A SPARSE FREQUENCY-AWARE SPIKING TRANSFORMER WITH HIERARCHICAL WAVELET PRUNING AND ADAPTIVE DUAL-PATH ROUTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Spiking neural networks (SNNs) offer significant energy efficiency advantages, yet existing compression methods perform uniform pruning on temporal feature maps, failing to exploit the substantially differentiated redundancy structures among different frequency components. Inspired by the M/P pathway frequency segregation mechanism in biological visual systems, this paper proposes the Sparse Frequency-Aware Transformer (SFAT), which achieves frequency-aware structured sparsity through hierarchical Haar wavelet decomposition: coefficient-wise soft thresholding distinguishes low/high-frequency subbands, while channel-wise energy gating adaptively prunes the four quadrant subbands. To compensate for high-frequency information loss induced by pruning, we design the Dual-Path MLP Routing mechanism (DPMR), comprising a frequency compensation path with subband-specific experts and global gating, supporting both a Parallel mode for high-precision scenarios and a Lightweight mode for edge deployment. Experiments on CIFAR-100, Tiny-ImageNet, and neuromorphic datasets demonstrate that SFAT(Parallel) achieves 79.70% on CIFAR-100 and 66.44% on Tiny-ImageNet, outperforming existing Spiking Transformer methods while reducing energy consumption by approximately 60%; SFAT(Lightweight) achieves comparable accuracy with nearly half the parameters, validating the synergistic optimization between frequency decomposition and structural compression. Code will be made available upon acceptance.

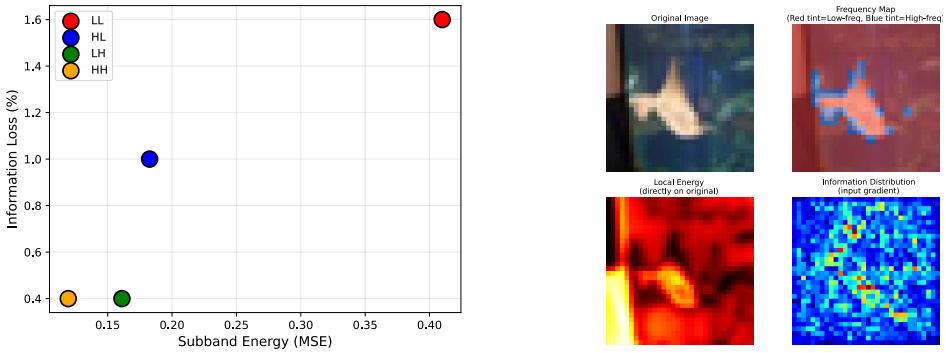
1 INTRODUCTION

Spiking neural networks (SNNs) emulate biological neuronal spike transmission, exchanging information via discrete spike sequences and triggering sparse computation only upon spike arrival, thereby drastically reducing resource consumption and offering significant energy-efficiency advantages Roy et al. (2019). With the rapid development of neuromorphic chips Merolla et al. (2014); Davies et al. (2018); Pei et al. (2019), the efficient hardware execution of SNNs has been validated. The maturation of direct training methods has further freed SNNs from the long latency constraints of ANN-to-SNN conversion Wu et al. (2018). In particular, Spikformer introduced the Transformer architecture into the SNN domain, pioneering research on spike-driven self-attention mechanisms Zhou et al. (2023); Yao et al. (2023).

However, existing SNN compression methods—such as the token-weight joint pruning of SparseSpikFormer Liu et al. (2024) and the NAS evolutionary search of Auto-SpikFormer Che et al. (2024)—all perform pruning on temporal-domain feature maps, failing to explicitly exploit structured redundancy in the frequency domain and thus limiting compression efficiency.

To explicitly exploit the structured redundancy residing in the frequency domain, this paper draws inspiration from the frequency-processing mechanisms of biological visual systems. The primate visual system achieves parallel frequency processing through the M/P pathways: the M pathway is sensitive to low spatial frequencies, rapidly transmitting global contours; the P pathway is sensitive to high spatial frequencies, slowly encoding fine details Livingstone & Hubel (1988); Tootell & Nasr (2017). The coarse-to-fine processing strategy further reduces neural encoding redundancy Skyberg et al. (2022). Inspired by these mechanisms, this paper adopts the Haar wavelet transform as

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.



(a) Subband-level analysis on SWFormer Block 2.

(b) Original-image granularity validation.

Figure 1: Frequency–energy–information consistency analysis at two complementary granularities.

a low/high-frequency decomposition tool. The orthogonality and binary coefficients of the Haar wavelet naturally match the discrete activation patterns of SNNs Luvizotto et al. (2012), and its multi-level subband structure (LL/HL/LH/HH) corresponds to the hierarchical frequency organization of the visual system.

To quantitatively verify the structured redundancy in the frequency domain, we conduct pre-experiments on CIFAR-100 using the Spiking Wavelet Transformer (SWFormer) Fang et al. (2024) at two complementary granularities. At the image level, we perform a frequency-energy-information consistency verification. For each 32×32 image, we apply the Haar transform and compare the reconstruction amplitudes of the LL subband versus the combined HL+LH+HH subbands, overlaying the low-/high-frequency attributes as semi-transparent tints onto the original image to visualize the frequency distribution. We further compute the pixel-wise squared energy and measure the information content via the gradient saliency of the correct-class logit with respect to input pixels. At the model feature level, we analyze the Haar forward transform output at Block 2 of SWFormer. We compute the energy (mean squared value) of each subband and measure the accuracy loss via a spatial-quadrant zero-out ablation experiment to quantify the information content of each subband. As shown in Fig. 1, both experiments consistently demonstrate that wavelet-domain energy positively correlates with information content: high-energy subbands (e.g., LL) correspond to greater information loss, while low-energy subbands (e.g., HH) have minimal impact, providing a quantifiable basis for differentiated sparsity.

To address the aforementioned challenges and findings, this paper proposes the Sparse Frequency-Aware Transformer (SFAT). The core contributions are as follows:

Contribution 1: Hierarchical Sparse Wavelet Layer (HSWL). We extend the single 2D transform into a horizontal-vertical separated two-stage decomposition, inserting specific pruning strategies after each level: Level 1 implements coefficient-level sparsity (soft-threshold denoising, distinguishing L/H subbands), and Level 2 implements channel-level sparsity (gated pruning, distinguishing the LL/HL/LH/HH four subbands). This progressive hierarchical processing explicitly mimics the biological coarse-to-fine strategy, achieving a controllable trade-off between computational efficiency and frequency representation.

Contribution 2: Dual-Path MLP Routing Mechanism (DPMR). We design a Frequency Compensation Pathway (FCP) to replace or augment the conventional MLP: features are split into four groups (corresponding to the four wavelet subbands), processed by independent subband experts, and dynamically weighted through global gating. Two operating modes are provided—Parallel mode (FCP fused with the base MLP residual, for high-precision scenarios) and Lightweight mode (FCP fully replaces the base MLP, reducing parameters to about one half, for edge-deployment scenarios).

Contribution 3: We validate the effectiveness of the proposed method on CIFAR-100, Tiny-ImageNet, and neuromorphic datasets (CIFAR10-DVS, DVS-Gesture). On Tiny-ImageNet, the Parallel mode achieves a 0.65% accuracy improvement over Spikformer while saving 65.9% en-

ergy; on CIFAR-100, the Lightweight mode achieves comparable accuracy to Spikformer (77.88% vs. 77.86%) while reducing parameters by 55.7%.

2 Related Work

Wavelet transform: bridging biological vision and computational models. Wavelet transform achieves time-frequency localization through multi-resolution analysis, and the spatial scales of its basis functions directly correspond to neuronal receptive fields Skriabine et al. (2026). The orthogonality and binary coefficient properties of the Haar wavelet make it particularly suitable for the discrete activation patterns of spiking neural networks Luvizotto et al. (2012). However, existing wavelet applications are limited to traditional ANN frameworks—WTConv Chen et al. (2025), WaveMixSR Jeevan et al. (2024), HWD Xu et al. (2023), and the frequency-domain Transformer GFNet Rao et al. (2021) all employ floating-point operations and dense activations, failing to achieve deep synergy with spike-driven sparse characteristics.

