BRIDGING EFFICIENCY–ACCURACY GAP IN SPIKING NEURAL NETWORKS VIA OPERATION-CONSTRAINED QUANTIZATION AND LOW-SPIKE SUPPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Spiking neural networks constitute a promising approach to energy-efficient intelligence by leveraging binary, event-driven computation. However, binary spike signals substantially constrain the channel capacity for information delivery, thereby limiting attainable model accuracy. To jointly improve representational capacity and inherit event-driven computation, we propose a novel multi-level spiking neuron termed **QuantLIF**, consisting of operation-constrained quantization and low-spike suppression. Specifically, we construct a dense quantization codebook whose values can be represented by a bounded number of signed power-of-two terms. Through shift-based re-parameterization, synaptic integration is replaced by several accumulations, without introducing excessive inference delay. To counteract the deterioration in event sparsity caused by multi-level spikes, the low-spike suppression stochastically filters low-contribution spikes during training and deterministically eliminates them during inference. We theoretically analyze QuantLIF from the perspectives of quantization and information theory, respectively. Comprehensive experiments demonstrate that QuantLIF achieves a favorable trade-off among predictive accuracy, firing sparsity, and energy consumption across both static and neuromorphic benchmarks.

1 INTRODUCTION

The human brain performs complex cognitive tasks with a power budget of merely 20 Watts (Siesjö, 1978). Motivated by this dominance, spiking neural networks (SNNs), often regarded as the third generation of neural networks (Maass, 1997), are extensively studied as a promising paradigm for energy-efficient intelligence. Spiking neuron models such as leaky integrate-and-fire (LIF) (Lapicque, 1907) provide computationally tractable abstractions of biological neurons, thereby enabling bio-inspired neural dynamics to be incorporated into deep learning at an affordable computational cost. Different from perceptrons in artificial neural networks (ANNs) that communicate through continuous activations, spiking neurons temporally integrate synaptic inputs into their membrane potentials and emit binary events. Information is consequently transmitted across network layers as sequences of binary spikes over multiple timesteps. Deployed on neuromorphic platforms such as TrueNorth (Merolla et al., 2014) and Loihi (Davies et al., 2018), the anticipated energy efficiency advantages of SNNs are preliminarily verified.

This energy efficiency primarily stems from two properties of binary spike communication. First, zero-valued spikes induce sparse, event-driven computation. At each timestep, only few neurons are active, allowing the corresponding synaptic operations to be skipped. Second, one-valued spikes eliminate the need for activation-weight multiplications, since synaptic integration can be computed by directly accumulating the associated weight. These computational advantages, however, come at a considerable representational loss. From a quantization perspective, the 0/1 sequence over limited timesteps is insufficient to preserve the information encoded in the original activations, resulting in substantial quantization error and ultimately restricting the performance of SNNs (Shen et al., 2024).

To overcome this bottleneck, multi-level spiking neurons (Feng et al., 2022; Guo et al., 2024; Luo et al., 2024) are proposed to transfer more information through spike amplitudes rather than across timesteps. Expanding the spike alphabet reduces the loss incurred when continuous membrane

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.

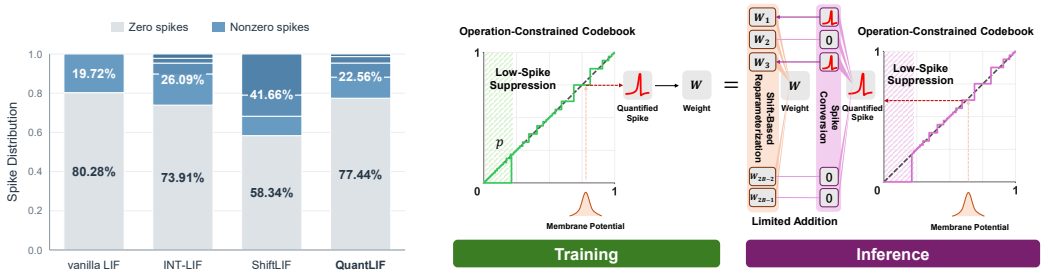


Figure 1: Left: Spike distribution comparison of SNNs. Right: The pipeline of QuantLIF.

potentials are mapped to discrete neuronal outputs, thereby improving the representational capacity of SNNs. However, multi-level spiking neurons often weaken the energy-efficiency property of binary spike computation. Representative multi-level firing approaches recover binary spike-driven inference by decomposing each integer spike into multiple binary events over extra timesteps. This decomposition increases the number of membrane potential accumulations and prolongs inference latency. As illustrated in Figure 1, introducing multiple spike levels substantially makes it difficult to preserve the event sparsity of conventional binary SNNs. Consequently, existing multi-level spiking neurons expose a fundamental dilemma between efficiency and accuracy.

To alleviate the tension between energy efficiency and predictive performance, we propose **QuantLIF**, a multi-level spiking neuron with **operation-constrained quantization** and **low-spike suppression** in Figure 1. To preserve multiplier-free synaptic computation, QuantLIF constructs a fractional quantization codebook by selecting values that can be represented using only a bounded number of power-of-two terms (Sakuma et al., 2021). The membrane potential is then quantized according to this codebook, which allows discrete amplitudes to fit continuous activation values. Since every candidate amplitude can be decomposed into accumulation of finite signed powers-of-two terms, its product with a synaptic weight can be implemented without invoking general-purpose multiplication. After shift-based re-parameterization, our network can perform inference with binary spike sequences, and alleviates latency within few simulation timesteps.

The expanded quantization codebook of multi-level spiking neurons substantially increase the likelihood of firing and thereby weaken event-driven sparsity. To mitigate this issue, we introduce a low-spike suppression mechanism. Specifically, during training, a predefined proportion of low-amplitude quantized spikes is randomly discarded. This stochastic suppression attenuates the contribution of weak events to subsequent synaptic integration and reduces the model's dependence on low-magnitude spikes, thereby promoting the redistribution of their informational content to more salient spikes. During inference, all quantized spikes with magnitude below a specified suppression threshold are deterministically removed, which decreases the number of transmitted events and the associated synaptic operations while incurring only a marginal degradation in performance. Our main contributions are summarized as follows:

- We introduce QuantLIF, a multi-level spiking neuron designed to reconcile predictive accuracy with energy efficiency through operation-constrained quantization and low-spike suppression, providing an alternative perspective on the quantization continuum between ANNs and SNNs.
- We design a fractional quantization codebook that fits activation values discretely along the amplitude dimension while implementing weight-spike products with limited additions; we also introduce low-spike suppression to reduce redundant events and preserve event-driven sparsity.
- We benchmark QuantLIF against representative spiking neurons on CIFAR10/100 (Krizhevsky et al., 2009), ImageNet-1K (Deng et al., 2009), and CIFAR10-DVS (Li et al., 2017), showing a consistent trade-off among accuracy, event sparsity, and inference energy.

2 Related Work

Spiking Neural Networks in Deep Learning. Bio-inspired SNNs are a long-standing network architecture (Maass, 1997). However, their non-differentiable firing and complex temporal dynamics in deep learning are key reasons why their development lagged behind that of ANNs. Current spik-

ing deep learning mainly follows two paradigms. *ANN-to-SNN Conversion* (ANN2SNN) converts continuous activations of pre-trained ANNs to discrete spike firing activities, enabling SNNs to inherit mature optimization and competitive accuracy of ANN (Rueckauer et al., 2017). A series of subsequent efforts expand the performance boundaries of this paradigm by reducing conversion errors (Bu et al., 2022; Liu et al., 2025), optimizing inference latency (Bu et al., 2025), and adapting to Transformers (You et al., 2024; Cao et al., 2026). Nevertheless, accurate conversion generally relies on temporal approximation of continuous ANN activations, which introduces inference latency and limit the exploitation of intrinsic neuronal dynamics. *Direct Training* preserves discrete spikes in network forward process and employs surrogate gradients during backpropagation to approximate the derivative of the non-differentiable firing function, enabling end-to-end optimization of spike-based representations (Wu et al., 2018; Neftci et al., 2019). By improving training strategies (Zheng et al., 2021; Deng et al., 2022) and scaling efficient network architectures such as residual networks (Zheng et al., 2021; Fang et al., 2021a; Hu et al., 2025) and Transformers (Zhou et al., 2023; Yao et al., 2023; 2024a), directly-trained SNNs expand the scope of research to more challenging tasks such as object detection (Su et al., 2023), semantic segmentation (Su et al., 2024), and 3D point-cloud recognition (Wu et al., 2025; Lu et al., 2026). Recently, Pan et al. (2025) and Xing et al. (2025) explore the potential applications of spiking neural networks in larger-scale data and models.

Spiking Neuron Improvements. The spiking neuron constitutes the fundamental unit of SNNs and serves as the primary mechanism underlying event-driven dynamics. Leaky integrate-and-fire neuron (Lapicque, 1907) is simple and efficient, but the limited firing constrains its temporal processing and representation capacity. Existing methods primarily optimize neuron modeling by refining neuronal dynamics and enriching spike firing states. *Neuronal dynamics refinement* methods preserve the binary firing characteristics of neurons, and improve the deep learning performance of SNNs through learnable membrane-related parameters (Fang et al., 2021b; Yao et al., 2022), parallel temporal processing (Fang et al., 2023; Yarga & Wood, 2023), and diverse neuronal dynamics (Huang et al., 2024; Hao et al., 2024). Addressing over-activation and instability in spiking neurons is also a major topic (Sun et al., 2025; Bai et al., 2026; Bu et al., 2026). *Spike state expansion* methods diversify the spike firing, enabling information to be transmitted over a wider spectrum. Real Spike (Guo et al., 2022b) and PIT (Yan et al., 2026) utilize learnable spike amplitudes with inference-time reparameterization. Another way is to extend the firing spikes from binary candidates to multi-level candidates, such as ternary spikes (Guo et al., 2024), integer spikes (Shen et al., 2024; Luo et al., 2024; Yao et al., 2025), and power-of-two spikes (Tang et al., 2026).

Our proposed QuantLIF follows the direct training manner and focuses on improving neuronal representation while preserving efficient inference. Different from existing multi-level spiking neurons, QuantLIF not only enables spike quantization with a limited budget, but also suppresses the firing of low-amplitude spikes.

