ShiftLIF: Efficient Multi-Level Spiking Neurons with Power-of-Two Quantization

Kaiwen Tang*
National University of Singapore

Di Yu*
Zhejiang University

Jiaqi Zheng
Sea AI Lab

Changze Lv
Fudan University

Qianhui Liu
Shandong University

Zhanglu Yan†
National University of Singapore

Weng-Fai Wong
National University of Singapore

Abstract

Spiking neural networks (SNNs) are promising for edge sensing due to their event-driven computation and temporal filtering capability. However, standard leaky integrate-and-fire (LIF) neurons communicate only through binary spikes, which severely limit representational capacity. Existing multi-level spiking neurons improve information transmission, but often rely on uniform quantization that mismatches membrane-potential distributions or introduces costly synaptic multiplications. In this paper, we propose *ShiftLIF*, a multi-level spiking neuron that maps membrane potentials to a logarithmically spaced power-of-two spike set. This design provides finer representation in the small-amplitude regime, where membrane potentials are densely concentrated, while enabling multiplier-free synaptic computation through bit-shift and accumulation operations. As a result, ShiftLIF improves spike-level expressiveness without sacrificing the hardware-friendly nature of standard SNN computation. We evaluate ShiftLIF on 10 datasets spanning wireless, acoustic, motion, and visual sensing tasks. Results show that ShiftLIF consistently matches or exceeds the accuracy of existing multi-level spiking neurons while maintaining synaptic energy consumption close to standard binary LIF. These results indicate that ShiftLIF provides a favorable accuracy-efficiency trade-off for cross-modal edge sensing.

CCS Concepts

• **Computing methodologies** → **Machine learning**; **Neural networks**.

Keywords

Spiking Neural Networks, Edge Computing, Multimodal Sensing, Energy-efficient Computing

1 Introduction

The rapid expansion of edge computing has driven an increasing demand for sensing applications across wireless, acoustic, and motion domains [20, 34, 44]. However, processing these continuous streams on edge devices poses a challenge to energy consumption [29]. *Spiking neural networks* (SNN) have emerged as a highly promising paradigm to address this [39, 51, 53]. The most widely adopted mechanism in these networks is the *leaky integrate-and-fire* (LIF) neuron [41]. Beyond providing event-driven energy efficiency, the leaky integration dynamics intrinsically act as a temporal low-pass filter [4, 11]. This enables the neuron to smooth out high-frequency environmental noise and effectively capture slowly varying temporal patterns in real-world physical signals.

Despite this advantage, standard LIF neurons suffer from a severe representational bottleneck [47]. Fundamentally, spiking neurons compute rich analog membrane potentials but encode them with only one bit. Because neurons communicate strictly through binary spikes, this extreme quantization discards much of the amplitude information preserved by temporal integration [16, 27]. As a result, the expressive capacity of the network is significantly constrained, limiting its ability to capture the full complexity of continuous sensory signals.

Recent efforts to address this limitation generally follow two directions. The first line of work enhances neuron expressiveness by modifying membrane dynamics, such as PLIF [9], GLIF [50], CLIF [19], and RPLIF [26], which introduce learnable time constants or additional state variables. While these approaches enrich temporal modeling, they increase computational overhead and still constrain neuron outputs to binary spikes, leaving the quantization bottleneck unresolved. The second line of work increases representational bandwidth by introducing multi-level spikes, such as LM-HT SNN [17] and INT-LIF [27]. However, these methods typically rely on uniformly spaced spike levels, which introduce two limitations: multi-bit spike amplitudes require synaptic multiplications, increasing hardware cost, and uniform quantization poorly matches the intrinsic distribution of membrane potentials, which are often concentrated near small values with a long tail toward larger ones. Consequently, uniform grids allocate excessive resolution to rare large activations while under-representing densely clustered small activations.

To address these challenges, we propose ShiftLIF, a novel spiking neuron that both improves representational capacity and hardware efficiency. ShiftLIF maps the continuous membrane potential to a logarithmically spaced set of spikes defined by powers of two, $\{0, 2^{-K}, 2^{-(K-1)}, \ldots, 2^{-1}, 1\}$. This design provides two key advantages. Algorithmically, the power-of-two grid is naturally denser near zero, which better matches the heavy-tailed distribution of membrane potentials and captures fine-grained variations in small activations that uniform quantization often misses [14, 31]. From a hardware perspective, because spike amplitudes are restricted

*Both authors contributed equally to this research.
†Corresponding author.

to powers of two and synaptic weights are represented as integers, spike–weight interactions can be implemented using simple bit-shift operations followed by integer accumulation [24, 40]. Consequently, ShiftLIF enables fully multiplier-free synaptic computation, preserving the core energy efficiency advantage of spiking neural networks. Moreover, since the bit-shift operation effectively reduces the bit-width of the shifted weights, the subsequent accumulation involves smaller operands, which can further reduce switching activity and energy consumption compared to standard accumulation [5, 36].

To evaluate ShiftLIF, we conduct extensive experiments across 11 diverse datasets encompassing wireless, acoustic, motion, and visual sensing modalities. Empirical results demonstrate that ShiftLIF consistently achieves competitive or superior classification accuracy compared to state-of-the-art multi-level spiking neurons. Furthermore, it accomplishes this while consuming lower synaptic energy comparable to that of standard binary LIF models.

The main contributions of this paper are:

(1) We identify the fundamental accuracy–efficiency dilemma in existing multi-level SNNs, where uniform quantization increases computational overhead and poorly matches the natural distribution of membrane potentials.
(2) We propose ShiftLIF, a novel spiking neuron that generates multi-level spikes constrained to powers of two, enabling distribution-aware representation while preserving multiplier-free synaptic computation.
(3) We establish a comprehensive cross-modal benchmark across wireless, acoustic, motion, and visual sensing tasks, demonstrating that ShiftLIF achieves state-of-the-art accuracy while significantly improving energy efficiency over other LIF variants.

2 Related Work

2.1 SNNs for Edge Sensing Data

Spiking Neural Networks have been increasingly explored for edge sensing applications, including acoustic recognition, wearable motion analysis, event-based vision, and wireless sensing, due to their event-driven sparsity and native temporal processing capability [3, 6, 25, 32]. Existing sensing-oriented SNN studies have primarily emphasized architectural adaptation, temporal feature extraction, encoding strategies, and surrogate gradient for direct training [13, 28, 30]. In comparison, the spike representation mechanism itself, especially the design of the neuron output alphabet beyond binary communication, remains relatively underexplored in SNNs for sensing applications design.

2.2 LIF Neurons and Their Variants

To overcome the representational limitations of standard LIF dynamics, numerous variants have been proposed, broadly falling into two categories. The first category enriches internal membrane dynamics while strictly retaining binary spike communication. For instance, PLIF [9] introduces learnable membrane time constants, GLIF [50] incorporates complex gating mechanisms, and CLIF [19] adds complementary paths for better gradient flow. While these variants significantly improve temporal modeling and optimization behavior, they ultimately compress the rich, continuous membrane state into a 1-bit binary spike, leaving the inter-layer communication bottleneck unresolved.

The second category directly expands the spike alphabet to mitigate this 1-bit limitation. Representative works include multi-threshold or ternary neurons, such as Ternary Spike [12], MT-SNN [45], I-LIF [38], and INT-LIF [27]. While these multi-level neurons increase the information carried per spike, they typically rely on uniformly spaced quantization grids or generic multi-bit integer amplitudes. This introduces two critical flaws: first, as demonstrated by recent studies [14, 15], uniform discretization severely mismatches the zero-concentrated, heavy-tailed distribution of membrane potentials, leading to suboptimal level utilization and high relative quantization error [24]. Second, utilizing generic multi-bit spikes inevitably reintroduces dense Multiply-Accumulate (MAC) operations at the synapses, fundamentally destroying the hardware-efficiency premise of SNNs.

ShiftLIF structurally advances this second category. By introducing a logarithmic, power-of-two quantization grid, it simultaneously achieves distribution-aware information maximization and strictly MAC-free, shift-based synaptic computation.

