BARS-FORMER: ROBUST SPIKING TRANSFORMER WITH BIPOLAR ACTIVATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Brain-inspired spiking neural networks (SNNs) are promising for energy-efficient computing, but their performance has historically lagged behind that of ANNs. Recent spiking Transformers have substantially narrowed this accuracy gap on large-scale vision benchmarks such as ImageNet-1K. However, in this work we show that a substantial robustness gap remains under corruption and distribution shift. To probe the underlying cause, we find that vanilla spiking self-attention suffers from significantly more severe gradient attenuation on augmented samples than on clean samples, which prevents spiking Transformers from fully exploiting data augmentation for robust representation learning. To address this limitation, we propose **Bipolar Activated Robust Spiking Transformer (BARS-Former)**, which integrates two key designs: Bipolar Accurate Spiking Self-Attention (BASA) and Ternary Pre-spike (Pre-TS). BASA adopts ternary-spike queries and real-valued keys to enlarge the effective gradient range and generate accurate bipolar attention scores, thereby improving gradient propagation from augmented samples. Pre-TS replaces the binary pre-spike activations in residual branches and downsampling modules with ternary neurons to alleviate activation truncation and gradient attenuation. Experiments on robustness benchmarks for image classification and segmentation demonstrate that BARS-Former substantially improves robustness against corruptions and distribution shifts. Further analyses show that BARS-Former effectively mitigates gradient attenuation on augmented samples, converges to flatter minima, and maintains more stable attention maps under input noise. Code is available at `https://anonymous.4open.science/r/BARS-Former-130F`.

1 INTRODUCTION

Spiking Neural Networks (SNNs) enable energy-efficient computation through sparse, event-driven processing and compatibility with neuromorphic hardware, which is particularly suitable for edge deployment and have attracted much research attention in the past few years (Xu et al., 2024; Che et al., 2024; Zheng et al., 2024; Zhu et al., 2024; Song et al., 2026; Sun et al., 2026; Hao et al., 2026). Particularly, while the performance of SNNs has historically lagged behind that of ANNs, recent spiking Transformers, such as Spikformer (Zhou et al., 2023b; 2024b), QKFormer (Zhou et al., 2024a), and Spike-driven Transformers (SDTs) (Yao et al., 2023; 2024; 2025), have made much progress in achieving accuracy comparable to artificial neural networks (ANNs) on large-scale vision recognition benchmarks. However, in this paper we show that a substantial robustness gap between spiking and ANN Transformers remains. As shown in Figure 1, despite achieving comparable clean-data accuracy on ImageNet-1K, SNN- and ANN-based Transformers can exhibit a gap of more than 10% in average accuracy on ImageNet-C. In many practical edge applications where SNNs are preferred, edge systems often encounter environmental noise, or distribution shift, which makes robustness as important as high accuracy on clean data.

Vision Transformers are generally more robust than convolutional neural networks (Paul & Chen, 2022), but often require strong data augmentation to achieve competitive performance under limited training data (Touvron et al., 2021). Data augmentation strategies, typically formulated as input perturbations, are also important for improving the robustness of spiking Transformers. However, spike activations are quantized and non-differentiable. Even with surrogate gradients, they remain susceptible to saturation and gradient attenuation, especially when learning from augmented samples. As the key component of spiking Transformers, Vanilla Spiking Self-Attention (VSSA) (Zhou

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.

et al., 2023b) represents queries, keys, and values with binary spike trains, which enables spike-compatible computation but restricts attention to low-precision unipolar interactions (Wang et al., 2025). Moreover, the cascaded derivatives of multiple binary spike neurons along the gradient path can further aggravate gradient attenuation of VSSA. Consequently, existing spiking Transformers cannot fully exploit data augmentation, which substantially limits their robustness.

In this paper, we investigate the underlying cause of the degraded robustness of spiking Transformers. By comparing the gradients on clean and augmented samples during training, we reveal that VSSA is more susceptible to gradient attenuation on augmented samples. To address this issue, we propose **Bipolar Accurate Spiking Self-Attention (BASA)**, which uses ternary-spike queries and real-valued keys to enlarge the effective gradient range and produce accurate bipolar attention scores. This design alleviates gradient attenuation on augmented samples while preserving spike-compatible addition/subtraction-based computation. Furthermore, we introduce **Ternary Pre-spike (Pre-TS)** by replacing the first binary spike activation layer with ternary spiking neurons (Guo et al., 2024) in residual branches and downsampling modules, which further enhances gradient propagation and finally improves performance. Based on BASA and Pre-TS, we build **Bipolar Activated Robust Spiking Transformer (BARS-Former)**, with substantially enhanced robustness. Our contributions are summarized as follows:

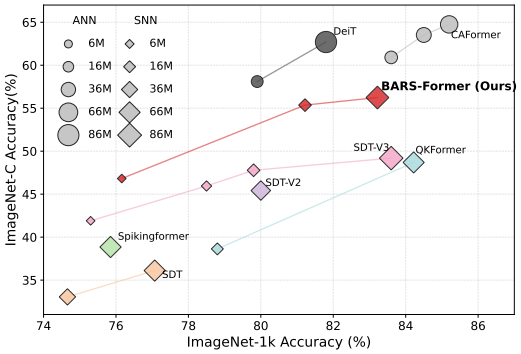


Figure 1: Robustness versus clean accuracy on ImageNet. The horizontal axis denotes ImageNet-1K accuracy, while the vertical axis denotes ImageNet-C accuracy averaged over 15 corruption types and 5 severity levels. The proposed BARS-Former substantially narrows the robustness gap between spiking Transformers and ANN Transformers, achieving state-of-the-art robustness among SNNs.