3 METHOD

3.1 PRELIMINARIES

Leaky Integrate-and-Fire Neuron. The generation and transmission of action potentials between biological neurons are extremely complex processes (Hodgkin & Huxley, 1952). LIF neuron provides a tractable abstraction that retains the essential temporal dynamics. Following the iterative formulation adopted in spatio-temporal backpropagation (STBP) (Wu et al., 2018; Neftci et al., 2019), the temporal dynamics of an LIF neuron can be decomposed into integration, firing and reset:

$$\mathbf{U}[t] = \beta \mathbf{H}[t-1] + \mathbf{X}[t], \tag{1}$$

$$\mathbf{S}[t] = \Theta(\mathbf{U}[t] - V_{\text{th}}), \tag{2}$$

$$\mathbf{H}[t] = \begin{cases} \mathbf{U}[t] \cdot (1 - \mathbf{S}[t]) + V_{\text{reset}} \cdot \mathbf{S}[t], & \text{hard reset}, \\ \mathbf{U}[t] - V_{\text{th}} \cdot \mathbf{S}[t], & \text{soft reset}, \end{cases} \tag{3}$$

where t is the timestep and β is the decay (leaky) factor. $\mathbf{U}[t]$ and $\mathbf{H}[t]$ denote the membrane potential before and after firing, respectively. $\mathbf{X}[t]$ and $\mathbf{S}[t]$ represent the spatial information from the preceding neuron and the signal transmitted to the next neuron, respectively.

In membrane potential integration, $\mathbf{U}[t]$ is formed by the decayed membrane potential from the preceding timestep and spatial information from the preceding neuron. $\Theta(\cdot)$ is the Heaviside function

that $\Theta(x) = 1$ if $x \geq 0$. Therefore, Eq. 2 implies that if the membrane potential exceeds the threshold V_{th}, a spike will be fired. After firing, the membrane potential of the neuron is hard-reset to the resting potential V_{reset} or soft-reset to a lower potential, awaiting subsequent integration.

At the same timestep t, the spatial synaptic contribution between LIF neurons can be denoted as

$$\mathbf{X}^l = \Phi(\mathbf{W}^{l-1}, \mathbf{S}^{l-1}), \tag{4}$$

where l is the layer index in SNN, and Φ denotes the feature aggregation block with weights $\mathbf{W}$. The binary spikes $\mathbf{S}$ can avoid the multiplication operations in Φ.

Multi-Level Spiking Neurons. Multi-level spiking neurons alleviate the representational bottleneck by allowing spikes to carry different amplitudes. Taking Integer Leaky Integrate-and-Fire Neuron (INT-LIF) (Luo et al., 2024) as an example, the main differences lie in the firing and reset processing:

$$\mathbf{S}[t] = \lfloor \text{Clip}(\mathbf{U}[t], 0, D) \rceil \tag{5}$$

$$\mathbf{H}[t] = \begin{cases} \mathbf{U}[t] \cdot (1 - \Theta(\mathbf{S}[t] - \epsilon)) + V_{\text{reset}} \cdot \Theta(\mathbf{S}[t] - \epsilon), & \text{hard reset}, \\ \mathbf{U}[t] - \mathbf{S}[t], & \text{soft reset}, \end{cases} \tag{6}$$

where ϵ is an arbitrarily small positive value, making $\Theta(x - \epsilon) = 1$ if $x > 0$. $\text{Clip}(\mathbf{U}[t], 0, D)$ truncates values outside the range from 0 to D, and $\lfloor \cdot \rceil$ represents the rounding function.

3.2 QUANTLIF

As analyzed in the Sec. 1, multi-level firing undermines the energy efficiency advantages inherent to binary spikes. To balance accuracy and energy efficiency, we propose QuantLIF, which consists of operation-constrained quantization in Section 3.2.1 and low-spike suppression in Section 3.2.2.

3.2.1 OPERATION-CONSTRAINED QUANTIZATION

Multi-Level Firing. Existing multi-level spiking neurons are often described with method-specific firing rules. To reveal the objective of discretizing continuous membrane states into a finite set of amplitudes, we reformulate spiking firing as a general expression from the quantitative perspective:

$$\mathbf{S}[t] = b + \sum_{(\alpha_i, \delta_i) \in \mathcal{A}} \big(\delta_i \cdot \Theta(\mathbf{U}[t] - \alpha_i)\big), \tag{7}$$

where quantization threshold α_i belongs to quantization codebook $\mathcal{A}$, and δ_i is the amplitude interval corresponding to α_i. The lower bound b of the quantization is set to 0 by default. Under Eq. 7, different firing mechanisms can be represented by choosing different codebooks. For example, the codebook of INT-LIF (Luo et al., 2024) can be defined as

$$\mathcal{A}_{\text{INT-LIF}} = \{(i - 2^{-1}, 1)\}, \quad i = 1, 2, \cdots, D. \tag{8}$$

Appendix C provides the diverse codebooks corresponding to other existing neurons.

Quantization Codebook. Eq. 7 reveals that the essential distinction among multi-level spiking neurons lies in how their codebooks are constructed. A denser codebook can reduce quantization error, but arbitrary fractional amplitudes reintroduce expensive multiplications during synaptic contributions. Therefore, codebook design should consider not only quantization precision but also the arithmetic cost of each candidate amplitude. Motivated by this requirement, we retain only values that can be implemented within a prescribed budget through shift and addition operations. To achieve this, we choose a finite number of signed terms from the B-bit power-of-two set to form the non-zero values in the codebook (Sakuma et al., 2021):

$$\{\alpha\} = \bigcup_{d=1}^{D} \Big\{ \sum_{j=1}^{d} (-1)^{c_j} 2^{-s_j} \mid 0 \leq s_1 < s_2 < \cdots < s_d \leq B - 1 \Big\}, \tag{9}$$

where D is the maximum number of power-of-two terms (the corresponding number of addition operations is $D-1$) and s_j denotes the term index of the power-of-two set. To ensure that the spikes are quantized between 0 and 1, $c_j \in \{0, 1\}$ must satisfy (1) $c_1 = 0$, and (2) $c_2 = 1$ if $s_1 = 0$.

To achieve non-scaling quantization, we set δ_i be exactly equal to the interval between adjacent quantization thresholds and α_0 is the zero value in $\mathcal{A}$.

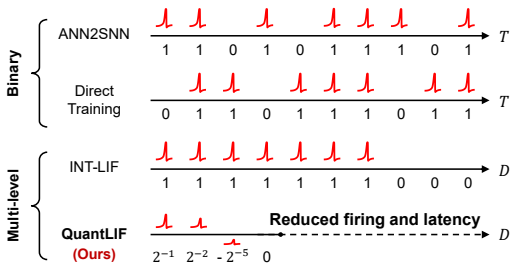


Figure 2: Spike firing patterns at an activation of 0.72 with different firing methods.

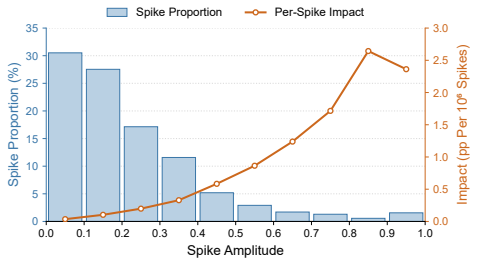


Figure 3: Spike proportion and percentage impact on the next layer at different amplitudes.

Shift-Based Reparameterization. The operation constraint is not only a restriction on the codebook, but also determines how each spike is processed by the subsequent neuron. To achieve multiplier-free synaptic contributions, we propose the shift-based re-parameterization technique in inference stage. Synaptic computation can be extended by shifting weights $\mathbf{W}^{l-1}$ up to $B-1$ times and changing their signs, thereby transforming the multiplication in Φ into a finite number of additions under $2B$-bit bandwidth binary communication across layers,

$$\left(\sum_{j=1}^{d}\mathbf{S}_j^{l-1}\right)\cdot\mathbf{W}^{l-1}=\sum_{j=1}^{d}\underbrace{(-1)^{c_j}(\mathbf{W}^{l-1}\gg s_j)}_{\text{Re-parameterization Weights}},\quad 0\le s_j\le B-1. \tag{10}$$

$d\le D$ is the actual number of terms in quantization. This process does not introduce any new parameters as reparameterization is based solely on the original network weights.

Analysis. In this part, we present the quantitative analysis of operation-constrained quantization, and provide the proofs in Appendix A.

Theorem 1 (Number of Quantization Levels)**.** *Let $L(B,D)$ denotes the total number of valid quantization levels in B-bit quantization codebook $\mathcal{A}$ with D power-of-two terms allowed. Within the quantitative range of 0 to 1, $L(B,D)$ satisfies*

$$L(B,D)=1+B+\sum_{d=2}^{D}\left(2^{d-1}\binom{B-d}{d}+2^{d-2}\binom{B-d}{d-1}\right), \tag{11}$$

where the combinatorial number $\binom{a}{b}=\frac{a!}{b!(a-b)!}$.

Theorem 2 (Quantization Precision)**.** *Let $\Delta(B,D)$ denotes the interval set between adjacent valid quantization levels in B-bit quantization codebook $\mathcal{A}$ with D power-of-two terms allowed. The maximum interval in $\mathcal{A}$ is*

$$\max\Delta(B,D)=\max_i\{\delta_i\}=2^{-\min(B-1,2D-1)}. \tag{12}$$

Theorem 1 and 2 quantify the available amplitude levels and maximum quantization interval under a prescribed operation budget, respectively. Assuming an approximation of 0.72, Fig. 2 illustrates how QuantLIF reduces inference latency and spike firing. Compared to other methods that require 10 or 7 timesteps to achieve a fitting error of 0.02, QuantLIF achieves a more accurate fit (with an error of 1.25×10^{-3}) in less inference firing and latency.

3.2.2 Low-Spike Suppression

Multi-level firing increases the likelihood of nonzero spikes, weakening event-driven sparsity of SNNs. We observe that not all nonzero spikes contribute equally to next-layer synaptic integration. As shown in Fig. 3, low-amplitude spikes contribute less while incurring more event-routing and synaptic-computing overhead as higher-amplitude spikes. We therefore propose a simple low-spike suppression to reduce firing activity and improve computational efficiency.