3 Method

3.1 Preliminary: Standard Leaky IF

The leaky integrate-and-fire (LIF) neuron is the most widely adopted computational unit in spiking neural networks [30, 42]. At each discrete time step t, the neuron integrates the incoming synaptic current into its membrane potential while gradually decaying previous states. Incorporating a soft-reset mechanism, the membrane dynamics can be formulated as

$$u_t = \lambda u_{t-1} + x_t - s_{t-1} V_{\text{th}}, \tag{1}$$

where u_t denotes the membrane potential, $\lambda \in [0, 1)$ is the leak factor, x_t is the input current, V_{th} is the firing threshold, and s_{t-1} represents the spike emitted at the previous time step. When the accumulated membrane potential exceeds the threshold, the neuron generates an output spike according to a Heaviside step function:

$$s_t = \begin{cases} 1, & \text{if } u_t \geq V_{\text{th}}, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

From a signal processing perspective, the leakage term λ endows the LIF neuron with the behavior of a first-order infinite impulse response low-pass filter [11]. It enables the neuron to temporally integrate recent inputs, effectively suppressing high-frequency environmental noise while preserving slowly varying signal components that are important for real-world sensory perception.

The binary firing mechanism also makes synaptic computation highly efficient, since the interactions between spikes and weights are reduced to simple accumulation without multiplications. However, the LIF neuron suffers from a representational bottleneck because its continuous membrane state is communicated only through binary spikes $s_t \in \{0, 1\}$ [12, 14]. This extreme quantization discards fine-grained amplitude information accumulated through temporal integration, thereby limiting the expressive capacity of the neuron, especially in continuous sensing scenarios where subtle signal variations are often informative.

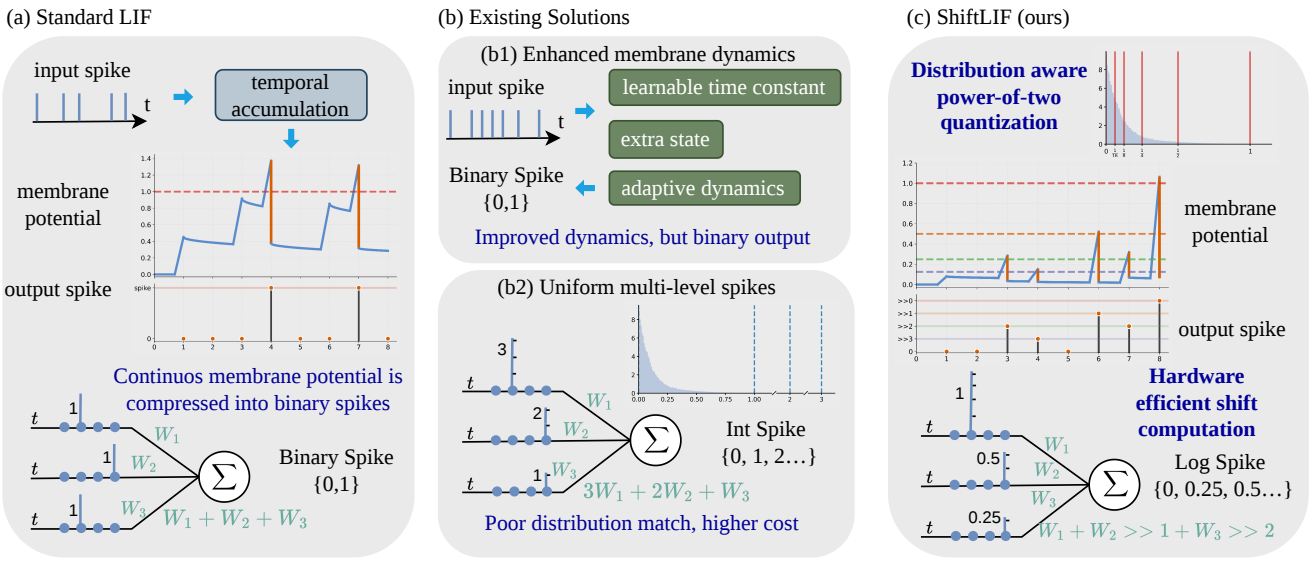


Figure 1: Comparison of standard LIF, existing solutions, and ShiftLIF. In (c), the colored dashed lines mark the thresholds of different power-of-two output levels, and the top inset shows a small-value-dominant membrane distribution with quantization points denser near zero. At the bottom, $W \gg k$ denotes a bitwise right shift by k bits, highlighting that ShiftLIF replaces multiplication with shift-and-accumulate computation.

To overcome this limitation, the next section introduces ShiftLIF, which replaces the binary spike output of conventional LIF neurons with a power-of-two multi-level spike representation while preserving hardware-friendly computation.

3.2 ShiftLIF

The design principle of ShiftLIF is to **break the binarized representation bottleneck while preserving the hardware efficiency of standard LIF neurons**. To this end, ShiftLIF retains the membrane integration dynamics of the conventional LIF neuron and modifies only the spike generation and reset processes.

3.2.1 Shift-Quantized Activation. The central challenge in reformulating spike generation is to increase spike-level representational bandwidth without sacrificing hardware efficiency. Existing multi-level spiking neurons typically rely on uniformly spaced quantization levels. While such designs improve amplitude representation, they generally require synaptic multiplications and therefore do not naturally preserve the multiplier-free nature of standard LIF computation. In addition, uniform quantization allocates equal resolution across the full dynamic range, which is not well matched to the non-uniform distribution of membrane potentials that typically cluster near small values [14, 15, 24, 31].

To address both issues, ShiftLIF replaces the standard Heaviside step function with a shift-quantization operator Q_{shift}, which maps the continuous membrane potential to a discrete power-of-two spike set. This design naturally produces a logarithmically spaced output space, providing finer resolution for small activations while preserving shift-friendly computation. Without loss of generality, we normalize the firing threshold to $V_{\text{th}} = 1.0$. The quantization operator is defined as

$$S_t = Q_{\text{shift}}(V_t), \quad Q_{\text{shift}}(v) = \begin{cases} 0, & \text{if } v < 2^{-(K+1)}, \\ 2^{-\lceil -\log_2 v \rceil}, & \text{if } 2^{-(K+1)} \le v \le 1, \end{cases} \tag{3}$$

where $v = \max(0, \min(V_t, 1))$ denotes the bounded membrane potential, and $K \in \mathbb{N}$ controls the smallest non-zero spike magnitude. The corresponding output alphabet is

$$\mathcal{S} = \{0,\ 2^{-K},\ 2^{-(K-1)},\ \ldots,\ 2^{-1},\ 1\}, \qquad |\mathcal{S}| = K + 2. \tag{4}$$

For example, setting $K = 7$ yields 9 discrete spike levels spanning multiple orders of magnitude, enabling substantially richer amplitude encoding than standard binary spikes.

3.2.2 Logarithmic Quantization. The quantization grid of Q_{shift} is logarithmically spaced as the bin corresponding to the output 2^{-k}, defined on the interval $[2^{-(k+1)}, 2^{-k})$, has width $2^{-(k+1)}$. This width shrinks exponentially as k increases, i.e., as activation values approach zero. In contrast, the uniform grid used in INT-LIF assigns a constant bin width $\Delta = 1/N$ across the full dynamic range.

This non-uniform spacing is motivated by the membrane-potential distribution in SNNs. Due to the continuous leakage dynamics, membrane potentials are typically concentrated near small values rather than being uniformly distributed over the full range. Prior studies on activation quantization and recent analyses of deep SNNs have shown that such non-uniform distributions are closely related to quantization error [14, 15, 24, 31]. As further illustrated by our empirical measurements across multiple sensing modalities

in Figure 2, the membrane potentials in trained models consistently cluster near zero.

For such non-uniform distributions, uniformly spaced quantization levels allocate equal resolution to both dense and sparse regions, which can under-represent the small-amplitude regime where membrane potentials occur most frequently. By contrast, the logarithmic power-of-two grid in ShiftLIF provides finer resolution near zero while maintaining coarser spacing for rare large values. This makes the quantization grid better aligned with the empirical membrane-potential distribution and naturally supports shift-based computation. Section 3.3 further analyzes the resulting representational properties.

After spike generation, the membrane potential must be reset. To retain the residual sub-threshold charge that is not encoded into the current spike, ShiftLIF employs a proportional soft-reset mechanism:

$$V_t \leftarrow V_t - S_t \cdot V_{\text{th}}.$$

Compared with the standard LIF soft reset, which subtracts a fixed threshold after each spike, ShiftLIF decrements the membrane potential proportionally to the emitted multi-level spike magnitude $S_t \in \mathcal{S}$. This preserves residual membrane charge across time steps and reduces information loss caused by hard state truncation.

The complete forward dynamics of a ShiftLIF neuron at a single time step are summarized in Algorithm 1.