- We evaluate the robustness of spiking Transformers and show a substantial robustness gap between spiking and ANN Transformers. We reveal the gradient attenuation problem of VSSA on augmented samples and identify it as a bottleneck for robust representation learning of spiking Transformers.
- We propose BARS-Former, which integrates BASA to produce accurate bipolar attention and Pre-TS to expand the effective gradient range, while maintaining spike-compatible computation.
- We validate BARS-Former on robustness benchmarks for image classification and segmentation tasks. The results demonstrate that it effectively mitigates gradient attenuation on augmented samples and yields substantially improved robustness.

2 Related Works

Recent spiking Transformers have achieved competitive performance on large-scale visual recognition. Spikingformer (Zhou et al., 2023a), Spikformer V2 (Zhou et al., 2024b), SpikingResformer (Shi et al., 2024), QKFormer (Zhou et al., 2024a), and the Spike-driven Transformer series (Yao et al., 2023; 2024; 2025) improve architectural design and scalability. However, these methods generally rely on VSSA with binary queries, keys, and values, resulting in low-precision, non-negative attention scores.

Several recent works have explored robustness-related designs for spiking Transformers. TDFormer introduces top-down attention control to enhance temporal information integration (Zhu et al., 2025), while SpiLiFormer employs lateral inhibition to suppress irrelevant responses and improve feature discrimination (Zheng et al., 2025). However, they do not evaluate robustness on large-scale corruption and distribution-shift benchmarks.

BSA (Wang et al., 2025) and A^{2}OS^{2}A (Guo et al., 2025) focus on modifying spiking self-attention. BSA adopts ternary neurons for both queries and keys to introduce bipolar interactions, but its attention matrix is still quantized and further normalized by Shiftmax, which may introduce additional approximation error. A^{2}OS^{2}A uses binary spikes, ReLU activations, and ternary neurons for the query, key, and value branches, respectively. Although it employs a floating-point key branch, ReLU

truncates negative responses, resulting in a non-negative attention matrix. In contrast, our BASA employs ternary queries and real-valued keys to produce high-precision bipolar attention scores without softmax or Shiftmax approximations, while preserving efficient addition/subtraction-based computation. Table 2 provides a comparison of different self-attention mechanisms.

3 PRELIMINARY

3.1 SPIKING NEURON LAYER

Inspired by the brain, SNNs operate in a spike-based processing paradigm that transmits information via temporally sparse binary spikes, where neural computation involves only accumulation operations. The charging equation of the membrane potential for neurons in the l-th layer at timestep t can be expressed as

$$U^l[t] = X^l[t]W^l + B^l, \tag{1}$$

where W^l and B^l represent the synaptic weights and bias of the l-th layer, respectively. $X^l[t]$ is the spike tensor of the l-th layer. For a **binary spike neuron**, the output spike is generated by

$$X^{l+1}[t] = \mathrm{H}\left(U^l[t] - V_{th}\right). \tag{2}$$

For a **ternary spike neuron**, the output spike is generated by

$$X^{l+1}[t] = \mathrm{Sign}\left(U^l[t]\right) \cdot \mathrm{H}\left(|U^l[t]| - V_{th}\right). \tag{3}$$

After spike generation, the membrane potential is reset by

$$U^l[t+1] = U^l[t] - V_{th} \cdot X^{l+1}[t]. \tag{4}$$

Compared with binary spike neurons, ternary spike neurons have a larger representational capacity (Guo et al., 2024; 2025). In the following sections, we denote the binary spike neuron layer and the ternary spike neuron layer as

$$X^{l+1} = \mathcal{BSN}\left(X^l W^l\right) \quad \text{and} \quad X^{l+1} = \mathcal{TSN}\left(X^l W^l\right), \tag{5}$$

respectively. Integer spiking neurons were introduced in Luo et al. (2024); Yao et al. (2025). During training, they emit integer-valued spikes, which can be interpreted as aggregating multiple spike events within a single timestep. This temporally compressed training scheme improves training efficiency and has been shown to achieve better performance (Luo et al., 2024; Yao et al., 2025). During inference, each integer-valued spike is decomposed into multiple standard binary or ternary spikes distributed over successive timesteps.

3.2 VANILLA SPIKING SELF-ATTENTION (VSSA)

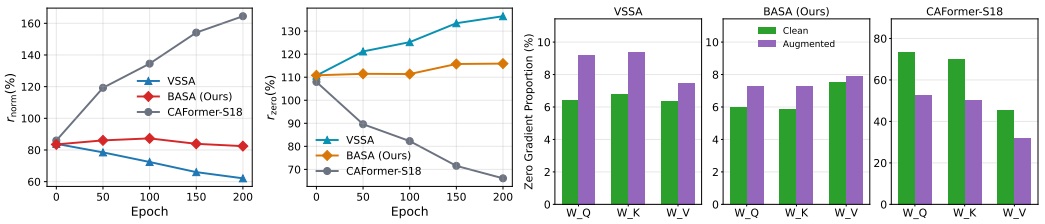
VSSA (Zhou et al., 2023b) uses a softmax-free, addition-only spike-compatible formulation. Given input feature X, the query, key, and value are all represented as binary spikes

$$Q = \mathcal{BSN}\left(\mathrm{BN}(XW_Q)\right), \quad K = \mathcal{BSN}\left(\mathrm{BN}(XW_K)\right), \quad V = \mathcal{BSN}\left(\mathrm{BN}(XW_V)\right), \tag{6}$$

where $\mathrm{BN}(\cdot)$ denotes batch normalization. The attention matrix is computed in a fully spiking manner $A = QK^\top$. Since both Q and K are binary spike tensors, A only contains non-negative integer values. A scaling factor s is used to regulate the magnitude of the attention output. The spike-efficient VSSA is formulated as

$$\mathrm{VSSA}(Q, K, V) = \mathcal{BSN}\left(s \cdot AV\right). \tag{7}$$

Although VSSA is effective for spiking Transformers, its attention matrix is restricted to quantized non-negative integers. Such a narrow and discrete value range limits the precision of attention scores and fails to capture bipolar relations between tokens. To address this limitation, we consider using ternary-spike queries and non-spiking real-valued keys, so that the resulting attention scores can be both positive and negative while preserving floating-point precision.