Stochastically Removal in Training Stage. Directly removing all low-amplitude spikes during training may discard useful information before the network has learned an alternative representation. Therefore, we randomly remove them with a probability p, and reformulate Eq. 7 as

$$\mathbf{S}[t] = \begin{cases} \sum_{(\alpha_i,\delta_i)\in\mathcal{A}} \big(\delta_i \cdot \Theta(\mathbf{U}[t]-\alpha_i)\big), & \mathbf{U}[t] \geq V_{\text{th}}, \\ m \cdot \sum_{(\alpha_i,\delta_i)\in\mathcal{A}} \big(\delta_i \cdot \Theta(\mathbf{U}[t]-\alpha_i)\big), & \mathbf{U}[t] < V_{\text{th}}, \end{cases} \tag{13}$$

where $m \sim \text{Bernoulli}(1-p)$, and V_{th} defines low-amplitude spikes. This stochastic mechanism modifies only the forward spike transmission and requires no explicit firing-rate regularizer. Therefore, it avoids introducing an additional optimization objective with a sensitive coefficient, while allowing the original task objective to guide the redistribution of information.

Deterministically Elimination in Inference Stage. Training-time stochastically removal prepares the network to remain robust when weak events are absent. Consequently, low-amplitude spikes can be deterministically removed during inference, directly reducing the firing number and the associated computation. The inference stage in low-spike suppression can be represented as

$$\mathbf{S}[t] = \Theta(\mathbf{U}[t]-V_{\text{th}}) \cdot \sum_{(\alpha_i,\delta_i)\in\mathcal{A}} \big(\delta_i \cdot \Theta(\mathbf{U}[t]-\alpha_i)\big). \tag{14}$$

Analysis. Let $\mathbf{S}_L$ and $\mathbf{S}_H$ denote the low and high spikes determined by V_{th}, respectively. Y denotes the prediction target. During training, $\mathbf{S}_L$ is transmitted through a mask vector,

$$\widetilde{\mathbf{S}}_L = \mathbf{M} \odot \mathbf{S}_L, \qquad \mathbf{M} \sim \text{Bernoulli}(1-p)^d. \tag{15}$$

Since the masking operation constitutes a stochastic information-degrading channel independent of Y, the conditional data-processing inequality gives $I(Y;\widetilde{\mathbf{S}}_L \mid \mathbf{S}_H) \leq I(Y;\mathbf{S}_L \mid \mathbf{S}_H)$. Therefore, stochastic suppression prevents the low-amplitude pathway from serving as a consistently reliable carrier of task-relevant information. When the model is optimized over different realizations of $\mathbf{M}$, minimizing the expected task loss encourages information required for predicting Y to be encoded by the consistently available $\mathbf{S}_H$. In the ideal redundancy limit, the relationship between mutual information I and entropy H satisfies

$$I(Y;\mathbf{S}_L \mid \mathbf{S}_H) = H(Y \mid \mathbf{S}_H) - H(Y \mid \mathbf{S}_L, \mathbf{S}_H) \to 0, \tag{16}$$

which is equivalently characterized by $H(Y \mid \mathbf{S}_H) \approx H(Y \mid \mathbf{S})$. Hence, once S_H is observed, S_L provides little task-relevant information and can be removed at inference with limited loss. Subsequent experimental results also corroborated this claim.

3.3 Analysis and Discussions

Conversion between ANN and SNN. ANN2SNN (Rueckauer et al., 2017) interprets the firing rate of an IF neuron as an unbiased estimator to the activation of `ReLU` over timesteps. In contrast, QuantLIF increases the output resolution, providing an alternative approximation along the amplitude dimension. Without taking low-spike suppression into account, for any activation x of `ReLU` that falls within the range specified in the codebook $\mathcal{A}$, the quantifying absolute error satisfies

$$|x - \mathbf{S}[t]| \leq 2^{-\min(B-1,2D-1)}. \tag{17}$$

When B and D both tend to infinity, $x = \mathbf{S}[t]$. Therefore, QuantLIF replaces the temporal approximation error of binary IF neurons with an amplitude-quantization error controlled by the local interval of its operation-constrained codebook.

Implementation on Neuromorphic Chips. On synchronous neuromorphic chips (Pei et al., 2019), each QuantLIF spike is decoded into at most D power-of-two components. These components are processed using additions, either in parallel or within a bounded number of local clock cycles. The network therefore consumes the same D timesteps during training. For chips supporting only binary events, a QuantLIF spike can instead be represented by at most D micro-events, each corresponds to a shift-based re-parameterized weight. On asynchronous neuromorphic chips (Yao et al., 2024b), a QuantLIF spike can be transmitted as either an amplitude-coded packet or a short burst of component events. Upon receiving an event, the synaptic core retrieves the corresponding shift-based re-parameterized weight, and accumulates the result into the postsynaptic membrane potential. A shared identifier or end marker can group components belonging to the same spike. Since the number of processed components is bounded, QuantLIF preserves event-driven execution while preventing unbounded communication overhead. Low-spike suppression can additionally discard events before routing, reducing both communication and synaptic computation.

4 EXPERIMENTAL ANALYSIS

We evaluate QuantLIF on the standard CIFAR10/100 datasets (Krizhevsky et al., 2009), the large-scale ImageNet-1K dataset (Deng et al., 2009), and the neuromorphic CIFAR10-DVS dataset (Li et al., 2017). Detailed configurations are provided in Appendix D.

4.1 COMPARATIVE EXPERIMENTS

Comparison of Different Spiking Neurons. To isolate the effect of neuron design, we compare spiking neurons on CIFAR10 using ResNet19, without any data augmentation or auxiliary loss functions. The results are summarized in Table 1. QuantLIF achieves the highest test accuracy while requiring the lowest estimated energy consumption of multi-level firing methods. LIAF-Net provides a reference as a continuous-valued limiting case of our implementation. With only two PoT terms (*i.e.*, $D = 2$), QuantLIF nearly matches the classification accuracy of the continuous-valued reference at less than one-quarter of its energy cost. Furthermore, QuantLIF exhibits competitive consistency across runs, with a reported accuracy variation of 0.15%.

Table 1: Different spiking neurons on CIFAR10. All experiments are repeated three times.

Method	T × D	Accuracy	Avg. Energy
LIF (Lapicque, 1907)	4×1	$91.36 \pm 0.03\%$	$0.85\,\mathrm{mJ} \pm 50.55\,\mu\mathrm{J}$
Real Spike (Guo et al., 2022b)	4×1	$91.61 \pm 0.24\%$	$1.99\,\mathrm{mJ} \pm 49.64\,\mu\mathrm{J}$
LIAF-Net (Wu et al., 2022)	1×23	$\mathbf{93.71} \pm 0.16\%$	$3.08\,\mathrm{mJ} \pm 40.31\,\mu\mathrm{J}$
Ternary Spike (Guo et al., 2024)	4×1	$86.88 \pm 0.39\%$	$2.58\,\mathrm{mJ} \pm 61.52\,\mu\mathrm{J}$
ShiftLIF (Tang et al., 2026)	4×1	$91.30 \pm 0.12\%$	$3.18\,\mathrm{mJ} \pm 72.56\,\mu\mathrm{J}$
INT-LIF (Luo et al., 2024)	1×4	$93.48 \pm 0.14\%$	$0.72\,\mathrm{mJ} \pm 3.03\,\mu\mathrm{J}$
MLF (Feng et al., 2022)	4×3	$93.13 \pm 0.24\%$	$2.39\,\mathrm{mJ} \pm 36.03\,\mu\mathrm{J}$
QuantLIF	1×2	$\mathbf{93.59} \pm 0.15\%$	$\mathbf{0.70}\,\mathrm{mJ} \pm 5.95\,\mu\mathrm{J}$

Table 2: Comparison with existing SNN methods on CIFAR10, CIFAR100, and ImageNet-1K.

Dataset	Method	Type	Architecture	Params.	T × D	Acc.
CIFAR10	TL (Wu et al., 2023)	Tandem learning	CIFARNet	44.48M	8×1	89.04%
	PLIF (Fang et al., 2021b)	SNN training	PLIFNet	37.31M	8×1	93.50%
	STBP-tdBN (Zheng et al., 2021)	SNN training	ResNet19	12.70M	2×1	92.34%
					4×1	92.92%
	TET (Deng et al., 2022)	SNN training	ResNet19	12.70M	2×1	94.16%
					4×1	94.44%
	Ternary Spike (Guo et al., 2024)	SNN training	ResNet19	12.70M	1×1	95.58%
					2×1	95.80%
	Real Spike (Guo et al., 2022b)	SNN training	ResNet19	12.70M	2×2	94.01%
					4×1	95.60%
	PIT (Yan et al., 2026)	SNN training	ResNet19	12.70M	1×2	96.31%
					1×4	96.72%
	QuantLIF	SNN training	ResNet19	12.70M	1×2	**96.97%**
					1×4	**96.98%**
CIFAR100	QCFS (Bu et al., 2022)	ANN2SNN	ResNet20	11.25M	64×1	70.49%
	LTL (Yang et al., 2022)	Tandem learning	ResNet20	11.25M	31×1	76.08%
	TET (Deng et al., 2022)	SNN training	ResNet19	12.70M	2×1	72.87%
					4×1	74.47%
	Ternary Spike (Guo et al., 2024)	SNN training	ResNet19	12.70M	1×1	78.45%
					2×1	80.20%
	Real Spike (Guo et al., 2022b)	SNN training	ResNet20	12.70M	5×1	66.60%
			VGG16	14.72M	5×1	70.62%
	PIT (Yan et al., 2026)	SNN training	ResNet19	12.70M	1×2	80.12%
					1×4	81.59%
	QuantLIF	SNN training	ResNet19	12.70M	1×2	**81.77%**
					1×4	**81.93%**
ImageNet-1K	TET (Deng et al., 2022)	SNN training	ResNet34	21.79M	6×1	64.79%
	TEBN (Duan et al., 2022)	SNN training	ResNet34	21.79M	4×1	68.28%
	SEW ResNet (Fang et al., 2021a)	SNN training	ResNet18	11.69M	4×1	63.18%
			ResNet34	21.79M	4×1	67.04%
	Ternary Spike (Guo et al., 2024)	SNN training	ResNet18	11.69M	4×1	67.68%
			ResNet34	21.79M	4×1	70.74%
	Real Spike (Guo et al., 2022b)	SNN training	ResNet18	11.69M	4×1	63.68%
			ResNet34	21.79M	4×1	67.69%
	E-SpikeFormer (Yao et al., 2025)	SNN training	E-SpikeFormer-S	5.11M	1×4	75.30%
	PIT (Yan et al., 2026)	SNN training	ResNet18	11.69M	1×4	69.39%
			ResNet34	21.79M	1×4	72.66%
			E-SpikeFormer-S	5.11M	1×4	76.00%
	QuantLIF	SNN training	ResNet18	11.69M	1×2	**70.10%**
			ResNet18	11.69M	1×4	**70.90%**
			ResNet34	21.79M	1×4	**73.67%**
			E-SpikeFormer-S	5.11M	1×4	**77.90%**

Evaluation on Static Datasets. Static image classification provides a useful test of whether SNNs can serve as effective alternatives to ANNs. As shown in Table 2, QuantLIF achieves the highest reported accuracy across all evaluated dataset–backbone combinations. With $T = 1$ and $D = 4$, it reaches 96.98% on CIFAR10 and 81.93% on CIFAR100. On ImageNet-1K, it exceeds state-of-the-art method PIT by 1.51%, 1.01%, and 1.90% under the same configurations.