Algorithm 1 ShiftLIF Forward Pass

Input: Current input X_t at timestep t, membrane potential V_{t-1}
Hyperparameters: Time constant τ, threshold $V_{\text{th}} = 1.0$, resting potential $V_{\text{reset}} = 0.0$, precision factor K

1. Leaky Integration
$V_t \leftarrow V_{t-1} + \frac{1}{\tau}(V_{\text{reset}} - V_{t-1}) + X_t$
2. Bounding (Clamp)
$v \leftarrow \max(0, \min(V_t, V_{\text{th}}))$
3. Shift-Quantized Activation
if $v < 2^{-(K+1)}$ **then**
  $S_t \leftarrow 0$
else
  $k \leftarrow \lceil -\log_2 v \rceil$
  $S_t \leftarrow 2^{-\min(k,K)}$
end if
4. Proportional Soft Reset
$V_t \leftarrow V_t - S_t \cdot V_{\text{th}}$
Return S_t, V_t

For continuous sequence processing, this forward pass is iteratively applied over $t = 1, \ldots, T$, updating the membrane state and accumulating the output spike train $\{S_t\}_{t=1}^{T}$.

3.3 Theoretical Analysis

We next compare ShiftLIF with INT-LIF [27], a representative multi-level spiking baseline based on uniformly spaced integer levels. Standard binary LIF is recovered as the special case $K = 0$ of INT-LIF, so the comparison with LIF is already covered by this formulation and is not discussed separately here. Let X denote a nonnegative random variable representing the membrane potential normalized by the firing threshold, and let v denote a generic realization of X. We use X in distribution-level expressions such as probabilities and expectations, and v in pointwise definitions of the quantizers. Let K be the precision factor introduced in Section 3.2.1. To match the same output budget, we compare the two level sets

$$\begin{aligned} \mathcal{S}_{\text{shift}} &= \{0, 2^{-K}, 2^{-(K-1)}, \ldots, 1\}, \\ \mathcal{S}_{\text{int}} &= \{0, 1, 2, \ldots, K+1\}, \end{aligned} \tag{5}$$

which both contain $K + 2$ discrete outputs. INT-LIF, and hence standard LIF as its $K = 0$ special case, follows the nearest-level rule on a uniform integer grid. ShiftLIF instead selects the largest admissible dyadic level not exceeding the membrane value; this preserves power-of-two spike magnitudes and can be implemented efficiently with bit-shift operations, making it more hardware-friendly. Accordingly, the ShiftLIF quantizer can be interpreted as

$$Q_{\text{shift}}(v) = \max\{s \in \mathcal{S}_{\text{shift}} : s \le v\}. \tag{6}$$

The INT-LIF quantizer is defined by

$$Q_{\text{int}}(v) = \begin{cases} 0, & 0 \le v < \frac{1}{2}, \\ j, & j - \frac{1}{2} \le v < j + \frac{1}{2}, \quad j = 1, \ldots, K, \\ K+1, & v \ge K + \frac{1}{2}. \end{cases} \tag{7}$$

3.3.1 Quantization Error Analysis. Assuming $\mathbb{E}[X] < \infty$, we first analyze the expected absolute quantization error

$$\mathcal{E}_A = \mathbb{E}[|X - Q_A(X)|], \qquad A \in \{\text{shift}, \text{int}\}. \tag{8}$$

For the sensing applications considered in this paper, the membrane distributions are not naturally uniform on the original linear scale. Instead, as illustrated by the empirical measurements in Figure 2, they are concentrated near zero while remaining spread across multiple logarithmic scales. In this regime, the dyadic grid of ShiftLIF is a more natural match to the underlying distribution, because it allocates finer resolution to the low-value region where most membrane mass is concentrated while still preserving multiple spike levels across scales. The lemma below gives a simple sufficient condition under which this allocation yields a smaller expected absolute error than INT-LIF.

Lemma 3.1. *The expected absolute quantization error $\mathcal{E}_{shift}$ is strictly less than $\mathcal{E}_{int}$ if*

$$2^{-K} \Pr(2^{-K} \le X < \tfrac{1}{2}) > \tfrac{1}{2} \Pr(\tfrac{3}{4} \le X < 1) + K \Pr(X \ge \tfrac{3}{2}). \tag{9}$$

Proof. Define

$$\Delta(v) = |v - Q_{\text{int}}(v)| - (v - Q_{\text{shift}}(v)), \quad v \ge 0. \tag{10}$$

Since $Q_{\text{shift}}(v) \le v$ for all $v \ge 0$, we have

$$|v - Q_{\text{shift}}(v)| = v - Q_{\text{shift}}(v). \tag{11}$$

A direct interval-by-interval comparison of the two quantizers yields the pointwise lower bound

$$\Delta(v) \ge 2^{-K} \mathbf{1}_{\{2^{-K} \le v < \frac{1}{2}\}} - \tfrac{1}{2} \mathbf{1}_{\{\frac{3}{4} \le v < 1\}} - K \mathbf{1}_{\{v \ge \frac{3}{2}\}}, \quad v \ge 0. \tag{12}$$

Indeed, on $[2^{-K}, \frac{1}{2})$, one has $\Delta(v) \ge 2^{-K}$. On $[\frac{3}{4}, 1)$, one has $\Delta(v) \ge -\frac{1}{2}$. On $[\frac{3}{2}, \infty)$, all cases satisfy $\Delta(v) \ge -K$. On the remaining

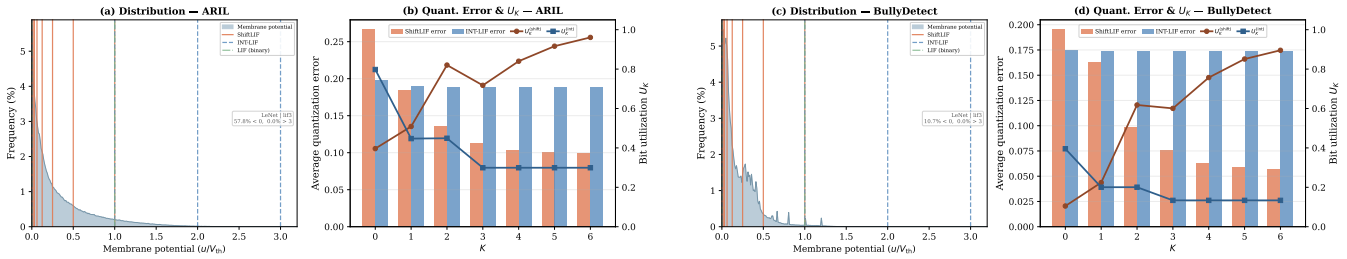


Figure 2: Membrane potential distribution, quantization error, and bit utilization metric on sensing datasets. Compared with INT-LIF, ShiftLIF yields lower quantization error and better bit utilization level as K increases.

region, the right-hand side is nonpositive while the inequality is immediate. Taking expectations yields

$$\begin{aligned}\mathbb{E}[\Delta(X)] \geq 2^{-K}\Pr(2^{-K} \leq X < \tfrac{1}{2}) \\ - \tfrac{1}{2}\Pr(\tfrac{3}{4} \leq X < 1) - K\Pr(X \geq \tfrac{3}{2}).\end{aligned} \tag{13}$$

Since

$$\mathcal{E}_{\text{shift}} - \mathcal{E}_{\text{int}} = -\mathbb{E}[\Delta(X)], \tag{14}$$

the stated condition implies $\mathbb{E}[\Delta(X)] > 0$, which is equivalent to $\mathcal{E}_{\text{shift}} < \mathcal{E}_{\text{int}}$. □

A direct comparison of the two quantizers shows that ShiftLIF is favored on $[2^{-K}, \frac{3}{4})$, the two quantizers are tied on $[0, 2^{-K}) \cup [1, \frac{3}{2})$, and INT-LIF is favored on $[\frac{3}{4}, 1) \cup [\frac{3}{2}, \infty)$. Consequently, the distribution of probability mass within the unit interval matters, not just the aggregate mass $\Pr(2^{-K} \leq X < 1)$. The advantage in quantization error of ShiftLIF on real sensing datasets is evident in Figure 2.

