Autaptic Synaptic Circuit Enhances Spatio-temporal Predictive Learning of Spiking Neural Networks

Lihao Wang [1 2] **Zhaofei Yu** [1]

Abstract

Spiking Neural Networks (SNNs) emulate the integrated-fire-leak mechanism found in biological neurons, offering a compelling combination of biological realism and energy efficiency. In recent years, they have gained considerable research interest. However, existing SNNs predominantly rely on the Leaky Integrate-and-Fire (LIF) model and are primarily suited for simple, static tasks. They lack the ability to effectively model long-term temporal dependencies and facilitate spatial information interaction, which is crucial for tackling complex, dynamic spatio-temporal prediction tasks. To tackle these challenges, this paper draws inspiration from the concept of autaptic synapses in biology and proposes a novel Spatio-Temporal Circuit (STC) model. The STC model integrates two learnable adaptive pathways, enhancing the spiking neurons' temporal memory and spatial coordination. We conduct a theoretical analysis of the dynamic parameters in the STC model, highlighting their contribution in establishing long-term memory and mitigating the issue of gradient vanishing. Through extensive experiments on multiple spatio-temporal prediction datasets, we demonstrate that our model outperforms other adaptive models. Furthermore, our model is compatible with existing spiking neuron models, thereby augmenting their dynamic representations. In essence, our work enriches the specificity and topological complexity of SNNs.

1. Introduction

Spiking Neural Networks (SNNs) (Maass, 1997) have emerged as a promising paradigm for brain-inspired computation, garnering considerable attention due to their biological plausibility and energy efficiency in comparison to Artificial Neural Networks (ANNs). Similar to real biological neurons, the fundamental computational units in SNNs, known as spiking neurons, exhibit diverse dynamics and possess the ability to process both temporal and spatial information concurrently. Information is encoded and transmitted between spiking neurons through binary spike signals. These spike signals are sparse and event-driven, endowing SNNs deployed on neuromorphic hardware with significant advantages in terms of computational efficiency and energy consumption (Roy et al., 2019; Pei et al., 2019; Shi et al., 2024). Moreover, SNNs offer greater flexibility in integrating relevant biological mechanisms and leveraging insights from neuroscience to address persistent challenges encountered in ANN models, such as catastrophic forgetting (Zhang et al., 2023b) and robustness (Chen et al., 2022).

Due to the non-differentiable nature of the spiking process, training SNNs directly through backpropagation has been a challenging task. However, the development of surrogate gradient approaches (Wu et al., 2018; Neftci et al., 2019) has significantly advanced the training of SNNs. Presently, it is feasible to train SNNs with hundreds of layers (Fang et al., 2021a), achieving performance comparable to state-of-the-art ANNs in tasks like image and speech classification (Deng et al., 2022; Guo et al., 2022; Yin et al., 2021). Nonetheless, these tasks primarily focus on static feature recognition. It has been observed that the widely adopted LIF model (Gerstner & Kistler, 2002), which serves as a standard spiking neuron model, exhibits limitations in capturing deep spatio-temporal information features and modeling spatio-temporal dependencies. As a result, SNNs face notable constraints when dealing with spatio-temporal prediction tasks.

In contrast, all organisms live in a dynamically changing world where we frequently encounter complex, dynamic, and uncertain tasks. The brain is highly specialized (Koch & Laurent, 1999), comprising numerous distinct types of neurons that exhibit diverse dynamic behaviors and characteristics (Harris & Shepherd, 2015; Yao et al., 2021), along with a variety of connectivity patterns (De Wit & Ghosh, 2016; Sanes & Zipursky, 2020). This intricate and complex topology has been demonstrated to allow the brain to re-

[1]Institute for Artificial Intelligence, Peking University, Peking, China. [2]Shenzhen X-Institute, Shenzhen, China. Correspondence to: Zhaofei Yu <yuzf12@pku.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

spond differently to various stimuli and tasks (Golomb & Rinzel, 1993; Rich et al., 2022). The crucial factor enabling the brain to excel in addressing dynamic problems lies in its capacity to integrate and analyze information at different levels and scales (Hutt et al., 2023; Gast et al., 2024).

In this paper, we propose the Spatio-Temporal Circuit Leaky Integrate-and-Fire (STC-LIF) model, a novel spatio-temporal self-connection model, inspired by biological autaptic synapses to enhance the specificity and topological complexity of SNNs, thereby improving their performance on spatio-temporal tasks. The STC-LIF model incorporates the axon-dendrite circuit and axon-soma circuit, which dynamically regulate the input current and historical information through proportional control. By leveraging this self-connection circuit, our model effectively promotes the interaction of spiking neurons in both spatial and temporal dimensions. The main contribution can be summarized as:

- We conduct a comprehensive analysis of LIF and its variants in spatio-temporal predictive learning, and identify the limitations of modeling long-term dependency and representing spatio-temporal information.
- We propose the STC-LIF model, which is inspired by biological autaptic synapses, to enhance the ability to extract spatio-temporal features and model spatio-temporal dependencies.
- Experimental results demonstrate the effectiveness of the spatio-temporal self-connected circuit in improving the learning of SNNs. Our method achieves remarkable improvements in various spatio-temporal prediction datasets, surpassing other adaptive neuron models.
- The concept of spatio-temporal self-connection can be easily transferred and scaled. We extend the application of this method to Parametric Leaky Integrate-and-Fire (PLIF) and Learnable Multi-Hierarchica (LM-H) models, achieving significant performance improvements in spatio-temporal prediction tasks.

2. Related Work

2.1. Learning of Spiking Neural Networks

The learning algorithms for SNNs can be divided into three main approaches: unsupervised learning, ANN-to-SNN conversion, and supervised learning. Unsupervised learning approaches primarily utilize bio-plausible learning rules, such as Hebbian learning (Hebb, 2005) and Spike-Timing-Dependent Plasticity (Bi & Poo, 1998). These optimization methods enable local plasticity learning based on the precise timing of spikes, aligning with the inherent characteristics of SNNs. However, they are typically limited to shallow networks (Diehl & Cook, 2015; Lee et al., 2018; Liu & Yue, 2018). When applied to deep networks, unsupervised learning methods necessitate combination with other techniques like clustering algorithms (Lu & Sengupta, 2023).

The second approach involves converting pre-trained ANNs into SNNs without the need for additional training. Early works (Diehl et al., 2015; Rueckauer et al., 2017) focused on approximating the ReLU function using the firing rate of spiking neurons. While these methods achieved lossless accuracy, they required long simulation time steps. More recent research focuses on reducing conversion error to enhance performance at low latency. These techniques include initial potential compensation (Bu et al., 2022a), revised activation function (Bu et al., 2022b; Deng & Gu, 2021), and spike calibrating (Hao et al., 2023a).

The supervised learning approach based on Backpropagation Through Time (BPTT) draws inspiration from the learning of recurrent neural networks. Due to the non-differentiability of the step function, a surrogate function is often used to approximate the gradient (Wu et al., 2018; Neftci et al., 2019). This method enhances the stability and effectiveness of SNN training, serving as a fundamental technique in SNN research (Zenke & Vogels, 2021; Stewart & Neftci, 2022; Lee et al., 2020). To further enhance the performance of SNN, Wang et al. (2023) and Deng et al. (2023) introduced adaptive gradient error computation to address the gradient mismatch between the surrogate and the original functions. Additionally, developing new normalization techniques (Wu et al., 2019; Zheng et al., 2021; Duan et al., 2022; Kim & Panda, 2021) and residual connections (Fang et al., 2021a) can significantly enhance the convergence speed and stability of deep SNNs. Xiao et al. (2022) and Zhu et al. (2024) implemented online training for SNNs, largely reducing the memory cost for GPU training.

2.2. Adaptive Spiking Neural Models

The fixed-parameter spiking neuron models commonly used for training and inference impose limitations on the representational capacity of SNNs. Bellec (2018) associated the firing threshold range with the spiking activity of the network during training, improving the dynamics of spiking neurons. To further enhance the adaptability of spiking neurons, researchers introduced learnable membrane leakage parameters and threshold parameters (Rathi & Roy, 2021; Fang et al., 2021b; Wu et al., 2021). GLIF (Yao et al., 2022) utilized gating factors to control the neuronal dynamical patterns, simulating different biological characteristics to expand the representation space and improve heterogeneity. To mitigate the issues of gradient vanishing and divergence, TC-LIF (Zhang et al., 2023a) and LM-H (Hao et al., 2023b) introduced dual membrane potentials and corresponding learnable parameters. These modifications significantly enhanced the adaptability of SNNs. However, despite enriching the representation space and adaptability, the model parameters of these approaches remained fixed

during the inference stage. This limitation evidently reduced the generalization capability of SNNs when confronted with dynamically changing data.