Architectural limitations of Spiking Transformers. Spikformer Zhou et al. (2023) pioneered spike-driven self-attention, but subsequent works—the hierarchical spatial sampling of QKFormer Zhou et al. (2024), the multi-scale residual fusion of SpikingResformer Shi et al. (2024), and the pure spike-driven residual learning of Spike-driven Transformer Yao et al. (2023)—all directly inherit ANN Transformer architectures, simply binarizing the self-attention mechanism without considering the essential characteristics of SNNs in processing high-frequency event data. SWFormer Fang et al. (2024) first introduced wavelets into the SNN-Transformer, but its fixed Haar transform plus block-diagonal weight structure does not optimize sparsity, leaving significant information redundancy; more critically, it employs the traditional MLP ($C \to 4C \to C$ dense fully connected), whose large capacity parameters fundamentally conflict with the sparsification objective.

Fundamental gap in SNN compression. Existing compression methods—the token-weight joint pruning of SparseSpikFormer Liu et al. (2024) and the NAS evolutionary search of AutoSpikFormer Che et al. (2024)—all perform pruning on temporal feature maps that have already lost high-frequency information. The low-pass characteristic of self-attention has already filtered out critical details, so subsequent compression can only optimize the storage of distorted features rather than achieving frequency-selective retention from the source. This limitation calls for a pruning mechanism deeply synergistic with frequency decomposition: implementing hierarchical sparsity immediately after wavelet transformation, completing selective retention and structural compression within the frequency domain.

3 Preliminaries

3.1 Spiking Neuron Model and Notation

Leaky Integrate-and-Fire (LIF) neuron. The LIF neuron simulates biological spike dynamics through a charge-leak-fire process Maass (1997); Stein (1965); Burkitt (2006):

$$v_i[t] = u_i[t] + \frac{1}{\tau}\left(I_i[t] - (u_i[t] - u_{\text{rest}})\right), \tag{1}$$

$$s_i[t] = H\left(v_i[t] - u_{\text{th}}\right), \tag{2}$$

$$u_i[t+1] = s_i[t]\, u_{\text{rest}} + (1 - s_i[t])\, v_i[t], \tag{3}$$

where $u_i[t]$ and $v_i[t]$ denote the membrane potential before and after charging at time t, respectively, τ is the membrane time constant, $I_i[t] = \sum_j w_{i,j} s_j[t]$ is the synaptic input current, u_{th} is the firing threshold, u_{rest} is the resting potential, and $H(\cdot)$ is the Heaviside step function.

3.2 Spike-Driven Haar Wavelet Basics

Ternary spikes and Haar transform. Traditional binary spikes $\{0, 1\}$ cannot represent negative values in wavelet coefficients, leading to severe distortion in frequency representation Fang et al.

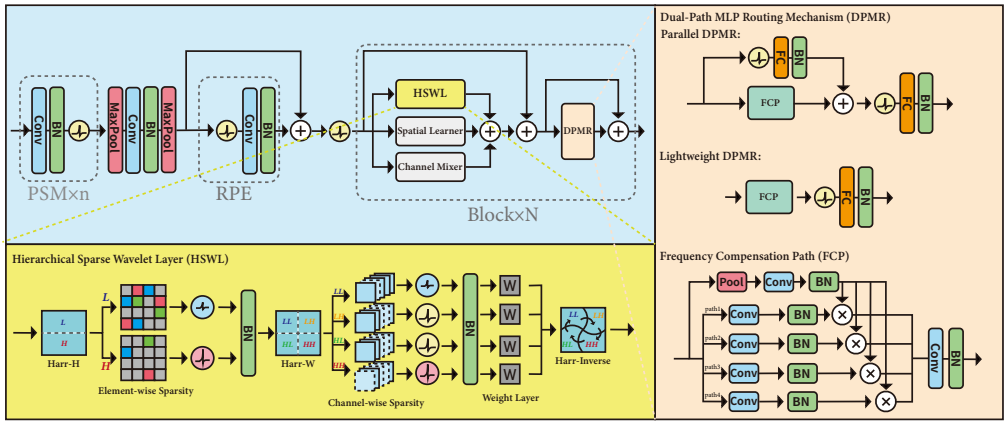


Figure 2: Overall architecture of SFAT. Top-left: SPS (PSM, RPE) and stacked encoder blocks; bottom-left: HSWL two-level decomposition-sparsity-spiking flow; right: DPMR Parallel/Lightweight dual-mode design.

(2024). SWFormer introduced the ternary neuron $\mathrm{Spk}_{\mathrm{sym}} : \mathbb{R} \to \{-1, 0, +1\}$:

$$\mathrm{Spk}_{\mathrm{sym}}(u; v_{\mathrm{th}}) = \begin{cases} +1, & u \geq v_{\mathrm{th}}, \\ -1, & u \leq -v_{\mathrm{th}}, \\ 0, & \text{otherwise}. \end{cases} \tag{4}$$

The Haar wavelet filter coefficients $[1, 1]/\sqrt{2}$ and $[1, -1]/\sqrt{2}$ contain negative values, naturally matching the ternary spikes Luvizotto et al. (2012). Following the spike-driven paradigm of SWFormer Fang et al. (2024), the forward/inverse transforms are spiked layer-by-layer through $\mathrm{Spk}_{\mathrm{sym}}$:

$$\mathbf{H}_f = \mathrm{Spk}_{\mathrm{sym}}\left(\mathbf{W}_{\mathrm{haar}} \cdot \mathrm{Spk}_{\mathrm{sym}}\left(\mathbf{I} \cdot \mathbf{W}_{\mathrm{haar}}^{\top}\right)\right), \tag{5}$$

$$\mathbf{I}_{\mathrm{rec}} = \mathrm{Spk}_{\mathrm{sym}}\left(\mathbf{W}_{\mathrm{haar}}^{\top} \cdot \mathrm{Spk}_{\mathrm{sym}}\left(\mathbf{H}_f \cdot \mathbf{W}_{\mathrm{haar}}\right)\right). \tag{6}$$

This paper extends the single 2D transform into a horizontal-vertical separated two-level decomposition, providing an implementation foundation for the hierarchical sparsity described in Section 4.2.

4 Method

4.1 Design Principles and Overall Architecture

Inspired by the M/P pathway separation mechanism and coarse-to-fine processing strategy of the human visual system, this paper proposes the Sparse Frequency-Aware Transformer (SFAT), constructing a dual-path Spiking Transformer architecture oriented toward edge intelligence. As shown in Figure 2, the network consists of a Spiking Patch Splitting (SPS) module, L SFAT encoder blocks cascaded together, and a classification head. Each encoder block contains: Sparse FATM layer $\rightarrow$ DPMR layer $\rightarrow$ residual connection, where the sparse FATM achieves frequency-aware spatial sparsification through the Hierarchical Sparse Wavelet Layer (HSWL, Section 4.2), mimicking the fast low-frequency response of the M pathway and the fine high-frequency processing of the P pathway; the DPMR selectively recovers high-frequency information lost to sparsification through the Frequency Compensation Pathway, and dynamically switches between Parallel/Lightweight dual modes according to the deployment scenario (Section 4.3), achieving channel-wise feature transformation and controllable trade-offs.

4.2 Hierarchical Sparse Wavelet Layer (HSWL)

Similar to SWFormer, our method uses FATM to replace the self-attention layer, comprising three parallel paths: a frequency-domain branch, a temporal-domain branch, and a cross-channel branch.

For the frequency-domain branch, we propose HSWL to achieve biologically inspired frequency-aware processing through progressive horizontal-vertical separated decomposition and hierarchical structured sparsity mechanisms. Specifically, HSWL contains two levels of sparsity strategies: Element-wise Sparsity (EWS) at Level 1 applies coefficient-level soft-threshold denoising to the **L/H** subbands after horizontal decomposition; Channel-wise Sparsity (CWS) at Level 2 applies channel-level energy pruning to the **LL/HL/LH/HH** four subbands after vertical decomposition.