(a) (b) (c)

Figure 2: Gradient analysis of different self-attentions. VSSA and BASA are evaluated using 5M BARS-Former with 8 timesteps, while ANN self-attention is evaluated using CAFormer-S18. (a) Evolution of the augmented-to-clean gradient norm ratio of $W_{Q/K/V}$ during training. (b) Evolution of the augmented-to-clean zero-gradient proportion ratio of $W_{Q/K/V}$ during training. (c) Zero-gradient proportion of the three attention mechanisms on clean and augmented data at the 150th epoch.

4 Gradient Attenuation of Spiking Self-Attention

We first empirically analyze the difference between gradients on clean and augmented samples for VSSA and ANN self-attention. Given a clean sample x_i and its augmented counterpart $\mathcal{A}(x_i)$, where $\mathcal{A}(\cdot)$ denotes the data augmentation operator, the gradients with respect to the parameter tensor W_j are defined as

$$g_{i,j}^{\text{clean}} = \frac{\partial \mathcal{L}(x_i)}{\partial W_j}, \quad g_{i,j}^{\text{aug}} = \frac{\partial \mathcal{L}(\mathcal{A}(x_i))}{\partial W_j}, \tag{8}$$

where W_j denotes the j-th parameter tensor, such as the query, key, and value projection matrices W_Q, W_K, or W_V. The mean gradient norm is computed as

$$\bar{g}^{\text{clean}} = \frac{1}{NL}\sum_{i=1}^{N}\sum_{j=1}^{L}\left\|g_{i,j}^{\text{clean}}\right\|_F, \quad \bar{g}^{\text{aug}} = \frac{1}{NL}\sum_{i=1}^{N}\sum_{j=1}^{L}\left\|g_{i,j}^{\text{aug}}\right\|_F, \tag{9}$$

where N denotes the number of samples and L denotes the total number of parameter tensors under analysis. We then define the augmented-to-clean gradient norm ratio as

$$r_{\text{norm}} = \bar{g}^{\text{aug}} / \bar{g}^{\text{clean}}. \tag{10}$$

Moreover, we define the zero-gradient proportion as the ratio of gradient elements whose absolute values are smaller than a threshold $\epsilon = 10^{-4}$:

$$\bar{p}^{\text{clean}} = \frac{1}{NL}\sum_{i=1}^{N}\sum_{j=1}^{L}\frac{1}{D_j}\sum_{k=1}^{D_j}\mathbb{I}\left(|g_{i,j,k}^{\text{clean}}| < \epsilon\right), \quad \bar{p}^{\text{aug}} = \frac{1}{NL}\sum_{i=1}^{N}\sum_{j=1}^{L}\frac{1}{D_j}\sum_{k=1}^{D_j}\mathbb{I}\left(|g_{i,j,k}^{\text{aug}}| < \epsilon\right), \tag{11}$$

where D_j denotes the number of elements in parameter tensor W_j, $g_{i,j,k}$ denotes the k-th element of gradient tensor $g_{i,j}$, and $\mathbb{I}(\cdot)$ is the indicator function. The augmented-to-clean zero-gradient proportion ratio is defined as

$$r_{\text{zero}} = \bar{p}^{\text{aug}} / \bar{p}^{\text{clean}}. \tag{12}$$

A larger r_{zero} indicates more severe gradient attenuation on augmented samples compared with clean samples. Figure 2 shows the evolution of r_{norm} and r_{zero} during training for different attention mechanisms, as well as the statistics of $\bar{p}^{\text{clean}}$ and $\bar{p}^{\text{aug}}$ at the 150th epoch. The analysis is performed only on the attention projection weights, including W_Q, W_K, and W_V. Specifically, CAFormer-S18 employs ANN self-attention, while VSSA and the proposed BASA are evaluated using 5M BARS-Former variants with 8 timesteps.

As shown in Figure 2a, CAFormer-S18 maintains a higher r_{norm} than VSSA, and the ratio increases during training. In contrast, VSSA exhibits a lower r_{norm} with a decreasing trend. As shown in Figure 2b, CAFormer-S18 has a lower r_{zero} that decreases during training, whereas VSSA shows a higher r_{zero} with an increasing trend. These observations suggest that ANN self-attention can effectively preserve gradient information from augmented samples, enabling more effective optimization with data augmentation. In contrast, VSSA suffers from gradient attenuation on

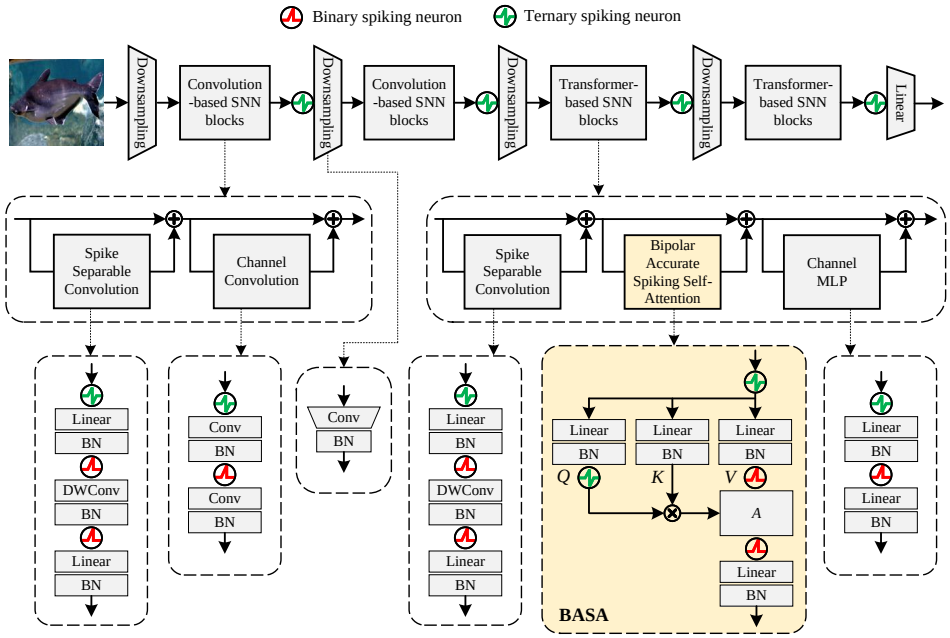


Figure 3: Overview of the proposed BARS-Former.