The benefit of increasing D depends on the scale of datasets. Increasing D from 2 to 4 results in a negligible performance improvement on CIFAR10 and CIFAR100, suggesting that $D = 2$ already provides sufficient quantization precision for these benchmarks. In contrast, the same change yields a 0.80% improvement on ImageNet-1K. This is consistent with the greater representational demands of large-scale classification, where finer quantization can preserve information that would otherwise be lost in a coarser representation. The comparison with PIT follows a similar pattern. The quantitative advantages of QuantLIF are more pronounced on large-scale dataset ImageNet-1K than that on CIFAR10/100.

Evaluation on Neuromorphic Datasets. Neuromorphic benchmarks directly assess a model's ability to process data captured by event-driven sensors. Following our observation on static benchmarks that a compact codebook is sufficient for smaller datasets, we set $D = 2$ and evaluate QuantLIF on CIFAR10-DVS using ResNet-19 and VGG-11, with $T = 10$.

Table 3: Comparisons on CIFAR10-DVS.

Method	Archi.	Params.	T × D	Acc.
STBP-tdBN (Zheng et al., 2021)	ResNet19	12.70M	10 × 1	67.80%
PLIF (Fang et al., 2021b)	PLIF Net	37.31M	20 × 1	74.80%
MPBN (Guo et al., 2023)	ResNet19	12.70M	10 × 1	74.40%
IM-Loss (Guo et al., 2022a)	ResNet19	12.70M	10 × 1	72.60%
Dspike (Li et al., 2021)	ResNet18	11.18M	10 × 1	75.40%
TET (Deng et al., 2022)	VGG11	9.33M	10 × 1	77.33%
Ternary Spike (Guo et al., 2024)	ResNet19	12.70M	10 × 1	79.80%
	VGG11	9.33M	10 × 1	76.60%
Real Spike (Guo et al., 2022b)	ResNet19	12.70M	10 × 1	72.85%
	VGG16	14.72M	10 × 1	74.58%
PIT (Yan et al., 2026)	VGG11	9.33M	10 × 1	80.50%
	VGG11	9.33M	10 × 4	81.50%
QuantLIF (Ours)	ResNet19	12.70M	10 × 2	**83.10%**
	VGG11	9.33M	10 × 2	**84.10%**

As shown in Table 3, QuantLIF achieves accuracies of 83.10% and 84.10% with ResNet-19 and VGG-11, respectively, outperforming all listed methods. On VGG-11, it surpasses PIT, the strongest competing method, by 2.60% under the same architecture and timesteps, despite using $D = 2$ rather than $D = 4$. On ResNet-19, QuantLIF improves upon Ternary Spike with the same backbone by 3.30%, demonstrating its effectiveness across both architectures. The gain over PIT is larger than those observed on the static benchmarks, suggesting that the proposed quantized representation is particularly well suited to event-based visual data.

4.2 Ablation Studies for QuantLIF

Analysis on Codebook Design. We investigate how D and B affect accuracy and energy in Figure 4. With $B = 16$, increasing D from 1 to 2 improves accuracy on both datasets, particularly CIFAR100, suggesting that SNNs benefit from richer quantized spike representations. Further increasing D yields little additional benefit, whereas energy consumption grows approximately linearly. This saturation indicates diminishing returns from additional terms and motivates our choice of $D = 2$.

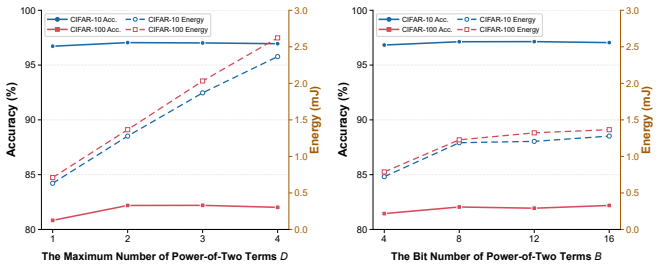


Figure 4: Ablation study for the codebook design.

With $D = 2$, increasing B from 4 to 8 improves accuracy on both datasets. Despite unchanged nominal quantization precision, the larger number of available quantization levels provides a richer representation. Beyond $B = 8$, accuracy largely saturates, with small fluctuations rather than consistent improvements. Meanwhile, energy consumption increases appreciably from $B = 4$ to $B = 8$, but grows only modestly thereafter. Nevertheless, larger B still incurs a communication cost, as it directly determines the inter-layer spike communication bandwidth after reparameterization.

Analysis on Low-Spike Suppression. Table 4 compares our stochastic low-spike suppression with training without suppression ($p = 0$) and with deterministic suppression ($p = 1.0$). Our method achieves the highest accuracy at every nonzero threshold and delays the sharp degradation observed

in the baselines as the threshold increases. In particular, at $V_{\text{th}} = 4 \times 2^{-4}$, it achieves 88.46% accuracy, outperforming the two baselines by 26.63% and 22.68%, respectively. These results suggest that stochastic suppression improves robustness to low-spike removal.

Energy consumption almost decreases monotonically with V_{th} for all three variants. Our method exhibits slightly smaller energy savings at some thresholds while retaining higher accuracy. This is consistent with the hypothesis that stochastic suppression encourages the network to encode more task-relevant information from low-spikes to high spikes, thereby reducing its reliance on low spikes.

Table 4: Ablation study for low-spike suppression on CIFAR10.

V_{th}	**No Suppression** $p = 0$		**Stochastic** $p = 0.5$		**Deterministic** $p = 1.0$	
($\times 2^{-4}$)	Acc. (%)	Eg. (mJ)	Acc. (%)	Eg. (mJ)	Acc. (%)	Eg. (mJ)
0	97.04	1.244	97.04	1.244	97.04	1.244
1	96.92	0.857	**96.97**	0.941	96.92	0.879
2	96.54	0.668	**96.72**	0.681	96.34	0.659
3	93.59	0.485	**94.22**	0.446	88.54	0.378
4	61.83	0.379	**88.46**	0.364	65.78	0.386
5	9.09	0.235	**51.58**	0.322	47.17	0.156

4.3 Further Analysis

Visualization of Spike Distributions. As illustrated in Figure 5, we visualize spike activity from the last layer of E-Spiking-Transformer-S on ImageNet. QuantLIF produces responses that broadly align with foreground objects, suggesting that its spike representations capture spatial structure relevant to category recognition. Compared with standard LIF, QuantLIF provides broader and more spatially coherent coverage. This difference is particularly apparent for the green snake, which closely resembles the surrounding foliage in color and texture. Even under this challenging situation, QuantLIF maintains responses along the object, suggesting improved sensitivity to its spatial structure.

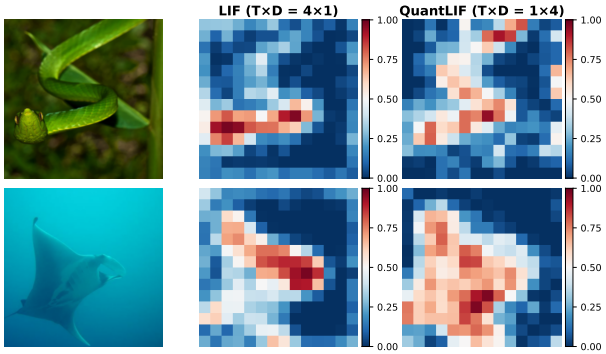


Figure 5: Visualization of spike distribution in images.

Robustness Against Corruptions and Perturbations. We evaluate robustness to distribution shifts without target-domain adaptation, which means SNNs are trained on CIFAR10/100 but are evaluated on CIFAR10-C and CIFAR100-C (Hendrycks & Dietterich, 2019). The results in Table 5 indicate that QuantLIF achieves higher classification accuracy on both benchmarks. It also exhibits a smaller accuracy drop from its corresponding clean-data performance than vanilla LIF and GLIF (Yao et al., 2022), indicating better preservation of predictive performance under input corruptions. Detailed comparisons across corruption types and severity levels are provided in Appendix E.5.

Table 5: Robustness study.

Method	Val. Set	Acc.	↓ Acc.
vanilla LIF	CIFAR10-C	82.49%	12.84%
	CIFAR100-C	57.63%	21.12%
GLIF	CIFAR10-C	75.23%	19.72%
	CIFAR100-C	49.11%	28.11%
QuantLIF	CIFAR10-C	**86.08%**	**10.89%**
	CIFAR100-C	**63.34%**	**18.43%**

5 Conclusion

We presented **QuantLIF**, a multi-level spiking neuron that combines operation-constrained quantization with low-spike suppression to improve the accuracy–efficiency trade-off in SNNs. By representing spike amplitudes with a bounded number of signed power-of-two terms, QuantLIF enables richer representations while retaining multiplier-free synaptic integration through shift-based reparameterization. Stochastic suppression during training reduces reliance on low-amplitude spikes, enabling their deterministic removal at inference with limited accuracy loss. Experiments demonstrate the effectiveness of this design across static and neuromorphic benchmarks, with favorable accuracy and estimated energy consumption. These results highlight the value of jointly designing spike representations and their computational cost for accurate and efficient spiking networks.

REFERENCES

Qianyi Bai, Haiteng Wang, and Qiang Yu. A brain-inspired gating mechanism unlocks robust computation in spiking neural networks. In *International Conference on Learning Representations*, 2026.

Tong Bu, Wei Fang, Jianhao Ding, Penglin Dai, Zhaofei Yu, and Tiejun Huang. Optimal ann-snn conversion for high-accuracy and ultra-low-latency spiking neural networks. In *International Conference on Learning Representations*, 2022.

Tong Bu, Maohua Li, and Zhaofei Yu. Inference-scale complexity in ann-snn conversion for high-performance and low-power applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24387–24397, 2025.

Tong Bu, Xinyu Shi, and Zhaofei Yu. Activity pruning for efficient spiking neural networks. *Advances in Neural Information Processing Systems*, 38:171134–171153, 2026.