Level 1: Element-wise Sparsity (EWS). For the input feature $\mathbf{S}' \in \mathbb{R}^{T\times B\times C\times H\times W}$, we perform Haar decomposition along the width dimension (see Section 3.2), obtaining $\mathbf{L}, \mathbf{H} \in \mathbb{R}^{T\times B\times C\times H\times W/2}$. Stacking them along the channel dimension yields $\mathbf{S}_1 \in \mathbb{R}^{T\times B\times 2C\times H\times W/2}$. We then apply learnable soft-threshold sparsification independently to each coefficient:

$$\tilde{\mathbf{S}}_1(t,b,c,h,w) = \operatorname{sign}(\mathbf{S}_1)\cdot \operatorname{ReLU}\left(|\mathbf{S}_1| - \tau_1(c)\right), \tag{7}$$

where the learnable threshold $\boldsymbol{\tau}_1 \in \mathbb{R}^{2C}$ is initialized differentially by subband: low-frequency branch $\tau_1^{\mathbf{L}} <$ high-frequency branch $\tau_1^{\mathbf{H}}$, and gradient propagation is achieved through the STE estimator.

After the first subband-aware spiking $\hat{\mathbf{S}}_1 = \mathrm{Spk}_{\mathrm{sym}}(\tilde{\mathbf{S}}_1; \mathbf{V}_{\mathrm{th}}^{(1)})$, the feature enters Level 2, where the threshold $\mathbf{V}_{\mathrm{th}}^{(1)} \in \mathbb{R}^{2C}$ distinguishes the **L/H** subbands.

Level 2: Channel-wise Sparsity (CWS). We perform vertical Haar decomposition on $\hat{\mathbf{S}}_1$ along the height dimension, obtaining the four subbands $\{\mathbf{LL}, \mathbf{HL}, \mathbf{LH}, \mathbf{HH}\}$ (each of dimension $\mathbb{R}^{T\times B\times C\times H/2\times W/2}$, see Section 3.2).

This paper adopts energy-driven channel sparsity (EnerChSparsity), which dynamically computes gating weights based on the channel energy response of each subband without additional learnable parameters. The subband energy is defined as:

$$\mathcal{E}_{\mathrm{sub}}(c) = \frac{1}{TB\cdot \frac{H}{2}\cdot \frac{W}{2}} \sum_{t,b,h,w} |\mathrm{sub}(t,b,c,h,w)|, \quad \mathrm{sub} \in \{\mathbf{LL}, \mathbf{HL}, \mathbf{LH}, \mathbf{HH}\}. \tag{8}$$

Energy-weighted fusion gating during training is

$$\alpha_{\mathrm{sub}}(c) = \sigma\left(\frac{\mathcal{E}_{\mathrm{sub}}(c) - E_{\mathrm{center}}}{\tau_E}\right), \quad z_{\mathrm{sub}}(c) = \alpha_{\mathrm{sub}}(c)\cdot z_{\mathrm{high}} + (1-\alpha_{\mathrm{sub}}(c))\cdot z_{\mathrm{low}}, \tag{9}$$

where $z_{\mathrm{low}} \sim \mathcal{N}(\mu_{\mathrm{low}}, \sigma^2)$, $z_{\mathrm{high}} \sim \mathcal{N}(\mu_{\mathrm{high}}, \sigma^2)$ are bimodal normal-distribution samples, with $\mu_{\mathrm{low}} = \ln(0.4/0.6)$ and $\mu_{\mathrm{high}} = \ln(0.6/0.4)$. Deterministic gating at inference is

$$g_{\mathrm{sub}}(c) = \alpha_{\mathrm{sub}}(c)\cdot 0.6 + (1-\alpha_{\mathrm{sub}}(c))\cdot 0.4, \quad \hat{\mathrm{sub}} = g_{\mathrm{sub}}(c)\cdot \mathrm{sub}. \tag{10}$$

The four subbands are stacked after channel-level sparsity into $\mathbf{S}_2 \in \mathbb{R}^{T\times B\times 4C\times H/2\times W/2}$, and then output through the second subband-aware spiking:

$$\hat{\mathbf{S}}_2 = \mathrm{Spk}_{\mathrm{sym}}(\mathbf{S}_2; \mathbf{V}_{\mathrm{th}}^{(2)}), \tag{11}$$

where the threshold $\mathbf{V}_{\mathrm{th}}^{(2)} \in \mathbb{R}^{4C}$ distinguishes the **LL/HL/LH/HH** four subbands, all jointly optimized through backpropagation. The forward flow is summarized in Algorithm 1.

4.3 Dual-Path MLP Routing Mechanism (DPMR)

HSWL achieves significant spike-activity sparsification through progressive wavelet decomposition, yet inevitably introduces high-frequency information loss. The traditional MLP ($C \to 4C \to C$) is insensitive to frequency information and overly large, failing to compensate for HSWL's high-frequency loss. To this end, we propose DPMR, achieving a controllable trade-off between error compensation and parameter efficiency.

DPMR contains two core components: the base MLP path provides standard feature transformation capability; the Frequency Compensation Pathway (FCP) selectively compensates for detail information lost to HSWL sparsification. Two operating modes are provided: Parallel mode fuses the base MLP and FCP residually for high-precision scenarios; Lightweight mode completely replaces the base MLP with FCP for edge-deployment scenarios.

Algorithm 1 HSWL Forward Propagation

Require: Input spike feature $\mathbf{S}' \in \mathbb{R}^{T\times B\times C\times H\times W}$, learnable threshold $\boldsymbol{\tau}_1$, spike thresholds $\mathbf{V}_{\text{th}}^{(1)}, \mathbf{V}_{\text{th}}^{(2)}$
Ensure: Sparsified spike feature $\hat{\mathbf{S}}_2 \in \mathbb{R}^{T\times B\times 4C\times H/2\times W/2}$
$\mathbf{L}, \mathbf{H} \leftarrow \text{Haar}_W(\mathbf{S}')$ {Section 3.2}
$\mathbf{S}_1 \leftarrow \text{Concat}(\mathbf{L}, \mathbf{H})$
$\tilde{\mathbf{S}}_1 \leftarrow \text{CoeffSparsity}(\mathbf{S}_1; \boldsymbol{\tau}_1)$ {Eq. (7)}
$\hat{\mathbf{S}}_1 \leftarrow \text{Spk}_{\text{sym}}(\tilde{\mathbf{S}}_1; \mathbf{V}_{\text{th}}^{(1)})$
$\mathbf{LL}, \mathbf{HL}, \mathbf{LH}, \mathbf{HH} \leftarrow \text{Slice}(\text{Haar}_H(\hat{\mathbf{S}}_1))$
$\tilde{\mathbf{sub}} \leftarrow \text{EnerChSparsity}(\mathbf{sub}), \forall \mathbf{sub}$ {Eqs. (8)–(10)}
$\mathbf{S}_2 \leftarrow \text{Concat}(\tilde{\mathbf{LL}}, \tilde{\mathbf{HL}}, \tilde{\mathbf{LH}}, \tilde{\mathbf{HH}})$
$\hat{\mathbf{S}}_2 \leftarrow \text{Spk}_{\text{sym}}(\mathbf{S}_2; \mathbf{V}_{\text{th}}^{(2)})$
return $\hat{\mathbf{S}}_2$

4.3.1 Frequency Compensation Pathway (FCP)

FCP simulates the parallel processing mechanism of cortical functional columns, splitting the input feature $\mathbf{x} \in \mathbb{R}^{T\times B\times C\times H\times W}$ into $K = 4$ groups along the channel dimension (corresponding to the four wavelet subband structures), with each group processed by an independent subband expert and dynamically weighted through a global gating network.

Subband partition and expert processing. The input is evenly divided into K groups along the channel dimension, and each group undergoes lightweight 1×1 convolution transformation:

$$\mathbf{x}^{(k)} = \mathbf{x}_{:,:,k\cdot C/K:(k+1)\cdot C/K,:,:}, \quad \mathbf{h}^{(k)} = \text{BN}\left(\text{Conv}_{1\times 1}\left(\mathbf{x}^{(k)}; \mathbf{W}_e^{(k)}\right)\right), \tag{12}$$

where $C_h = \lfloor C_{\text{hidden}} \cdot r/K \rfloor$, and r is the frequency-expert ratio (default 0.25).