3.3.2 Information-Theoretic Capacity. We now analyze how effectively the two quantizers use the same $K + 2$ output levels. We measure this through the normalized output entropy relative to the bit budget required to encode these levels:

$$U_K^{(A)} = \frac{H(Q_A(X))}{\lceil \log_2(K+2) \rceil}, \qquad A \in \{\text{shift}, \text{int}\}. \tag{15}$$

Here, $\lceil \log_2(K + 2) \rceil$ is the number of bits needed to encode the $K + 2$ quantization levels. Thus, $U_K^{(A)}$ measures how effectively those bits are utilized: larger values, especially those closer to 1, indicate better use of the available coding capacity and are therefore preferred. For a discrete random variable Y with probability mass function p_Y, we define

$$H(Y) = -\sum_y p_Y(y) \log_2 p_Y(y), \tag{16}$$

with the convention $0 \log_2 0 := 0$. For a finite probability vector such as R, V, or T, $H(R), H(V)$, and $H(T)$ are defined by the same formula. For the entropy comparison, the key split occurs at $\frac{1}{2}$: ShiftLIF preserves variation in the low-amplitude region by distributing the mass below $\frac{1}{2}$ across dyadic shells, which can increase output entropy when membrane values are concentrated near zero. By contrast, INT-LIF collapses all mass below $\frac{1}{2}$ to the single output 0, and only distributes mass across uniform cells above $\frac{1}{2}$. The next lemma makes this entropy decomposition explicit.

Lemma 3.2. *Let $r = \Pr(X < \frac{1}{2})$. Let R, V, and T denote the conditional output distributions of ShiftLIF below $\frac{1}{2}$, of ShiftLIF on $[\frac{1}{2}, \infty)$, and of rounded INT-LIF on $[\frac{1}{2}, \infty)$, respectively. Then*

$$\begin{aligned}H(Q_{shift}(X)) &= h_2(r) + rH(R) + (1-r)H(V), \\ H(Q_{int}(X)) &= h_2(r) + (1-r)H(T),\end{aligned} \tag{17}$$

where $h_2(r) = -r\log_2 r - (1-r)\log_2(1-r)$ is the binary entropy. Consequently,

$$U_K^{(shift)} > U_K^{(int)} \iff rH(R) + (1-r)H(V) > (1-r)H(T). \tag{18}$$

The same formulas remain valid in the degenerate cases $r = 0$ and $r = 1$, with the absent conditional-entropy term interpreted as zero.

The proof is provided in the supplementary materials. This criterion shows that ShiftLIF is favored when a substantial fraction of the mass lies below $\frac{1}{2}$ and is well spread across the dyadic shells there. INT-LIF is only favored when the mass above $\frac{1}{2}$ is broadly spread across the nearest-integer cells. In Figure 2, we see that ShiftLIF behaves well in terms of bit utilization in sensing datasets.

3.4 Training Procedure

Surrogate Gradients. The shift-quantization operator Q_{shift} is piecewise constant and therefore has zero derivative almost everywhere, which prevents direct backpropagation. To enable end-to-end training, we adopt a straight-through estimator (STE) that is matched to the support of the ShiftLIF quantizer.

In the forward pass, the spike output is generated by

$$s_t = Q_{\text{shift}}(u_t). \tag{19}$$

Since Q_{shift} is defined on the bounded interval $[0, 1]$, we approximate its derivative in the backward pass by

$$\frac{\partial s_t}{\partial u_t} \approx \mathbf{1}_{[0 \leq u_t \leq 1]}. \tag{20}$$

Accordingly, the gradient of the loss is propagated as

$$\frac{\partial \mathcal{L}}{\partial u_t} \approx \frac{\partial \mathcal{L}}{\partial s_t} \cdot \mathbf{1}_{[0 \leq u_t \leq 1]}. \tag{21}$$

That is, gradients are passed through unchanged within the quantization support and set to zero outside it.

This choice is consistent with the STE framework widely used in quantized and spiking neural networks, while being specifically adapted to the multi-level support of ShiftLIF. In contrast to binary surrogate functions that primarily provide a gradient around a single firing threshold, the proposed STE supplies a non-zero gradient

over the full activation range of the quantizer, allowing all active quantization regions to participate in optimization.

Spike rate regularization. To further control the firing activity during training, we add a spike rate regularizer to the task loss:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{sr}}\mathcal{L}_{\text{sr}}, \tag{22}$$

where

$$\mathcal{L}_{\text{sr}} = \frac{1}{L}\sum_{\ell=1}^{L}\left[\text{mean}\Big(|\boldsymbol{S}^{(\ell)}|\Big) - r^*\right]^+. \tag{23}$$

This term penalizes only layers whose average spike magnitude exceeds the target r^*, thereby reducing excessive firing activity without suppressing already sparse responses. We use an L1-style penalty so that the regularization signal remains stable near the target rate.

3.5 Hardware Computation

A key motivation for adopting a power-of-two spike alphabet is to avoid the general multiplication overhead introduced by conventional multi-level spiking neurons. In standard synaptic computation, the pre-synaptic input to neuron i at time step t is given by

$$x_i^{(t)} = \sum_j W_{ij} s_j^{(t)}, \tag{24}$$

where W_{ij} denotes the synaptic weight and $s_j^{(t)}$ is the pre-synaptic spike. For conventional multi-level SNNs, $s_j^{(t)}$ typically takes arbitrary integer or floating-point values, so evaluating $W_{ij}s_j^{(t)}$ generally requires standard multiplication.

In ShiftLIF, all non-zero spikes are constrained to the form

$$s_j^{(t)} = 2^{-k}, \qquad k \in \{0, 1, \ldots, K\}. \tag{25}$$

Under integer or fixed-point weight representations, the corresponding synaptic product can be implemented as a right-shift operation,

$$W_{ij}s_j^{(t)} = W_{ij}\cdot 2^{-k} \quad\longrightarrow\quad W_{ij} \gg k, \tag{26}$$

where $\gg$ denotes arithmetic right shift. When $s_j^{(t)} = 0$, the synaptic update is skipped entirely, preserving the event-driven sparsity of spiking computation.

As a result, ShiftLIF replaces MAC(multiply-accumulate) operations in multi-level synaptic interactions with shift-based accumulation. This preserves the hardware efficiency of standard binary SNN computation while enabling richer spike representations. In addition, the reduced effective bit width of shifted operands can make the subsequent accumulation more efficient in practical integer implementations. Compared with conventional multi-level spike designs, ShiftLIF therefore provides a more favorable trade-off between representational capacity and hardware efficiency.

4 Experiment

4.1 Setup

4.1.1 Dataset. To evaluate ShiftLIF, we benchmark across 10 datasets with four distinct sensing modalities: (1) **Wireless Sensing:** ARIL [43], UT-HAR [49], Fi-HumanID [49], and BullyDetect [21] for WiFi-based human activity and identification tasks; (2) **Acoustic Sensing:** UrbanSound8K [33] (10-fold cross-validation) and Google Speech Commands v2 (GSC, 35-class) [46]; (3) **Motion Sensing:** UCI-HAR [2] and HHAR [37] for IMU-based activity recognition; and (4) **Neuromorphic Vision:** CIFAR10-DVS [23] and DVS128-Gesture [1]. These datasets include continuous signals that require initial encoding as well as data from native event sensors. This variety allows us to evaluate the proposed neuronal dynamics under different spatial and temporal distributions. Continuous motion signals are segmented using a sliding window with a size of 128 and a stride of 64, while acoustic waveforms are transformed into 128-dimensional Mel-spectrograms.

4.1.2 Besealine Neurons. We compare ShiftLIF against nine spiking neuron models: the standard binary LIF [30], PLIF [9], GLIF [50], CLIF [19], PSN [10], TLIF [12], I-LIF [38], RPLIF [26], and INT-LIF [27]. Among these, INT-LIF is the most direct comparator as it also provides multi-level spike outputs, while the remaining baselines emit binary $\{0, 1\}$ or ternary $\{-1, 0, 1\}$ spikes.

4.1.3 Implementation. The comparison of different neurons is done with an identical spiking LeNet [22] backbone. Models are trained from scratch for 150 epochs using the Adam optimizer with the learning rate of 10^{-3} and a cosine annealing schedule. The temporal simulation window is fixed to $T = 4$ time steps across all tasks. The variants of spike neurons are implemented with spikingjelly [7].

For ShiftLIF, we set the precision factor to $K = 2$, defining a 2-level spike alphabet $\mathcal{S} = \{0, 2^{-2}, 2^{-1}, 1\}$. To ensure a perfectly fair comparison, the fundamental dynamic parameters are consistently aligned across all evaluated neuron models: the firing threshold is $V_{\text{th}} = 1.0$, the reset potential is $V_{\text{reset}} = 0.0$, and the membrane time constant is $\tau = 2.0$. All experiments are executed on a single NVIDIA GPU with a fixed random seed.

