EXPLOITING TEMPORAL HIGH FREQUENCIES IN SPIKING TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing Spiking Transformers flatten the temporal dimension into the batch axis and average over time, which is equivalent to low-pass filtering each pixel's temporal signal. We find that the temporal high-frequency components discarded by this paradigm carry discriminative information: although they account for only 18.1% of the signal energy, masking them at the input layer reduces accuracy by up to 6.2%. We further reveal a compensatory amplification mechanism: as spiking dynamics attenuate these components with depth, learnable frequency gates partially offset this decay, raising their effective utilization from 17.9% to 45.7% on CIFAR10-DVS. We propose a Spiking Spatial-Temporal Wavelet Transformer (SSTWformer), which embeds spatial-temporal Haar transforms into Spiking Transformers to explicitly decompose features into eight spatial-temporal subbands for structured learning across multiple frequency bands.Experiments on six datasets spanning object recognition (CIFAR10-DVS, N-Caltech101, NCARS), gesture recognition (DVS128),and speech recognition (SHD, SSC) demonstrate that SSTWformer consistently outperforms spatial-only baselines and the attention-based Spikformer, achieving a 1.81% average accuracy gain and state-of-the-art (SOTA) performance on multiple tasks.The code is available in the supplementary material.

1 INTRODUCTION

Spiking Neural Networks (SNNs)Maass (1997); Guo et al. (2024); Kim et al. (2020); Ding et al. (2024); Cao et al. (2024), recognized as the third generation of neural networks, fundamentally depart from Artificial Neural Networks (ANNs)Agatonovic-Kustrin & Beresford (2000), which rely on continuous real-valued activations for information transmission. In contrast, SNNs encode and transmit information through sparse, asynchronous discrete spikes. This biologically plausible architecture not only better mimics neural system operation but also inherently provides superior temporal dynamics and ultra-low computational cost, conferring distinct advantages in event-based vision tasks.In recent years, Spiking TransformersZhou et al. (2023b;a); Yao et al. (2023); Lee et al. (2025) have made remarkable progress in this domain. However, when examined from a frequency-analysis perspective, existing Spiking Transformers suffer from structural modeling deficiencies in the spatiotemporal frequency domain when processing neuromorphic data.

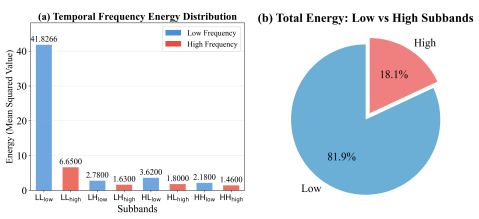


Figure 1: The average time-frequency energy distribution of multiple samples in the event stream dataset.

At the spatial frequency level, self-attention inherently acts as a low-pass filter, retaining low-frequency spatial information while responding inadequately to high-frequency components—a property fundamentally at odds with event-camera data, whose brightness changes correspond precisely to high-frequency motion edges. At the temporal frequency level, existing methods suffer from a more fundamental structural deficiency. Mainstream Spiking TransformersZhou et al. (2024a;b); Zheng et al. (2025); Shi et al. (2024), exemplified by SpikformerZhou et al. (2023b), flatten the temporal dimension into the batch dimension, applying spatial attention independently at each timestep followed by simple averaging. From a temporal frequency perspective, this is equivalent to low-pass

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

filtering each pixel's temporal signal: high-frequency fluctuations across time steps cancel out during averaging, structurally filtering out dynamic evolution information. The model thus captures little more than "the averaged features of a sequence of static frames" rather than "a dynamic process evolving over time"—a loss that is structural and cannot be remedied by simply increasing network depth or width.

This raises a fundamental question: what is lost when the spatiotemporal frequency dimension, especially temporal high frequencies, is neglected? The event stream is a sparse temporal sampling of a continuous process, where static backgrounds yield low-frequency triggers and fast-moving edges excite high-frequency responses. Figure 1 shows that although energy concentrates in the spatiotemporal low-frequency band, the temporal high-frequency subband still retains 18.1% of the total energy. Masking these frequencies degrades SSTWformer's accuracy across datasets, and replacing them with energy-matched noise is far more harmful than removing them entirely—indicating that the network actively reads their structured content rather than treating them as residual noise. In contrast, SWFormeFang et al. (2025), which models only the spatial frequency domain, suffers a much smaller drop, revealing that existing architectures exploit this information only implicitly and incompletely. Once high frequencies are removed, the accuracy gap between the two models narrows, showing that the advantage of explicit spatiotemporal decomposition is largely mediated by temporal high-frequency information that existing architectures structurally waste. We further find a compensatory amplification mechanism: spiking dynamics attenuate temporal high frequencies with depth, while learned frequency gates re-amplify them in deeper blocks, improving their effective utilization. Together, these results suggest that temporal high frequencies are a scarce, structured, and causally important resource that existing Spiking Transformers underutilize; explicit decoupling and independent addressing of frequency components offer an effective means of exploiting it.

Inspired by these findings, we propose SSTWformer (Spiking Spatial-Temporal Wavelet Transformer), a neuromorphic vision model that integrates the spatial-temporal Haar wavelet transform into the Spiking Transformer architecture. Its core is an explicit spatiotemporal frequency decomposition of the input event tensor: by coupling the 2D spatial Haar transform with the 1D temporal Haar transform, the tensor is decomposed into eight physically meaningful subbands (4 spatial orientations $\times$ 2 temporal frequencies). This converts the "spatiotemporal mixed representation" of the event stream into a "structured frequency-domain representation," allowing the network to directly perform differentiated learning across frequency components—systematically capturing spatiotemporal evolution patterns from static backgrounds to rapid transient motions—instead of implicitly fitting complex dependencies in the raw event stream. The main contributions of this paper are as follows:

- We provide the first causal evidence that temporal high-frequency components of event streams carry structured discriminative information—a scarce yet critical resource that mainstream architectures, due to their temporal low-pass paradigm, exploit only implicitly and incompletely.
- We reveal a compensatory amplification mechanism: as spiking dynamics intrinsically attenuate temporal high frequencies with depth, learned frequency gates partially offset this attenuation and defend their effective utilization back toward parity, offering a mechanistic basis for frequency-aware architecture design.
- We propose SSTWformer, the first spatiotemporal wavelet transform model for Spiking Transformers, enabling explicit modeling of spatiotemporal spectral structure through independent decoupling of spatial and temporal frequencies.
- SSTWformer consistently outperforms both the spatial-domain-only decomposition baseline and the attention-based Spikformer across six datasets spanning visual and audio modalities, achieving SOTA or competitive performance on the majority of benchmarks, with an average performance gain of 1.81%.