3. Preliminaries

3.1. Vanilla LIF Model

The vanilla Leaky Integrate-and-Fire (LIF) model is currently the most commonly used spiking neuron model in SNNs. The neural dynamics of the LIF model incorporate three processes of biological neurons (Gerstner et al., 2014), including integration, firing, and leakage. These processes can be described by the following equations (Figure 1(b)):

$$\boldsymbol{x}^l[t] = \boldsymbol{W}^l \boldsymbol{s}^{l-1}[t] \tag{1}$$

$$\boldsymbol{m}^l[t] = \alpha^l \boldsymbol{v}^l[t-1] + (1-\alpha^l)\boldsymbol{x}^l[t] \tag{2}$$

$$\boldsymbol{s}^l[t] = H(\boldsymbol{m}^l[t] - v_{th}) \tag{3}$$

$$\boldsymbol{v}^l[t] = \boldsymbol{m}^l[t] - v_{th}\boldsymbol{s}^l[t] \tag{4}$$

Here, $\boldsymbol{x}^l[t]$ represents the synaptic current input to layer l at time-step t from presynaptic neurons, and $\boldsymbol{s}^l[t]$ represents the spike firing of the neurons in layer l at time-step t, the element of which equals 1 if the corresponding neuron fires and 0 otherwise. The parameter $\boldsymbol{W}^l$ denotes the synaptic weight between layer l and the previous layer, $l-1$. $\boldsymbol{m}^l[t]$ and $\boldsymbol{v}^l[t]$ represent the membrane potential of neurons in layer l before and after firing, respectively. $\alpha^l$is the membrane constant that controls the decay ratio of input and historical information, and it ranges from 0 to 1. When Eq. (2) changes to $\boldsymbol{m}^l[t] = \boldsymbol{v}^l[t-1] + (1-\alpha^l)\boldsymbol{x}^l[t]$, the vanilla LIF model degenerates to simple Integrate-and-Fire (IF) model. $H(\cdot)$ represents the Heaviside step function. Note that this paper employs a soft reset mechanism to keep more information, as suggested by (Rueckauer et al., 2017; Han et al., 2020). Specifically, when a spike is fired, the final membrane potential value is reduced by a threshold.

3.2. Autapses

An autapse (Van Der Loos & Glaser, 1972) is a unique synaptic structure that refers to a synaptic connection formed between a neuron's axon and its own dendrites or soma (Figure 1(a)). Initially observed in laboratory-cultured neurons, autapses were regarded as erroneous or redundant synaptic connectivity patterns. However, subsequent studies have revealed the presence of autapses in various types of neurons within the central nervous system, including those in the striatum (Park et al., 1980; Preston et al., 1980; Shi & Rayport, 1994), neocortex (Lübke et al., 1996), and hippocampal regions (Thomson et al., 1996; Tamas et al., 1997; Cobb et al., 1997). In terms of functionality, inhibitory activation of autapses can suppress the generation of subsequent action potentials (Bacci et al., 2003), thereby regulating the precision of neuronal discharge (Bacci & Huguenard, 2006). Conversely, excitatory activation of autapses can enhance neuronal firing, influence neuronal responses, and facilitate coincidence detection (Yin et al., 2018). These findings highlight the significant regulatory role of autapses in neuronal discharge and emphasize their status as functional structures within the cortical circuits of the brain.

3.3. Spatio-temporal Predictive Learning

Spatio-temporal predictive learning aims to learn future frames based on historical frames. Given a video sequence $X_{t,T} = \{y_i\}_{t-T+1}^{t}$ from time t to the past T frames, our objective is to predict the subsequent T' frames $Y_{t+1,T'} = \{y_i\}_{t+1}^{t+T'}$, where $y_i \in \mathbb{R}^{C\times H\times W}$ represents an image with channel number C, height H, and width W. The model with learnable parameters Θ learns a mapping $\mathcal{F}_\Theta : X_{t,T} \longmapsto Y_{t+1,T'}$, and the optimization objective is:

$$\Theta^* = \arg\min_{\Theta} \mathcal{L}(\mathcal{F}_\Theta(X_{t,T}), Y_{t+1,T'}) \tag{5}$$

Here $\mathcal{L}$ is the loss function. In our experiments, we consistently adopt the widely used mean squared error (MSE) function as the loss function.

4. Methodology

In this section, we systematically analyze three primary limitations inherent in the vanilla LIF model when applied to spatio-temporal prediction tasks. To overcome these challenges, we propose a spatio-temporal self-connection circuit model inspired by biological autaptic structures. We elucidate the design principles of the model and provide a comprehensive analysis of its dynamic characteristics.

4.1. Limitations of the Vanilla LIF Model in Spatio-temporal Prediction Tasks

(a): Limited capacity to model long-term dependencies. Spatio-temporal prediction tasks involve capturing temporal features and dependencies to predict future states. The vanilla LIF model employs leakage and discharge processes to regulate historical information, thereby exhibiting certain sequential modeling capabilities. However, its discharge process follows a fixed linear decay, while the leakage process exhibits a fixed exponential decay. To be more specific, according to Eqs. (1)-(4), the membrane potential of the vanilla LIF model at time-step $t+T$, derived from $\boldsymbol{v}^l[t]$, can be obtained after T iterations and is represented as follows:

$$\boldsymbol{v}^l[t+T] = (\alpha^l)^T \boldsymbol{v}^l[t] + \sum_{i=1}^{T} (\alpha^l)^{T-i}(1-\alpha^l)\boldsymbol{x}^l[t+i] - v_{th}\sum_{i=1}^{T}(\alpha^l)^{T-i}\boldsymbol{s}^l[t+i] \tag{6}$$

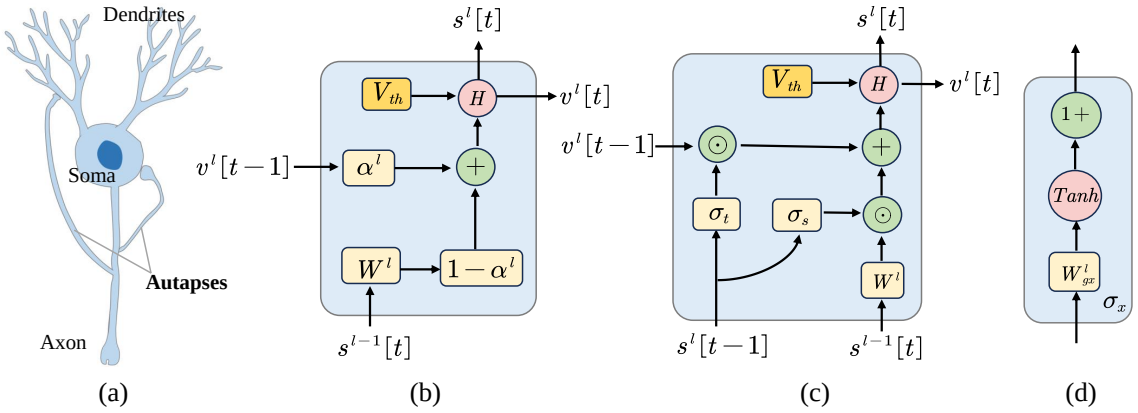


Figure 1. (a) The structure of Autapses in neuroscience, which includes axon-soma and axon-dendrite circuits. (b) The structure of the vanilla LIF model. (c) The structure of the STC-LIF model. (d) The structure of dynamic spatio-temporal circuit module $\boldsymbol{\sigma}$ in Figure 1(c).

The detailed derivation is provided in the Appendix. According to Eq. (6), we observe an inevitable decay in the historical membrane potential information $\boldsymbol{v}_t^l$. Specifically, $\boldsymbol{v}_t^l$ at time-step t undergoes T steps of exponential decay and $\sum_{i=1}^{T} \boldsymbol{s}^l[t+i]$ steps of linear decay after T iterations. As α^l approaches 0, the decay of membrane potential information becomes more pronounced. Conversely, as α^l approaches 1, it leads to a weaker input current. Due to the presence of leakage, maintaining membrane potential balance and outputting spike information heavily relies on the current input at the current moment. This monotonic leakage mechanism poses challenges for the vanilla LIF model in capturing long-term dependencies. Moreover, previous studies (Zhang et al., 2023a) have highlighted the issue of gradient vanishing when employing BPTT to optimize SNNs with LIF neurons for long-term modeling tasks, further restricting the model's optimization capabilities.

(b): Limited capacity to represent spatio-temporal information. Spatio-temporal prediction necessitates the deep interactions and joint learning of temporal and spatial information, which poses a challenge for LIF model. To see this, we rewrite the output of the vanilla LIF model at time t as:

$$\boldsymbol{s}^l[t] = f(\boldsymbol{x}^l[t], \boldsymbol{v}^l[t-1]) = f(\boldsymbol{W}^l \boldsymbol{s}^{l-1}[t], \boldsymbol{v}^l[t-1]) \quad (7)$$

Here, function $f(\cdot)$ denotes the dynamics of the LIF neuron (Eqs. (2)-(3)). For the LIF model, the spiking neurons in the l-th layer only receive inputs from the spike outputs of neurons in previous layer, $\boldsymbol{s}^{l-1}[t]$, while the incorporation of historical information is solely based on the neuron's own input at the previous time-step. Additionally, the network's learnable parameters primarily focus on extracting spatial information, lacking effective parameter training for feature extraction in the temporal domain. This simple connectivity pattern and topology restrict the network's capacity to attain a profound representation of spatial-temporal information.

(c): Weak adaptability to different datasets. Spatio-temporal prediction requires modeling dynamic spatio-temporal data. However, the parameters of the LIF model remain unchanged, whether during training or inference. Although some variants of the LIF model optimize parameters like membrane time constants and thresholds in training (Fang et al., 2021b; Yao et al., 2022), these parameters remain fixed in inference. The distribution and correlation of spatio-temporal data can change over time and space, and fixed parameters are insufficient to capture these variations, resulting in reduced generalization ability in prediction.