Honglin Cao, Shuai Wang, Zijian Zhou, Ammar Belatreche, Wenjie Wei, Yu Liang, Yu Yang, Rui Xi, Malu Zhang, and Haizhou Li. Towards training-free and accurate ann-to-snn conversion via activation-aware redistribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1703–1711, 2026.

Mike Davies, Narayan Srinivasa, 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.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Shikuang Deng, Yuhang Li, Shanghang Zhang, and Shi Gu. Temporal efficient training of spiking neural network via gradient re-weighting. In *International Conference on Learning Representations*, 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.

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

Wei Fang, Zhaofei Yu, Yanqi Chen, Timothée Masquelier, Tiejun Huang, and Yonghong Tian. Incorporating learnable membrane time constant to enhance learning of spiking neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2661–2671, 2021b.

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

Lang Feng, Qianhui Liu, Huajin Tang, De Ma, and Gang Pan. Multi-level firing with spiking DS-ResNet: Enabling better and deeper directly-trained spiking neural networks. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 2471–2477, 2022.

Yufei Guo, Yuanpei Chen, Liwen Zhang, Xiaode Liu, Yinglei Wang, Xuhui Huang, and Zhe Ma. Im-loss: Information maximization loss for spiking neural networks. In *Advances in Neural Information Processing Systems*, volume 35, 2022a.

Yufei Guo, Liwen Zhang, Yuanpei Chen, Xinyi Tong, Xiaode Liu, Yinglei Wang, Xuhui Huang, and Zhe Ma. Real spike: Learning real-valued spikes for spiking neural networks. In *European Conference on Computer Vision*, pp. 52–68. Springer, 2022b.

Yufei Guo, Yuhan Zhang, Yuanpei Chen, Weihang Peng, Xiaode Liu, Liwen Zhang, Xuhui Huang, and Zhe Ma. Membrane potential batch normalization for spiking neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19420–19430, 2023.

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

Zecheng Hao, Xinyu Shi, Zihan Huang, Tong Bu, Zhaofei Yu, and Tiejun Huang. A progressive training framework for spiking neural networks with learnable multi-hierarchical model. In *International Conference on Learning Representations*, 2024.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Alan L Hodgkin and Andrew F Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4):500, 1952.

Mark Horowitz. 1.1 computing's energy problem (and what we can do about it). In *2014 IEEE international solid-state circuits conference digest of technical papers (ISSCC)*, pp. 10–14. Ieee, 2014.

Yifan Hu, Lei Deng, Yujie Wu, Man Yao, and Guoqi Li. Advancing spiking neural networks toward deep residual learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2): 2353–2367, 2025. doi: 10.1109/TNNLS.2024.3355393.

Yulong Huang, Xiaopeng Lin, Hongwei Ren, Haotian Fu, Yue Zhou, Zunchang Liu, Biao Pan, and Bojun Cheng. Clif: Complementary leaky integrate-and-fire neuron for spiking neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 19949–19972. PMLR, 2024.

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

Louis Lapicque. Recherches quantitatives sur l'excitation electrique des nerfs traitee comme une polarization. *Journal de physiologie et de pathologie générale*, 9:620–635, 1907.

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, Yufei Guo, Shanghang Zhang, Shikuang Deng, Yongqing Hai, and Shi Gu. Differentiable spike: Rethinking gradient-descent for training spiking neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 23426–23439, 2021.

Chang Liu, Jiangrong Shen, Xuming Ran, Mingkun Xu, Qi Xu, Yi Xu, and Gang Pan. Efficient ann-snn conversion with error compensation learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 39482–39495. PMLR, 2025.

Yijie Lu, Zhiyi Pan, Renrui Zhang, Yanhao Jia, Ronggang Wang, and Zhaokun Zhou. Spiking discrepancy transformer for point cloud analysis. In *International Conference on Learning Representations*, volume 2026, pp. 109747–109769, 2026.

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

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

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.

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

Yuqi Pan, Yupeng Feng, Jinghao Zhuang, Siyu Ding, Han Xu, Zehao Liu, Bohan Sun, Yuhong Chou, Xuerui Qiu, Anlin Deng, et al. Spikingbrain: Spiking brain-inspired large models. *arXiv preprint arXiv:2509.05276*, 2025.

Priyadarshini Panda, Sai Aparna Aketi, and Kaushik Roy. Toward scalable, efficient, and accurate deep spiking neural networks with backward residual connections, stochastic softmax, and hybridization. *Frontiers in Neuroscience*, 14:653, 2020.

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.

Bodo Rueckauer, Iulia-Alexandra Lungu, Yuhuang Hu, Michael Pfeiffer, and Shih-Chii Liu. Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. *Frontiers in neuroscience*, 11:682, 2017.

Yuiko Sakuma, Hiroshi Sumihiro, Jun Nishikawa, Toshiki Nakamura, and Ryoji Ikegaya. n-hot: Efficient bit-level sparsity for powers-of-two neural network quantization. *arXiv preprint arXiv:2103.11704*, 2021.

Guobin Shen, Dongcheng Zhao, Tenglong Li, Jindong Li, and Yi Zeng. Are conventional snns really efficient? a perspective from network quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27538–27547, 2024.

Bo K Siesjö. Brain energy metabolism. *(No Title)*, 1978.

Qiaoyi Su, Yuhong Chou, Yifan Hu, Jianing Li, Shijie Mei, Ziyang Zhang, and Guoqi Li. Deep directly-trained spiking neural networks for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6555–6565, 2023.

Qiaoyi Su, Weihua He, Xiaobao Wei, Bo Xu, and Guoqi Li. Multi-scale full spike pattern for semantic segmentation. *Neural Networks*, 176:106330, 2024. doi: 10.1016/j.neunet.2024.106330.

Kai Sun, Peibo Duan, Levin Kuhlmann, Beilun Wang, and Bin Zhang. Ilif: Temporal inhibitory leaky integrate-and-fire neuron for overactivation in spiking neural networks. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 4200–4208, 2025.

Kaiwen Tang, Di Yu, Jiaqi Zheng, Changze Lv, Qianhui Liu, Zhanglu Yan, and Weng-Fai Wong. Shiftlif: Efficient multi-level spiking neurons with power-of-two quantization. *arXiv preprint arXiv:2605.01866*, 2026.

Jibin Wu, Yansong Chua, Malu Zhang, Guoqi Li, Haizhou Li, and Kay Chen Tan. A tandem learning rule for effective training and rapid inference of deep spiking neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 34(1):446–460, 2023. doi: 10.1109/TNNLS.2021.3095724.

Peixi Wu, Bosong Chai, Hebei Li, Menghua Zheng, Yansong Peng, Zeyu Wang, Xuan Nie, Yueyi Zhang, and Xiaoyan Sun. Spiking point transformer for point cloud classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21563–21571, 2025.

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.

Zhenzhi Wu, Hehui Zhang, Yihan Lin, Guoqi Li, Meng Wang, and Ye Tang. Liaf-net: Leaky integrate and analog fire network for lightweight and efficient spatiotemporal information processing. *IEEE Transactions on Neural Networks and Learning Systems*, 33(11):6249–6262, 2022.

Xingrun Xing, Boyan Gao, Zheng Liu, David A Clifton, Shitao Xiao, Wanpeng Zhang, Li Du, Zheng Zhang, Guoqi Li, and Jiajun Zhang. Spikellm: Scaling up spiking neural network to large language models via saliency-based spiking. In *International Conference on Learning Representations*, 2025.

Yinsong Yan, Yujie Wu, and Jibin Wu. Advancing spatiotemporal representations in spiking neural networks via parametric invertible transformation. In *International Conference on Learning Representations*, 2026.

Qu Yang, Jibin Wu, Malu Zhang, Yansong Chua, Xinchao Wang, and Haizhou Li. Training spiking neural networks with local tandem learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 12662–12676, 2022.

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, 2023.

Man Yao, Jiakui Hu, Tianxiang Hu, Yifan Xu, Zhaokun Zhou, Yonghong Tian, Bo Xu, and Guoqi Li. Spike-driven transformer v2: Meta spiking neural network architecture inspiring the design of next-generation neuromorphic chips. In *International Conference on Learning Representations*, 2024a.

Man Yao, Ole Richter, Guangshe Zhao, Ning Qiao, Yannan Xing, Dingheng Wang, Tianxiang Hu, Wei Fang, Tugba Demirci, Michele De Marchi, et al. Spike-based dynamic computing with asynchronous sensing-computing neuromorphic chip. *Nature communications*, 15(1):4464, 2024b.

Man Yao, Xuerui Qiu, Tianxiang Hu, Jiakui Hu, Yuhong Chou, Keyu Tian, Jianxing Liao, Luziwei Leng, Bo Xu, and Guoqi Li. Scaling spike-driven transformer with efficient spike firing approximation training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4): 2973–2990, 2025.

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

Sidi Yaya Arnaud Yarga and Sean U N Wood. Accelerating snn training with stochastic parallelizable spiking neurons. In *2023 International Joint Conference on Neural Networks*, pp. 1–8. IEEE, 2023.

Kang You, Zekai Xu, Chen Nie, Zhijie Deng, Qinghai Guo, Xiang Wang, and Zhezhi He. Spikezip-tf: Conversion is all you need for transformer-based snn. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 57367–57383. PMLR, 2024.

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.

Zhaokun Zhou, Yuesheng Zhu, Chao He, Yaowei Wang, Shuicheng Yan, Yonghong Tian, and Li Yuan. Spikformer: When spiking neural network meets transformer. In *International Conference on Learning Representations*, 2023.

APPENDIX

A PROOFS FOR QUANTLIF

In this part, we provide the proofs of Theorem 1 and Theorem 2. Let $\mathcal{Q}$ denote the set of distinct amplitudes in $[0, 1]$ that can be represented by at most D signed terms from the B-bit power-of-two basis. When multiple signed representations correspond to the same amplitude, we consider its simplest representation. Such a representation can be expressed in non-adjacent form (NAF), in which no two nonzero digits are adjacent. We define the convention $\binom{n}{k} = 0$ for $k < 0$ or $k > n$.

A.1 PROOF OF THEOREM 1

Proof. $\forall q \in \mathcal{Q}$, amplitude q admits a minimum-weight signed representation

$$q = \sum_{i=0}^{B-1} z_i 2^{-i}, \qquad z_i \in \{-1, 0, 1\}, \tag{18}$$

which can be expressed in NAF such that $z_i z_{i+1} = 0$. Therefore, counting distinct QuantLIF amplitudes with at most D power-of-two terms is equivalent to counting positive NAF representations with at most D nonzero digits.