Global compensation gating and weighted fusion. Compensation weights for each subband are generated based on global spatial information, and after weighted concatenation the output is spiked:

$$\boldsymbol{\gamma} = \sigma\left(\text{BN}\left(\text{Conv}_{1\times 1}\left(\text{GAP}(\mathbf{x}); \mathbf{W}_g\right)\right)\right) \in \mathbb{R}^{T\times B\times K}, \tag{13}$$

$$\mathbf{h}_{\text{comp}} = \text{Spk}_{\text{sym}}\left(\text{BN}\left(\text{Conv}_{1\times 1}\left(\text{Concat}\left[\boldsymbol{\gamma}^{(k)}\mathbf{h}^{(k)}\right]_{k=0}^{K-1}; \mathbf{W}_f\right)\right); v_{\text{th}}^{\text{fcp}}\right), \tag{14}$$

where $v_{\text{th}}^{\text{fcp}} = 0.2$ is a reduced spike threshold to mitigate spike-induced information loss.

Learnable compensation strength. A learnable scalar β controls the fusion ratio of the FCP output:

$$\mathbf{x}_{\text{out}} = \mathbf{x}_{\text{base}} + \sigma(\beta) \cdot \mathbf{h}_{\text{comp}}, \tag{15}$$

where $\mathbf{x}_{\text{base}}$ is the base MLP path output, and $\sigma(\beta)$ is constrained to $(0, 1)$ through Sigmoid.

4.3.2 Two Operating Modes

DPMR switches modes by configuring the `skip_base_mlp` Boolean parameter. The two modes share the same FCP frequency compensation pathway, differing only in whether the base MLP is retained and in the output fusion method, enabling flexible trade-offs between accuracy and efficiency.

Parallel mode (`skip_base_mlp=False`). The base MLP retains the standard structure ($\mathbf{x}_{\text{base}} = \text{BaseMLP}(\mathbf{x})$), and FCP serves as a residual branch ($\mathbf{h}_{\text{comp}} = \text{FCP}(\mathbf{x})$) to compensate for high-frequency information. The output is fused via Eq. (15).

Lightweight mode (`skip_base_mlp=True`). The two large-scale convolution layers of the base MLP are removed, retaining only the FCP path and an output projection ($\mathbf{x}_{\text{out}} = \text{BN}_{\text{out}}(\text{Conv}_{\text{out}}(\mathbf{h}_{\text{comp}}))$), reducing parameters to about one half of the Parallel mode.

Table 1: Comparison of architectures on CIFAR-100. $*$ indicates results trained by us. Results are reported as mean$\pm$std over three seeds when available.

Method	**Arch.**	T	Top-1 (%)	Param (M)	SOPs (G)	Energy (mJ)
ANN ResNet-19 Zheng et al. (2021)	ResNet-19	1	75.35	12.63	–	–
ANN Transformer-4-384 Zhou et al. (2023)	Transformer	1	81.02	9.32	–	–
STBP-tdBN Zheng et al. (2021)	ResNet-19	4	70.86	12.63	–	–
TET Deng et al. (2022)	ResNet-19	4	74.47	12.63	–	–
TEBN Duan et al. (2022)	ResNet-19	6	76.41	12.63	–	–
Real Spike Guo et al. (2022)	ResNet-19	10	71.24	39.9	–	–
DSR Meng et al. (2022)	ResNet-19	20	78.50	11.2	–	–
Spikformer Zhou et al. (2023)	T-4-256	4	75.96	4.15	–	–
	T-2-384	4	76.95	5.76	–	–
	T-4-384	4	77.86	9.32	1.458	1.318
SDformer Yao et al. (2023)	T-4-384	4	78.40	9.36	0.564	0.513
SWFormer Fang et al. (2024)	T-4-384	4	79.30	7.51	1.04	0.942
Ours (Parallel)$*$	T-4-256	4	78.12$\pm$0.07	4.50	0.294	0.268
	T-2-384	4	78.00$\pm$0.45	6.15	0.369	0.338
	T-4-384	4	**79.70$\pm$0.18**	10.06	0.574	0.522
Ours (Lightweight)$*$	T-4-256	4	75.40$\pm$0.17	1.86	0.181	0.167
	T-2-384	4	76.29$\pm$0.04	3.19	0.253	0.233
	T-4-384	4	77.88$\pm$0.24	4.13	0.375	0.342

5 Experiments

5.1 Experimental Setup

Datasets. Static datasets include CIFAR-100 Krizhevsky et al. (2009) (32×32) and Tiny-ImageNet Le & Yang (2015) (64×64, 200 classes); neuromorphic datasets include CIFAR10-DVS Li et al. (2017) (128×128) and DVS-Gesture Amir et al. (2017) (128×128, 11 classes).

Implementation details. Implemented in PyTorch/SpikingJelly with LIF neurons, AdamW optimizer, cosine annealing (base LR 0.0015, weight decay 0.06), and RandAugment+Mixup+label smoothing. CIFAR-100/Tiny-ImageNet: 400/200 epochs, batch size 128, $T = 4$; DVS datasets: 100 epochs, batch size 16, $T = 16$. HSWL thresholds $\tau_1^L = 0.01$, $\tau_1^H = 0.05$; DPMR gate β initialized to 0 and clamped by Sigmoid to $(0, 1)$. All CIFAR-100 results are averaged over three independent random seeds.

Evaluation metrics. Top-1 accuracy, parameter count (Params), synaptic operations (SOPs), and energy consumption (Energy, based on 45nm technology $E_{\text{MAC}} = 4.6$ pJ, $E_{\text{AC}} = 0.9$ pJ).

5.2 Comparison Experiments

CIFAR-100. Table 1 summarizes the comparison results on CIFAR-100. Ours (Parallel) reaches a three-seed mean of **79.70%** under the Transformer-4-384 configuration, surpassing Spikformer (77.86%), SDformer (78.40%), and SWFormer (79.30%). While achieving higher accuracy, Ours (Parallel) reduces SOPs and energy consumption to 55.2% and 55.4% of SWFormer, respectively, demonstrating superior computational efficiency. Ours (Lightweight), designed for edge-deployment scenarios, achieves 77.88% with approximately 55% of the parameters of SWFormer, with SOPs and energy further reduced to about 37% of SWFormer, trading a marginal accuracy change for substantial efficiency gains.

Tiny-ImageNet. Table 2 summarizes the main comparison on Tiny-ImageNet ($T = 4$, Transformer-4-384). Ours (Parallel) achieves **66.44%**, surpassing all comparison methods. In particular, the energy consumption of Ours (Parallel) is only 34.1% of Spikformer and 63.0% of SWFormer, while accuracy leads by +0.65% and +1.70%, respectively. Ours (Lightweight) achieves comparable performance (62.68%) with approximately 55% of the parameters, with energy further reduced to 20.0% of Spikformer and 37.0% of SWFormer.

Table 2: Main comparison results on Tiny-ImageNet (T-4-384). $*$ indicates our retrained results.

Method	T	Top-1 (%)	Param (M)	SOPs (G)	Energy (mJ)
Spikformer$*$ Zhou et al. (2023)	4	65.79	9.40	5.98	5.40
SDformer$*$ Yao et al. (2023)	4	65.63	9.39	2.15	1.96
SWFormer$*$ Fang et al. (2024)	4	64.74	7.55	3.22	2.92
Ours (Parallel)$*$	4	**66.44**	10.10	2.02	1.84
Ours (Lightweight)$*$	4	62.68	4.17	1.18	1.08

No Spiking Transformer results on Tiny-ImageNet were publicly reported prior to this work; all baselines were retrained from official code under identical settings.