4.2. Spatio-Temporal Circuit Leaky Integrate-and-Fire (STC-LIF) Model

To overcome the challenges associated with spatio-temporal data learning in vanilla LIF and its variant models, we propose a novel spatio-temporal self-connection circuit model called the STC-LIF model, which can be described as:

$$\boldsymbol{x}^l[t] = \boldsymbol{W}^l \boldsymbol{s}^{l-1}[t] \quad (8)$$

$$\boldsymbol{\beta}^l[t] = \mathrm{Tanh}(\boldsymbol{W}^l_{gt} \boldsymbol{s}^l[t-1]) \quad (9)$$

$$\boldsymbol{\gamma}^l[t] = \mathrm{Tanh}(\boldsymbol{W}^l_{gs} \boldsymbol{s}^l[t-1]) \quad (10)$$

$$\boldsymbol{m}^l[t] = \boldsymbol{v}^l[t-1] \odot (1 + \boldsymbol{\beta}^l[t]) + \boldsymbol{x}^l[t] \odot (1 + \boldsymbol{\gamma}^l[t]) \quad (11)$$

$$\boldsymbol{s}^l[t] = H(\boldsymbol{m}^l[t] - v_{th}) \quad (12)$$

$$\boldsymbol{v}^l[t] = \boldsymbol{m}^l[t] - v_{th} \boldsymbol{s}^l[t] \quad (13)$$

Here, $\odot$ denotes the Hadamard product. Compared to the vanilla LIF model presented in Eqs. (1)-(4), the STC-LIF model retains the integration, spiking, and leakage processes while introducing axon-dendrite and axon-soma circuits inspired by biological autaptic synapses to dynamically regulate the input current and historical information (Figure 1(c)).

It replaces the fixed leakage mechanism with two dynamic regulatory factors, $\boldsymbol{\beta}^l[t]$ and $\boldsymbol{\gamma}^l[t]$. $\boldsymbol{W}^l_{gt}$ and $\boldsymbol{W}^l_{gs}$ represent the synaptic weights of the temporal and the spatial circuits.

We further elaborate on the design principles of STC-LIF. Previous experiments (Bacci et al., 2003; Bacci & Huguenard, 2006; Yin et al., 2018) have demonstrated that autaptic synapses can regulate the precision and stability of action potential firing by forming delayed feedback that either inhibits or enhances subsequent action potential generation. Therefore, we introduce the regulatory factors, $\boldsymbol{\beta}^l[t]$ and $\boldsymbol{\gamma}^l[t]$, to simulate the axon-soma and axon-dendrite circuits found in autaptic structures. These factors are within the range of [-1, 1] and determined by the spike output of the previous time step, $\boldsymbol{s}^l[t-1]$. In the axon-dendrite circuit, if $\boldsymbol{\gamma}^l[t] > 0$, it indicates the dominance of excitatory synapses, resulting in an enhancement of the input current to the pre-synaptic neuron and promoting spiking. Conversely, if $\boldsymbol{\gamma}^l[t] < 0$, it signifies the dominance of inhibitory synapses, leading to an attenuation of the input current to the pre-synaptic neuron and suppressing spiking. Similarly, in the axon-soma circuit, if $\boldsymbol{\beta}^l[t] > 0$, it amplifies the membrane potential, thereby preserving significant historical information. Conversely, if $\boldsymbol{\beta}^l[t] < 0$, it weakens the membrane potential, facilitating the forgetting of irrelevant historical information.

The spatio-temporal circuit enables spiking neurons to regulate the current input and membrane potential at the current time step based on their spike output from the previous time step, aligning with the functionality of autaptic synaptic in biology. However, relying solely on self-spike output for regulation has limitations and inflexibility, potentially constraining the expressive power of SNNs. To address this, we extend the concept of self-connection beyond individual neurons to include connections within the layer of neighboring cells. We use group convolution operation to characterize the influence of the spike output from the neuron itself as well as its adjacent neurons on a spiking neuron. This approach is not only computationally efficient but also more biologically plausible, as it aligns with the sparse connectivity observed in biological networks, where neurons primarily engage in local information exchange rather than global interactions (Guo et al., 2017).

4.3. Dynamic Analysis of STC-LIF Model

Relate STC-LIF model to LIF and IF models. The STC-LIF model is a natural extension of the vanilla LIF model. Here we establish the relationship between STC-LIF and the LIF and IF models. According to Eqs. (8)-(13), when $\boldsymbol{\gamma}^l[t]$ is less than zero and $\boldsymbol{\beta}^l[t] + \boldsymbol{\gamma}^l[t]$ equals -1, the STC-LIF model is equivalent to the LIF model. When $\boldsymbol{\beta}^l[t] = \boldsymbol{\gamma}^l[t] = 0$, the STC-LIF model is equivalent to the IF model.

We further analyze the advantages of the STC-LIF model compared to the LIF and IF models. The derivation process follows a similar approach to that of Eq. (6). According to Eqs. (8)-(13), the membrane potential of the STC-LIF model at time-step $t+T$, derived from $\boldsymbol{v}^l[t]$, can be obtained after T iterations and is represented as follows:

$$\begin{aligned}\boldsymbol{v}^l[t+T] &= \boldsymbol{v}^l[t]\prod_{j=1}^{T}(1+\boldsymbol{\beta}^l[t+j]) \\ &+ \sum_{i=1}^{T}(1+\boldsymbol{\gamma}^l[t+i])\boldsymbol{x}^l[t+i]\prod_{j=1}^{T-i}(1+\boldsymbol{\beta}^l[t+i+j]) \\ &- \boldsymbol{v}_{th}\sum_{i=1}^{T}\boldsymbol{s}^l[t+i]\prod_{j=1}^{T-i}(1+\boldsymbol{\beta}^l[t+i+j])\end{aligned} \tag{14}$$

The detailed derivation is provided in the Appendix. From Eq. (14), the STC-LIF model has the following advantages:

(a): STC-LIF model can capture long-term dependency. From Eq. (14), it is evident that the historical membrane potential information is determined by the dynamic modulation factors $\boldsymbol{\beta}^l[t]$ and $\boldsymbol{\gamma}^l[t]$. For the membrane potential $\boldsymbol{v}^l[t]$ at time-step t, we can effectively transmit it to the membrane potential $\boldsymbol{v}^l[t+T]$ at T steps later by employing appropriate dynamic modulation. The factor $\boldsymbol{\gamma}^l[t]$ also plays a vital role in regulating the current input at the current time-step, thereby ensuring a relative balance between the input and the historical membrane potential. Moreover, the spatio-temporal circuit plays a critical role in preserving the propagation of gradient errors, which is fundamental for optimizing SNNs using the BPTT method. The temporal flow of gradient information in the STC-LIF model can be represented by the following recursive form:

$$\begin{aligned}\frac{\partial \boldsymbol{v}^l[t+T]}{\partial \boldsymbol{v}^l[t]} &= \prod_{i=1}^{T}\frac{\partial \boldsymbol{v}^l[t+i]}{\partial \boldsymbol{m}^l[t+i]}\frac{\partial \boldsymbol{m}^l[t+i]}{\partial \boldsymbol{v}^l[t+i-1]} \\ &= \prod_{i=1}^{T}(1-v_{th}H'(\boldsymbol{m}^l[t+i]-v_{th}))(1+\boldsymbol{\beta}^l[t+i])\end{aligned} \tag{15}$$

The derivative $H'(\cdot)$ is often approximated by the derivative of a continuous surrogate function. As T increases, it is possible that $\prod_{i=1}^{T}(1-v_{th}H'(\boldsymbol{m}^l[t+i]-v_{th})) \to 0$ or $\prod_{i=1}^{T}(1-v_{th}H'(\boldsymbol{m}^l[t+i]-v_{th})) \to \infty$, which may lead to gradient vanishing and exploding issues. However, by appropriately assigning the values of $\boldsymbol{\beta}^l[t]$, we can effectively alleviate the problems, thereby preserving the propagation of gradient flow information.

(b): STC-LIF model enhances spatio-temporal representation capability. The spike output of the STC-LIF model for layer l can be represented as follows:

$$\boldsymbol{s}^l[t] = f(\boldsymbol{W}^l\boldsymbol{s}^{l-1}[t], \boldsymbol{W}^l_{gs,gt}\boldsymbol{s}^l[t-1], \boldsymbol{v}^l[t-1]) \tag{16}$$

Here, the function $f(\cdot)$ represents the dynamics of the STC-LIF neuron, as depicted in Eqs. (9)-(12). The spike output

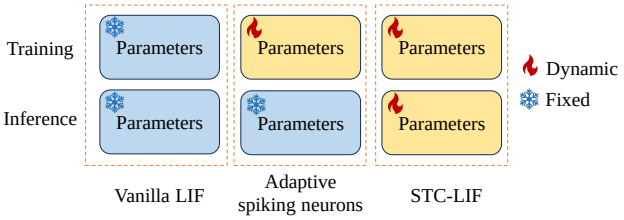


Figure 2. Comparison of parameter changes of different models.

of the STC-LIF model is influenced by both asynchronous signals within the layer and synchronous signals between layers. The spatio-temporal circuit forms new synaptic connections, thereby enhancing the interaction of spatio-temporal information. The learnable synaptic weights $\boldsymbol{W}_{gs}^{l}$ and $\boldsymbol{W}_{gt}^{l}$ strengthen the capability to extract spatial and temporal features, consequently advancing the representational capacity of spiking neurons.

