023.

[Yao *et al.*, 2023b] Man Yao, Hengyu Zhang, Guangshe Zhao, Xiyu Zhang, Dingheng Wang, Gang Cao, and Guoqi Li. Sparser spiking activity can be better: Feature refine-and-mask spiking neural network for event-based visual recognition. *Neural Networks*, 166:410–423, 2023.

[Zeng *et al.*, 2023] Yi Zeng, Dongcheng Zhao, Feifei Zhao, Guobin Shen, Yiting Dong, Enmeng Lu, Qian Zhang, Yinqian Sun, Qian Liang, Yuxuan Zhao, et al. Braincog: A spiking neural network based, brain-inspired cognitive intelligence engine for brain-inspired ai and brain simulation. *Patterns*, 4(8), 2023.

[Zhang *et al.*, 2021] 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.

[Zhao *et al.*, 2020] Dongcheng Zhao, Yi Zeng, Tielin Zhang, Mengting Shi, and Feifei Zhao. Glsnn: A multi-layer spiking neural network based on global feedback alignment and local stdp plasticity. *Frontiers in Computational Neuroscience*, 14:576841, 2020.

[Zhao *et al.*, 2023] Dongcheng Zhao, Guobin Shen, Yiting Dong, Yang Li, and Yi Zeng. Improving stability and performance of spiking neural networks through enhancing temporal consistency. *arXiv preprint arXiv:2305.14174*, 2023.

[Zheng *et al.*, 2023] Xu Zheng, Yexin Liu, Yunfan Lu, Tongyan Hua, Tianbo Pan, Weiming Zhang, Dacheng Tao, and Lin Wang. Deep learning for event-based vision: A comprehensive survey and benchmarks. *arXiv preprint arXiv:2302.08890*, 2023.

[Zhou *et al.*, 2022] Zhaokun Zhou, Yuesheng Zhu, Chao He, Yaowei Wang, Shuicheng Yan, Yonghong Tian, and Li Yuan. Spikformer: When spiking neural network meets transformer. *arXiv preprint arXiv:2209.15425*, 2022.

[Zhou *et al.*, 2023a] 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, 2023.

[Zhou *et al.*, 2023b] Yue Zhou, Jiawei Fu, Zirui Chen, Fuwei Zhuge, Yasai Wang, Jianmin Yan, Sijie Ma, Lin Xu, Huanmei Yuan, Mansun Chan, et al. Computational event-driven vision sensors for in-sensor spiking neural networks. *Nature Electronics*, pages 1–9, 2023.

[Zhu *et al.*, 2024] Rui-Jie Zhu, Malu Zhang, Qihang Zhao, Haoyu Deng, Yule Duan, and Liang-Jian Deng. Tcja-snn: Temporal-channel joint attention for spiking neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.