Neuromorphic datasets. To verify generalization on event-driven data, we conducted supplementary experiments on CIFAR10-DVS and DVS-Gesture. Because the high-frequency dynamic characteristics of neuromorphic data differ structurally from the low-frequency dominant prior established on static images, SFAT performs slightly below Spikformer and SDformer on this domain (CIFAR10-DVS: 78.7% vs 80.9%; DVS-Gesture: 97.9% vs 99.3%). This result reflects that the frequency-selective sparsity mechanism of HSWL achieves optimal energy-efficiency trade-offs on low-frequency-dominant static images, but faces domain mismatch on high-frequency event data; nevertheless, SFAT maintains comparable accuracy.

5.3 Ablation Study

Table 3 presents the core ablation results for the Parallel and Lightweight modes on CIFAR-100. All experiments use three seeds; accuracy is mean$\pm$std.

First, removing HSWL has negligible impact on the accuracy of the Parallel mode (79.84% vs. 79.70%), with the difference falling within statistical fluctuations. However, SOPs and energy increase by 22.6% and 22.4%, respectively, demonstrating that HSWL effectively compresses computational cost and energy expenditure. For the Lightweight mode, removing HSWL leads to a 0.37% accuracy drop (77.51% vs. 77.88%), accompanied by 23.5% and 23.4% increases in SOPs and energy, indicating that HSWL contributes positively to both efficiency and accuracy of the lightweight architecture.

Second, replacing the proposed FCP with the classic MLP reduces Parallel accuracy by 0.40% (79.30% vs. 79.70%) while SOPs and energy remain nearly unchanged, suggesting that the gated subband experts achieve superior feature representation under the same computational budget. In the Lightweight mode, classic MLP increases SOPs and energy by 52.3% and 51.8%, respectively, indicating that FCP trades a controllable accuracy cost for substantial reductions in computation and energy consumption, which is critical for resource-constrained scenarios.

Finally, removing the global compensation gate causes a 0.26% accuracy drop in the Parallel mode (79.44% vs. 79.70%) with virtually unchanged SOPs and energy, demonstrating that the gate effectively selects key features without incurring additional computational overhead. In the Lightweight mode, removing the gate leads to a 0.55% accuracy drop (77.33% vs. 77.88%) along with slight increases in SOPs and energy, further validating the necessity of the compensation gate for maintaining lightweight-mode performance.

5.4 Energy Efficiency Analysis

Figure 3 shows representative SOPs dynamics on CIFAR-100 (Block 0 as a representative example). SWFormer's FATM and MLP SOPs increase monotonically with epoch, indicating continuous spike-rate inflation. After introducing HSWL, FATM SOPs drop to 35–50% of SWFormer, suppressing inflation, and this benefit cascades to MLP layers, making Parallel DPMR's MLP SOPs lowest and most stable. Lightweight DPMR reduces FCP SOPs by 47% relative to Parallel mode in Block 0 and remains lower throughout training. These dynamics demonstrate that frequency-aware structured sparsity reduces computational cost and stabilizes training while suppressing redundant high-frequency spikes.

Table 3: Core ablation results on CIFAR-100 (accuracy and efficiency).

Mode	**Variant**	T	Top-1 (%)	SOPs (G)	Energy (mJ)
Parallel	base	4	79.70±0.18	0.574	0.522
	w/o HSWL	4	79.84±0.14	0.704	0.639
	w/ Classic MLP	4	79.30±0.21	0.571	0.519
	w/o CompGate	4	79.44±0.20	0.573	0.522
Lightweight	base	4	77.88±0.24	0.375	0.342
	w/o HSWL	4	77.51±0.18	0.463	0.422
	w/ Classic MLP	4	79.30±0.21	0.571	0.519
	w/o CompGate	4	77.33±0.30	0.387	0.353

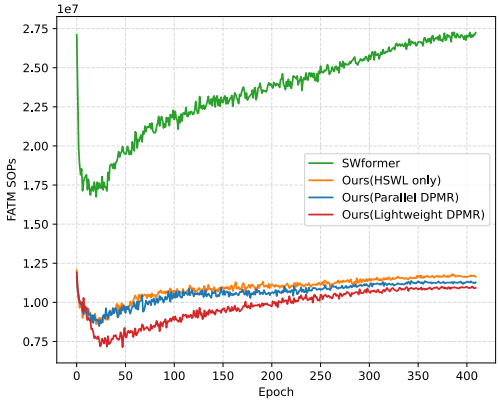


Figure 3: Representative FATM/HSWL SOPs dynamics on CIFAR-100 (Block 0).

6 Conclusion

This paper proposes SFAT, which achieves frequency-aware structured sparsity via HSWL and flexible accuracy–efficiency trade-offs via DPMR. HSWL combines hierarchical Haar decomposition with coefficient-wise soft thresholding and channel-wise energy gating; DPMR compensates high-frequency loss through subband experts and global gating, with Parallel mode for high precision and Lightweight mode for edge deployment. On CIFAR-100, Tiny-ImageNet, and neuromorphic datasets, SFAT(Parallel) attains strong accuracy while reducing energy by about 60%, and Lightweight attains comparable accuracy with nearly half the parameters. Future work will explore adaptive frequency thresholds and three-dimensional spatio-temporal Haar transforms to mitigate domain mismatch on neuromorphic data, as well as hardware co-design to further exploit structured sparsity on neuromorphic chips.

AI USE STATEMENT

Generative AI tools were used only for language polishing and LaTeX formatting assistance during manuscript preparation. No generative AI tool was used to design the method, run experiments, produce numerical results, or generate figures. All AI-assisted text was manually checked by the authors for factual accuracy, citation correctness, and potential plagiarism. The authors take full responsibility for the final content of this work, including text, claims, and artifacts.

ETHICS STATEMENT

This work studies efficient spiking neural network compression using public image and neuromorphic benchmarks. It does not involve human subjects, personally identifiable data, or new dataset releases. The main societal impact is improved energy efficiency of visual inference, which is generally beneficial; potential risks are limited to general misuse of efficient recognition models. No additional ethical concerns are identified.

REPRODUCIBILITY STATEMENT

The model architecture and training objectives are described in Sections 4 and 5, with datasets and preprocessing in Appendix C.1, complete hyperparameters in Appendix C.2, and extended results in Appendices C.3–C.6. Anonymized source code and configuration files are included as a single self-contained supplementary zip file; no external anonymous repository link is used. All CIFAR-100 results are averaged over three seeds, and retrained baselines follow the same setting described in Appendix C.2.

REFERENCES

Arnon Amir, Brian Taba, David J Berg, Timothy Melano, Jeffrey L McKinstry, Carmelo Di Nolfo, Tapan K 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 (CVPR)*, pp. 7243–7252, 2017.

Anthony N. Burkitt. A review of the integrate-and-fire neuron model: I. homogeneous synaptic input. *Biological Cybernetics*, 95(1):1–19, 2006.

Kaiwei Che, Zhaokun Zhou, Jun Niu, Zhengyu Ma, Wei Fang, Yanqi Chen, Shuaijie Shen, Li Yuan, and Yonghong Tian. Auto-spikformer: Spikformer architecture search. *Frontiers in Neuroscience*, 18:1372257, 2024.

Hailong Chen, Qingqing Song, and Guantong Chen. A lightweight multi-frequency feature fusion network with efficient attention for breast tumor classification in pathology images. *Information*, 16(7):579, 2025.

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

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

Chaoteng Duan, Jianhao Ding, Shiyan Chen, Zhaofei Yu, and Tiejun Huang. Temporal effective batch normalization in spiking neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 34377–34390, 2022.

Yuetong Fang, Zihan Wang, Lei Zhang, Jianhao Cao, Haotian Chen, and Rui Xu. Spiking wavelet transformer. In *European Conference on Computer Vision*, pp. 19–37, 2024.

Yufei Guo, Li Zhang, Yansen Chen, Xiaode Tong, Xiang Liu, Yuhang Wang, Xiang Huang, and Zhe Ma. Real spike: Learning real-valued spikes for spiking neural networks. In *European Conference on Computer Vision*, pp. 52–68, Cham, 2022. Springer. doi: 10.1007/978-3-031-19775-8_4.