augmented samples, preventing spiking Transformers from fully exploiting data augmentation for robustness improvement. Consistent with this observation, CAFormer-S18 achieves substantially stronger robustness than the VSSA-based spiking Transformer. The proposed BASA effectively alleviates gradient attenuation. As shown in Figure 2c, BASA has lower zero-gradient proportion on augmented samples ($\bar{p}^{\text{aug}}$) compared with VSSA, indicating that more gradient information is retained during training with augmented data.

5 Proposed Method

5.1 Overview of BARS-Former

As shown in Figure 3, the proposed BARS-Former follows a hierarchical multi-stage architecture, where the spatial resolution is reduced and the channel dimension is increased across stages. Previous studies have empirically shown that several architectural designs, including overlapping patch embedding, hierarchical feature representations, replacing linear layers with convolutional operations in MLP blocks, and depth-wise convolution (DWConv), are beneficial for improving the robustness of visual Transformers (Mao et al., 2022; Tian et al., 2022). Among existing spiking Transformers, SDT-V3 (Yao et al., 2025) adopts a similar multi-stage design and provides a strong robust baseline, as shown in Figure 1. Therefore, we build BARS-Former upon the SDT-V3 architecture and modify its attention and activation components. Moreover, we adopt integer-valued spiking neurons (Luo et al., 2024; Yao et al., 2025) for BARS-Former. Additional experiments in Appendix B show that using integer-valued spiking neurons can achieve better performance.

5.2 Bipolar Accurate Spiking Self-Attention (BASA)

To overcome the limitation of VSSA, we propose Bipolar Accurate Spiking Self-Attention (BASA). Different from VSSA, BASA replaces the binary spike activation in the query branch with a ternary spike neuron, removes the spike activation in the key branch, and keeps binary spike activation in the value branch:

$$Q = \mathcal{TSN}\left(\mathrm{BN}(XW_Q)\right), \quad K = \mathrm{BN}(XW_K), \quad V = \mathcal{BSN}\left(\mathrm{BN}(XW_V)\right). \tag{13}$$

The attention matrix is computed as $A = QK^\top$. Unlike vanilla SSA, whose attention scores are restricted to non-negative integers, the attention matrix of BASA can theoretically cover the whole real

domain, i.e., $A \in \mathbb{R}$. This enables BASA to represent both positive and negative token correlations with higher numerical precision. The BASA output is formulated as

$$\mathrm{BASA}(Q, K, V) = \mathcal{BSN}\left(s \cdot AV\right). \tag{14}$$

BASA still preserves multiplication-free computation, as all operations can be performed using spike-compatible addition and subtraction. Note that the scaling factor s can be absorbed into the key projection during re-parameterization.

5.3 Ternary Pre-spike (Pre-TS)

Membrane Shortcut (MS) is a widely used residual connection in spiking Transformers, as it ensures identity mapping and has achieved strong performance in recent architectures (Zhou et al., 2023a; Yao et al., 2023; 2024; 2025; Guo et al., 2025; Fang et al., 2025; Hu et al., 2024). In the residual branch of MS, the input feature is first passed through a spike activation layer before the subsequent transformation, analogous to the pre-activation design in ANNs (He et al., 2016). We refer to this operation as *pre-spike*. However, pre-spike inevitably truncates the feature values and may hinder gradient propagation, especially when binary spike neurons are used.

To alleviate this issue, we propose Ternary Pre-spike (Pre-TS), which replaces the original binary pre-spike with a ternary spiking layer. Compared with binary pre-spike, Pre-TS provides bipolar responses and a larger representational range, while still maintaining spike-compatible computation.

Experimental results show that Pre-TS complements BASA and improves accuracy on both clean and robustness benchmarks. Moreover, BARS-Former equipped with Pre-TS and BASA can converge to a flatter local minima.

5.4 Gradient Comparison between VSSA and BASA

Based on equation 6 and equation 7, the gradients of VSSA with respect to the projection weights W_Q, W_K, and W_V can be written as follows. For simplicity, the batch normalization operation is absorbed into the corresponding projection weights. Let $G_{\mathrm{VSSA}} = \frac{\partial \mathcal{L}_{\mathrm{VSSA}}}{\partial (AV)}$ denote the gradient propagated from the VSSA output. Then we have

$$\frac{\partial \mathcal{L}_{\mathrm{VSSA}}}{\partial W_Q} = X^\top \left[(G_{\mathrm{VSSA}} \mathcal{BSN}^\top (XW_V) \mathcal{BSN}(XW_K)) \odot \mathcal{BSN}'(XW_Q) \right], \tag{15}$$

$$\frac{\partial \mathcal{L}_{\mathrm{VSSA}}}{\partial W_K} = X^\top \left[(\mathcal{BSN}(XW_V) G_{\mathrm{VSSA}}^\top \mathcal{BSN}(XW_Q)) \odot \mathcal{BSN}'(XW_K) \right], \tag{16}$$

$$\frac{\partial \mathcal{L}_{\mathrm{VSSA}}}{\partial W_V} = X^\top \left[(\mathcal{BSN}(XW_K) \mathcal{BSN}^\top (XW_Q) G_{\mathrm{VSSA}}) \odot \mathcal{BSN}'(XW_V) \right]. \tag{17}$$