4.2 Comparison Among Spiking Neurons

Table 1 compares ShiftLIF with standard LIF, dynamic enhanced variants, and existing multi-level spiking neurons on ten sensing benchmarks. ShiftLIF achieves the best average accuracy, reaching 89.34%, surpassing the 88.74% and 88.63% of the strongest competing baselines, CLIF and I-LIF, and improving substantially over 82.00% of the standard LIF.

Its advantage is most evident on continuous sensing tasks. Across the eight benchmarks in wireless, acoustic, and motion sensing, ShiftLIF ranks first on ARIL, Fi-Humanid, GSC, UCIHAR, and HHAR, and ties for first on UT-HAR. It also achieves the second-best result on BullyDetect and UrbanSound8K. On neuromorphic sensing benchmarks, it is less dominant but remains competitive, reaching 63.65% on CIFAR10-DVS and 92.36% on DVS-Gesture. Overall, these results show that ShiftLIF is particularly effective on continuous sensing data while maintaining competitive performance on the other event-based tasks.

4.3 Comparison Across Backbones

We further compare ShiftLIF and INT-LIF across a diverse set of backbones, including CNN, ResNet, and Transformer-style architectures [8, 18, 35, 52, 54]. As shown in Table 2, ShiftLIF outperforms INT-LIF on the majority of backbone–dataset pairs. The improvement is especially consistent on ARIL, UT-HAR, and BullyDetect,

Table 1: Comparison of different spiking neuron models across low-filtering wireless sensing, acoustic sensing, motion sensing, and neuromorphic sensing benchmarks. Results are reported as accuracy (%). Best and second-best results are highlighted in bold and underlined, respectively.

Method	Low-filtering Wireless Sensing				Acoustic Sensing		Motion Sensing		Neuromorphic Sensing		Avg.
	ARIL	UT-HAR	Fi-Humanid	BullyDetect	UrbanSound8K	GSC	UCIHAR	HHAR	CIFAR10-DVS	DVS-Gesture	
CLIF	<u>92.09</u>	<u>99.20</u>	<u>99.45</u>	81.38	**83.06**	89.54	<u>93.86</u>	<u>91.77</u>	<u>63.65</u>	<u>93.40</u>	<u>88.74</u>
GLIF	89.93	98.80	98.17	78.26	79.01	90.31	92.74	91.28	63.15	<u>93.40</u>	87.50
INTLIF	89.57	**99.40**	98.17	79.69	79.67	90.51	93.28	91.68	62.15	**95.14**	87.93
LIF	91.01	97.09	99.08	67.28	79.37	84.90	83.71	81.02	46.65	89.93	82.00
PLIF	91.73	98.90	98.72	79.91	82.76	90.39	93.25	91.71	**64.35**	93.06	88.48
PSN	85.61	96.89	99.27	63.39	76.29	89.31	92.91	90.50	56.65	91.32	84.21
TLIF	88.13	96.39	<u>99.45</u>	15.63	78.52	52.39	91.52	69.80	22.85	84.03	69.87
ILIF	91.37	**99.40**	99.08	**82.37**	81.97	<u>90.70</u>	93.69	91.36	**64.35**	92.01	88.63
RPLIF	87.77	97.39	99.27	67.68	80.94	85.59	91.25	81.71	45.05	88.19	82.48
ShiftLIF	**92.81**	**99.40**	**99.63**	<u>82.14</u>	<u>83.00</u>	**91.31**	**94.30**	**94.75**	<u>63.65</u>	92.36	**89.34**

Table 2: Comparison between ShiftLIF and INT-LIF across different backbones on wireless sensing benchmarks. For each dataset, we report the accuracy (%) of ShiftLIF, the accuracy (%) of INT-LIF, and their difference Δ = ShiftLIF − INT-LIF. The better result between the two neurons is highlighted in bold.

Backbone	ARIL			UT-HAR			Fi-Humanid			BullyDetect		
	ShiftLIF	INT-LIF	Δ	ShiftLIF	INT-LIF	Δ	ShiftLIF	INT-LIF	Δ	ShiftLIF	INT-LIF	Δ
SpikingLeNet	**92.01**	89.57	+2.44	**99.40**	**99.40**	0.00	**99.63**	98.17	+1.46	**82.14**	79.69	+2.45
SpikingVGG9	**97.84**	94.60	+3.24	**99.90**	99.60	+0.30	**99.63**	99.45	+0.18	**80.36**	30.62	+49.74
SEWResNet34	89.21	**89.57**	-0.36	**99.80**	99.30	+0.50	**99.45**	**99.45**	0.00	**79.55**	58.53	+21.02
MSResNet34	**93.17**	88.85	+4.32	**99.60**	99.40	+0.20	**99.45**	99.08	+0.37	**80.13**	77.86	+2.27
Spikformer256	**94.96**	89.57	+5.39	**99.70**	98.80	+0.90	99.63	**99.82**	-0.19	**85.94**	55.54	+30.40
MetaSpikformer256	**94.24**	92.45	+1.79	**99.80**	99.70	+0.10	**99.63**	99.08	+0.55	**77.59**	67.99	+9.60
QKformer256	**96.04**	89.21	+6.83	**99.90**	99.10	+0.80	99.63	**99.82**	-0.19	**74.82**	59.51	+15.31

where ShiftLIF is better for nearly all available backbones. On Fi-Humanid, the two neurons are close, but ShiftLIF still achieves equal or higher accuracy in most cases. Only a small number of cases show no gain or a slight decrease, such as SEWResNet34 on ARIL and QKformer256 on Fi-Humanid.

Overall, these results indicate that the advantage of ShiftLIF is not tied to a particular architecture. Instead, it transfers across different backbone families, suggesting that the proposed spike representation is broadly compatible with existing SNN designs.

4.4 Energy Analysis

The computational cost of an SNN depends on both the temporal window length T and the average spike rate s. A larger T repeats synaptic processing over more timesteps, while a larger s increases the number of effective events that trigger computation and data movement. The energy of a spiking layer is modeled as $E_{\text{total}} = T \cdot s \cdot \left(E_{\text{ACC}} + E_{\text{move}} + E_{\text{weight}}\right)$, where E_{ACC} denotes the accumulation energy, E_{move} denotes the spike movement energy, and E_{weight} denotes the SRAM weight access energy. All reported energy values are derived from measurements on a commercial 22 nm process [48].

Since the temporal window is fixed to $T = 4$ for all tasks, the energy differences in Table 4 are driven mainly by spike activity and synaptic efficiency. Training with spike rate regularization, ShiftLIF consistently achieves a stronger accuracy–energy trade-off than INT-LIF on the four wireless sensing benchmarks. At the same time, ShiftLIF delivers better accuracy on three of the four datasets. Compared with standard LIF, ShiftLIF also improves accuracy on three datasets while maintaining similar or lower energy in most cases.

4.5 Ablation Study

Effect of precision factor. We first study the effect of the precision factor K, which determines the smallest non-zero spike magnitude 2^{-K} and thus the granularity of the spike alphabet. As shown in Figure 3, the best performance is achieved at moderate values of K rather than increasing monotonically with finer quantization. Specifically, the optimum appears at $K = 2$ on ARIL and $K = 3$ on BullyDetect, while the coarsest setting $K = 1$ gives the worst result on both datasets. This indicates that the improvement of ShiftLIF does not come from simply increasing the number of spike levels. When K becomes larger, the additional bits bring limited benefit.

Table 3: Ablation study on quantization strategy. We compare the proposed logarithmic power-of-two quantization (ShiftLIF) with a uniform multi-level quantization under the same number of spike levels. Results are reported as accuracy (%). The better result in each column is highlighted in bold.

Method	Wireless Sensing				Acoustic Sensing		Motion Sensing		Neuromorphic Sensing		Avg.
	ARIL	UT-HAR	Fi-HumanID	BullyDetect	UrbanSound8K	GSC	UCIHAR	HHAR	CIFAR10-DVS	DVS-Gesture	
ShiftLIF	**92.81**	99.40	**99.63**	**82.14**	**83.00**	91.31	94.30	**94.75**	63.65	92.36	**89.34**
Uniform	89.21	**99.50**	99.27	81.25	82.21	**91.63**	**94.37**	93.51	**64.40**	**95.14**	89.05
Δ (ShiftLIF − Uniform)	+3.60	-0.10	+0.36	+0.89	+0.79	-0.32	-0.07	+1.24	-0.75	-2.78	+0.29

Table 4: Spike Rate and Accuracy Comparison Across Datasets and Neuron Types. ShiftLIF* indicates training with spike rate regularization.