(c): STC-LIF model has dynamic parameters. As shown in Figure 2, the parameters of the STC-LIF model undergo dynamic changes during both training and inference. The current parameters of the neuron are influenced by the feedback modulation from its output in the previous time step. In essence, the parameters of the STC-LIF model are adaptive and evolve with the input sequence. This enables the generation of diverse dynamics parameters, catering to different input sequences and thereby enhancing the network's adaptability to various tasks and scenarios. Consequently, the model exhibits improved self-adaptiveness.

5. Experiments

In this section, we conducted experiments on multiple spatio-temporal prediction benchmark tasks to validate the effectiveness of our proposed method. These tasks include the Moving MNIST (Srivastava et al., 2015) dataset, the TaxiBJ (Zhang et al., 2017) traffic flow dataset, and the KTH (Schuldt et al., 2004) human action dataset. We compare our method with the Vanilla LIF model, and adaptive spiking neural models, including PLIF (Fang et al., 2021b), GLIF (Yao et al., 2022), TC-LIF (Zhang et al., 2023a) and LM-H (Hao et al., 2023b).

Datasets. Moving MNIST is a video sequence dataset generated from the MNIST dataset, each frame of which contains two randomly moving digits. The objective of this dataset is to input the first 10 frames and predict the positions and shapes of the digits in the subsequent 10 frames. TaxiBJ is a dataset that contains GPS data from taxis in Beijing. The objective is to predict future urban population flow based on historical data. KTH is a dataset designed for human action recognition. The goal of this dataset is to recognize human action categories based on video sequences. The statistical summaries of these datasets are provided in Table. 1.

Table 1. The statistics of datasets. The training/testing set has N_{train}/N_{test} samples, composed of T and T' images.

	N_{train}	N_{test}	(C, H, W)	T	T'
Moving MNIST	10000	10000	(1,64,64)	10	10
TaxiBJ	19627	1334	(2,32,32)	4	4
KTH	5200	3167	(1,128,128)	10	20 or 40

Table 2. Quantitative results on the Moving MNIST dataset.

Method	MSE↓	MAE↓	SSIM↑
Vanilla LIF	102.8	246.2	0.640
PLIF (Fang et al., 2021b)	99.2	233.3	0.691
GLIF (Yao et al., 2022)	79.7	225.3	0.685
TCLIF (Zhang et al., 2023a)	105.1	257.9	0.634
LM-H (Hao et al., 2023b)	93.2	244.5	0.661
STC-LIF	**47.0**	**136.4**	**0.863**

Evaluation. We evaluated the prediction performance using commonly used metrics in spatio-temporal prediction tasks, namely mean squared error (MSE), mean absolute error (MAE), structural similarity index measure (SSIM), and peak signal-to-noise ratio (PSNR).

Detailed Setup. Our network architecture consisted of four layers, with the first three layers composed of spiking neurons and channel dimensions set to $[256, 256, 256]$. The feature map resolution remained constant throughout networks. The final layer was a convolutional layer designed to restore the predicted image dimensions. In order to address covariate shift (Wu & He, 2018) and enhance training stability, each convolutional layer, except for the last one, was followed by a group normalization layer. For more detailed information regarding the experiments, please refer to the Appendix.

5.1. Evaluation on the Moving MNIST Dataset

Moving MNIST is a benchmark dataset for spatio-temporal prediction.It poses challenges due to the random motion trajectories of the digits, which encompass complex scenarios like collisions, occlusions, and boundary rebounds. Therefore, predicting future frames necessitates modeling and inferring information such as digit shape, speed, direction, and position. Additionally, the dataset's low resolution introduces blurriness, further complicating the prediction.

Quantitative results for Moving MNIST are presented in Table 2. It is evident that the STC-LIF model outperforms other spiking neuron models in terms of MSE, MAE, and SSIM metrics. For instance, our model reduces the MSE by 55.8 compared to the vanilla LIF model and by 32.7

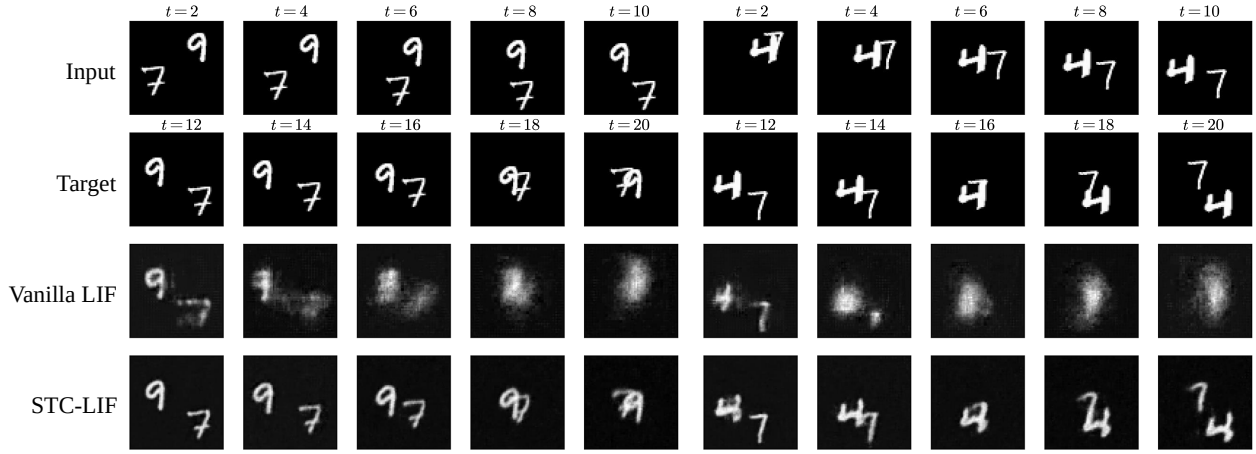


Figure 3. Qualitative visualization of the prediction results of vanilla LIF and STC-LIF models on the Moving MNIST dataset.

Table 3. Quantitative results on the TaxiBJ dataset.

Method	MSE($\times 100$)↓	MAE↓	SSIM↑
Vanilla LIF	70.4	22.1	0.966
PLIF (Fang et al., 2021b)	70.8	21.9	0.967
GLIF (Yao et al., 2022)	62.9	20.6	0.971
TCLIF (Zhang et al., 2023a)	59.2	19.3	0.973
LM-H (Hao et al., 2023b)	63.3	20.1	0.971
STC-LIF	**56.4**	**19.5**	**0.973**

Table 4. Per-frame MSE($\times 100$) on the TaxiBJ dataset.

Method	Frame1	Frame2	Frame3	Frame4
Vanilla LIF	66.7	68.1	68.5	78.5
PLIF (Fang et al., 2021b)	66.8	66.9	71.5	78.0
GLIF (Yao et al., 2022)	55.3	60.5	65.5	78.4
TCLIF (Zhang et al., 2023a)	49.0	56.8	62.2	69.2
LM-H (Hao et al., 2023b)	59.7	62.5	64.4	67.6
STC-LIF	**48.1**	**54.8**	**58.9**	**63.0**

compared to the best-performing PLIF model among other spiking neuron models. Qualitative prediction results are depicted in Figure 3. The images generated by the vanilla LIF model are notably blurry, particularly in later time steps. Conversely, the STC-LIF model produces higher-quality images, clearly depicting the trajectories and contours of each digit. Therefore, it can be concluded that the STC-LIF model significantly enhances the extraction of spatio-temporal features and the modeling of long-term dependencies. Additional visualization results can be found in the Appendix.

5.2. Evaluation on the TaxiBJ Dataset

TaxiBJ is a real and representative urban traffic flow dataset that holds significant value in urban planning and traffic management. Traffic patterns are known to be complex and highly dynamic, with strong dependencies between adjacent timestamps. In the TaxiBJ dataset, each data sequence consists of 8 consecutive frames, and the task involves predicting the remaining 4 frames given the initial 4 frames as input. We evaluated the performance of our model on this dataset, and the quantitative results are reported in Table 3. Additionally, we assessed the MSE differences ($\times 100$) for each frame, as presented in Table 4. It is evident that among all the spiking neuron models, our proposed STC-LIF model demonstrated the best predictive performance, exhibiting the smallest prediction errors.

Table 5. Quantitative results on the KTH dataset.

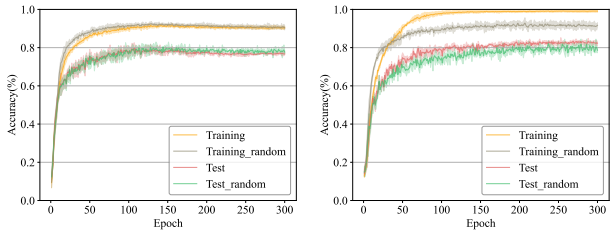
Method	KTH($10 \rightarrow 20$)		KTH($10 \rightarrow 40$)	
	SSIM↑	PSNR↑	SSIM↑	PSNR↑
Vanilla LIF	0.767	21.62	0.754	20.89
PLIF (Fang et al., 2021b)	0.796	21.79	0.785	21.17
GLIF (Yao et al., 2022)	0.805	22.41	0.795	21.90
TCLIF (Zhang et al., 2023a)	0.786	22.08	0.746	20.0
LM-H (Hao et al., 2023b)	0.710	17.13	0.708	16.91
STC-LIF	**0.822**	**23.14**	**0.815**	**22.63**

5.3. Evaluation on the KTH Dataset

Recursive models offer the advantage of flexible output length expansion. We evaluate the dynamic generalization of our model for predicting variable frame numbers on the KTH dataset. The task involved predicting the next 20 frames or 40 frames given a 10-frame input. During training, we made predictions for 20 frames, but during evaluation, we extended it to predict 40 frames, distinguishing it from the previous two datasets. We used SSIM and PSNR as evaluation metrics, where higher values indicate higher

Table 6. Comparison of quantitative results of the adaptive spiking neural models and the enhanced models.