2 Related Works

2.1 Spiking Vision Transformers

In recent years, Transformer architectures have achieved remarkable success across various domains of deep learning Chen et al. (2021); Yuan et al. (2021); Carion et al. (2020); Ji et al. (2024);

Chen et al. (2024); Ma et al. (2024); Yao et al. (2024). Their exceptional capability in modeling long-range dependencies via self-attention has inspired researchers to introduce this paradigm into SNNs. Spikformer Zhou et al. (2023b) stands as one of the most representative efforts in this direction, which designs spike-form Query, Key, and Value to enable self-attention computation while preserving spike sparsity, significantly improving SNN performance on classification tasks. Subsequently, Spikingformer Zhou et al. (2023a) introduces a pre-activation shortcut design that effectively reduces neuronal firing rates and avoids floating-point multiplication operations, while QKformer Zhou et al. (2024a) and Spiking ResFormer Shi et al. (2024) further explore the integration of hierarchical structures with spike-based self-attention. In addition, TIM Shen et al. (2024) is introduced to boost temporal processing, while TCJA Zhu et al. (2024) employs a temporal-channel joint attention mechanism for spatiotemporal feature evaluation. STAtten Lee et al. (2025) processes information in spatiotemporal blocks, achieving comprehensive feature capture at the same cost as purely spatial methods. Despite accuracy gains, these models remain spatially oriented and offer limited modeling of spatiotemporal frequency dynamics in event streams.

2.2 Frequency-Domain Representations in SNNs

Integrating spatiotemporal frequency analysis into SNNs is of significant importance. Neuromorphic data essentially represents continuous variations in the logarithm of light intensity, whose transient mutations manifest both as high-frequency edges and texture details in the spatial domain and as high-frequency dynamic changes in the temporal domain. However, the inherent low-pass filtering characteristics of spiking neurons progressively attenuate high-frequency information during layer-wise propagation Fang et al. (2026; 2025). This dilemma of "high-frequency data dominance versus low-frequency model bias" renders explicit spatiotemporal frequency modeling in SNNs a critical endeavor.

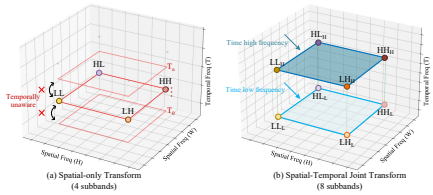


Figure 2: Schematic diagram of spatial wavelet transform and spatial-temporal wavelet transform.

In existing explorations, López-Randulfe et al. (2021; 2022) attempted to implement the Fourier transform via time-to-value mapping, yet the computational overhead of approximately 1024 time steps renders this approach impractical. SWFormer Fang et al. (2025) leverages the inherent compatibility between the wavelet transform and the binary sparsity of SNNs, introducing wavelet decomposition into SNNs for spatial frequency feature extraction. However, this method suffers from several limitations. On the one hand, it treats the Haar transform as a fixed projection operator that implicitly encodes frequency-domain information into feature maps through matrix multiplication, so that the spatial subbands are implicitly mixed within the feature maps after transformation, preventing explicit differentiation and processing of different spatial frequency components. On the other hand, this method primarily focuses on wavelet decomposition in the spatial domain, lacks explicit modeling of the temporal dynamics of spike sequences, and struggles to effectively characterize feature evolution patterns across different time scales. To address the aforementioned gaps in spatiotemporal modeling of existing frequency analysis methods, this paper proposes a spatiotemporal wavelet decomposition strategy (as shown in Figure 2), which first decomposes event streams into multiple spatiotemporal frequency bands via spatial and temporal wavelet transforms, and then performs independent modeling on each frequency band, thereby achieving systematic capture of spatiotemporal frequency information in event streams.

3 Methods

3.1 Spatial-Temporal Wavelet Transform

3.1.1 Transform Strategy

Let $x \in \mathbb{R}^{T\times H\times W}$ denote the input spatiotemporal spike tensor, with T, H, and W representing the timestep, height, and width, respectively. Given the separability of the Haar transform across

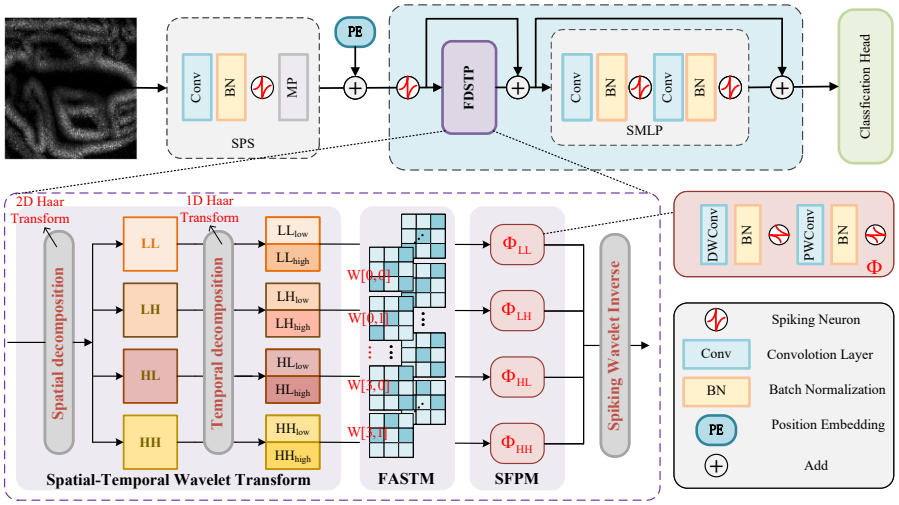


Figure 3: Overall architecture of SSTWformer. A spatiotemporal wavelet transform decomposes the input into 4 spatial-orientation $\times$ 2 temporal-scale subbands to capture the spatiotemporal information in event streams.

the temporal and spatial dimensions, the spatiotemporal decomposition can be implemented in a cascaded fashion as follows:

$$\mathcal{H}(x) = \mathcal{H}_t(\mathcal{H}_s(x)) \tag{1}$$

where $\mathcal{H}_s()$ and $\mathcal{H}_t()$ denote the spatial 2D Haar transform and the temporal 1D Haar transform, respectively.