Let $A_d(B)$ denote the number of positive amplitudes whose minimum representation contains exactly d nonzero terms. For $d = 1$, the valid amplitudes are $\{1, 2^{-1}, \ldots, 2^{-(B-1)}\}$, giving

$$A_1(B) = B. \tag{19}$$

For $d \geq 2$, we consider two cases. First, suppose that the leading nonzero position satisfies $s_1 \geq 1$. Selecting d non-adjacent positions from $\{1, \ldots, B-1\}$ gives $\binom{B-d}{d}$ choices. Since the represented amplitude is positive, the leading sign must be positive, while the remaining $d-1$ signs can be chosen independently. Hence,

$$A_d^{(1)}(B) = 2^{d-1}\binom{B-d}{d}. \tag{20}$$

Second, consider $s_1 = 0$. The leading term is 1, and the second nonzero term must be negative; otherwise, the represented amplitude would exceed one. The remaining $d-1$ nonzero positions are selected from $\{2, \ldots, B-1\}$, resulting in $\binom{B-d}{d-1}$ choices. The sign of the second term is fixed, while the remaining $d-2$ signs are unrestricted. Therefore,

$$A_d^{(2)}(B) = 2^{d-2}\binom{B-d}{d-1}. \tag{21}$$

Combining the two cases gives

$$A_d(B) = 2^{d-1}\binom{B-d}{d} + 2^{d-2}\binom{B-d}{d-1}, \qquad d \geq 2. \tag{22}$$

Including the zero amplitude and all representations containing at most D nonzero terms, we obtain

$$\begin{aligned} L(B, D) &= 1 + \sum_{d=1}^{D} A_d(B) \\ &= 1 + B + \sum_{d=2}^{D} \left(2^{d-1}\binom{B-d}{d} + 2^{d-2}\binom{B-d}{d-1} \right), \end{aligned} \tag{23}$$

which is exactly Eq. 11. □

A.2 Proof of Theorem 2

Proof. Let the elements of $\mathcal{Q}$ be ordered as $0 = q_0 < q_1 < \cdots < q_M = 1$, and define $m = \min(B-1, 2D-1)$. We first establish the upper bound on the maximum interval. Consider the uniform grid

$$\mathcal{G}_m = \left\{ k2^{-m} \mid k = 0, 1, \ldots, 2^m \right\}. \tag{24}$$

Every element of $\mathcal{G}_m$ admits an NAF representation whose nonzero positions are contained in $\{0, \ldots, m\}$. Since adjacent nonzero digits are prohibited, such a representation contains at most

$$\left\lceil \frac{m+1}{2} \right\rceil \leq D \tag{25}$$

nonzero terms, where the inequality follows from $m \leq 2D-1$. Moreover, $m \leq B-1$ ensures that all required powers of two are contained in the available basis. Thus, $\mathcal{G}_m \subseteq \mathcal{Q}$, which implies

$$\max \Delta(B, D) \leq 2^{-m}. \tag{26}$$

We next show that the bound is attainable. When $B-1 \leq 2D-1$, we have $m = B-1$. Every representable amplitude is an integer multiple of $2^{-(B-1)}$, while every point on this grid belongs to $\mathcal{Q}$. Therefore,

$$\max \Delta(B, D) = 2^{-(B-1)}. \tag{27}$$

When $2D-1 < B-1$, we have $m = 2D-1$. Consider the two valid amplitudes

$$q_D^- = \sum_{r=1}^{D} 2^{-(2r-1)}, \qquad q_D^+ = 1 - \sum_{r=1}^{D-1} 2^{-2r}. \tag{28}$$

Both are valid NAF representations containing at most D nonzero terms, and their difference is

$$q_D^+ - q_D^- = 2^{-(2D-1)}. \tag{29}$$

For any positive NAF representation with $s_1 \geq 1$, the largest amplitude obtainable using at most D nonzero terms is achieved by selecting the earliest admissible positions $\{1, 3, \ldots, 2D-1\}$ with positive signs, which gives q_D^-. Hence, all such amplitudes satisfy $q \leq q_D^-$. Conversely, among representations with $s_1 = 0$, the smallest valid amplitude is obtained by selecting $+1$ followed by negative terms at positions $\{2, 4, \ldots, 2D-2\}$, which gives q_D^+. Thus, all such amplitudes satisfy $q \geq q_D^+$. It follows that

$$(q_D^-, q_D^+) \cap \mathcal{Q} = \varnothing, \tag{30}$$

so q_D^- and q_D^+ are adjacent quantization levels. Therefore,

$$\max \Delta(B, D) \geq 2^{-(2D-1)}. \tag{31}$$

Combining the upper and lower bounds for the two regimes yields

$$\max \Delta(B, D) = 2^{-\min(B-1, 2D-1)}, \tag{32}$$

which proves Eq. 12. □

A.3 Quantization Error Bound

Theorem 2 directly provides a worst-case error bound for the downward quantization used by QuantLIF. Let $Q(x)$ denote the quantized output for an input $x \in [0, 1]$. For two adjacent quantization levels satisfying $q_i \leq x < q_{i+1}$, we have $Q(x) = q_i$. Therefore,

$$0 \leq x - Q(x) < q_{i+1} - q_i, \quad q_{i+1} - q_i \in \Delta(B, D). \tag{33}$$

Consequently,

$$|x - Q(x)| \leq 2^{-\min(B-1, 2D-1)}. \tag{34}$$

This result shows that increasing D by one improves the guaranteed effective resolution by two binary positions until the finest resolution supported by the B-bit power-of-two basis is reached. In particular, $D \geq \lceil B/2 \rceil$ is sufficient to recover the complete grid with spacing $2^{-(B-1)}$.

Algorithm 1 Operation-constrained codebook construction

Require: Bit width $B \geq 1$; maximum number of terms $D \geq 1$
Ensure: Codebook $\mathcal{A}$; decomposition lookup $\mathcal{E}$

$\mathcal{Q} \leftarrow \{0\}$; $\mathcal{E}(0) \leftarrow \varnothing$
for $d = 1, \ldots, \min(D, B)$ **do**
  for all $(s_1, \ldots, s_d)$ with $0 \leq s_1 < \cdots < s_d \leq B - 1$ **do**
    for all $(c_1, \ldots, c_d) \in \{0, 1\}^d$ with $c_1 = 0$ **do**
      $q \leftarrow \sum_{j=1}^{d} (-1)^{c_j} 2^{-s_j}$
      if $0 < q \leq 1$ and $q \notin \mathcal{Q}$ **then**
        $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{q\}$
        $\mathcal{E}(q) \leftarrow \{(c_j, s_j)\}_{j=1}^{d}$ ▷ Store a minimum-term decomposition
Sort $\mathcal{Q}$ as $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_K = 1$
$\delta_i \leftarrow \alpha_i - \alpha_{i-1}$ for $i = 1, \ldots, K$
$\mathcal{A} \leftarrow \{(\alpha_i, \delta_i)\}_{i=1}^{K}$
return $\mathcal{A}, \mathcal{E}$

Algorithm 2 QuantLIF firing and synaptic transmission

Require: $\mathcal{A}, \mathcal{E}$ from Algorithm 1; suppression threshold V_{th}; probability p
mode $\in \{\text{train}, \text{infer}\}$

function FIRING($\mathbf{U}[t], \mathcal{A}, V_{\text{th}}, p, \text{mode}$)
  $q[t] \leftarrow \sum_{(\alpha_i, \delta_i) \in \mathcal{A}} \delta_i \Theta(\mathbf{U}[t] - \alpha_i)$ ▷ Quantization
  if mode $=$ train **then**
    $m \leftarrow 1$
    if $\mathbf{U}[t] < V_{\text{th}}$ **then**
      $m \sim \text{Bernoulli}(1 - p)$ ▷ Independent mask
    $\mathbf{S}[t] \leftarrow m\, q[t]$
  else
    $\mathbf{S}[t] \leftarrow \Theta(\mathbf{U}[t] - V_{\text{th}})\, q[t]$ ▷ Deterministic suppression
  return $\mathbf{S}[t]$

function SYNAPTICCONTRIBUTION($\mathbf{S}[t], \mathbf{W}, \mathcal{E}, \text{mode}$)
  if mode $=$ train **then**
    return $\mathbf{W}\, \mathbf{S}[t]$
  $\mathbf{w}_s \leftarrow \mathbf{W} \gg s$ for $s = 0, 1, \cdots, B - 1$ ▷ Shift-based reparameterization
  $z \leftarrow 0$
  for all $(c, s) \in \mathcal{E}(\mathbf{S}[t])$ **do**
    if $c = 0$ **then**
      $z \leftarrow z + \mathbf{w}_s$
    else
      $z \leftarrow z - \mathbf{w}_s$
  return z

B ALGORITHMS OF QUANTLIF

Algorithms 1 and 2 summarize codebook construction and the forward computation of QuantLIF. The codebook and its signed power-of-two decompositions are constructed once for each configuration (B, D). The firing routine takes the pre-firing membrane potential $\mathbf{U}[t] = \beta \mathbf{H}[t-1] + \mathbf{X}[t]$ as input and is evaluated once per neuron and time step. The resulting spike is shared by all outgoing synapses. During inference, each nonzero spike is implemented using at most D signed, shifted-weight contributions.

Here, $q[t]$ denotes the quantized amplitude before suppression, and $S[t]$ denotes the transmitted spike. The lookup $\mathcal{E}(q)$ stores a minimum-term signed power-of-two decomposition of amplitude q. Since $\mathcal{E}(0) = \varnothing$, a suppressed spike contributes no synaptic operations in the inference routine. Contributions from incoming synapses are accumulated to form the input to the next neuron. The

notation $w \gg s$ follows the shift-based reparameterization in the main text; the shifted weights can be precomputed for inference without introducing additional learnable parameters.

C CODEBOOKS OF SNN NEURONS

Table 6: Codebooks of representative SNN neurons. $\star$ indicates that the lower bound for quantization of this method is set to -1.