	Moving MNIST			TaxiBJ			KTH($10 \rightarrow 20$)		KTH($10 \rightarrow 40$)	
Method	MSE↓	MAE↓	SSIM↑	MSE×100↓	MAE↓	SSIM↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑
PLIF (Fang et al., 2021b)	99.2	233.3	0.691	70.8	21.9	0.967	0.796	21.79	0.785	21.17
STC-PLIF	**48.4**	**139.4**	**0.860**	**58.6**	**20.0**	**0.972**	**0.819**	**23.02**	**0.806**	**22.32**
LM-H (Hao et al., 2023b)	93.2	244.5	0.661	63.3	20.1	0.971	0.710	17.13	0.708	16.91
STC-LM-H	**45.9**	**125.8**	**0.881**	**54.6**	**19.3**	**0.975**	**0.827**	**23.16**	**0.818**	**22.61**



(a) Vanilla LIF (b) STC-LIF

Figure 4. Comparison of vanilla LIF (a) and STC-LIF (b) when inputting sequential sequences or random sequences from the DVS128 Gesture dataset. The experiments were repeated three times, and the solid curve represents the average value.

quality predictions. The results are presented in Table 5. Notably, our model exhibited superior performance compared to other spiking neuron models in both scenarios: predicting 20 frames and extending the prediction to 40 frames.

5.4. Generalize to Other Spiking Neuron Models

The concept of the spatio-temporal self-connection circuit is highly adaptable and scalable, making it compatible with various spiking neural models (Yu et al., 2018a;b). Here we extend its application to the PLIF and LM-H models, denoting the enhanced versions as STC-PLIF and STC-LM-H, respectively. The detailed derivations can be found in the Appendix. Table 6 presents the results, clearly indicating that both STC-PLIF and STC-LM-H outperform their respective base models, PLIF and LM-H, by a significant margin. Notably, the improvement is particularly pronounced for the LM-H model, which has more complex dynamics. These results demonstrate the generality of our model and its effectiveness in enhancing spiking neurons' ability to extract spatio-temporal features and model long-term dependencies.

5.5. Rethinking Neuromorphic Data Recognition

The neuromorphic dataset contains asynchronous, event-driven, and sparse spatio-temporal information, making it naturally suitable for processing with spiking neural networks. However, can SNNs truly learn the rich spatio-temporal information in neuromorphic data? Here we consider the commonly used DVS128 Gesture (Amir et al., 2017) dataset to evaluate the performance of SNNs. The dataset provides information such as timestamps, spatial coordinates, and polarity, encompassing both spatial (gesture positions) and temporal (gesture order) aspects. During training, we use the normal gesture sequence as input, but during prediction, we deliberately disrupt the gesture order to discern the information SNNs rely on for classification. As shown in Figure 4, the vanilla LIF model achieves comparable performance on randomly shuffled input and sequential input, implying that it primarily learns spatial positional relationships. In contrast, the STC-LIF model exhibits a significant decrease in accuracy for randomly shuffled input across both training and testing sets. This indicates that the STC-LIF model is capable of simultaneously learning both temporal and spatial feature in neuromorphic data.

Table 7. Ablation study of our proposed method.

Method	MSE↓	MAE↓	Parameters(M)	Flops(G)
Vanilla LIF (Baseline)	102.8	246.2	3.305	64.346
STC-LIF w/o TC	62.5	168.1	3.614	70.398
STC-LIF w/o SC	58.3	163.8	3.614	70.398
STC-LIF	47.0	136.4	3.922	76.449

5.6. Computational Cost and Ablation Study

We perform ablative experiments on the STC-LIF model and present the results in Table 7.The findings indicate that both the temporal circuit (TC) and spatial circuit (SC) contribute to the model's performance, with the temporal circuit playing a more prominent role. In terms of computational cost, the spatio-temporal circuit exhibits a modest increase in parameter count and computational complexity per module, thanks to the utilization of grouped convolutions, with respective increments of only 9.3% and 9.4%. Considering the noticeable improvement in performance, the increase in computational cost is clearly acceptable.

6. Conclusion

This paper analyzes the limitations of the vanilla LIF model in spatio-temporal prediction tasks. Inspired by biological

autaptic synapses, we propose the STC-LIF model, which incorporates a spatio-temporal self-connection circuit to enhance the extraction of spatio-temporal features and model spatio-temporal dependencies. Our method achieves state-of-the-art performance among all spiking neuronal models, showcasing the immense potential of temporal processing capabilities in SNNs.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62176003, 62088102) and the Beijing Nova Program (20230484362).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Amir, A., Taba, B., Berg, D., Melano, T., McKinstry, J., Di Nolfo, C., Nayak, T., Andreopoulos, A., Garreau, G., Mendoza, M., et al. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7243–7252, 2017.

Bacci, A. and Huguenard, J. R. Enhancement of spike-timing precision by autaptic transmission in neocortical inhibitory interneurons. *Neuron*, 49(1):119–130, 2006.

Bacci, A., Huguenard, J. R., and Prince, D. A. Functional autaptic neurotransmission in fast-spiking interneurons: a novel form of feedback inhibition in the neocortex. *Journal of Neuroscience*, 23(3):859–866, 2003.

Bellec, G., Salaj, D., Subramoney, A., Legenstein, R., and Maass, W. Long short-term memory and learning-to-learn in networks of spiking neurons. *Advances in Neural Information Processing Systems*, 31, 2018.

Bi, G.-q. and Poo, M.-m. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of Neuroscience*, 18(24):10464–10472, 1998.

Bu, T., Ding, J., Yu, Z., and Huang, T. Optimized potential initialization for low-latency spiking neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 11–20, 2022a.

Bu, T., Fang, W., Ding, J., Dai, P., Yu, Z., and Huang, T. Optimal ANN-SNN conversion for high-accuracy and ultra-low-latency spiking neural networks. In *International Conference on Learning Representations*, pp. 1–19, 2022b.

Chen, G., Scherr, F., and Maass, W. A data-based large-scale model for primary visual cortex enables brain-like robust and versatile visual processing. *Science Advances*, 8(44):eabq7592, 2022.

Cobb, S., Halasy, K., Vida, I., Nyıri, G., Tamás, G., Buhl, E., and Somogyi, P. Synaptic effects of identified interneurons innervating both interneurons and pyramidal cells in the rat hippocampus. *Neuroscience*, 79(3):629–648, 1997.

De Wit, J. and Ghosh, A. Specification of synaptic connectivity by cell surface interactions. *Nature Reviews Neuroscience*, 17(1):4–4, 2016.

Deng, S. and Gu, S. Optimal conversion of conventional artificial neural networks to spiking neural networks. In *International Conference on Learning Representations*, 2021.

Deng, S., Li, Y., Zhang, S., and Gu, S. Temporal efficient training of spiking neural network via gradient re-weighting. *arXiv preprint arXiv:2202.11946*, 2022.

Deng, S., Lin, H., Li, Y., and Gu, S. Surrogate module learning: Reduce the gradient error accumulation in training spiking neural networks. In *International Conference on Machine Learning*, pp. 7645–7657. PMLR, 2023.

Diehl, P. U. and Cook, M. Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Frontiers in Computational Neuroscience*, 9:99, 2015.

Diehl, P. U., Neil, D., Binas, J., Cook, M., Liu, S.-C., and Pfeiffer, M. Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. ieee, 2015.

Duan, C., Ding, J., Chen, S., Yu, Z., and Huang, T. Temporal effective batch normalization in spiking neural networks. *Advances in Neural Information Processing Systems*, 35: 34377–34390, 2022.

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

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

Fang, W., Chen, Y., Ding, J., Yu, Z., Masquelier, T., Chen, D., Huang, L., Zhou, H., Li, G., and Tian, Y. Spikingjelly: An open-source machine learning infrastructure platform for spike-based intelligence. *Science Advances*, 9(40): eadi1480, 2023.

Gast, R., Solla, S. A., and Kennedy, A. Neural heterogeneity controls computations in spiking neural networks. *Proceedings of the National Academy of Sciences*, 121(3): e2311885121, 2024.

Gerstner, W. and Kistler, W. M. *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge university press, 2002.

Gerstner, W., Kistler, W. M., Naud, R., and Paninski, L. *Neuronal Dynamics: From single neurons to networks and models of cognition*. Cambridge University Press, 2014.

Golomb, D. and Rinzel, J. Dynamics of globally coupled inhibitory neurons with heterogeneity. *Physical Review E*, 48(6):4810, 1993.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Guo, S., Yu, Z., Deng, F., Hu, X., and Chen, F. Hierarchical Bayesian inference and learning in spiking neural networks. *IEEE Transactions on Cybernetics*, 49(1):133–145, 2017.

Guo, Y., Zhang, L., Chen, Y., Tong, X., Liu, X., Wang, Y., Huang, X., and Ma, Z. Real spike: Learning real-valued spikes for spiking neural networks. In *European Conference on Computer Vision*, pp. 52–68. Springer, 2022.