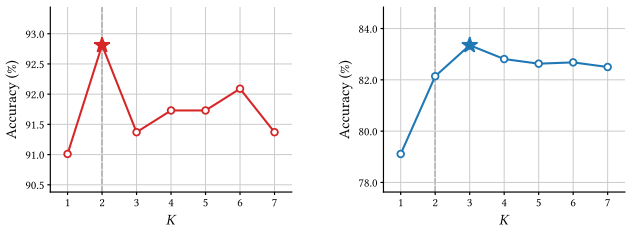
Dataset	Neuron	Acc (%)	Spike Rate	Energy(mJ)
UTHar	ShiftLIF*	98.49	0.030	3.32
	LIF	97.09	0.040	4.27
	INTLIF	99.40	0.046	5.44
FiHumanID	ShiftLIF*	98.35	0.021	2.51
	LIF	99.08	0.109	10.58
	INTLIF	98.17	0.203	25.71
BullyDetect	ShiftLIF*	82.54	0.038	4.04
	LIF	67.28	0.027	3.08
	INTLIF	79.69	0.080	7.97
ARIL	ShiftLIF*	92.45	0.049	5.03
	LIF	91.01	0.047	4.91
	INTLIF	89.57	0.151	15.19

Based on this observation, we use $K = 2$ as the default setting in the main experiments.

Logarithmic vs. uniform quantization. We next compare the proposed logarithmic power-of-two quantization with a uniform multi-level quantization under the same number of spike levels. Table 3 shows that ShiftLIF achieves better overall performance, with clearer gains on continuous sensing tasks such as ARIL, Fi-HumanID, BullyDetect, UrbanSound8K, and HHAR. This result is consistent with our design motivation: membrane potentials in trained SNNs are typically concentrated in the low-amplitude regime, where the logarithmic grid provides finer resolution than a uniform one. The advantage is less consistent on event-based vision benchmarks, suggesting that the benefit of quantization depends on the underlying signal statistics. Overall, these results support the central design of ShiftLIF: the gain comes not only from using multi-level spikes, but from using a quantization scheme that is better matched to SNN membrane dynamics.

5 Discussion

The results show a clear pattern that ShiftLIF is most effective on continuous sensing tasks, including wireless, acoustic, and motion data, while its advantage is smaller on event-based vision benchmarks. A likely reason is that continuous signals contain useful amplitude variations over time, and these variations are not well



(a) ARIL **(b) BullyDetect**

Figure 3: Ablation on the precision factor K of ShiftLIF on ARIL and BullyDetect. Moderate K values achieve the best accuracy, while increasing K further brings no consistent gain.

preserved by binary spikes. By using a multi-level spike set with finer resolution near small values, ShiftLIF can retain more of this information during neuron-to-neuron communication.

This observation is also consistent with the membrane-potential distributions analyzed in our study. Since a large fraction of membrane states lie in the low-amplitude regime, a logarithmic quantization grid is better matched to the underlying signal statistics than a uniform grid. At the same time, the power-of-two constraint keeps synaptic computation efficient, since spike-weight interactions can still be implemented through shift and accumulation rather than general multiplication. The main benefit of ShiftLIF therefore, comes from combining richer spike representation with a lightweight computation path.

6 Conclusion

In this paper, we propose ShiftLIF, a simple multi-level spiking neuron that improves both accuracy and hardware efficiency. By using a logarithmic power-of-two spike set, ShiftLIF better matches the membrane potential distribution and preserves more useful information, especially for small activations. Experiments on ten datasets across four edge-sensing modalities show that ShiftLIF consistently achieves a better accuracy–energy trade-off than existing neuron designs. The gains are especially clear on continuous sensing tasks, where fine signal changes matter. We also show that logarithmic quantization works better than uniform quantization, and that a moderate precision factor K is enough to reach strong performance. Overall, ShiftLIF provides a simple and effective way to improve SNNs.

References

[1] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, et al. 2017. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7243–7252.

[2] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, Jorge Luis Reyes-Ortiz, et al. 2013. A public domain dataset for human activity recognition using smartphones.. In *Esann*, Vol. 3. 3–4.

[3] Suwhan Baek and Jaewon Lee. 2024. Snn and sound: a comprehensive review of spiking neural networks in sound. *Biomedical Engineering Letters* 14, 5 (2024), 981–991.

[4] William M Connelly, Michael Laing, Adam C Errington, and Vincenzo Crunelli. 2016. The thalamus as a low pass filter: filtering at the cellular level does not equate with filtering at the network level. *Frontiers in neural circuits* 9 (2016), 89.

[5] Samuel Coward, Theo Drane, Emiliano Morini, and George A Constantinides. 2024. Combining power and arithmetic optimization via datapath rewriting. In *2024 IEEE 31st Symposium on Computer Arithmetic (ARITH)*. IEEE, 24–31.

[6] Shuiguang Deng, Di Yu, Changze Lv, Xin Du, Linshan Jiang, Xiaofan Zhao, Wentao Tong, Xiaoqing Zheng, Weijia Fang, Peng Zhao, et al. 2025. Edge intelligence with spiking neural networks. *arXiv preprint arXiv:2507.14069* (2025).

[7] Wei Fang, Yanqi Chen, Jianhao Ding, Zhaofei Yu, Timothée Masquelier, Ding Chen, Liwei Huang, Huihui Zhou, Guoqi Li, and Yonghong Tian. 2023. Spikingjelly: An open-source machine learning infrastructure platform for spike-based intelligence. *Science Advances* 9, 40 (2023), eadi1480.

[8] Wei Fang, Zhaofei Yu, Yanqi Chen, Tiejun Huang, Timothée Masquelier, and Yonghong Tian. 2021. Deep residual learning in spiking neural networks. *Advances in neural information processing systems* 34 (2021), 21056–21069.

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

[10] Wei Fang, Zhaofei Yu, Zhaokun Zhou, Ding Chen, Yanqi Chen, Zhengyu Ma, Timothée Masquelier, and Yonghong Tian. 2023. Parallel spiking neurons with high efficiency and ability to learn long-term dependencies. *Advances in Neural Information Processing Systems* 36 (2023), 53674–53687.

[11] Yuetong Fang, Deming Zhou, Ziqing Wang, Hongwei Ren, ZeCui Zeng, Lusong Li, Renjing Xu, et al. [n. d.]. Spiking Neural Networks Need High-Frequency Information. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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

[13] Yufei Guo, Xuhui Huang, and Zhe Ma. 2023. Direct learning-based deep spiking neural networks: a review. *Frontiers in Neuroscience* 17 (2023), 1209795.

[14] Yufei Guo, Xiaode Liu, Yuanpei Chen, Liwen Zhang, Weihang Peng, Yuhan Zhang, Xuhui Huang, and Zhe Ma. 2023. Rmp-loss: Regularizing membrane potential distribution for spiking neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 17391–17401.

[15] Yufei Guo, Xinyi Tong, Yuanpei Chen, Liwen Zhang, Xiaode Liu, Zhe Ma, and Xuhui Huang. 2022. Recdis-snn: Rectifying membrane potential distribution for directly training spiking neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 326–335.

[16] Bing Han, Gopalakrishnan Srinivasan, and Kaushik Roy. 2020. Rmp-snn: Residual membrane potential neuron for enabling deeper high-accuracy and low-latency spiking neural network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13558–13567.

[17] Zecheng Hao, Xinyu Shi, Yujia Liu, Zhaofei Yu, and Tiejun Huang. 2024. Lm-ht snn: Enhancing the performance of snn to ann counterpart through learnable multi-hierarchical threshold model. *Advances in Neural Information Processing Systems* 37 (2024), 101905–101927.

[18] Yifan Hu, Lei Deng, Yujie Wu, Man Yao, and Guoqi Li. 2024. Advancing spiking neural networks toward deep residual learning. *IEEE transactions on neural networks and learning systems* 36, 2 (2024), 2353–2367.