Step 1: Spatial decomposition. For each temporal frame, we independently apply the 2D spike-based Haar transform (Eq.19-22 in the Appendix) to obtain four spatial subbands:

$$\mathrm{S}^t \xrightarrow{\mathcal{H}_s} \{\mathrm{LL}^t, \mathrm{LH}^t, \mathrm{HL}^t, \mathrm{HH}^t\}, t = 0, 1, ..., T-1 \tag{2}$$

At this stage, each subband retains the full temporal resolution T , while the spatial dimensions are halved to $\frac{H}{2} \times \frac{W}{2}$.

Step 2: Temporal decomposition. For each spatial subband $S \in \{\mathrm{LL}, \mathrm{LH}, \mathrm{HL}, \mathrm{HH}\}$, we apply the 1D spike-based Haar transform (Eq. 16-17 in the Appendix) along the temporal dimension at each spatial location.

$$S_s[k, i, j]_{k=0}^{T-1} \xrightarrow{\mathcal{H}_t} (\mathrm{Low}_s[i, j], \mathrm{High}_s[i, j]) \tag{3}$$

The low-frequency and high-frequency components each have a temporal length of $\frac{T}{2}$. After the above two steps, the transform yields a total of 8 subbands, each with dimensions $\frac{T}{2} \times \frac{H}{2} \times \frac{W}{2}$.

3.1.2 Ordered Subband Organization and Indexing Scheme

To guarantee the invertibility of the transform and the consistency of frequency-domain operations, we introduce a standardized subband indexing mechanism that establishes a fixed correspondence between the forward and inverse transforms via a predefined subband order. This eliminates the adverse effects of subband permutation ambiguity on network learning. The eight subbands are arranged in a fixed order according to the design principles of spatial dominance, low-frequency priority, and temporal alignment as follows:

$$\Omega = [\mathrm{LL}_l, \mathrm{LL}_h, \mathrm{LH}_l, \mathrm{LH}_h, \mathrm{HL}_l, \mathrm{HL}_h, \mathrm{HH}_l, \mathrm{HH}_h] \tag{4}$$

The corresponding index mapping $M : \Omega \to \{0, 1, \ldots, 7\}$ is defined as:

$$M(s,l) = 2 \times \text{idx}(s) + \text{idx}(l) \tag{5}$$

where $\text{idx}(s) \in \{0,1,2,3\}$ denotes the frequency order of the spatial subbands , and $\text{idx}(l) \in \{0,1\}$ denotes the frequency order of the temporal components.

3.1.3 Reconstruction Property

We employ the spatiotemporal Haar wavelet transform, from whose eight subbands the original spatiotemporal tensor x can be perfectly reconstructed. The inverse transform is given by:

$$x = \mathcal{H}^{-1}(\mathcal{H}(x)) \tag{6}$$

3.2 Spiking Spatial-Temporal Wavelet Transformer

Based on the spatiotemporal Haar wavelet transform described above, we design an end-to-end spike-driven feature extraction module, termed Frequency-Decoupled Spatial-Temporal Processor (FDSTP), which can be seamlessly integrated into arbitrary SNN frameworks in a plug-and-play manner.

3.2.1 Overall Architecture

Figure 3 illustrates the overall architecture of SSTWformer, which comprises a Spike Patch Splitting (SPS) module, Position Encoding (PE), FDSTP , and a linear classification head(CH). The SPS and PE follow SpikformerZhou et al. (2023b).For an input 2D image sequence $x \in \mathbb{R}^{T\times H\times W}$, the SPS module and PE generate a feature map, which is then passed through the FDSTP and SMLPZhou et al. (2023b) modules with residual connections. The FDSTP module employs a spike frequency representation mechanism for spatial-temporal feature extraction. Finally, Global Average Pooling (GAP) compresses the output into a D-dimensional vector, which is fed into the CH to produce the prediction Y. The overall architecture of SSTWformer is:

$$X_0 = \mathcal{SN}(\text{SPS}(x) + \text{PE}), X_0 \in \mathbb{R}^{T\times H\times D} \tag{7}$$

$$X_l^{'} = \text{FDSTP}(X_{l-1}) + X_{l-1}, l = 0,\ldots,M-1 \tag{8}$$

$$X_l = \text{SMLP}(X_l^{'}) + X_l^{'}, l = 0,\ldots,M-1 \tag{9}$$

$$Y = \text{CH}(\text{GAP}(X_l)) \tag{10}$$

where $\mathcal{SN}$ denotes the Leaky-Integrate-and-Fire (LIF) modelZhou et al. (2023b),X_l and $X_l^{'}$ denote the membrane potential and spike output of FDSTP at the l-th layer.M refers to the total number of layers.

Frequency-Aware Spatial-Temporal Modulation (FASTM). The spatial-temporal Haar wavelet transform decomposes the input spike tensor into eight spatial-temporal subbands, with each subband contributing distinctively to the downstream task. The low-frequency subbands carry the global structural information of the input, while the high-frequency subbands contain local texture details along with potential noise components, whose effectiveness often varies dynamically with data distribution and is difficult to predefine. To address this, we introduce a learnable FASTM to adaptively recalibrate the importance of each subband. Specifically, we define the gating weights as a three-dimensional tensor $\mathbf{W} \in \mathbb{R}^{4\times 2\times C}$, where the first dimension $p \in \{\text{LL},\text{LH},\text{HL},\text{HH}\}$ indexes the spatial-frequency type, the second dimension $q \in \{\text{low},\text{high}\}$ indexes the temporal frequency, and C denotes the embedding dimension. For each input subband $\tilde{S}_{p,q}$, the process of the FASTM module is formulated as:

$$\hat{S}_{p,q} = \tilde{S}_{p,q} \odot \mathbf{W}[p,q,:], \mathbf{W} \in \mathbb{R}^{4\times 2\times C} \tag{11}$$

where $\odot$ denotes the element-wise product, and $\mathbf{W}$ is broadcast across the temporal and spatial dimensions to align with the full spatial-temporal resolution of each subband tensor.