Neurons	**Output Domain**	**Codebooks**	
		Threshold α_i	**Increment** δ_i
LIF (Wu et al., 2018)	$\{0, 1\}$	V_{th}	1
Ternary Spike$\star$ (Guo et al., 2024)	$\{-1, 0, 1\}$	$-V_{\text{th}}, V_{\text{th}}$	1
RealSpike (Guo et al., 2022b)	$\{0, a\}$	V_{th}	a
ShiftLIF (Tang et al., 2026)	$\{1, 2^{-1}, 2^{-2}, \cdots, 2^{-(B-1)}, 0\}$	2^{i-B}	$\begin{cases} 2^{1-B}, & i = 1 \\ 2^{i-B-1}, & i > 1 \end{cases}$
INT-LIF (Luo et al., 2024)	$\{0, 1, 2, \cdots, D\}$	$i - 2^{-1}$	1
MLF (Yao et al., 2025)	$\{0, 1, 2, \cdots, D\}$	V_i	1
PIT (Yan et al., 2026)	$\{0, a, 2a, \cdots, Da\}$	$\left(i - 2^{-1}\right) a$	a

Table 6 illustrates the generality of our quantization formulation by expressing the firing rules of representative spiking neurons through a common codebook parameterization, $\mathcal{A} = \{(\alpha_i, \delta_i)\}$, together with the base level b. For thresholds arranged in ascending order, crossing the first k thresholds yields an output of $b + \sum_{i=1}^{k} \delta_i$. Thus, the thresholds specify the decision boundaries, while the cumulative increments determine the output amplitudes. Within this formulation, LIF and RealSpike correspond to a single threshold with unit and scaled increments, respectively, whereas Ternary Spike uses two thresholds with $b = -1$. INT-LIF and MLF generate integer-valued outputs through unit increments, with their differences captured by threshold placement. PIT scales both the thresholds and increments, while ShiftLIF adopts power-of-two output levels. These correspondences show that the firing rules of the methods in Table 6 can be represented as specific instances of Eq. 7. At the level of the instantaneous membrane-to-spike mapping, these methods can therefore be understood as different codebook designs that specify how continuous membrane potentials are partitioned and mapped to discrete spike amplitudes. This perspective also motivates QuantLIF: we explicitly incorporate an arithmetic operation budget into codebook construction to balance quantization precision with the computational cost of the resulting spike amplitudes.

D DETAILED EXPERIMENTAL SETTINGS

D.1 DATASETS

CIFAR10. CIFAR10 consists of 32×32 color images distributed across 10 classes, with 50,000 images for training and 10,000 images for testing. During training, we apply random horizontal flipping with a probability of 0.5, TrivialAugmentWide, and random erasing with a probability of 0.1. At the batch level, either MixUp ($\alpha = 0.2$) or CutMix ($\alpha = 1.0$) is randomly selected and applied. The images are normalized using a mean of $(0.4914, 0.4822, 0.4465)$ and a standard deviation of $(0.2023, 0.1994, 0.2010)$. During evaluation, only tensor conversion and normalization are applied, without test-time augmentation. ResNet-19 is employed as the backbone.

CIFAR100. CIFAR100 follows the same image resolution and train-test split as CIFAR10, containing 50,000 training images and 10,000 test images distributed across 100 classes. We use the same augmentation pipeline as for CIFAR10, including random horizontal flipping with a probability of 0.5, TrivialAugmentWide, and random erasing with a probability of 0.1. At the batch level, either MixUp ($\alpha = 0.2$) or CutMix ($\alpha = 1.0$) is randomly selected for each batch. The images are normalized using a mean of $(0.5071, 0.4867, 0.4408)$ and a standard deviation of $(0.2675, 0.2565, 0.2761)$. During evaluation, only tensor conversion and normalization are applied. ResNet-19 is used as the backbone.

Table 7: Training configurations for different datasets and backbones.

Configuration	CIFAR10 ResNet19	CIFAR100 ResNet19	ImageNet-1K ResNet18/34	ImageNet-1K E-SpikeFormer-S	CIFAR10-DVS Resnet-19	CIFAR10-DVS VGG11
Optimizer	SGD	SGD	SGD	LAMB	SGD	SGD
Epoch	200	200	200	200	300	300
LR	0.1	0.1	0.01	0.0048	0.1	0.05
Batch	128	128	2048	2048	128	128
WD	5×10^{-5}	5×10^{-5}	1×10^{-4}	5×10^{-2}	5×10^{-4}	5×10^{-4}
Momentum	0.9	0.9	0.9	-	0.9	0.9
LR Schedule	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine
Loss	CE	CE	CE	CE	TET	TET
Number of Bits B	16	16	16	16	16	16
Timestep T	1	1	1	1	10	10
Number of Terms D	$\{2, 4\}$	$\{2, 4\}$	$\{2, 4\}$	4	2	2
Threshold V_{th}	2^{-4}	2^{-4}	2^{-4}	2^{-4}	2^{-4}	2^{-4}
Probability p	0.5	0.5	0.5	0.5	0.5	0.5

ImageNet-1K. ImageNet-1K contains 1,281,167 training images and 50,000 validation images distributed across 1,000 object categories. For the SEW-ResNet18 and SEW-ResNet34 experiments, the training images are augmented using random resized cropping to 224×224 pixels and random horizontal flipping with a probability of 0.5. During validation, images are resized to 256 pixels and subsequently center-cropped to 224×224 pixels. For E-SpikeFormer-S, we follow the timm ImageNet training pipeline, using RandomResizedCrop(224), random horizontal flipping, RandAugment (`rand-m9-mstd0.5-inc1`), and random erasing with a probability of 0.25 in pixel mode. Its validation pipeline uses bicubic resizing followed by center cropping to 224×224 pixels. For all ImageNet-1K experiments, images are normalized using the standard ImageNet mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$.

CIFAR10-DVS. CIFAR10-DVS is a neuromorphic dataset converted from CIFAR10 using a Dynamic Vision Sensor. It contains 10,000 event-stream samples, with 9,000 samples used for training and 1,000 samples for testing. Each sample is represented by 10 temporal frames with two event-polarity channels. Each frame is resized from 128×128 to 48×48 pixels. During training, random horizontal flipping with a probability of 0.5 and random spatial translation are applied. The horizontal and vertical translation offsets are independently sampled from $[-5, 5]$ pixels, and the same spatial transformation is consistently applied to all temporal frames of each sample. During evaluation, the temporal frames are only resized to 48×48, without random augmentation or normalization. VGG11 and ResNet-19 are employed as the backbone architectures.

D.2 Implementation Setup

Our experiments are designed to answer four questions. First, we evaluate whether the complete QuantLIF design generalizes across different dataset scales, network architectures, and input modalities. Second, we isolate the effect of the neuron design and examine whether QuantLIF improves accuracy while maintaining competitive inference energy. Third, we disable low-spike suppression and investigate how the codebook parameters B and D affect the trade-off between quantization precision and computational cost. Finally, with the quantization configuration fixed to $B = 16$ and $D = 2$, we study whether low-amplitude spikes can be safely removed and whether stochastic suppression during training is more effective than inference-only removal or deterministic suppression during training.

Unless otherwise specified, the QuantLIF results in the main comparisons use the complete method, including operation-constrained quantization and low-spike suppression applied to quantized spike amplitudes with $V_{\text{th}} = 2^{-4}$. During training, these low-amplitude spikes are randomly suppressed with probability $p = 0.5$, while during inference they are deterministically removed. In contrast, the accuracy–energy comparison adopts a controlled setting with a shared backbone and training recipe, so that only the neuronal representation is changed. The B–D ablation disables low-spike suppression entirely, while the suppression ablation fixes $B = 16$ and $D = 2$ and varies only the suppression strategy and threshold. Therefore, the absolute accuracies reported in different

tables correspond to different controlled experimental settings and should be interpreted within their respective protocols.

Low-spike suppression is applied in the forward pass, while the local Jacobian of the unsuppressed output is preserved during backpropagation through a straight-through estimator. Thus, suppressing a spike does not directly mask its local surrogate gradient. For CIFAR10-DVS, we set the membrane decay factor to $\beta = 0.25$ and use the input-scaled integration rule $\mathbf{U}[t] = \beta\mathbf{H}[t-1] + (1-\beta)\mathbf{X}[t]$. We apply a hard reset with $V_{\text{reset}} = 0$: the membrane potential is reset to zero wherever the post-suppression spike output is nonzero, and otherwise $\mathbf{H}[t] = \mathbf{U}[t]$. Therefore, spikes suppressed to zero do not trigger a reset.

E More Results

E.1 Learning Curves of QuantLIF

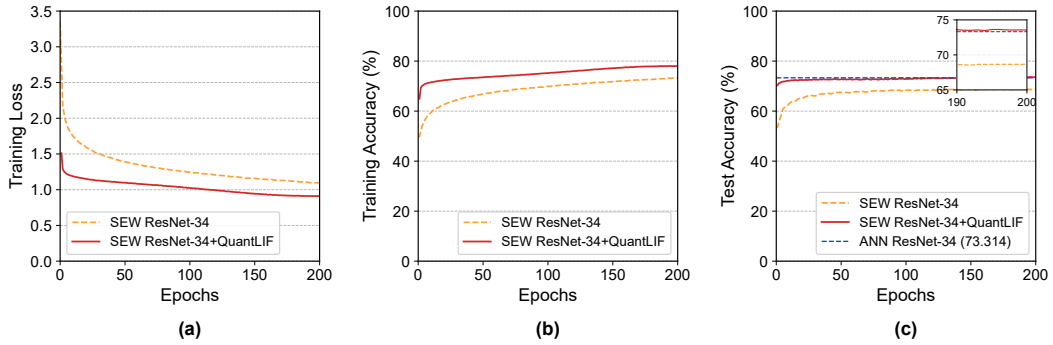


Figure 6: Learning curves of (a) training loss, (b) training accuracy, and (c) test accuracy on ImageNet-1K.

Figure 6 compares the learning dynamics of SEW ResNet-34 with and without QuantLIF on ImageNet-1K. QuantLIF consistently achieves lower training loss and higher training accuracy across the 200-epoch training schedule, indicating more effective optimization. These improvements emerge early and persist throughout training, yielding a lower final loss and a higher final training accuracy. The improved training fit is accompanied by a sustained gain in test accuracy. QuantLIF outperforms the baseline by approximately five percentage points at the end of training and reaches a level comparable to the ANN ResNet-34 reference (73.3%)[1]. The enlarged view of the final epochs further shows that this performance is maintained near convergence.

E.2 The Impact of Signed PoT Terms on Codebook Design

Table 8: Classification accuracy (%) of codebooks composed with unsigned or signed power-of-two terms.