Pranav Jeevan, Akella Srinidhi, Pasunuri Prathiba, and Amit Sethi. Wavemixsr: Resource-efficient neural network for image super-resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024*, pp. 5872–5880, 2024.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Hongmin Li, Hanchao Liu, Xiangyang Ji, Guoqi Li, and Luping Shi. Cifar10-dvs: an event-stream dataset for object classification. *Frontiers in Neuroscience*, 11:309, 2017.

Yuhang Li, Youngeun Kim, Hyoungseob Park, Tobias Geller, and Priyadarshini Panda. Neuromorphic data augmentation for training spiking neural networks. In *European Conference on Computer Vision*, pp. 631–649, 2022.

Yue Liu, Shanlin Xiao, Bo Li, and Zhiyi Yu. Sparsespikformer: A co-design framework for token and weight pruning in spiking transformer. In *Proceedings of the 49th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2024*, pp. 6410–6414, 2024.

Margaret Livingstone and David Hubel. Segregation of form, color, movement, and depth: Anatomy, physiology, and perception. *Science*, 240:740–749, 1988.

Andre Luvizotto, Cesar Renno-Costa, and Paul F. M. J. Verschure. A wavelet-based neural model to optimize and read out a temporal population code. *Frontiers in Computational Neuroscience*, 6:21, 2012.

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

Qingyan Meng, Mingqing Xiao, Shen Yan, Yisen Wang, Zhouchen Lin, and Zhi-Quan Luo. Training high-performance low-latency spiking neural networks by differentiation on spike representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12444–12453, 2022.

Paul A. Merolla, John V. Arthur, Rodrigo Alvarez-Icaza, Andrew S. Cassidy, Jun Sawada, Filipp Akopyan, Bryan L. Jackson, Nabil Imam, Chen Guo, Yutaka Nakamura, et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science*, 345 (6197):668–673, 2014.

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

Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 980–993, 2021.

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

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.

Sophie Skriabine, Maxwell Shinn, Samuel Picard, Kenneth D. Harris, and Matteo Carandini. Mapping the visual cortex with zebra noise and wavelets. *Journal of Vision*, 26(1):1–16, 2026.

Rolf Skyberg, Seiji Tanabe, Hui Chen, and Jianhua Cang. Coarse-to-fine processing drives the efficient coding of natural scenes in mouse visual cortex. *Cell Reports*, 38(13):110606, 2022.

Richard B. Stein. A theoretical analysis of neuronal variability. *Biophysical Journal*, 5(2):173–194, 1965.

Roger B. H. Tootell and Shahin Nasr. Columnar segregation of magnocellular and parvocellular streams in human extrastriate cortex. *Journal of Neuroscience*, 37(33):8014–8032, 2017.

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

Guoping Xu, Wei Liao, Xiang Zhang, Chenxi Li, Xing He, and Xiang Wu. Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation. *Pattern Recognition*, 143:109819, 2023.

Man Yao, Jiakui Hu, Zhaokun Zhou, Li Yuan, Yonghong Tian, Bo Xu, and Guoqi Li. Spike-driven transformer. In *Advances in Neural Information Processing Systems*, volume 36, pp. 64043–64058, 2023.

Hanle Zheng, Yujie Wu, Lei Deng, Yifan Hu, and Guoqi Li. Going deeper with directly-trained larger spiking neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11062–11070, 2021.

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 q-k attention. In *Advances in Neural Information Processing Systems*, volume 37, pp. 13074–13098, 2024.

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*, 2023.

A Extended Related Work

The parallel processing architecture of the primate visual system provides rich biological constraints for computational model design. The M/P pathway separation begins at the retinal level: magnocellular (M) cells have large receptive fields, transient responses, and high conduction velocity, being sensitive to low spatial frequency and high temporal frequency; parvocellular (P) cells have small receptive fields, sustained responses, and low conduction velocity, being sensitive to high spatial frequency and color contrast Livingstone & Hubel (1988). Tootell and Nasr's fMRI study demonstrated in the human brain that M/P information is segregated and processed through specific columnar structures in the extrastriate cortex (V2, V4), forming parallel functional channels Tootell & Nasr (2017). At the computational theory level, coarse-to-fine processing has been shown to be a key mechanism for efficient encoding of natural scenes: neuronal responses in mouse V1 exhibit a rapid dynamic shift in spatial frequency preference from low to high frequencies, counteracting the $1/f^2$ power spectrum statistics of natural images and reducing neural encoding redundancy Skyberg et al. (2022). This suggests that progressive frequency processing should be adaptive and task-driven, rather than a fixed-resolution pyramid structure.

Wavelet transform provides a mathematical implementation framework for the above biological mechanisms. Unlike the global basis functions of the Fourier transform, wavelet transform achieves time-frequency localization through multi-resolution analysis, and the spatial scales of its basis functions correspond to neuronal receptive fields. Prior work systematically compared Gabor wavelets, Haar wavelets, and V1 neuronal receptive fields, finding that wavelet basis functions can effectively capture spatial frequency tuning and location selectivity Skriabine et al. (2026). At the neural coding level, the orthogonality and binary coefficient properties of the Haar wavelet make it suitable for the discrete activation patterns of spiking neural networks, forming a natural match with binary/ternary spike mechanisms Luvizotto et al. (2012).

Wavelet applications in ANNs have proven engineering value. WTConv embeds 2D discrete wavelet transform into convolutional layers, achieving an equivalent large receptive field while parameters grow only linearly Chen et al. (2025). WaveMixSR uses the Haar wavelet as a basic unit of lightweight visual architectures for image super-resolution Jeevan et al. (2024). Haar wavelet downsampling replaces traditional pooling and better preserves high-frequency detail information Xu et al. (2023). In Transformer architectures, GFNet uses fast Fourier transform to replace self-attention for global perception, but its complex-valued operations and dense activations are unsuitable for spike-driven constraints Rao et al. (2021). These successes are limited to ANN frameworks; their floating-point operations, dense matrix multiplications, and continuous activations fail to achieve deep synergy with the sparsity and locality of wavelet coefficients.

Introducing Transformer architectures into SNNs is an important recent advance. SpikFormer first implemented spike-driven self-attention through spike-based Query-Key inner products Zhou et al. (2023). QKFormer introduced hierarchical spatial sampling to improve efficiency Zhou et al. (2024); SpikingResformer combined ResNet multi-scale features with Transformer global modeling Shi et al. (2024); Spike-driven Transformer achieved pure spike-driven residual learning Yao et al. (2023). SWFormer first attempted to introduce wavelets into SNN Transformers, replacing self-attention with a single 2D Haar wavelet transform and performing frequency-domain feature transformation through block-diagonal weight matrices Fang et al. (2024). This validated feasibility but did not optimize sparsity, and its traditional MLP design conflicts with the sparsification objective.

For compression, SparseSpikFormer proposed joint token and weight pruning through learnable masks Liu et al. (2024). Auto-SpikFormer adopts one-shot NAS with evolutionary algorithms to explore architectures on the energy-accuracy Pareto frontier Che et al. (2024). These methods are designed for standard SpikFormer-style architectures, and pruning is performed on feature maps that have already lost high-frequency information. This fundamental limitation calls for completing selective retention and structural compression within the frequency domain, which is the goal of SFAT.

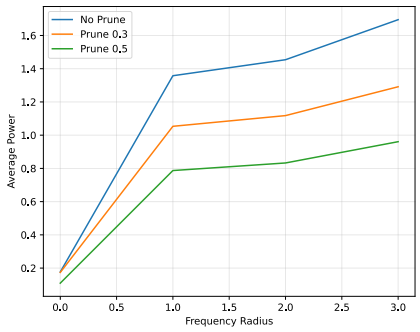


Figure 4: Radial frequency power distribution under different channel-pruning ratios.