Datasets	Methods	Architecture	Param (M)	T	Acc(%)
CIFAR10-DVS	TIM Shen et al. (2024)	Spikformer-2-256	2.60	10	81.6
	STAtten+STLee et al. (2025)	ST-2-256	-	16	81.1
	SpikingformerZhou et al. (2026)	Spikingformer-2-256	2.57	16	81.4
	SpikformerZhou et al. (2023b)	Spikformer-2-256	2.57	16	80.6†
	SWFormerFang et al. (2025)	SWFormer-2-256	2.10	16	82.2†
	SSTWformer	SSTWformer-2-256	2.59	10	**82.2**
	SSTWformer	SSTWformer-2-256	2.59	16	**84.9(+2.7)**
N-Caltech101	TIM Shen et al. (2024)	Spikformer-2-256	2.60	10	79
	TCJAZhu et al. (2024)	VGGSNN	-	10	82.5
	STAtten+STLee et al. (2025)	ST-2-256	-	16	83.15
	SpikformerZhou et al. (2023b)	Spikformer-2-256	2.57	16	84.2†
	SWFormerFang et al. (2025)	SWFormer-2-256	2.10	16	84.93†
	SSTWformer	SSTWformer-2-256	2.59	10	**85.78**
	SSTWformer	SSTWformer-2-256	2.59	16	**87.48 (+2.55)**
NCARS	VMV-GCNXie et al. (2022)	VMV-GCN	-	-	93.2
	TIMShen et al. (2024)	Spikformer-2-256	2.60	16	94.41†
	SpikformerZhou et al. (2023b)	Spikformer-2-256	2.57	16	94.78†
	SWFormerFang et al. (2025)	SWFormer-2-256	2.10	16	93.17†
	SSTWformer	SSTWformer-2-256	2.59	10	**95.00**
	SSTWformer	SSTWformer-2-256	2.59	16	**95.74 (+2.57)**
DVS128	TCJAZhu et al. (2024)	VGGSNN	-	20	98.2
	ESTSformerLu et al. (2025)	ESTSformer-2-256	-	16	98.26
	SpikingformerZhou et al. (2026)	Spikingformer-2-256	2.57	16	98.6
	SpikformerZhou et al. (2023b)	Spikformer-2-256	2.57	16	98.13†
	SWFormerFang et al. (2025)	SWFormer-2-256	2.10	16	97.57†
	SSTWformer	SSTWformer-2-256	2.59	10	97.2
	SSTWformer	SSTWformer-2-256	2.59	16	**98.61(+1.04)**

Table 1: Experimental Results on Various Neuromorphic Datasets."Acc" = Accuracy (%), "T" = Time Step, "model-L-D" denotes a model with L encoder blocks and D channels. We label the reproduced implementation data as "†". Results are averaged over three seeds (std $< 0.28\%$).

Spiking Frequency Processing Module(SFPM). Although the spatial-temporal wavelet transform yields eight subbands, the same spatial-frequency type shares similar structural priors across temporal-low and temporal-high branches. To avoid parameter redundancy, we adopt parameter sharing with only four spatial feature extractors, one per spatial-frequency type, covering subbands from both temporal branches.The architecture of the feature extractor adopts a cascaded dual-convolutional-block structure.Given the input features, each convolutional module can be formulated as:

$$\Phi = \mathcal{SN}\left(\mathrm{BN}\left(\mathrm{PW}\left(\mathcal{SN}\left(\mathrm{BN}\left(\mathrm{DW}(X)\right)\right)\right)\right)\right) \tag{12}$$

where DW() denotes a depthwise convolution with a kernel size of 3, PW() denotes a pointwise convolution, and BN() denotes Batch Normalization. The four feature extractors Φ_{LL}, Φ_{LH},Φ_{HL},Φ_{HH} share the same structure but have independent parameters to fit different subband characteristics.

4 Experiments and Results

To evaluate the effectiveness of our model, we conduct extensive comparative experiments on multiple neuromorphic vision datasets (CIFAR10-DVSLi et al. (2017), N-Caltech101Orchard et al. (2015), NCARSSironi et al. (2018), and DVS128Amir et al. (2017)) that present distinct challenge levels, along with two audio datasets (SHDCramer et al. (2020) and SSCCramer et al. (2020)). All experiments are conducted on an A800 GPU under a consistent configuration: the timestep is set to 16, the AdamW optimizer is employed with an initial learning rate of 0.005(0.001 for DVS128), and a cosine annealing schedule is utilized for learning rate adjustment. The number of training epochs is 106 for CIFAR10-DVS, N-Caltech101, and NCARS; 200 for DVS128; and 106 for both SHD and SSC.All experiments are conducted with three random seeds.

Datasets	Methods	T	Acc(%)
SHD	TIMShen et al. (2024)	10	86.3
	SpikformerZhou et al. (2023b)	16	88.96†
	SWFormerFang et al. (2025)	16	90.06†
	SSTWformer	10	**87.59**
	SSTWformer	16	**91.17(+1.11)**
SSC	SpikformerZhou et al. (2023b)	16	89.53†
	SWFormerFang et al. (2025)	16	89.20†
	SSTWformer	10	**86.79**
	SSTWformer	16	**90.07 (+0.87)**

Table 2: Experimental Results on Spiking Audio Datasets.

Datasets	Model	T	Acc(%)
CIFAR10-DVS	NoS&F	16	83.1
	NoS	16	83.5
	NoF	16	83.8
	NoTime	16	83.6
	base	16	**84.9**
N-Caltech101	NoS&F	16	85.57
	NoS	16	86.05
	NoF	16	86.28
	NoTime	16	86.35
	base	16	**87.48**

Table 3: Ablation Study of SSTWformer Components.