Dataset	D	**Unsigned**	**Signed**	Δ (%)
CIFAR10	2	96.80	**96.97**	+0.17
	4	96.87	**96.98**	+0.11
CIFAR100	2	82.04	81.77	−0.27
	4	81.38	**81.93**	+0.55
Mean	–	89.27	**89.41**	+0.14

Table 8 evaluates the effect of including signed power-of-two terms in the quantization codebook on CIFAR10 and CIFAR100 for $D \in \{2, 4\}$. Averaging equally over the two values of D, the

[1]Refer to `https://pytorch.org/hub/pytorch_vision_resnet/`

signed codebook improves accuracy by 0.14 percentage points on each dataset, yielding the same mean gain across all four configurations. Improvements are observed in three of the four settings, supporting a modest average advantage of signed codebooks within the operation budgets.

E.3 Analysis of Confusion Matrix

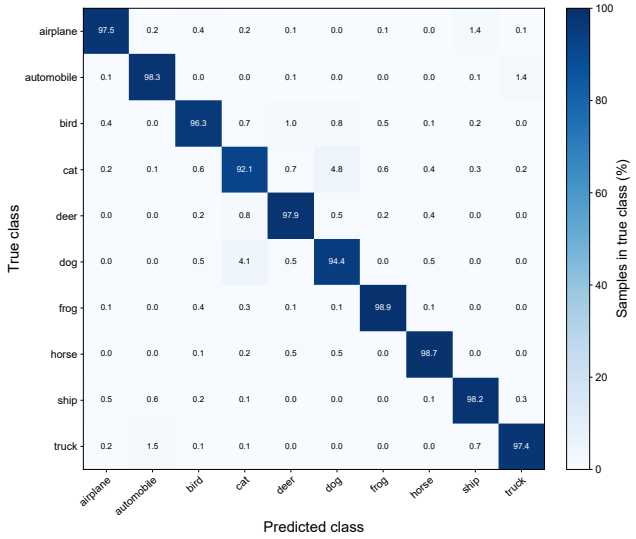


Figure 7: Visualization of Confusion matrix on CIFAR10.

Figure 7 presents the confusion matrix of QuantLIF on CIFAR-10, where each entry denotes the percentage of samples from a given true class assigned to a predicted class. The strong diagonal concentration indicates high classification performance across all ten categories, with per-class recall ranging from 92.1% to 98.9% and exceeding 96% for eight classes. QuantLIF achieves particularly high recall for frogs (98.9%), horses (98.7%), and automobiles (98.3%). The largest off-diagonal entries occur between cats and dogs: 4.8% of cat samples are classified as dogs, while 4.1% of dog samples are classified as cats. This confusion accounts for a substantial portion of the errors in these two categories, which have the lowest recall. Smaller confusions occur between automobiles and trucks, at 1.4% and 1.5% in the respective directions. Overall, the matrix demonstrates strong discrimination across the CIFAR-10 categories, while identifying closely related animal classes as the primary source of the most pronounced classification errors.

E.4 More visualization of spike distribution

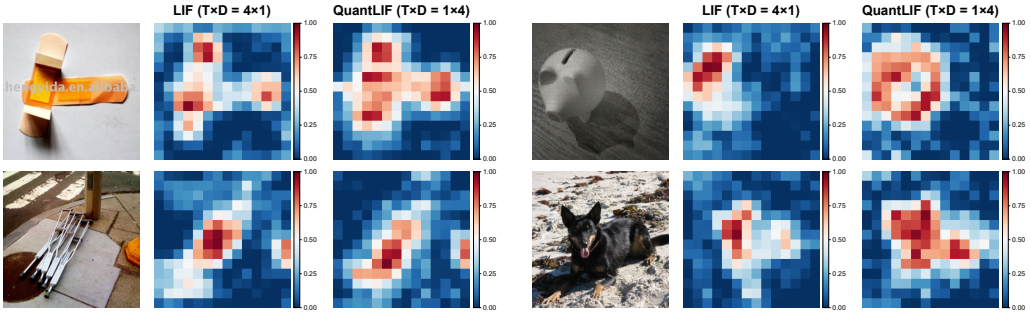


Figure 8: More visualization comparisons of spike distributions between vanilla LIF ($T \times D = 4 \times 1$) and QuantLIF ($T \times D = 1 \times 4$).

Figure 8 visualizes additional examples comparing the spatial distributions of spike activity between vanilla LIF ($T \times D = 4 \times 1$) and QuantLIF ($T \times D = 1 \times 4$). Across the displayed examples, QuantLIF generally produces broader and more spatially coherent responses over foreground objects, whereas the responses of vanilla LIF are more concentrated on individual object regions. This difference is particularly apparent in the piggy-bank and dog examples, where QuantLIF responses extend across larger portions of the objects and exhibit greater continuity between activated regions. The additional examples thus reinforce the observation in the main text that QuantLIF captures object-related spatial structure across diverse appearances and backgrounds. The visualizations provide qualitative evidence that QuantLIF can retain informative spatial responses with reduced temporal depth, consistent with the motivation for enriching spike representations through finer amplitude quantization.

E.5 Detailed Performance on Distribution-Shift Data

Table 9: Classification accuracy (%) of QuantLIF on CIFAR10/100-C under 15 corruption types at five severity levels. For each dataset, the last column averages over corruption types, the Mean row averages over severity levels, and their intersection reports the overall accuracy.

Dataset	Severity	Gaussian noise	Shot noise	Impulse noise	Defocus blur	Glass blur	Motion blur	Zoom blur	Snow	Frost	Fog	Brightness	Contrast	Elastic transform	Pixelate	JPEG	**Mean**
CIFAR10-C	1	91.11	93.59	96.01	96.92	71.94	95.58	95.16	95.53	94.85	96.76	96.86	96.67	95.18	95.29	88.67	**93.34**
	2	82.54	90.57	94.22	96.55	73.16	93.11	94.55	92.39	92.72	96.50	96.61	96.14	94.97	90.39	82.46	**91.13**
	3	70.04	79.62	91.90	95.55	74.47	89.20	92.78	92.82	88.48	96.11	96.39	95.81	93.65	85.01	79.53	**88.09**
	4	63.39	73.71	83.06	92.89	56.75	89.14	90.40	91.05	87.27	95.12	96.13	95.38	85.71	55.35	76.11	**82.10**
	5	56.37	63.65	68.21	84.03	60.32	84.93	85.36	88.78	81.92	88.09	95.56	93.33	76.86	37.93	70.97	**75.75**
	Mean	**72.69**	**80.23**	**86.68**	**93.19**	**67.33**	**90.39**	**91.65**	**92.11**	**89.05**	**94.52**	**96.31**	**95.47**	**89.27**	**72.79**	**79.55**	**86.08**
CIFAR100-C	1	69.86	74.79	79.98	81.56	30.13	78.00	77.17	77.56	74.70	81.10	81.68	81.43	76.24	77.63	62.23	**73.60**
	2	58.50	69.82	77.14	80.33	32.08	73.46	75.56	68.41	68.08	79.57	81.22	79.79	76.94	69.39	51.78	**69.47**
	3	46.64	56.20	73.73	77.42	34.71	67.94	71.98	70.37	58.25	77.60	80.37	78.81	73.81	62.28	47.79	**65.19**
	4	40.09	49.49	64.99	72.33	20.72	68.46	68.05	68.14	56.29	72.80	79.21	77.21	62.32	34.60	44.58	**58.62**
	5	34.94	39.67	51.83	58.47	23.71	61.40	62.13	61.96	47.53	57.35	76.62	71.84	48.02	11.63	39.86	**49.80**
	Mean	**50.01**	**57.99**	**69.53**	**74.02**	**28.27**	**69.85**	**70.98**	**69.29**	**60.97**	**73.68**	**79.82**	**77.82**	**67.47**	**51.11**	**49.25**	**63.34**

Table 9 provides a detailed breakdown of the QuantLIF results summarized in Table 5 across 15 corruption types and 5 severity levels on CIFAR10-C and CIFAR100-C. Following the evaluation protocol in the main text, the models are trained on clean CIFAR10/100 and evaluated without target-domain adaptation. Averaged over all corruption types and severity levels, QuantLIF achieves accuracies of 86.08% and 63.34%, respectively. As severity increases from level 1 to level 5, the accuracy averaged over corruption types decreases from 93.34% to 75.75% on CIFAR10-C and from 73.60% to 49.80% on CIFAR100-C. The breakdown also reveals substantial variation across corruption types. Brightness and contrast yield the highest severity-averaged accuracies on both datasets, indicating comparatively strong tolerance to these appearance changes. Glass blur remains particularly challenging, with corresponding mean accuracies of 67.33% and 28.27%, while pixelation causes pronounced degradation at the highest severity level. These results complement the aggregate evaluation by characterizing where QuantLIF maintains predictive performance under distribution shifts and identifying severe spatial distortions as a remaining robustness challenge.

F Theoretical Energy Consumption

Following established conventions in the spiking neural network literature (Panda et al., 2020), we assume that multiply–accumulate (MAC) and accumulate (AC) operations are implemented on 45 nm hardware (Horowitz, 2014). The corresponding per-operation energy costs are taken to be

$EC_{\text{MAC}} = 4.6\,\text{pJ}$ and $EC_{\text{AC}} = 0.9\,\text{pJ}$, respectively. For QuantLIF, we count the actual number of signed power-of-two terms used by each transmitted spike, rather than assigning the maximum budget D to every nonzero spike. Let $d(\mathbf{S})$ denote the actual number for value $\mathbf{S}$, with $d(0) = 0$. Each term requires one signed accumulation per outgoing synaptic connection after shift-based reparameterization. The number of AC operations in layer l is defined as

$$\text{ACs}^l = \sum_{t=1}^{T} \sum_{i} a_i^l \, d\big(\mathbf{S}_i^{l-1}[t]\big)\,, \tag{35}$$

where a_i^l is the number of synaptic contributions associated with input element i. Both statistics are computed over all timesteps after deterministic low-spike suppression. If no spike is transmitted, the AC count is zero.

The estimated synaptic arithmetic energy per inference is

$$EC_{\text{QuantLIF}} = EC_{\text{MAC}}\, N_{\text{MAC}} + EC_{\text{AC}} \sum_{l \in \text{SNN}} \text{ACs}^l, \tag{36}$$

where N_{MAC} counts all remaining MAC operations over the complete inference. This formulation accounts for both the removal of low-amplitude spikes and the variation in decomposition length across transmitted amplitudes. Reported energy values are obtained by averaging the per-sample estimates over the test set.

G The Assistance of Large Language Models

Large language models are used solely to assist in reviewing code for potential logical errors and refining the clarity and fluency of the manuscript. The authors retain full responsibility for the implementation, experimental results, and scientific claims presented in this work.