B Preliminary Experiments: Configurations and Results

B.1 Subband-level Frequency–Energy–Information Correlation

To verify that the LL/HL/LH/HH energy distribution after vertical decomposition is uneven and that high-energy subbands contain more information, this experiment selects the Haar forward output of Block 2 in a converged SWFormer as the object, computes the energy (mean squared value) of each subband, and measures the accuracy loss through a spatial-quadrant zero-out ablation experiment as the information content of that subband. The averaged scatter statistics are visualized in Figure 1a. The result shows a positive correlation between wavelet-domain energy and information content: zeroing high-energy subbands causes larger accuracy degradation, while zeroing low-energy subbands has minimal impact.

B.2 Original-image Granularity Consistency Validation

To prove that the frequency-domain distribution, energy distribution, and information distribution are consistent at the original-image granularity, this experiment outputs four views for CIFAR-100 images while maintaining the original 32×32 resolution. The first is the original image; the second is the frequency distribution map, obtained by performing Haar transform on the original image and inverse-transform reconstructing each subband separately (other subbands zeroed), comparing LL reconstruction magnitude with HL+LH+HH reconstruction magnitude at each pixel location, and overlaying low-/high-frequency attributes as semi-transparent tints on the original image; the third is the energy distribution map, directly computing pixel-level squared energy on the original grayscale image and applying Gaussian smoothing; the fourth is the information distribution map, i.e., the gradient saliency map of the correct-class logit with respect to input pixels. Figure 1b summarizes the consistency trend used to motivate frequency-aware sparsity.

B.3 High-frequency Semantic Loss from Channel Pruning

To support the necessity of DPMR compensation, we compare feature maps of Block 2 under three settings: no pruning, zeroing 30% of low-energy channels, and zeroing 50% of low-energy channels. The 2D Fourier power spectrum (log scale) and its radially averaged power distribution curve are visualized to observe the trend of high-frequency component decay as the pruning ratio increases. As shown in Figure 4, stronger low-energy channel pruning attenuates high radial frequencies, motivating a lightweight compensation path rather than a uniformly enlarged MLP.

C Supplementary Experimental Materials

C.1 Dataset Details and Preprocessing

CIFAR-100. 50,000 training images and 10,000 test images, 100 classes in total. Preprocessing: normalized to mean $[0.5071, 0.4867, 0.4408]$ and standard deviation $[0.2675, 0.2565, 0.2761]$; Ran-

dAugment (magnitude=9, $n = 1$), Mixup ($\alpha = 0.5$), and Random Erasing (probability 0.25) are applied during training.

Tiny-ImageNet. 200 classes, 100,000 training images (500 per class), and 10,000 validation images. Image resolution is 64×64; we directly input at this resolution with patch size 4, yielding $16 \times 16 = 256$ tokens. Data preprocessing is consistent with CIFAR-100, with additional color jittering and scale transformations.

DVS-Gesture. A gesture recognition dataset collected with the IBM DVS128 camera, 11 classes in total, containing 1,172 training samples and 288 test samples. Event stream time step $T = 16$, spatial resolution 128×128. Integral event frame representation is adopted: the event stream is uniformly divided into 16 intervals along time, and events in each interval are accumulated to generate a grayscale frame as network input. Neuromorphic Data Augmentation (NDA) Li et al. (2022) is applied during training, including temporal jittering and spatial transformations.

CIFAR10-DVS. An event-stream dataset converted from CIFAR-10 through an event camera, containing 10 classes, 9,000 training samples and 1,000 test samples. Event stream time step $T = 16$, spatial resolution 128×128. Integral event frame representation and NDA are used in the same way as DVS-Gesture.

C.2 Complete Training Hyperparameters

Table 4 lists the unified hyperparameter configuration across all datasets. HSWL-specific hyperparameters: Level-1 learnable threshold initialization $\tau_1^L = 0.01$ (low-frequency branch), $\tau_1^H = 0.05$ (high-frequency branch); Level-2 energy temperature coefficient $\tau_E = 1.0$, energy center $E_{\text{center}} = C/2$, gating range $[0.4, 0.6]$, sampling standard deviation $\sigma = 0.2$; subband-aware spike thresholds $V_{\text{th}}^{LL} = 0.9$, $V_{\text{th}}^{LH} = 1.0$, $V_{\text{th}}^{HL} = 1.0$, $V_{\text{th}}^{HH} = 1.2$; Level-1 thresholds $V_{\text{th}}^{L} = 0.9$, $V_{\text{th}}^{H} = 1.1$. DPMR-specific hyperparameters: frequency expert ratio $r = 0.25$, FCP spike threshold $v_{\text{th}}^{\text{FCP}} = 0.2$, compensation strength scalar β initialized to 0.0 and constrained to $(0, 1)$ through Sigmoid.

Table 4: Complete training hyperparameter configuration.

Hyperparameter	**CIFAR-100**	**Tiny-ImageNet**	**CIFAR10-DVS**	**DVS-Gesture**
Optimizer	AdamW	AdamW	AdamW	AdamW
Base LR	0.0015	0.0015	0.0015	0.0015
Weight Decay	0.06	0.06	0.06	0.06
LR Schedule	Cosine	Cosine	Cosine	Cosine
Warmup Epochs	20	20	20	20
Total Epochs	400	200	100	100
Batch Size	128	128	16	16
Patch Size	4	4	4	4
Time Steps T	4	4	16	16
Mixup α	0.5	0.5	0.5	0.5
Label Smoothing	0.1	0.1	0.1	0.1
RandAugment	m9-n1	m9-n1	–	–

C.3 Detailed Tiny-ImageNet Results

Table 5 gives the complete results for all tested architectures on Tiny-ImageNet. No Spiking Transformer results on Tiny-ImageNet were publicly reported prior to this work; all baselines were retrained from official code under identical settings.

C.4 Neuromorphic Dataset Results and Analysis

Table 6 presents the comparison results on neuromorphic datasets. SFAT's performance is slightly lower than Spikformer and SDformer. On CIFAR10-DVS, Ours (Parallel, 2-256) reaches 78.7%, which is 2.2 percentage points lower than Spikformer (80.9%); on DVS-Gesture, Ours (Parallel, 2-256) reaches 97.9%, which is 1.4 percentage points lower than SDformer (99.3%).

Table 5: Complete comparison results on Tiny-ImageNet.

Method	T-4-256			T-2-384			T-4-384		
	Top-1 (%)	SOPs (G)	Energy (mJ)	Top-1 (%)	SOPs (G)	Energy (mJ)	Top-1 (%)	SOPs (G)	Energy (mJ)
Spikformer∗	64.46	2.84	2.57	63.27	5.46	4.93	65.79	5.98	5.40
SDformer∗	63.82	1.12	1.02	65.07	1.53	1.40	65.63	2.15	1.96
SWFormer∗	64.46	1.61	1.47	63.21	1.84	1.68	64.74	3.22	2.92
Ours (Parallel)∗	64.69	1.01	0.92	63.58	1.23	1.13	**66.44**	2.02	1.84
Ours (Lightweight)∗	60.66	0.61	0.56	61.60	0.78	0.72	62.68	1.18	1.08

This gap is consistent with the design prior of SFAT. The core assumption—low-frequency subbands carry more semantic information and high-frequency subbands have higher redundancy—originates from pre-experimental observations on static image datasets. Based on this assumption, HSWL implements more aggressive suppression of high-frequency components: in Level-1, the learnable threshold of the high-frequency branch τ_1^H is higher than the low-frequency threshold τ_1^L; in Level-2, the channel retention rate of HH is much lower than that of LL. This frequency-selective sparsity effectively eliminates redundant textures and noise on static images, but faces structural mismatch on neuromorphic data, whose essence is a sparse encoding of brightness changes with rich high-frequency temporal dynamics. Although FCP is designed to recover high-frequency details, its experts are trained under the energy distribution prior of static images, and global average pooling may dilute key local dynamic signals in event data.

Promising directions include adaptive frequency threshold mechanisms, temporal-frequency joint sparsity via 3D Haar transforms, event-density-aware compensation gating, and mixed static/neuromorphic training with domain-adaptive losses. These directions aim to preserve the energy-efficiency advantage on low-frequency-dominant static vision while relaxing high-frequency suppression for event-driven domains.

Table 6: Comparison results on neuromorphic datasets.