4.1 Results on Neuromorphic Vision Datasets

The experimental results are presented in Table 1, which validates the effectiveness of the proposed spatial-temporal Haar wavelet transform across multiple datasets. Compared to the baselineFang et al. (2025), our method achieves performance improvements of 2.7%, 2.55%, 2.57%, and 1.04% on CIFAR10-DVS, N-Caltech101, NCARS, and DVS128, respectively. These results demonstrate that the spatial-temporal Haar transform, by explicitly decoupling the spatial structure and temporal dynamics in event streams, effectively enhances the model's representation and generalization capabilities on diverse event data, confirming its broad applicability in event-driven vision tasks.Detailed training settings are provided in Appendix C.

4.2 Results on Neuromorphic Audio Datasets

To thoroughly assess the generalization capability of our method, we additionally conduct experiments on the spiking audio benchmarks SHD and SSC, with quantitative results summarized in Table 2. As the raw inputs are one-dimensional temporal spike sequences while our framework is inherently designed for two-dimensional spatial structures, we reshape the feature vector at each timestep into a 2×350 pseudo-spatial grid, enabling the reuse of the identical architecture for fair cross-modal comparisonShen et al. (2024). These artificially constructed spatial dimensions carry no physical meaning and are merely employed as a structured prior for feature organization. Under this configuration, the performance gain can be primarily attributed to the modeling capability of the spatiotemporal Haar transform on the internal structure of spike sequences. SSTWformer attains 91.17% accuracy on SHD, surpassing Spikformer (88.96%)Zhou et al. (2023b) and SWFormer (90.06%)Fang et al. (2025); on SSC, it exceeds both counterparts by 0.54% and 0.87%, respectively. These findings suggest that the advantage of our approach is not attributable to overfitting to visual domains, but rather to its fundamental capability in modeling generic spike sequences. Coupled with the superior results achieved on four visual datasets, the spatial-temporal Haar transform demonstrates consistent and compelling performance across both visual and non-visual spike-based processing tasks, demonstrating its potential as a general-purpose spike feature extraction backbone.

Config	Datasets(%)	
	CIFAR10-DVS	N-Caltech101
S → F	83.2	86.39
F → S	**84.9**	**87.48**

Table 4: Effect of the cascading order of FASTM(F) and SFPM(S).

TS	N-Caltech101(%)		CIFAR10-DVS(%)	
	SWFormer	SSTWformer	SWFormer	SSTWformer
T=6	84.81	85.54	78.8	79.7
T=10	85.17	85.78	80.7	82.2
T=16	84.93	**87.48**	82.2	**84.9**
T=20	85.43	86.87	81.9	84.9

Table 5: Performance of models under different Time Steps.

4.3 Ablation Experiments

Table 3 presents an ablation study that systematically quantifies the contribution of each module to SSTWformer performance on the CIFAR10-DVS and N-Caltech101 datasets. The full model (base)

attains accuracies of 84.9% and 87.48% on the two datasets, respectively. The removal of both SFPM and FASTM (NoS&F) incurs the most substantial degradation, with performance dropping by 1.8% and 1.91% , underscoring the critical importance of their cooperative interaction. Removing SFPM (NoS), FASTM (NoF), and temporal wavelet decomposition (NoTime) individually yields declines of 1.4%, 1.1%, and 1.3% on CIFAR10-DVS, and 1.43%, 1.20%, and 1.13% on N-Caltech101, respectively. It is worth noting that while the variant without temporal wavelet decomposition (NoTime) already attains a respectable accuracy, the incorporation of temporal-frequency decomposition to provide temporal priors further enhances performance at no additional parameter cost, strongly highlighting the essential role of temporal information in spike-based modeling.

In the architectural design for spatial-temporal feature extraction from event streams, we compare the impact of different cascade orders of the FASTM and the SFPM on recognition performance, with results reported in Table 4. Experimental results demonstrate that the "FASTM-first, SFPM-second" configuration (denoted as FASTM $\rightarrow$ SFPM) achieves consistent performance advantages across all tested datasets, attaining accuracies of 84.9% and 87.48% on CIFAR10-DVS, and N-Caltech101, respectively, outperforming the reverse order by 1.7% and 1.09%.

Model	**CIFAR10-DVS**			**N-Caltech101**		
	Full	Mask T-High	Noise T-High	Full	Mask T-High	Noise T-High
SWFormer	82.20	78.30 (-3.90)	44.50 (-37.70)	84.93	84.20 (-0.73)	49.89 (-34.31)
SSTWformer	84.90	78.70 (-6.20)	41.20 (-43.70)	87.48	85.05 (-2.43)	46.66 (-40.83)

Table 6: Performance comparison of SWFormer and SSTWformer on CIFAR10-DVS and N-Caltech101 under full and degraded (Mask/Noise T-High) inputs.

Datasets	**Block**	$\boldsymbol{\rho^E}$	**H/L**	$\boldsymbol{\rho^F}$
CIFAR10-DVS	$0 \rightarrow 1$	$29.3\% \rightarrow 15.5\%$	$0.59 \rightarrow 1.25$	$17.9\% \rightarrow 45.7\%$
N-Caltech101	$0 \rightarrow 1$	$27.0\% \rightarrow 16.6\%$	$0.56 \rightarrow 1.11$	$18.4\% \rightarrow 42.6\%$
DVS128	$0 \rightarrow 1$	$21.7\% \rightarrow 14.1\%$	$0.96 \rightarrow 1.02$	$30.1\% \rightarrow 34.2\%$

Table 7: Per-block temporal high-frequency profile of the trained SSTWformer.

Table 5 presents a comparative evaluation of SSTWformer and SWFormerFang et al. (2025) on the N-Caltech101 and CIFAR10-DVS datasets in four time-step configurations: T = 6, 10, 16, and 20. Extending the temporal window generally improves accuracy, confirming the benefit of temporal depth for spike feature learning. However, performance saturates or decreases slightly at T = 20, suggesting that overly long sequences can introduce redundancy. SSTWformer consistently outperforms SWFormer in all settings, demonstrating its robustness for dynamic spike-based recognition.