Han, B., Srinivasan, G., and Roy, K. Rmp-snn: Residual membrane potential neuron for enabling deeper high-accuracy and low-latency spiking neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13558–13567, 2020.

Hao, Z., Ding, J., Bu, T., Huang, T., and Yu, Z. Bridging the gap between anns and snns by calibrating offset spikes. In *International Conference on Learning Representations*, 2023a.

Hao, Z., Shi, X., Huang, Z., Bu, T., Yu, Z., and Huang, T. A progressive training framework for spiking neural networks with learnable multi-hierarchical model. In *The Twelfth International Conference on Learning Representations*, 2023b.

Harris, K. D. and Shepherd, G. M. The neocortical circuit: themes and variations. *Nature Neuroscience*, 18(2):170–181, 2015.

Hebb, D. O. *The organization of behavior: A neuropsychological theory*. Psychology Press, 2005.

Hutt, A., Rich, S., Valiante, T. A., and Lefebvre, J. Intrinsic neural diversity quenches the dynamic volatility of neural networks. *Proceedings of the National Academy of Sciences*, 120(28):e2218841120, 2023.

Kim, Y. and Panda, P. Revisiting batch normalization for training low-latency deep spiking neural networks from scratch. *Frontiers in Neuroscience*, 15:773954, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Koch, C. and Laurent, G. Complexity and the nervous system. *Science*, 284(5411):96–98, 1999.

Lee, C., Panda, P., Srinivasan, G., and Roy, K. Training deep spiking convolutional neural networks with stdp-based unsupervised pre-training followed by supervised fine-tuning. *Frontiers in Neuroscience*, 12:435, 2018.

Lee, C., Sarwar, S. S., Panda, P., Srinivasan, G., and Roy, K. Enabling spike-based backpropagation for training deep neural network architectures. *Frontiers in Neuroscience*, pp. 119, 2020.

Liu, D. and Yue, S. Event-driven continuous stdp learning with deep structure for visual pattern recognition. *IEEE Transactions on Cybernetics*, 49(4):1377–1390, 2018.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Lu, S. and Sengupta, A. Deep unsupervised learning using spike-timing-dependent plasticity. *arXiv preprint arXiv:2307.04054*, 2023.

Lübke, J., Markram, H., Frotscher, M., and Sakmann, B. Frequency and dendritic distribution of autapses established by layer 5 pyramidal neurons in the developing rat neocortex: comparison with synaptic innervation of adjacent neurons of the same class. *Journal of Neuroscience*, 16(10):3209–3218, 1996.

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

Neftci, E. O., Mostafa, H., and Zenke, F. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63, 2019.

Park, M. R., Lighthall, J. W., and Kitai, S. T. Recurrent inhibition in the rat neostriatum. *Brain Research*, 194(2): 359–369, 1980.

Pei, J., Deng, L., Song, S., Zhao, M., Zhang, Y., Wu, S., Wang, G., Zou, Z., Wu, Z., He, W., et al. Towards artificial general intelligence with hybrid tianjic chip architecture. *Nature*, 572(7767):106–111, 2019.

Preston, R., Bishop, G., and Kitai, S. Medium spiny neuron projection from the rat striatum: an intracellular horseradish peroxidase study. *Brain Research*, 183(2): 253–263, 1980.

Rathi, N. and Roy, K. Diet-snn: A low-latency spiking neural network with direct input encoding and leakage and threshold optimization. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

Rich, S., Chameh, H. M., Lefebvre, J., and Valiante, T. A. Loss of neuronal heterogeneity in epileptogenic human tissue impairs network resilience to sudden changes in synchrony. *Cell Reports*, 39(8), 2022.

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

Rueckauer, B., Lungu, I.-A., Hu, Y., Pfeiffer, M., and Liu, S.-C. Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. *Frontiers in Neuroscience*, 11:682, 2017.

Sanes, J. R. and Zipursky, S. L. Synaptic specificity, recognition molecules, and assembly of neural circuits. *Cell*, 181(3):536–556, 2020.

Schuldt, C., Laptev, I., and Caputo, B. Recognizing human actions: a local svm approach. In *International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pp. 32–36. IEEE, 2004.

Shi, W.-X. and Rayport, S. Gaba synapses formed in vitro by local axon collaterals of nucleus accumbens neurons. *Journal of Neuroscience*, 14(7):4548–4560, 1994.

Shi, X., Ding, J., Hao, Z., and Yu, Z. Towards energy efficient spiking neural networks: An unstructured pruning framework. In *International Conference on Learning Representations*, 2024.

Srivastava, N., Mansimov, E., and Salakhudinov, R. Unsupervised learning of video representations using lstms. In *International Conference on Machine Learning*, pp. 843–852. PMLR, 2015.

Stewart, K. M. and Neftci, E. O. Meta-learning spiking neural networks with surrogate gradient descent. *Neuromorphic Computing and Engineering*, 2(4):044002, 2022.

Tamas, G., Buhl, E. H., and Somogyi, P. Massive autaptic self-innervation of gabaergic neurons in cat visual cortex. *Journal of Neuroscience*, 17(16):6352–6364, 1997.

Tan, C., Li, S., Gao, Z., Guan, W., Wang, Z., Liu, Z., Wu, L., and Li, S. Z. Openstl: A comprehensive benchmark of spatio-temporal predictive learning. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Thomson, A. M., West, D. C., Hahn, J., and Deuchars, J. Single axon ipsps elicited in pyramidal cells by three classes of interneurones in slices of rat neocortex. *The Journal of Physiology*, 496(1):81–102, 1996.

Van Der Loos, H. and Glaser, E. M. Autapses in neocortex cerebri: synapses between a pyramidal cell's axon and its own dendrites. *Brain Research*, 48:355–360, 1972.

Wang, Z., Jiang, R., Lian, S., Yan, R., and Tang, H. Adaptive smoothing gradient learning for spiking neural networks. In *International Conference on Machine Learning*, pp. 35798–35816. PMLR, 2023.

Wu, Y. and He, K. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, 2018.

Wu, Y., Deng, L., Li, G., Zhu, J., and Shi, L. Spatio-temporal backpropagation for training high-performance spiking neural networks. *Frontiers in Neuroscience*, 12: 331, 2018.

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

Wu, Z., Zhang, H., Lin, Y., Li, G., Wang, M., and Tang, Y. Liaf-net: Leaky integrate and analog fire network for lightweight and efficient spatiotemporal information processing. *IEEE Transactions on Neural Networks and Learning Systems*, 33(11):6249–6262, 2021.

Xiao, M., Meng, Q., Zhang, Z., He, D., and Lin, Z. Online training through time for spiking neural networks. *Advances in Neural Information Processing Systems*, 35: 20717–20730, 2022.

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

Yao, Z., van Velthoven, C. T., Nguyen, T. N., Goldy, J., Sedeno-Cortes, A. E., Baftizadeh, F., Bertagnolli, D., Casper, T., Chiang, M., Crichton, K., et al. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell*, 184(12):3222–3241, 2021.

Yin, B., Corradi, F., and Bohté, S. M. Accurate and efficient time-domain classification with adaptive spiking recurrent neural networks. *Nature Machine Intelligence*, 3(10): 905–913, 2021.

Yin, L., Zheng, R., Ke, W., He, Q., Zhang, Y., Li, J., Wang, B., Mi, Z., Long, Y.-s., Rasch, M. J., et al. Autapses enhance bursting and coincidence detection in neocortical pyramidal cells. *Nature Communications*, 9(1):4890, 2018.

Yu, Z., Guo, S., Deng, F., Yan, Q., Huang, K., Liu, J. K., and Chen, F. Emergent inference of hidden markov models in spiking neural networks through winner-take-all. *IEEE Transactions on Cybernetics*, 50(3):1347–1354, 2018a.

Yu, Z., Tian, Y., Huang, T., and Liu, J. K. Winner-take-all as basic probabilistic inference unit of neuronal circuits. *arXiv preprint arXiv:1808.00675*, 2018b.

Zenke, F. and Vogels, T. P. The remarkable robustness of surrogate gradient learning for instilling complex function in spiking neural networks. *Neural Computation*, 33(4): 899–925, 2021.

Zhang, J., Zheng, Y., and Qi, D. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Zhang, S., Yang, Q., Ma, C., Wu, J., Li, H., and Tan, K. C. Tc-lif: A two-compartment spiking neuron model for long-term sequential modelling. *arXiv preprint arXiv:2308.13250*, 2023a.

Zhang, T., Cheng, X., Jia, S., Li, C. T., Poo, M.-m., and Xu, B. A brain-inspired algorithm that mitigates catastrophic forgetting of artificial and spiking neural networks with low computational cost. *Science Advances*, 9(34): eadi2947, 2023b.

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

Zhu, Y., Ding, J., Huang, T., Xie, X., and Yu, Z. Online stabilization of spiking neural networks. In *International Conference on Learning Representations*, 2024.