	CIFAR10-DVS			DVS-Gesture		
Method	Arch.	T	Top-1 (%)	Arch.	T	Top-1 (%)
Spikformer	2-256	10	78.9	2-256	10	96.9
	2-256	16	80.9	2-256	16	98.3
SDformer	2-256	16	80.0	2-256	16	99.3
Ours (Parallel)	2-256	16	78.7	2-256	16	97.9
	2-512	16	79.8	2-512	16	97.6
Ours (Lightweight)	2-256	16	78.8	2-256	16	97.6
	2-512	16	79.6	2-512	16	97.2

C.5 Dynamic Visualization of HSWL Sparsity Mechanisms

Figures 5 and 6 show the dynamic changes of HSWL two-level sparsity during training (averaged over four layers). The sparsity of the high-frequency branch is consistently higher than that of the low-frequency branch; in Parallel mode, high-frequency converges from 0.82 to 0.71 and low-frequency from 0.54 to 0.37, while in Lightweight mode they converge to 0.76 and 0.46, respectively. The high-frequency sparsity is about 1.6–1.8$\times$ that of the low-frequency branch, indicating that the initialization strategy $\tau_1^L < \tau_1^H$ is maintained and reinforced during training. The channel retention rates of the four subbands follow LL $>$ LH $>$ HL $\approx$ HH, consistent with the energy-information analysis in the pre-experiments.

C.6 Layer-wise SOPs Dynamics Analysis

Figures 7, 8, and 9 show the per-layer SOPs dynamics during CIFAR-100 training for Blocks 0–3. SWFormer's FATM SOPs are significantly higher than other methods in all layers and increase monotonically with epoch, indicating continuous spike-rate inflation. After introducing HSWL, SOPs immediately drop to 40–50% of SWFormer and remain stable. The MLP curves reveal a

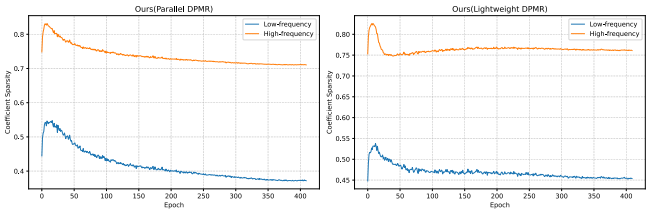


Figure 5: Level-1 coefficient-level sparsity dynamics.

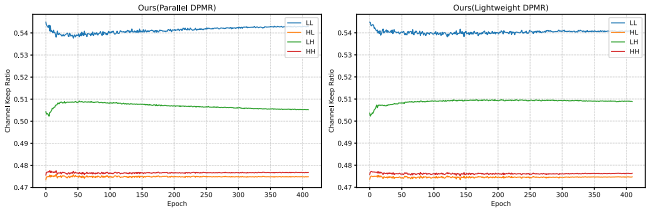


Figure 6: Level-2 channel retention-rate dynamics.

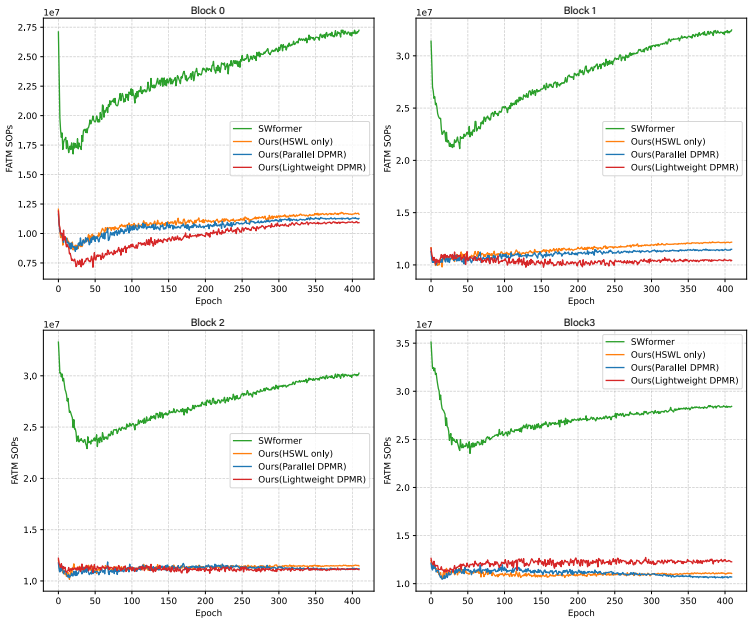


Figure 7: Layer-wise SOPs dynamics of FATM / HSWL on CIFAR-100.

cascaded sparsity effect: HSWL reduces the number of spikes input to the MLP, and DPMR further lowers MLP load by sharing feature transformation through the sparse FCP pathway. Lightweight DPMR has lower FCP SOPs than Parallel DPMR in all layers, because FCP becomes the main feature transformation pathway and learns a more aggressive sparsity strategy.

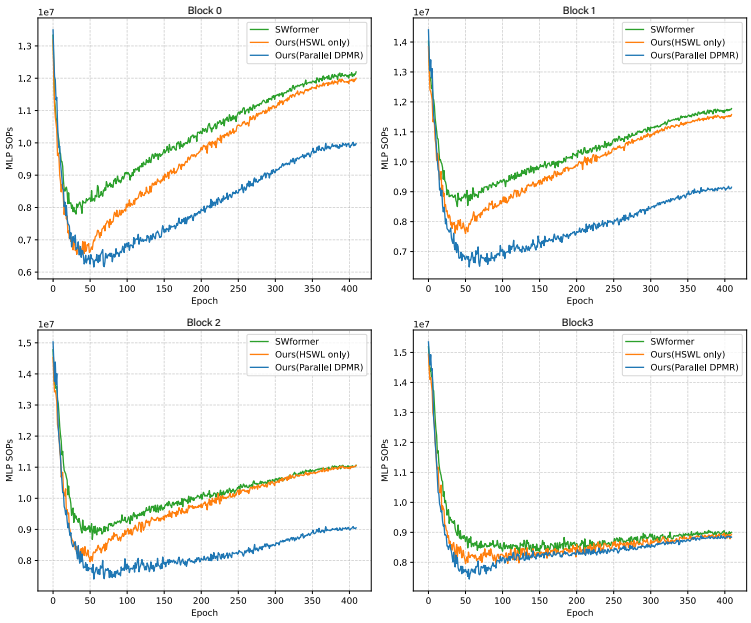


Figure 8: Layer-wise SOPs dynamics of MLP on CIFAR-100.

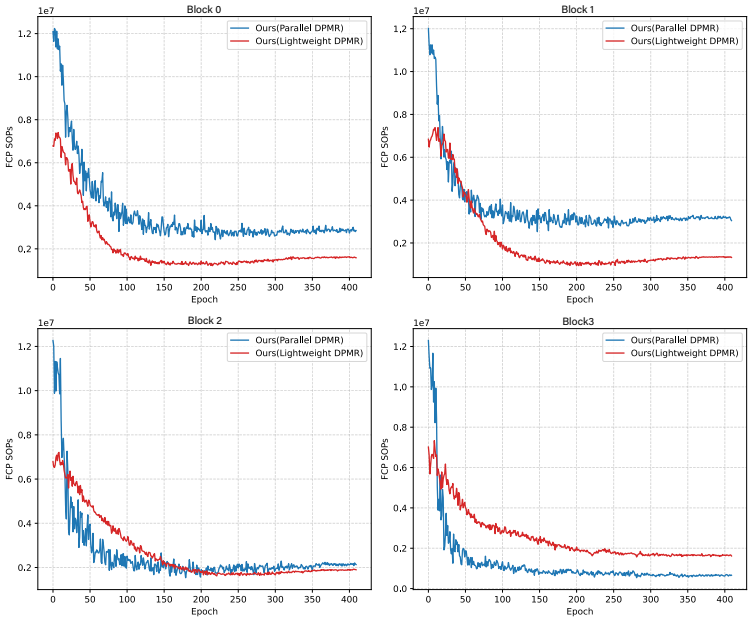


Figure 9: Layer-wise SOPs dynamics of FCP on CIFAR-100.