4.4 Input-level subband intervention experiments

As shown in Table 6, masking temporal high-frequency components reduces the accuracy of the frequency-aware model by 6.2% and 2.43% on CIFAR10-DVS and N-Caltech101, respectively; replacing this component with energy-matched noise causes 7× and 17× greater damage than complete removal, indicating that the network actively reads the structured content in this channel rather than treating it as residual noise. In contrast, removing temporal high frequencies reduces SWFormer's accuracy by only 3.9% and 0.73%, which is 1.6× and 3.3× smaller than the drop for SSTWformer, suggesting that its utilization of this information is merely implicit and incomplete. After removal, the accuracy gap between the two models narrows: 85% and 67% of the advantage of explicit spatiotemporal decomposition is mediated by temporal high-frequency information that existing architectures structurally waste.

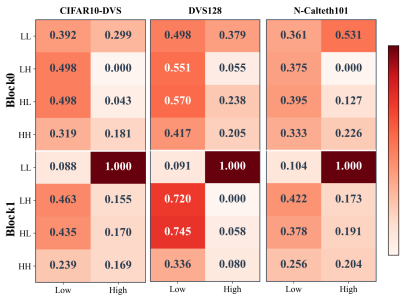


Figure 4: Visualization of spatiotemporal subband weights.

4.5 COMPENSATORY AMPLIFICATION OF TEMPORAL HIGH-FREQUENCY COMPONENTS

Figure 4 presents the spatiotemporal subband weight distributions of the model on three visual datasets. The model adopts a two-stage cascaded structure, where each spatial subband contains both temporal low-frequency and high-frequency components, totaling 8 spatiotemporal subbands. Through Min-Max normalization, the relative contribution of each subband is quantified.As the network depth increases, the weights of temporal high-frequency subbands increase markedly.

Models	IT	TT	MM
SWFormer	0.277s	1.24s	4235M
SSTWformer	0.073s	0.51s	4064M

Table 8: Comparison of inference time (IT), training time (TT) per batch, along with maximum GPU memory usage (MM) on N-Caltech101 at T=16.

To quantify this behavior, we profile three quantities at each block l over the full test set with the frozen model:

$$\rho_l^E = \frac{\sum_{k\in\mathcal{T}} E_{l,k}}{\sum_{k=1}^{8} E_{l,k}}, \qquad E_{l,k} = \frac{1}{N}\sum_{n=1}^{N} \frac{\left\| S_{l,k}^{(n)} \right\|_2^2}{\left| S_{l,k}^{(n)} \right|}, \qquad \rho_l^F = \frac{\sum_{k\in\mathcal{T}} \sqrt{E_{l,k}}\, \bar{w}_{l,k}}{\sum_{k=1}^{8} \sqrt{E_{l,k}}\, \bar{w}_{l,k}} \tag{13}$$

where $\mathcal{T}$ indexes the four temporal high-frequency subbands, $S_{l,k}^{(n)}$ is the subband-k tensor of the n-th test sample at block l, $|\cdot|$ denotes the number of elements, and N is the number of test samples.ρ^E: proportion of temporal high-frequency subbands in total subband energy. $\bar{w}_{l,k}$ is the channel-averaged FASTM gating weight of subband k at block l, $\mathrm{H/L} = \bar{w}_{\text{high}}/\bar{w}_{\text{low}}$, and ρ_l^F measures how much the network actually uses the information.

Statistical results point to a compensatory amplification of temporal high frequencies with depth (Table 7). On one hand, the energy share ρ^E of temporal high frequencies decreases with depth, confirming the inherent low-pass filtering property of spiking neural networks. On the other hand, the frequency-aware network does not passively accept this attenuation: on CIFAR10-DVS and N-Caltech101, the gating ratio H/L rises from 0.59 and 0.56 to 1.25 and 1.11, respectively—relative to the frequency-neutral initialization, shallow layers learn to down-weight temporal high frequencies, whereas deep layers learn to up-weight them. Combining the gates with the attenuated amplitudes, the effective utilization ρ^F rises from 17.9% and 18.4% to 45.7% and 42.6%. Motion-dominated tasks such as DVS128 inherently rely on high frequencies, so their compensatory amplification is less pronounced, though the overall trend remains consistent. These results portray temporal high frequencies as a scarce yet critical resource: deep spiking networks must actively defend them against their intrinsic low-pass nature. This defense requires frequency components to be explicitly decoupled and independently addressable—without a separately gatable temporal high-frequency channel, attenuated components cannot be selectively re-amplified.

4.6 EFFICIENCY ANALYSIS

As shown in Table 8, SSTWformer achieves a 3.8× inference speedup (0.073 s vs. 0.277 s per batch) and a 2.4× training speedup (0.51 s vs. 1.24 s per batch) over SWFormer on N-Caltech101, with slightly lower peak GPU memory (4064 MB vs. 4235 MB). The spatial-temporal Haar decomposition is nearly cost-free—parameter-free, matrix-free, and computed in place—while SWFormer's signed ternary neurons cannot leverage CuPy-accelerated kernels, whereas SSTWformer retains standard LIF neurons throughout.

5 CONCLUSION

Existing Spiking Transformers largely restrict frequency modeling to the spatial domain, overlooking temporal dynamics. Through input-level causal interventions, we reveal the critical role of temporal high-frequency information and identify a compensatory amplification mechanism in deep networks. This demonstrates the necessity of explicit spatiotemporal frequency modeling. We therefore propose SSTWformer, a spiking Transformer built on a spatial-temporal Haar wavelet transform with learnable subband weights.We hope this causal and mechanistic perspective provides useful insights for designing frequency-aware spiking neural networks.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and polishing of the manuscript. We have not used generative AI tools for generating synthetic data, developing theoretical models, formulating mathematical claims, proposing hypotheses, designing experiments, implementing methods, or interpreting results. We have reviewed all AI-assisted work and take full responsibility for the final content of this paper, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Snezana Agatonovic-Kustrin and Rosemary Beresford. Basic concepts of artificial neural network (ann) modeling and its application in pharmaceutical research. *Journal of pharmaceutical and biomedical analysis*, 22(5):717–727, 2000.

Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, 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.