These equations show that all three projection branches in VSSA are gated by binary spike activation and their surrogate gradient $\mathcal{BSN}'(\cdot)$. As a result, the effective gradient range is small, which may increase the risk of gradient attenuation during direct training. According to equation 13 and equation 14, the gradients of BASA can be expressed as

$$\frac{\partial \mathcal{L}_{\mathrm{BASA}}}{\partial W_Q} = X^\top \left[(G_{\mathrm{BASA}} \mathcal{BSN}^\top (XW_V)(XW_K)) \odot \mathcal{TSN}'(XW_Q) \right], \tag{18}$$

$$\frac{\partial \mathcal{L}_{\mathrm{BASA}}}{\partial W_K} = X^\top [\mathcal{BSN}(XW_V) G_{\mathrm{BASA}}^\top \mathcal{TSN}(XW_Q)], \tag{19}$$

$$\frac{\partial \mathcal{L}_{\mathrm{BASA}}}{\partial W_V} = X^\top [((XW_K) \mathcal{TSN}^\top (XW_Q) G_{\mathrm{BASA}}) \odot \mathcal{BSN}'(XW_V)], \tag{20}$$

where $G_{\mathrm{BASA}} = \frac{\partial \mathcal{L}_{\mathrm{BASA}}}{\partial (AV)}$. Compared with VSSA, BASA provides a wider numerical and gradient range. The ternary query branch preserves both positive and negative responses, while removing the spike activation from the key branch enables gradients to propagate through real-valued keys without being gated by the surrogate gradient. Consequently, the gradients of $W_{Q/K/V}$ retain higher precision and a larger dynamic range, making BASA less susceptible to gradient attenuation.

6 EXPERIMENTS

6.1 MAIN RESULTS ON IMAGENET

We compare BARS-Former with representative ANN and SNN Transformer models. Following the original SDT-V3 training recipe (Yao et al., 2025), we additionally adopt Mixup with $\alpha = 0.8$ and Color Jitter, while keeping all other settings unchanged, since we find that Mixup is beneficial for improving robustness. The final data augmentation strategy includes AutoAugment, Label Smoothing, Random Erasing, Color Jitter, and Mixup. All experiments are conducted on NVIDIA V100 GPUs. All models are trained on the ImageNet-1K (Deng et al., 2009) training set and evaluated on the ImageNet validation set, ImageNet-C (Hendrycks & Dietterich, 2019), ImageNet-A (Hendrycks et al., 2021b), and ImageNet-R (Hendrycks et al., 2021a). Inference input resolution is 224×224 for all datasets. More training details are given in Appendix D. The results are reported in Table 1. For ImageNet-C, we report the average accuracy over 15 corruption types and 5 severity levels. BARS-Former significantly outperforms other spiking Transformers on robustness benchmarks. We further train SDT-V3 using the same data augmentation recipe as BARS-Former (Appendix C). The results show that SDT-V3 still underperforms BARS-Former even with the same training settings.

6.2 COMPARISON WITH SPIKING SELF-ATTENTION VARIANTS

We investigate the attention design in BARS-Former by replacing BASA with representative spiking self-attention variants. All models use the same BARS-Former architecture with 5M parameters and are trained under the same settings with 8 timesteps and Pre-TS. The only difference lies in the spiking self-attention mechanism.

As shown in Table 2, BARS-Former equipped with BASA achieves the best performance on both clean ImageNet and robustness benchmarks. This improvement mainly comes from the accurate bipolar attention scores produced by BASA. Compared with A^2OS^2A (Guo et al., 2025), whose attention scores are restricted to the non-negative domain, BASA can model both positive and negative token correlations. Compared with BSA (Wang et al., 2025), which uses Shiftmax for attention normalization, BASA avoids additional normalization approximation and obtains better performance.

6.3 COMPARISON OF TRAINING LOSS CURVE AND LOSS LANDSCAPE

Training Loss Curve Comparison. We further analyze the optimization behavior of different designs under the 5M model setting with 8 timesteps. As shown in Figure 4a, we compare the training loss dynamics of variants with different attention mechanisms and pre-spike designs, including VSSA, A^2OS^2A, BASA, VSSA+Pre-TS, and BASA+Pre-TS. Among these variants, BASA consistently achieves lower training loss than VSSA, indicating that accurate bipolar attention improves the fitting ability. After introducing Pre-TS, the training loss is further reduced. It is worth noting that the reported training loss is computed with data augmentation. Therefore, lower training loss suggests that BASA and Pre-TS can better fit the augmented training data.

Loss Landscape Comparison. We visualize the loss landscapes of the SDT-V3 baseline and BARS-Former on the ImageNet-1K under the same 5M-parameter and 8-timestep setting according to Li et al. (2018). As shown in Figure 4b, the SDT-V3 baseline with VSSA converges to a sharper local minimum, where the loss increases rapidly around the converged solution. In contrast, BARS-Former converges to a flatter local minimum, as shown in Figure 4c. This indicates that BARS-Former is less sensitive to parameter perturbation, which contributes to its improved generalization and robustness.