[19] Yulong Huang, Xiaopeng Lin, Hongwei Ren, Haotian Fu, Yue Zhou, Zunchang Liu, Biao Pan, and Bojun Cheng. 2024. Clif: Complementary leaky integrate-and-fire neuron for spiking neural networks. *arXiv preprint arXiv:2402.04663* (2024).

[20] Meng-Huang Lai and Kang-Shuo Chang. 2023. Ai sensor applications in edge computing. *IEEE Nanotechnology Magazine* 17, 6 (2023), 23–28.

[21] Bo Lan, Fei Wang, Lekun Xia, Fan Nai, Shiqiang Nie, Han Ding, and Jinsong Han. 2024. Bullydetect: Detecting school physical bullying with wi-fi and deep wavelet transformer. *IEEE Internet of Things Journal* 12, 5 (2024), 5160–5169.

[22] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 2002. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (2002), 2278–2324.

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

[24] Yuhang Li, Xin Dong, and Wei Wang. 2019. Additive powers-of-two quantization: An efficient non-uniform discretization for neural networks. *arXiv preprint arXiv:1909.13144* (2019).

[25] Yuhang Li, Ruokai Yin, Youngeun Kim, and Priyadarshini Panda. 2023. Efficient human activity recognition with spatio-temporal spiking neural networks. *Frontiers in Neuroscience* 17 (2023), 1233037.

[26] Yang Li, Xinyi Zeng, Zhe Xue, Pinxian Zeng, Zikai Zhang, and Yan Wang. 2025. Incorporating the Refractory Period into Spiking Neural Networks through Spike-Triggered Threshold Dynamics. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 10876–10885.

[27] Xinhao Luo, Man Yao, Yuhong Chou, Bo Xu, and Guoqi Li. 2024. Integer-valued training and spike-driven inference spiking neural network for high-performance and energy-efficient object detection. In *European Conference on Computer Vision*. Springer, 253–272.

[28] Kai Malcolm and Josue Casco-Rodriguez. 2023. A comprehensive review of spiking neural networks: Interpretation, optimization, efficiency, and best practices. *arXiv preprint arXiv:2303.10780* (2023).

[29] Mattia Merluzzi, Nicola Di Pietro, Paolo Di Lorenzo, Emilio Calvanese Strinati, and Sergio Barbarossa. 2021. Discontinuous computation offloading for energy-efficient mobile edge computing. *IEEE Transactions on Green Communications and Networking* 6, 2 (2021), 1242–1257.

[30] Emre O Neftci, Hesham Mostafa, and Friedemann Zenke. 2019. Surrogate gradient learning in spiking neural networks. *IEEE Signal Processing Magazine* 36, 6 (2019), 51–63.

[31] Dominika Przewlocka-Rus, Syed Shakib Sarwar, H Ekin Sumbul, Yuecheng Li, and Barbara De Salvo. 2022. Power-of-two quantization for low bitwidth and hardware compliant neural networks. *arXiv preprint arXiv:2203.05025* (2022).

[32] Hemanth Sabbella, Archit Mukherjee, Thivya Kandappu, Sounak Dey, Arpan Pal, Archan Misra, and Dong Ma. 2025. The Promise of Spiking Neural Networks for Ubiquitous Computing: A Survey and New Perspectives. *arXiv preprint arXiv:2506.01737* (2025).

[33] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. 2014. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*. 1041–1044.

[34] Weisong Shi, Jie Cao, Quan Zhang, Youhuizi Li, and Lanyu Xu. 2016. Edge computing: Vision and challenges. *IEEE internet of things journal* 3, 5 (2016), 637–646.

[35] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[36] Neeraj Solanki, Sepehr Tabrizchi, Samin Sohrabi, Jason Schmidt, and Arman Roohi. 2025. ATM-Net: Adaptive Termination and Multi-Precision Neural Networks for Energy-Harvested Edge Intelligence. *arXiv preprint arXiv:2502.09822* (2025).

[37] Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen. 2015. Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition. In *Proceedings of the 13th ACM conference on embedded networked sensor systems*. 127–140.

[38] Kai Sun, Peibo Duan, Levin Kuhlmann, Beilun Wang, and Bin Zhang. 2025. Ilif: Temporal inhibitory leaky integrate-and-fire neuron for overactivation in spiking neural networks. *arXiv preprint arXiv:2505.10371* (2025).

[39] Kaiwen Tang, Jiaqi Zheng, Yuze Jin, Yupeng Qiu, Guangda Sun, Zhanglu Yan, and Weng-Fai Wong. 2026. SpikySpace: A Spiking State Space Model for Energy-Efficient Time Series Forecasting. *arXiv preprint arXiv:2601.02411* (2026).

[40] Hokchhay Tann, Soheil Hashemi, R Iris Bahar, and Sherief Reda. 2017. Hardware-software codesign of accurate, multiplier-free deep neural networks. In *Proceedings of the 54th Annual Design Automation Conference 2017*. 1–6.

[41] Corinne Teeter, Ramakrishnan Iyer, Vilas Menon, Nathan Gouwens, David Feng, Jim Berg, Aaron Szafer, Nicholas Cain, Hongkui Zeng, Michael Hawrylycz, et al. 2018. Generalized leaky integrate-and-fire models classify multiple neuron types. *Nature communications* 9, 1 (2018), 709.

[42] Michael Voudaskas, Jack Iain MacLean, Neale AW Dutton, Brian D Stewart, and Istvan Gyongy. 2025. Spiking neural networks in imaging: A review and case study. *Sensors* 25, 21 (2025), 6747.

[43] Fei Wang, Jianwei Feng, Yinliang Zhao, Xiaobin Zhang, Shiyuan Zhang, and Jinsong Han. 2019. Joint activity recognition and indoor localization with WiFi fingerprints. *Ieee Access* 7 (2019), 80058–80068.

[44] Tian Wang, Yuzhu Liang, Xuewei Shen, Xi Zheng, Adnan Mahmood, and Quan Z Sheng. 2023. Edge computing and sensor-cloud: Overview, solutions, and directions. *ACM computing surveys* 55, 13s (2023), 1–37.

[45] Xiaoting Wang and Yanxiang Zhang. 2023. MT-SNN: enhance spiking neural network with multiple thresholds. *arXiv preprint arXiv:2303.11127* (2023).

[46] Pete Warden. 2018. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209* (2018).

[47] Yongjun Xiao, Xianlong Tian, Yongqi Ding, Pei He, Mengmeng Jing, and Lin Zuo. 2024. Multi-bit mechanism: A novel information transmission paradigm for spiking neural networks. *arXiv preprint arXiv:2407.05739* (2024).

[48] Zhanglu Yan, Zhenyu Bai, and Weng-Fai Wong. 2024. Reconsidering the energy efficiency of spiking neural networks. *arXiv preprint arXiv:2409.08290* (2024).

[49] Jianfei Yang, Xinyan Chen, Han Zou, Chris Xiaoxuan Lu, Dazhuo Wang, Sumei Sun, and Lihua Xie. 2023. SenseFi: A library and benchmark on deep-learning-empowered WiFi human sensing. *Patterns* 4, 3 (2023).

[50] Xingting Yao, Fanrong Li, Zitao Mo, and Jian Cheng. 2022. Glif: A unified gated leaky integrate-and-fire neuron for spiking neural networks. *Advances in Neural Information Processing Systems* 35 (2022), 32160–32171.

[51] Guanlei Zhang, Lei Feng, Fanqin Zhou, Zhixiang Yang, Qiyang Zhang, Alaa Saleh, Praveen Kumar Donta, and Chinmaya Kumar Dehury. 2024. Spiking neural networks in intelligent edge computing. *IEEE Consumer Electronics Magazine* 14, 4 (2024), 66–75.

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

[53] Yue Zhou, Jiawei Fu, Zirui Chen, Fuwei Zhuge, Yasai Wang, Jianmin Yan, Sijie Ma, Lin Xu, Huanmei Yuan, Mansun Chan, et al. 2023. Computational event-driven vision sensors for in-sensor spiking neural networks. *Nature Electronics* 6, 11 (2023), 870–878.