Jiahang Cao, Hanzhong Guo, Ziqing Wang, Deming Zhou, Hao Cheng, Qiang Zhang, and Renjing Xu. Spiking diffusion models. *IEEE Transactions on Artificial Intelligence*, 2024.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

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

Jun Chen, Heye Zhang, Mingming Gong, and Zhifan Gao. Collaborative compensative transformer network for salient object detection. *Pattern Recognition*, 154:110600, 2024.

Benjamin Cramer, Yannik Stradmann, Johannes Schemmel, and Friedemann Zenke. The heidelberg spiking data sets for the systematic evaluation of spiking neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

Yongqi Ding, Lin Zuo, Mengmeng Jing, Pei He, and Yongjun Xiao. Shrinking your timestep: Towards low-latency neuromorphic object recognition with spiking neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11811–11819, 2024.

Yuetong Fang, Ziqing Wang, Lingfeng Zhang, Jiahang Cao, Honglei Chen, and Renjing Xu. Spiking wavelet transformer. In *European Conference on Computer Vision*, pp. 19–37. Springer, 2025.

Yuetong Fang, Deming Zhou, Ziqing Wang, Hongwei Ren, Lusong Li, Renjing Xu, et al. Spiking neural networks need high-frequency information. *Advances in Neural Information Processing Systems*, 38:103404–103425, 2026.

Yufei Guo, Yuanpei Chen, Xiaode Liu, Weihang Peng, Yuhan Zhang, Xuhui Huang, and Zhe Ma. Ternary spike: Learning ternary spikes for spiking neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 12244–12252, 2024.

Xun Ji, Shijie Chen, Li-Ying Hao, Jingchun Zhou, and Long Chen. Fbdpn: Cnn-transformer hybrid feature boosting and differential pyramid network for underwater object detection. *Expert Systems with Applications*, 256:124978, 2024.

Seijoon Kim, Seongsik Park, Byunggook Na, and Sungroh Yoon. Spiking-yolo: spiking neural network for energy-efficient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 11270–11277, 2020.

Donghyun Lee, Yuhang Li, Youngeun Kim, Shiting Xiao, and Priyadarshini Panda. Spiking transformer with spatial-temporal attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13948–13958, 2025.

Hongmin Li, Hanchao Liu, Xiangyang Ji, Guoqi Li, and Luping Shi. Cifar10-dvs: an event-stream dataset for object classification. *Frontiers in neuroscience*, 11:244131, 2017.

Javier López-Randulfe, Tobias Duswald, Zhenshan Bing, and Alois Knoll. Spiking neural network for fourier transform and object detection for automotive radar. *Frontiers in Neurorobotics*, 15: 688344, 2021.

Javier López-Randulfe, Nico Reeb, Negin Karimi, Chen Liu, Hector A Gonzalez, Robin Dietrich, Bernhard Vogginger, Christian Mayr, and Alois Knoll. Time-coded spiking fourier transform in neuromorphic hardware. *IEEE Transactions on Computers*, 71(11):2792–2802, 2022.

Chengzhuo Lu, Huilin Du, Wenjie Wei, Qian Sun, Yuchen Wang, Dingyi Zeng, Wenyu Chen, Malu Zhang, and Yang Yang. Estsformer: Efficient spatio-temporal spiking transformer. *Neural Networks*, pp. 107786, 2025.

Xianping Ma, Xiaokang Zhang, Man-On Pun, and Ming Liu. A multilevel multimodal fusion transformer for remote sensing semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024.

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

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.

Sicheng Shen, Dongcheng Zhao, Guobin Shen, and Yi Zeng. Tim: an efficient temporal interaction module for spiking transformer. *arXiv preprint arXiv:2401.11687*, 2024.

Xinyu Shi, Zecheng Hao, and Zhaofei Yu. Spikingresformer: Bridging resnet and vision transformer in spiking neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5610–5619, 2024.

Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. Hats: Histograms of averaged time surfaces for robust event-based object classification. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 1731–1740. IEEE, 2018.

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.

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

Min Yao, Yaozu Zhang, Guofeng Liu, and Dongdong Pang. Ssnet: A novel transformer and cnn hybrid network for remote sensing semantic segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:3023–3037, 2024.

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 558–567, 2021.

Zeqi Zheng, Yanchen Huang, Yingchao Yu, Zizheng Zhu, Junfeng Tang, Zhaofei Yu, and Yaochu Jin. Spiliformer: Enhancing spiking transformers with lateral inhibition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24539–24548, 2025.

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

Chenlin Zhou, Han Zhang, Zhaokun Zhou, Liutao Yu, Liwei Huang, Xiaopeng Fan, Li Yuan, Zhengyu Ma, Huihui Zhou, and Yonghong Tian. Qkformer: Hierarchical spiking transformer using qk attention. *Advances in Neural Information Processing Systems*, 37:13074–13098, 2024a.

Chenlin Zhou, Liutao Yu, Zhaokun Zhou, Han Zhang, Jiaqi Wang, Huihui Zhou, Zhengyu Ma, and Yonghong Tian. Spikingformer: A key foundation model for spiking neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2236–2244, 2026.

Shibo Zhou, Bo Yang, Mengwen Yuan, Runhao Jiang, Rui Yan, Gang Pan, and Huajin Tang. Enhancing snn-based spatio-temporal learning: A benchmark dataset and cross-modality attention model. *Neural Networks*, 180:106677, 2024b.

Zhaokun Zhou, Yuesheng Zhu, Chao He, Yaowei Wang, Shuicheng YAN, Yonghong Tian, and Li Yuan. Spikformer: When spiking neural network meets transformer. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=frE4fUwz_h.

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*, 36(3):5112–5125, 2024.

A PRELIMINARY

A.1 LIF SPIKING NEURON MODEL

The spiking neuron, as the fundamental unit of SNNs, receives input currents and accumulates membrane potential, which is then compared with a threshold to determine whether to generate a spike. In our study, we uniformly adopt the Leaky Integrate-and-Fire (LIF) spiking neuron model. The dynamics of the LIF model are described as follows:

$$H[t] = V[t-1] + \frac{1}{\tau}(X[t] - (V[t-1] - V_{reset})) \tag{14}$$

$$S[t] = \Theta(H[t] - V_t h), V[t] = H[t](1 - S[t]) + V_{reset} S[t] \tag{15}$$

where τ denotes the membrane time constant, and $X[t]$ represents the input current at timestep t. When the membrane potential $H[t]$ exceeds the firing threshold Vth, the neuron emits a spike $S[t]$. $\Theta(v)$ is a step function that equals 1 when $v \geq 0$ and 0 otherwise. $V[t]$ indicates the post-spiking membrane potential: it equals $H[t]$ if no spike is generated, otherwise it resets to the reset potential V_{reset}.