6.4 COMPARISON OF ATTENTION MAP

Inspired by Guo et al. (2023), we analyze the stability of attention maps under different corruption severities. We conduct the comparison under the BARS-Former setting with 5M parameters and 8 timesteps, and evaluate three configurations: VSSA, BASA, and BASA+Pre-TS. The VSSA configuration is equivalent to the SDT-V3 baseline, while BASA+Pre-TS corresponds to the complete BARS-Former. For each configuration, we compute the cosine similarity between the attention maps of clean images from the ImageNet-1K validation set and their corresponding Gaussian noise corrupted images from ImageNet-C. The attention maps are computed from rows of $QK^\top$ across all

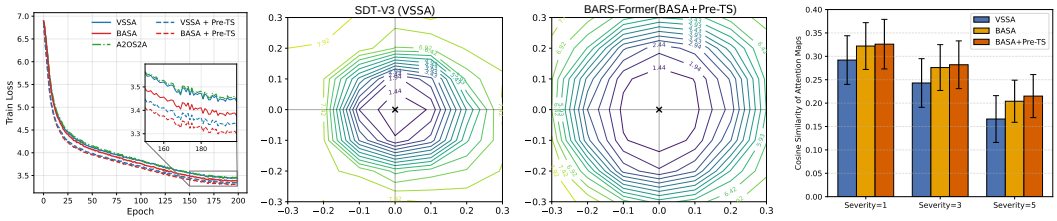
Table 1: Accuracy (%) comparison on ImageNet-1K (IN-1k) and robustness benchmarks including ImageNet-C (IN-C), ImageNet-A (IN-A) and ImageNet-R (IN-R). † denotes models trained using the same data augmentation recipe as BARS-Former. Power is the average theoretical energy consumption for predicting a sample from IN-1k test set (Zhou et al., 2024a).

Model	Param(M)	Timestep	Power(mJ)	IN-1k	IN-C	IN-A	IN-R
DeiT (Touvron et al., 2021)	22	–	21.4	79.90	58.10	18.90	42.20
	87	–	80.5	81.80	62.70	27.40	44.90
CAFormer (Yu et al., 2024)	26	–	18.9	83.60	60.91	36.59	48.12
	39	–	36.7	84.50	63.52	44.12	50.61
	56	–	60.7	85.20	64.75	49.37	50.75
ResNet-34 (Zhang et al., 2025)	22	4	16.1	71.30	40.37	2.08	31.26
Max-Former (Fang et al., 2025)	16	4	4.9	77.82	40.67	7.65	39.66
	29	4	7.5	79.86	44.46	12.32	41.80
	64	4	14.9	82.39	49.06	20.07	42.41
QKFormer (Zhou et al., 2024a)	16	4	15.1	78.80	38.63	5.52	38.20
	65	4	38.9	84.22	48.69	16.53	44.88
SpikingResformer (Shi et al., 2024)	11	4	2.5	74.34	38.65	5.37	35.08
	18	4	3.4	75.95	41.35	7.47	37.57
	36	4	5.5	77.24	42.80	10.11	38.70
	60	4	8.8	78.77	46.16	14.01	40.18
Spikingformer (Zhou et al., 2023a)	66	4	13.7	75.85	38.85	10.96	41.57
SpiLiFormer (Zheng et al., 2025)	69	4	44.2	85.82	49.59	14.68	44.10
SDT (Yao et al., 2023)	36	4	5.5	74.66	33.05	5.80	33.27
	66	4	6.1	77.07	36.11	7.92	33.79
SDT-V2 (Yao et al., 2024)	55	4	52.4	80.00	45.42	9.99	40.07
SDT-V3 (Yao et al., 2025)	5	4	1.7	75.30	41.90	7.39	36.19
	10	4	3.0	78.50	45.95	12.11	39.49
	19	4	5.9	79.80	47.78	16.01	40.85
	83	4	19.1	83.60	49.19	18.41	41.87
	83	8	30.9	84.00	45.59	11.35	37.68
SDT-V3† (Yao et al., 2025)	5	8	3.2	74.74	44.27	8.51	38.66
	19	4	6.0	79.30	50.90	16.48	43.89
	19	8	9.6	80.19	53.03	20.85	45.02
BARS-Former (Ours)	5	8	3.0	76.16	46.82	11.20	40.07
	19	4	6.6	80.59	53.97	21.59	44.75
	19	8	11.3	81.22	55.36	24.04	**46.36**
	77	8	43.4	83.23	**56.04**	**25.23**	45.42

Table 2: Comparison of spiking self-attention mechanisms and their accuracy (%). $\mathcal{B} = \{0, 1\}$ denotes binary spikes, $\mathcal{T} = \{-1, 0, +1\}$ denotes ternary spikes, $\mathbb{R}^+$ denotes ReLU activation, and $\mathbb{R}$ denotes non-spike real-value.

Method	Q	K	V	Attention score($QK^\top$)	IN-1k	IN-C	IN-A	IN-R
VSSA (Zhou et al., 2023b)	$\mathcal{B}$	$\mathcal{B}$	$\mathcal{B}$	$\mathbb{Z}^+$	75.60	45.30	9.73	39.16
A^{2}OS^{2}A (Guo et al., 2025)	$\mathcal{B}$	$\mathbb{R}^+$	$\mathcal{T}$	$\mathbb{R}^+$	75.52	44.85	9.89	38.92
BSA (Wang et al., 2025)	$\mathcal{T}$	$\mathcal{T}$	$\mathbb{R}$	$\mathbb{Z}$	74.85	45.47	10.33	39.42
BASA (Ours)	$\mathcal{T}$	$\mathbb{R}$	$\mathcal{B}$	$\mathbb{R}$	**76.16**	**46.82**	**11.20**	**40.07**

layers, without averaging over attention heads. As shown in Figure 4d, BASA consistently achieves higher attention map similarity than VSSA across different severity levels. When BASA is further combined with Pre-TS, the complete BARS-Former achieves the highest similarity. These results show that BARS-Former can preserve more stable attention patterns under input noise.



(a) (b) (c) (d)

Figure 4: Analysis under the 5M model setting with 8 timesteps. (a) Training loss comparison of variants with different attention mechanisms and pre-spike designs. (b) and (c) visualize the loss landscapes of SDT-V3 baseline and BARS-Former on ImageNet-1K validation set, respectively. (d) Cosine similarity of attention maps between clean and noisy images across all layers.