A. Experimental Details

Datasets. Moving MNIST (Srivastava et al., 2015) serves as a benchmark dataset for spatio-temporal prediction. It is a video sequence dataset generated from the MNIST dataset. It comprises 10,000 sequences, each consisting of 20 frames. Each frame is a grayscale image with dimensions 64x64 and contains two randomly moving digits. The objective of this dataset is to input the first 10 frames and predict the positions and shapes of the digits in the subsequent 10 frames. TaxiBJ (Zhang et al., 2017) is a dataset that contains GPS data from taxis in Beijing, including both inflow and outflow channels. The objective is to predict future urban population flow based on historical data. KTH (Schuldt et al., 2004) is a dataset designed for human action recognition. It includes six different actions (walking, jogging, running, boxing, waving, and clapping) performed by 25 individuals in four distinct scenarios (outdoor, outdoor with scale variation, outdoor with clothing variation, and indoor). The goal of this dataset is to recognize human action categories based on video sequences. We used individuals 1-16 for training and individuals 17-25 for prediction.

Evaluation. We evaluated the prediction performance using commonly used metrics in spatio-temporal prediction tasks, namely mean squared error (MSE), mean absolute error (MAE), structural similarity index measure (SSIM), and peak signal-to-noise ratio (PSNR). MSE and MAE are image-based metrics that measure the similarity of content in the spatial domain. They calculate the average of squared or absolute pixel differences between the ground truth and predicted images, respectively. SSIM is a feature-based metric that evaluates category/texture similarity. It mimics the theory of structural similarity in the human visual system (HVS) and demonstrates sensitivity to local structural variations in images. SSIM quantifies image attributes such as luminance, contrast, and structure by estimating the mean for luminance, variance for contrast, and covariance for structural similarity. SSIM values range from 0 to 1, with higher values indicating greater image similarity. One advantage of SSIM is its better alignment with human subjective evaluation. PSNR is a widely used measure for evaluating image quality. It represents the reference value of image quality between the maximum signal and background noise. PSNR is expressed in decibels (dB), with higher values indicating lesser image distortion.

Decoupling of spike output and self-regulation. In the STC-LIF model, the spiking signals play a dual role in information output and self-regulation. Therefore, it is necessary to decouple these two functions to strike a balance between the output and regulation of spiking neurons. During backpropagation, we truncate the gradient information of the spatio-temporal circuit (Figure 5). This ensures that the additional learnable weight parameters, $\boldsymbol{W}_{gt}^{l}$ and $\boldsymbol{W}_{gs}^{l}$, only affect the optimization of the dynamic modulation factors, $\boldsymbol{\beta}^{l}[t]$ and $\boldsymbol{\gamma}^{l}[t]$, without impacting the output of the spiking signals.

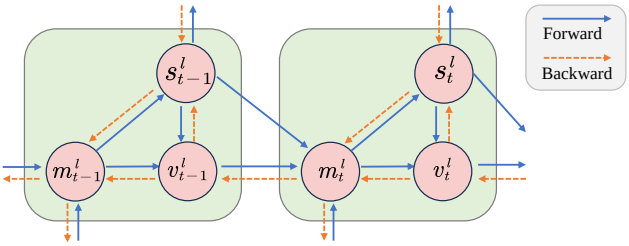


Figure 5. The unfolded computation graph of STC-LIF model.

Implementation Details. The experimental settings of different datasets are as shown in Table 8. For the Moving MNIST and KTH datasets, we set the patchsize to 2×2 and 4×4, which means that each 64×64 frame and 128×128 frame are resized into $4 \times 32 \times 32$ and $16 \times 32 \times 32$ tensors for the representation, following (Srivastava et al., 2015). The kernel size of all convolutional layers is set to 5×5, stride to 1 and padding to 2. For experiments on the DVS128 Gesture dataset, we used a 3Conv-1FC classification architecture. The Conv layer consisted of a convolutional layer, a spiking neuron layer, and a Maxpool layer, with dimensions of [64,64,64]. The FC layer contained an AvgPool and a linear fully connected layer. Each Maxpool layer and Avgpool layer performed a downsampling of the feature map by a factor of 2.The Adam (Kingma & Ba, 2014) optimizer was employed for all experiments, and we utilized a combination of warm-up (Goyal et al., 2017) and cosine learning rate scheduling (Loshchilov & Hutter, 2016). The other hyperparameters are shown in Table 9. The hyperparameters of the PLIF (Fang et al., 2021b), GLIF (Yao et al., 2022), TC-LIF (Zhang et al., 2023a) and LM-H (Hao et al., 2023b) models follow the settings in their papers or public codes. Our codes are available https:\\github.com\wangtianyi1874\stclif. The implementation is based on the OpenSTL (Tan et al., 2023) and SpikingJelly (Fang et al., 2023) frameworks, utilizing a single NVIDIA-3090 GPU.

Table 8. Experimental settings of different datasets.

Dataset	Architecture	Resize	Optimizer(weight decay)	Batchsize	Epochs	Gradient Detach
Moving MNIST	4Conv	(4,32,32)	adam(0)	16	500	N
TaxiBJ	4Conv	(2,32,32)	adam(0)	16	200	Y
KTH	4Conv	(16,32,32)	adam(0)	16	200	Y
DVS128 Gesture	3Conv-FC	(2,128,128)	sgd(0.05)	8	300	Y

Table 9. Hyperparameter settings for ours experimental implementation.

Parameters	Descriptions	Value
α^l	membrane constant	0.5
v_{th}	firing threshold	1.0
-	initial learning rate	1e-3
-	final learning rate	1e-5
-	learning rate schedule	cosine decay
-	warmup epochs	10
γ	sgd optimizer momentum	0.9
β_1, β_2	adam optimizer momentum	0.9,0.999
k	number of groups for group convolution and group normalization	16
n	number of frames for dvs128 gesture dataset sliced along time axis	20

B. Proof of Theorem

Proof of Eq. (6). According to Eqs. (1)-(4),We derived the relationship between the membrane potential $\boldsymbol{v}^l[t]$ at time t and the membrane potential $\boldsymbol{v}^l[t+T]$ after T iterations of the vanilla LIF model. The iterative derivation process is represented as follows:

$$\begin{aligned}\boldsymbol{v}^l[t+T] &= \alpha^l \boldsymbol{v}^l[t+T-1] + (1-\alpha^l)\boldsymbol{x}^l[t+T] - v_{th}\boldsymbol{s}^l[t+T] \\ &= (\alpha^l)^2 \boldsymbol{v}^l[t+T-2] + \alpha^l(1-\alpha^l)\boldsymbol{x}^l[t+T-1] - v_{th}\alpha^l \boldsymbol{s}^l[t+T-1] + (1-\alpha^l)\boldsymbol{x}^l[t+T] - v_{th}\boldsymbol{s}^l[t+T] \\ &= (\alpha^l)^T \boldsymbol{v}^l[t] + \sum_{i=1}^{T}(\alpha^l)^{T-i}(1-\alpha^l)\boldsymbol{x}^l[t+i] - v_{th}\sum_{i=1}^{T}(\alpha^l)^{T-i}\boldsymbol{s}^l[t+i] \end{aligned} \tag{17}$$

Proof of Eq. (14). Similarly,for the STC-LIF model, according to Eqs. (8)-(13), we derived the relationship between the membrane potential $\boldsymbol{v}^l[t]$ at time t and the membrane potential $\boldsymbol{v}^l[t+T]$ after T iterations. The iterative derivation process is represented as follows:

$$\begin{aligned}\boldsymbol{v}^l[t+T] &= (1+\boldsymbol{\beta}^l[t+T])\boldsymbol{v}^l[t+T-1] + (1+\boldsymbol{\gamma}^l[t+T])\boldsymbol{x}^l[t+T] - v_{th}\boldsymbol{s}^l[t+T] \\ &= (1+\boldsymbol{\beta}^l[t+T])(1+\boldsymbol{\beta}^l[t+T-1])\boldsymbol{v}^l[t+T-2] + (1+\boldsymbol{\beta}^l[t+T])(1+\boldsymbol{\gamma}^l[t+T-1])\boldsymbol{x}^l[t+T-1] \\ &- \boldsymbol{v}_{th}(1+\boldsymbol{\beta}^l[t+T])\boldsymbol{s}^l[t+T-1] + (1+\boldsymbol{\gamma}^l[t+T])\boldsymbol{x}^l[t+T] - v_{th}\boldsymbol{s}^l[t+T] \\ &= \boldsymbol{v}^l[t]\prod_{j=1}^{T}(1+\boldsymbol{\beta}^l[t+j]) + \sum_{i=1}^{T}(1+\boldsymbol{\gamma}^l[t+i])\boldsymbol{x}^l[t+i]\prod_{j=1}^{T-i}(1+\boldsymbol{\beta}^l[t+i+j]) \\ &- \boldsymbol{v}_{th}\sum_{i=1}^{T}\boldsymbol{s}^l[t+i]\prod_{j=1}^{T-i}(1+\boldsymbol{\beta}^l[t+i+j]) \end{aligned} \tag{18}$$

C. Enhanced Versions of PLIF and LM-H Models

The STC-PLIF Model. We extended the spatio-temporal circuit to the PLIF model and the LM-H model. The enhanced version for the PLIF model, STC-PLF model, is represented as follows:

$$\boldsymbol{x}^l[t] = \boldsymbol{W}^l \boldsymbol{s}^{l-1}[t] \tag{19}$$

$$\boldsymbol{\beta}^l[t] = \mathrm{Tanh}(\boldsymbol{W}^l_{gt} \boldsymbol{s}^l[t-1]) \tag{20}$$

$$\boldsymbol{\gamma}^l[t] = \mathrm{Tanh}(\boldsymbol{W}^l_{gs} \boldsymbol{s}^l[t-1]) \tag{21}$$

$$\boldsymbol{m}^l[t] = \alpha^l \boldsymbol{v}^l[t-1] \odot (1 + \boldsymbol{\beta}^l[t]) + (1-\alpha^l)\boldsymbol{x}^l[t] \odot (1 + \boldsymbol{\gamma}^l[t]) \tag{22}$$

$$\boldsymbol{s}^l[t] = H(\boldsymbol{m}^l[t] - v_{th}) \tag{23}$$

$$\boldsymbol{v}^l[t] = \boldsymbol{m}^l[t] - v_{th}\boldsymbol{s}^l[t] \tag{24}$$

Here the black part is the original formula of the PLIF model, and the red part is the improvement for the PLIF model. It is worth noting that the membrane constant α^l is different from the vanilla LIF model in that it is a learnable parameter, and its range is [0,1]. The PLIF model enhances the adaptability of spiking neurons by learnable membrane constant. Our improvement for the PLIF model is similar to the vanilla LIF model, which adds the regulatory factors $\boldsymbol{\beta}^l[t]$ and $\boldsymbol{\gamma}^l[t]$ on the basis of the original dynamics of the PLIF model.