B SPIKING HAAR TRANSFORMER

B.1 SPIKING 1D HAAR TRANSFORM

In SNNs, the input signal is not a continuous amplitude value but a discrete sequence of events—spikes. The sole information carrier of a spike is the presence or absence of a spike, rather than its amplitude. Therefore, for a spike sequence $x = s[n]_{n=0}^{N-1}, s[n] \in \{0, 1\}$, the Haar transform no longer carries the physical meaning of "numerical decomposition" but rather a statistical analysis of the distribution of spike events across different spatial/temporal scales. The spiking Haar transform is defined as follows:

$$a[i] = s[2i] + s[2i+1] \tag{16}$$

$$d[i] = s[2i] - s[2i+1] \tag{17}$$

The approximation coefficient is given by Eq.16, and is computed as the number of spike events occurring within a local window, rather than the conventional amplitude average. This coefficient directly quantifies the spike firing density within that window. The detail coefficient is defined by

Eq. 17. Unlike the traditional Haar transform, where the detail coefficient reflects the magnitude of amplitude differences between adjacent samples, the detail coefficient in the spike Haar transform only determines whether a transition in spike state occurs between adjacent time instants. The detail coefficient of the spike Haar transform does not care about "how much the amplitude has changed," but only about "whether a change has occurred" and "in which direction it occurs". This is precisely in line with the event-driven nature of SNNs—each spike carries only the information that "an event has occurred."

B.2 Spiking 2D Haar Transform

Consider a two-dimensional spike image $S \in \{0,1\}^{H\times W}$, where each pixel $s_{ij} \in \{0,1\}$ indicates whether a spike event exists at that location, and both H and W are even integers. For any 2×2 spike block:

$$S_{ij} = \begin{bmatrix} s_{00} & s_{01} \\ s_{10} & s_{11} \end{bmatrix}, \quad i,j = 0,1,\ldots,\frac{H}{2}-1, \tag{18}$$

The conventional two-dimensional Haar transform performs weighted linear combinations of pixel values in the four quadrants to extract magnitude-difference information in the horizontal, vertical, and diagonal directions. However, in the spike-based visual representation, an individual pixel carries only the binary semantics of event presence or absence, with no amplitude dimension involved. Therefore, the two-dimensional Haar transform needs to be reinterpreted as a statistical measure of the spatial distribution of spike events across various directions:

$$\mathrm{LL}_{i,j} = s_{00} + s_{01} + s_{10} + s_{11} \tag{19}$$
$$\mathrm{LH}_{i,j} = (s_{00} + s_{10}) - (s_{01} + s_{11}) \tag{20}$$
$$\mathrm{HL}_{i,j} = (s_{00} + s_{01}) - (s_{10} + s_{11}) \tag{21}$$
$$\mathrm{HH}_{i,j} = (s_{00} + s_{11}) - (s_{01} + s_{10}) \tag{22}$$

C Experimental Details

C.1 Datasets

C.1.1 Spiking Vision Datasets

CIFAR10-DVSLi et al. (2017) is a neuromorphic dataset that provides 9,000 training samples and 1,000 test samples by converting static image samples into a format recorded by DVS cameras. N-Caltech101Orchard et al. (2015), derived from the traditional computer vision dataset Caltech101, is an event-camera-based version specifically designed for neuromorphic vision tasks. The DVS128 Gesture datasetAmir et al. (2017) comprises 11 hand gesture categories, capturing the movements of 29 participants under three different lighting conditions.NCARSSironi et al. (2018)is a dataset collected from real-world driving scenarios, with the task of distinguishing between cars and non-cars (background). The data was recorded by an ATIS camera mounted behind a car's windshield, and samples were extracted from continuous driving video streams in a semi-automated manner, making it highly representative of real-world application scenarios. It contains more than 24,000 samples, each 100 ms in length, and serves as a valuable resource for event-based vision in autonomous driving.

C.1.2 Spiking Audio Datasets

SHD(Spiking Heidelberg Digits)Cramer et al. (2020) is a dataset of spike trains generated by processing audio signals through a cochlear model. The task is to classify spoken digits (0–9) in both English and German, totaling 20 categories. The SSC(Spiking Speech Commands)Cramer et al. (2020) dataset is a spike-based dataset generated from Google's Speech Commands audio dataset by processing audio signals through a cochlear model. The dataset comprises 35 word categories and is stored in HDF5 format as spike trains with 700 input channels. Together with the SHD dataset, it belongs to the Heidelberg Spiking Datasets series, providing a standardized benchmark for evaluating SNN performance in auditory information processing.

C.2 TRAINING DETAILS

For all neuromorphic vision datasets used in our experiments, the model employs a two-stage configuration. The datasets are split into training and test sets at a 9:1 ratio, with inputs resized to 128 × 128. SHD and SSC use the official data splits.

In our experiments, we train models on various datasets using a single NVIDIA A800 GPU with 80 GB of memory. For direct training of SNN models, we employ surrogate gradient(Eq. 23) methods for backpropagation.We select the Sigmoid function $\sigma(x)$ as the surrogate function with $\alpha = 4$ during the backpropagation of direct training in all experiments.

$$\sigma(x) = \frac{1}{1 + exp(-\alpha x)} \tag{23}$$

For all neuromorphic datasets, the same data augmentation strategies are adopted, including horizontal flip, SNNAugmentWide, and Mixup. The AdamW optimizer is used, and the learning rate is scheduled via cosine decay. On the CIFAR10-DVS , N-Caltech101,NCARS,SHD, and SSC datasets, the number of training epochs is set to 106, with an initial learning rate of 0.005. For the DVS128 dataset, the training epochs are set to 200, and the initial learning rate is 0.001. The number of time steps is uniformly set to 16, and the batch size is set to 16.