6.5 Ablation Study

We conduct ablation studies to analyze the effect of different configurations in the query, key, value branches and pre-spike operation. As shown in Table 3, the proposed configuration, i.e., ternary queries, real-valued keys, binary values, and ternary pre-spike, achieves the best performance on both clean ImageNet and robustness benchmarks.

The results show that using bipolar attention score is generally beneficial. Compared with the vanilla VSSA baseline, whose attention scores are restricted to non-negative integers due to binary Q and binary K, variants with bipolar $QK^\top$ usually have better robustness. Pre-TS consistently improves performance. These results indicate that BASA and Pre-TS are complementary, for BASA improves the attention representation, while Pre-TS enhances feature propagation in the residual branch.

Experiments reported in Appendix B show that BASA and Pre-TS are also effective when applied to the Spikingformer (Zhou et al., 2023a) backbone.

Table 3: Ablation study based on the 5M BARS-Former with 8 timesteps. All results are reported as accuracy (%). $\mathcal{B}$, $\mathcal{T}$, $\mathbb{R}$, and $\mathbb{R}^+$ denote binary spikes, ternary spikes, real-valued activation, and ReLU activation, respectively. IN-C (S=5) denotes the accuracy on ImageNet-C with severity level 5.

Q	K	V	Pre-spike	IN-1k	IN-C(S=5)	IN-A	IN-R
$\mathcal{B}$	$\mathcal{B}$	$\mathcal{B}$	$\mathcal{B}$	74.74	24.90	8.55	38.69
$\mathcal{B}$	$\mathcal{B}$	$\mathcal{B}$	$\mathcal{T}$	75.60	26.40	9.73	39.16
$\mathcal{B}$	$\mathbb{R}^+$	$\mathcal{B}$	$\mathcal{B}$	74.55	25.27	8.79	38.82
$\mathcal{B}$	$\mathbb{R}$	$\mathcal{B}$	$\mathcal{B}$	74.90	25.75	9.96	38.73
$\mathcal{T}$	$\mathcal{T}$	$\mathcal{B}$	$\mathcal{B}$	74.77	26.93	9.91	38.45
$\mathcal{T}$	$\mathbb{R}^+$	$\mathcal{B}$	$\mathcal{B}$	75.06	26.51	10.39	39.22
$\mathcal{T}$	$\mathbb{R}$	$\mathcal{B}$	$\mathcal{B}$	74.87	26.61	10.44	38.65
$\mathcal{T}$	$\mathbb{R}$	$\mathcal{B}$	$\mathcal{T}$	**76.16**	**28.62**	**11.20**	**40.07**

6.6 Results on Semantic Segmentation Task

Following Yao et al. (2024; 2025), we use checkpoints pretrained on ImageNet as the backbone and initialization, equip the model with Semantic FPN (Kirillov et al., 2019), train it on the Cityscapes (Cordts et al., 2016), and evaluate its robustness on Cityscapes-C (Michaelis et al., 2019). As shown in Table 7, the proposed BARS-Former achieves higher mIoU than SDT-V3 on both clean and corrupted data.

Table 4: Comparison of mIoU (%) on Cityscapes (City) and Cityscapes-C (City-C) using 19M-parameter models with 8 timesteps.

Backbone	City	City-C
SDT-V3 (Yao et al., 2025)	70.26	35.18
BARS-Former (Ours)	**72.43**	**41.53**

7 Conclusion

We reveal the robustness gap between spiking and ANN Transformers and identify gradient attenuation as the key bottleneck. We propose BARS-Former with BASA and Pre-TS to improve gradient propagation and robust representation learning. BARS-Former substantially improves the performance and achieves state-of-the-art robustness among SNN models on robustness benchmarks.

AI USE STATEMENT

In this work, we used generative AI tools for only to aid and polish the writing. We have not used generative AI tools for retrieval and discovery of related work, research ideation or execution, drafting sections of the paper, generating synthetic datasets, proving mathematical claims, or any other research-related tasks. We have reviewed all AI-assisted work. We checked that all LLM-modified sentences do not alter the intended meaning or technical content of the manuscript. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of our results. Codes for reproducing all the results in our paper are available at `https://anonymous.4open.science/r/BARS-Former-130F`. Details of our experimental setup, including model architectures, data preprocessing, and hyperparameter settings, are provided in Section 6 and Appendix D. Extensive results are provided in Appendix B and C.

REFERENCES

Kaiwei Che, Zhengyu Ma, Yifan Huang, Peng Xue, Li Yuan, Timothée Masquelier, Wei Fang, and Yonghong Tian. Efficiently training time-to-first-spike spiking neural networks from scratch. *arXiv preprint arXiv:2410.23619*, 2024.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Yuetong Fang, Deming Zhou, Ziqing Wang, Hongwei Ren, ZeCui Zeng, Lusong Li, Shibo Zhou, and Renjing Xu. Spiking transformers need high frequency information. *arXiv preprint arXiv:2505.18608*, 2025.

Yong Guo, David Stutz, and Bernt Schiele. Robustifying token attention for vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17557–17568, 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.

Yufei Guo, Xiaode Liu, Yuanpei Chen, Weihang Peng, Yuhan Zhang, and Zhe Ma. Spiking transformer: Introducing accurate addition-only spiking self-attention for transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24398–24408, 2025.

Zecheng Hao, Yifan Huang, Zijie Xu, Wenxuan Liu, Yuanhong Tang, Zhaofei Yu, and Tiejun Huang. Rethinking snn online training and deployment: Gradient-coherent learning via hybrid-driven lif model. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016.

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

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021b.

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, 2024.

Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6399–6408, 2019.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Neural Information Processing Systems*, 2018.