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

A Theoretical Analysis

Lemma 3.2. *Let $r = \Pr(X < \frac{1}{2})$. Let R, V, and T denote the conditional output distributions of ShiftLIF below $\frac{1}{2}$, of ShiftLIF on $[\frac{1}{2}, \infty)$, and of rounded INT-LIF on $[\frac{1}{2}, \infty)$, respectively. Then*

$$\begin{aligned} H(Q_{shift}(X)) &= h_2(r) + rH(R) + (1-r)H(V), \\ H(Q_{int}(X)) &= h_2(r) + (1-r)H(T), \end{aligned} \tag{27}$$

where $h_2(r) = -r\log_2 r - (1-r)\log_2(1-r)$ is the binary entropy. Consequently,

$$U_K^{(shift)} > U_K^{(int)} \iff rH(R) + (1-r)H(V) > (1-r)H(T). \tag{28}$$

The same formulas remain valid in the degenerate cases $r = 0$ and $r = 1$, with the absent conditional-entropy term interpreted as zero.

Proof. Assume first that $0 < r < 1$, and define

$$\begin{aligned} m &= \Pr(\tfrac{1}{2} \le X < 1), \\ c &= \Pr(X \ge 1), \end{aligned} \tag{29}$$

so that $r + m + c = 1$. Let $R = (R_0, R_2, \ldots, R_K)$ be the dyadic-shell distribution below $\frac{1}{2}$, where

$$\begin{aligned} R_0 &= \Pr(X < 2^{-K} \mid X < \tfrac{1}{2}), \\ R_k &= \Pr(2^{-k} \le X < 2^{-k+1} \mid X < \tfrac{1}{2}), \quad k = 2, \ldots, K, \end{aligned} \tag{30}$$

let

$$V = \left(\frac{m}{1-r}, \frac{c}{1-r}\right), \tag{31}$$

and let $T = (T_1, T_2, \ldots, T_{K+1})$ be the rounded-INT distribution on $[\frac{1}{2}, \infty)$, where

$$\begin{aligned} T_j &= \Pr(j - \tfrac{1}{2} \le X < j + \tfrac{1}{2} \mid X \ge \tfrac{1}{2}), \quad j = 1, \ldots, K, \\ T_{K+1} &= \Pr(X \ge K + \tfrac{1}{2} \mid X \ge \tfrac{1}{2}). \end{aligned} \tag{32}$$

For ShiftLIF, the outputs below $\frac{1}{2}$ have total mass r and conditional distribution R, while the two outputs on $[\frac{1}{2}, \infty)$ have total mass $1-r$ and conditional distribution V. By the chain rule for Shannon entropy,

$$H(Q_{\text{shift}}(X)) = h_2(r) + rH(R) + (1-r)H(V). \tag{33}$$

For rounded INT-LIF, the output 0 corresponds to the single event $\{X < \frac{1}{2}\}$ of probability r, while the remaining $K+1$ outputs have total mass $1-r$ and conditional distribution T. Another application of the chain rule gives

$$H(Q_{\text{int}}(X)) = h_2(r) + (1-r)H(T). \tag{34}$$

Because both utilizations are normalized by the same denominator $\lceil \log_2(K+2) \rceil$,

$$U_K^{(\text{shift})} > U_K^{(\text{int})} \iff H(Q_{\text{shift}}(X)) > H(Q_{\text{int}}(X)). \tag{35}$$

Substituting the two entropy identities above yields

$$U_K^{(\text{shift})} > U_K^{(\text{int})} \iff rH(R) + (1-r)H(V) > (1-r)H(T), \tag{36}$$

which is the desired criterion.

It remains to consider the edge cases. If $r = 1$, then $Q_{\text{int}}(X) \equiv 0$ and $H(Q_{\text{int}}(X)) = 0$, while $Q_{\text{shift}}(X)$ is distributed according to R; this agrees with the stated formula after setting $(1-r)H(V) = (1-r)H(T) = 0$. If $r = 0$, then ShiftLIF has only the two outputs $\frac{1}{2}$ and 1, distributed according to V, while rounded INT-LIF is distributed according to T; again the same formula holds with $rH(R) = 0$. □

Algorithm 2 SNN Training with Spike Rate Regularization

Require: Training set $\mathcal{D}_{\text{train}}$, pretrained weights $\boldsymbol{\theta}_0$, number of epochs E, learning rate η, spike rate penalty weight λ_{sr}, target spike magnitude r^*

Ensure: Trained model parameters $\boldsymbol{\theta}^*$

Initialize SNN model $f_{\boldsymbol{\theta}}$ with L spiking layers; load $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_0$
$\text{acc}_{\text{best}} \leftarrow 0$
for $e = 1$ **to** E **do**
  for each mini-batch $(\mathbf{X}, \mathbf{y}) \in \mathcal{D}_{\text{train}}$ **do**
    // Forward pass over T timesteps
    for $t = 1$ **to** T **do**
      for $\ell = 1$ **to** L **do**
        $\mathbf{v}_t^{(\ell)} \leftarrow \mathbf{v}_{t-1}^{(\ell)} + \left(v_{\text{reset}} - \mathbf{v}_{t-1}^{(\ell)} + \mathbf{z}_t^{(\ell)}\right)/\tau$ ▷ *LIF charging*
        $\mathbf{s}_t^{(\ell)} \leftarrow Q_{\text{shift}}\left(\mathbf{v}_t^{(\ell)}\right)$ ▷ *Multi-level spike: $\{0, 2^{k_{\min}}, \ldots, 2^0\}$*
        $\mathbf{v}_t^{(\ell)} \leftarrow \mathbf{v}_t^{(\ell)} - \mathbf{s}_t^{(\ell)} \cdot v_{\text{th}}$ ▷ *Soft reset*
        Collect $\mathbf{s}_t^{(\ell)}$ into spike buffer $\mathcal{S}^{(\ell)}$
      end for
    end for
    $\hat{\mathbf{y}} \leftarrow f_{\boldsymbol{\theta}}(\mathbf{X})$
    // Task loss
    $\mathcal{L}_{\text{CE}} \leftarrow \text{CrossEntropy}(\hat{\mathbf{y}}, \mathbf{y})$
    // Spike rate regularization (differentiable L1 penalty)
    $\mathcal{L}_{\text{sr}} \leftarrow \frac{1}{L}\sum_{\ell=1}^{L}\left[\text{mean}\left(|\mathcal{S}^{(\ell)}|\right) - r^*\right]^+$
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{CE}} + \lambda_{\text{sr}} \cdot \mathcal{L}_{\text{sr}}$
    // Backward with surrogate gradients and update
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}$
    Clear all spike buffers $\mathcal{S}^{(\ell)}$
  end for
  Evaluate on $\mathcal{D}_{\text{test}}$; **if** $\text{acc}_{\text{test}} > \text{acc}_{\text{best}}$ **then** $\boldsymbol{\theta}^* \leftarrow \boldsymbol{\theta}$, $\text{acc}_{\text{best}} \leftarrow \text{acc}_{\text{test}}$
end for
return $\boldsymbol{\theta}^*$

B Training with Spike Activity Regularization

To control the overall spiking activity during training, we augment the task loss with a spike activity regularization term. The training objective is defined as

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{sr}}\mathcal{L}_{\text{sr}}, \tag{37}$$

where $\mathcal{L}_{\text{CE}}$ denotes the task loss, $\mathcal{L}_{\text{sr}}$ is the spike activity regularizer, and λ_{sr} controls its strength. For the ℓ-th spiking layer, we collect its output spikes over all timesteps in the current mini-batch into $\mathcal{S}^{(\ell)}$, and define

$$\mathcal{L}_{\text{sr}} = \frac{1}{L}\sum_{\ell=1}^{L}\left[\text{mean}\left(\left|\mathcal{S}^{(\ell)}\right|\right) - r^*\right]^+, \tag{38}$$

where L is the number of spiking layers, r^* is the target spike activity, and $[\cdot]^+ = \max(0, \cdot)$. This term penalizes only the amount by which the average spike activity exceeds the target level, thereby encouraging sparse responses without unnecessarily suppressing already low-activity layers. Since our neuron emits multi-level

spikes, mean($|\mathcal{S}^{(\ell)}|$) measures the average spike magnitude rather than a strictly binary firing count.

The training process is shown as Algorithm 2, where Q_{shift} denotes the ShiftQuant surrogate function that maps membrane potentials to multi-level spikes in $\{0, 2^{k_{\min}}, \ldots, 2^0\}$, $\mathcal{S}^{(\ell)}$ denotes the spike buffer of the ℓ-th spiking layer that collects layer outputs over all timesteps in the current mini-batch, L is the number of spiking layers, T is the number of simulation timesteps, r^* is the target spike activity, λ_{sr} is the regularization weight, and $[\cdot]^+$ denotes the positive-part operator $\max(0, \cdot)$.