The STC-LM-H Model. The enhanced version for the LM-H model, STC-LM-H model, is represented as follows:

$$\boldsymbol{x}^l[t] = \boldsymbol{W}^l \boldsymbol{s}^{l-1}[t] \tag{25}$$

$$\boldsymbol{\beta}^l[t] = \mathrm{Tanh}(\boldsymbol{W}^l_{gt} \boldsymbol{s}^l[t-1]) \tag{26}$$

$$\boldsymbol{\gamma}^l[t] = \mathrm{Tanh}(\boldsymbol{W}^l_{gs} \boldsymbol{s}^l[t-1]) \tag{27}$$

$$\boldsymbol{v}^l_D[t] = \mu^l_D \boldsymbol{v}^l_D[t-1] + \mu^l_S \boldsymbol{v}^l_S[t-1] + \boldsymbol{x}^l[t] \odot (1 + \boldsymbol{\gamma}^l[t]) \tag{28}$$

$$\boldsymbol{m}^l[t] = \lambda^l_S \boldsymbol{v}^l_S[t-1] \odot (1 + \boldsymbol{\beta}^l[t]) + \lambda^l_D \boldsymbol{v}^l_D[t] \tag{29}$$

$$\boldsymbol{s}^l[t] = H(\boldsymbol{m}^l[t] - v_{th}) \tag{30}$$

$$\boldsymbol{v}^l_S[t] = \boldsymbol{m}^l[t] - v_{th}\boldsymbol{s}^l[t] \tag{31}$$

Similarly, the black part is the original formula of the LM-H model, and the red part is the improvement for the LM-H model. The LM-H model has more complex neural dynamics than the vanilla LIF model and the PLIF model. It is a two-compartment model, which contains two membrane potentials $\boldsymbol{v}^l_D[t]$ and $\boldsymbol{v}^l_S[t]$. Moreover, it contains four learnable parameters μ^l_D,μ^l_S,λ^l_D and λ^l_S. The LM-H model significantly enhances the representation ability and adaptability by regulating the dual membrane potentials and multiple learnable parameters. Following the similar idea of STC-LIF and STC-PLIF, $\boldsymbol{\gamma}^l[t]$ regulates the input $\boldsymbol{x}^l[t]$, while $\boldsymbol{\beta}^l[t]$ regulates the key membrane potential $\boldsymbol{v}^l_S[t]$. In addition, in the soft reset process of the LM-H model, the gradient backpropagation of $\boldsymbol{s}^l[t]$ is truncated. We find that this increases the gradient calculation error, so we choose not to truncate the gradient information of $\boldsymbol{s}^l[t]$.

Analysis of Enhanced Models. For the enhanced models, it is obvious that they have more complex dynamic characteristics than the original models. When $\boldsymbol{\beta}^l[t]$ and $\boldsymbol{\gamma}^l[t]$ equals 0, the enhanced models are equivalent to the original models, Therefore, the spatio-temporal circuit is an improvement in the basic representation ability of the spiking neuron models. The comparison of the qualitative visualization results of the enhanced models and the original models on Moving MNIST is shown in Figure 6, which shows that the spatio-temporal circuit significantly improves the models' performance in capturing deep spatio-temporal information features and modeling spatio-temporal dependencie.

D. Additional Experiments

Experiments on static classification tasks. We have conducted experiments on the CIFAR100 dataset using the VGGSNN (64C3-128C3-MP2-256C3-256C3-MP2-512C3-512C3-MP2-512C3-512C3-MP2-FC) architecture. The results are presented in Table 10. It is evident that the STC-LIF model is also beneficial in improving the performance of classification tasks. The accuracy of the STC-LIF model surpasses that of the vanilla LIF model at the same time step. Notably, this advantage becomes more pronounced as the time step increases. Consequently, we believe that spiking neural networks, with their remarkable spatial and temporal processing capabilities, possess substantial potential for advancements in the domain of complex dynamic task processing.

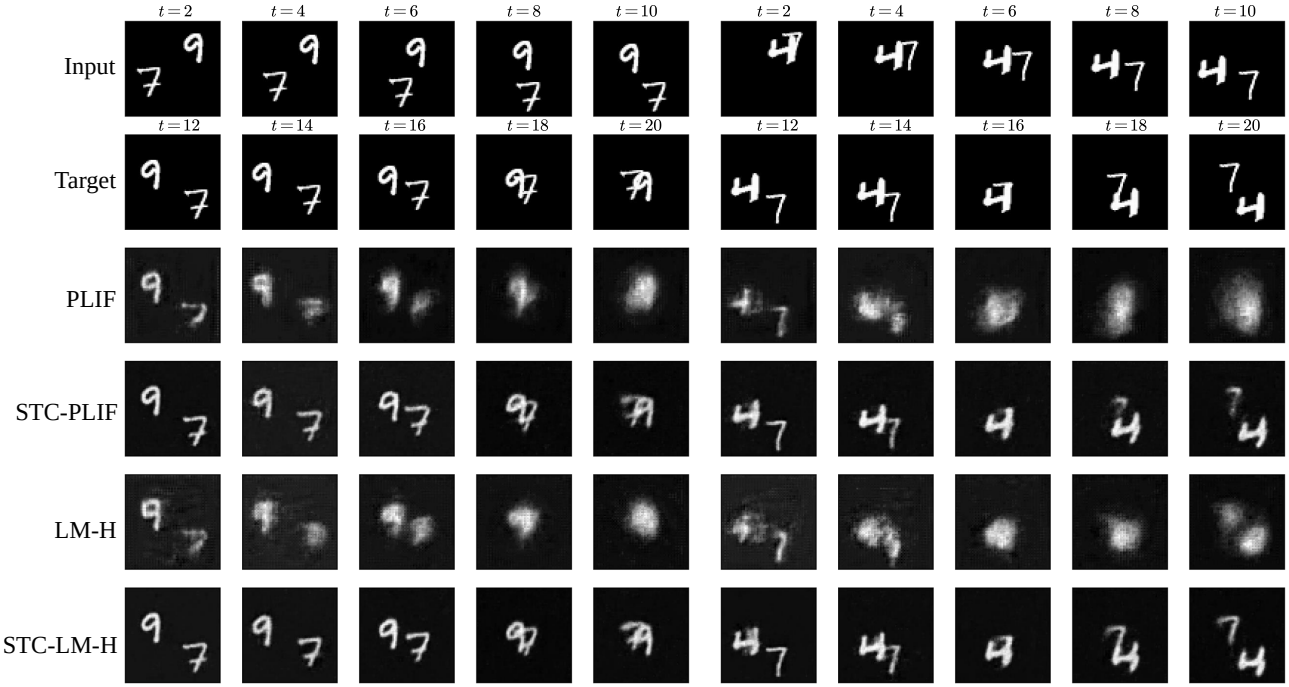


Figure 6. Comparison of the qualitative visualization prediction results of original models and enhanced models on the Moving MNIST dataset.

Table 10. Comparsion of different methods on the CIFAR100 dataset.

Method	Time-steps	Accuracy(%)
Vanilla LIF	4	63.40
STC-LIF	4	66.64
Vanilla LIF	6	63.99
STC-LIF	6	68.93

Experiments with different self-connection methods. We have conducted more comprehensive experiments on the Moving MNIST dataset, and the corresponding results are presented in Table 11. It has been observed that employing self-connections alone can effectively enhance the dynamic representation of spiking neurons, resulting in a significant improvement compared to the vanilla LIF model. However, this approach has certain drawbacks. Firstly, it lacks flexibility and limits the representational capacity of spiking neurons, thereby leading to inferior performance compared to the STC-LIF model with group convolution or global convolution. Secondly, it entails a higher memory cost due to the assignment of weight parameters to each neuron. While global convolution yields the best performance, it comes at the expense of substantial computational requirements. Therefore, considering the overall trade-off, utilizing group convolution emerges as a favorable choice. It significantly enhances the model's representational capacity without imposing a significant increase in computational cost.

Table 11. Comparison of models with self-connection, group convolution and global convolution on the Moving MNIST dataset.

Model	Operation	MSE↓	MAE↓	Parameters(M)
Vanilla LIF	-	102.8	246.2	3.305
STC-LIF	Only self-connection	64.5	177.6	6.451
STC-LIF	Group convolution(groups=16)	47.0	136.4	3.922
STC-LIF	Global convolution	38.5	108.1	13.138