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.

Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 12042–12051, 2022.

Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019.

Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 36, pp. 2071–2081, 2022.

Xinyu Shi, Zecheng Hao, and Zhaofei Yu. Spikingresformer: Bridging resnet and vision transformer in spiking neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Ruiqi Song, Dujun Nie, Siyu Teng, Baiyong Ding, Xiaotong Zhang, Dong Li, Chenming Zhang, Yuchen Li, Hangbin Wu, and Long Chen. Spikevla: Vision-language-action models with spiking neural networks. *arXiv preprint arXiv:2606.27807*, 2026.

Pengfei Sun, Zhe Su, Jascha Achterberg, Giacomo Indiveri, Dan F. M. Goodman, and Danyal Akarca. Algorithm–hardware co-design of neuromorphic networks with dual memory pathways. *Nature Machine Intelligence*, 8(6):901–912, June 2026. ISSN 2522-5839. doi: 10.1038/s42256-026-01255-3. URL https://doi.org/10.1038/s42256-026-01255-3.

Rui Tian, Zuxuan Wu, Qi Dai, Han Hu, and Yu-Gang Jiang. Deeper insights into the robustness of vits towards common corruptions. *arXiv preprint arXiv:2204.12143*, 2022.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Shuai Wang, Malu Zhang, Jingya Wang, Dehao Zhang, Yimeng Shan, Jieyuan Zhang, Yichen Xiao, Honglin Cao, Haonan Zhang, Zeyu Ma, Yang Yang, and Haizhou Li. Bipolar self-attention for spiking transformers. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Mengting Xu, De Ma, HuaJin Tang, Qian Zheng, and Gang Pan. Feel-snn: Robust spiking neural networks with frequency encoding and evolutionary leak factor. *Advances in Neural Information Processing Systems*, 37:91930–91950, 2024.

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

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1SIBN5Xyw7.

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.

Weihao Yu, Chenyang Si, Pan Zhou, Mi Luo, Yichen Zhou, Jiashi Feng, Shuicheng Yan, and Xinchao Wang. Metaformer baselines for vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):896–912, 2024. doi: 10.1109/TPAMI.2023.3329173.

Jiahong Zhang, Kexin Wang, Man Yao, Han Xu, Peng Zhou, Bo Xu, and Guoqi Li. Enhancing robustness of spiking neural networks through retina-like coding and memory-based neurons. *Neural Networks*, pp. 107950, 2025.

Hanle Zheng, Zhong Zheng, Rui Hu, Bo Xiao, Yujie Wu, Fangwen Yu, Xue Liu, Guoqi Li, and Lei Deng. Temporal dendritic heterogeneity incorporated with spiking neural networks for learning multi-timescale dynamics. *Nature Communications*, 15(1):277, 2024.

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

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

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

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

Zhaokun Zhou, Kaiwei Che, Wei Fang, Keyu Tian, Yuesheng Zhu, Shuicheng Yan, Yonghong Tian, and Li Yuan. Spikformer v2: Join the high accuracy club on imagenet with an snn ticket. *arXiv preprint arXiv:2401.02020*, 2024b.

Yaoyu Zhu, Jianhao Ding, Tiejun Huang, Xiaodong Xie, and Zhaofei Yu. Online stabilization of spiking neural networks. In *The Twelfth International Conference on Learning Representations*, 2024.

Zizheng Zhu, Yingchao Yu, Zeqi Zheng, Zhaofei Yu, and Yaochu Jin. Tdformer: A top-down attention-controlled spiking transformer. *arXiv preprint arXiv:2505.15840*, 2025.

A Block Structures of BASA and Pre-TS

We provide the block structures of the proposed BASA and Pre-TS modules in Figure 5 and Figure 6, respectively. Figure 5 illustrates the differences between VSSA and BASA. Figure 6 presents the integration of Pre-TS into a typical spiking Transformer block with membrane shortcut.

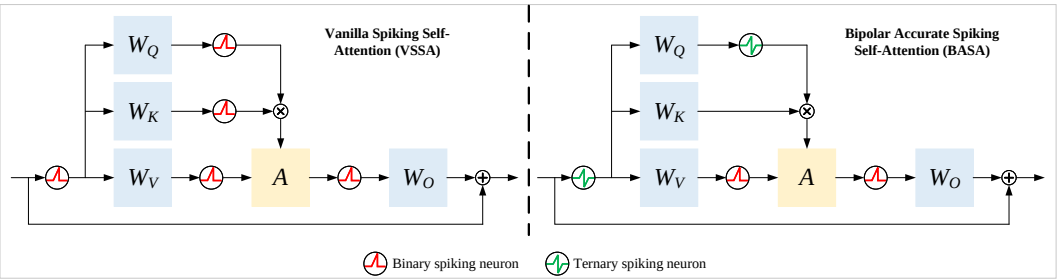


Figure 5: Comparison between VSSA and the proposed BASA. VSSA adopts binary spike neurons for all Q, K, and V branches, leading to non-negative integer attention score. In contrast, BASA employs ternary spike neurons for Q, removes spike activation in K, and keeps binary spikes for V, enabling real-valued bipolar attention score. BN layers have been re-parameterized into the corresponding linear projections.

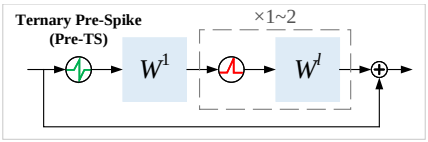


Figure 6: Spiking Transformer block with membrane shortcut (MS) and ternary pre-spike (Pre-TS). The typical residual branch employing MS can be viewed as a stack of pre-spike + linear operation (W^l) layers, where W^l can be a convolutional layer + BN or a fully connected layer + BN. Pre-TS replaces the first spike layer in the branch with a ternary spike layer.

B Evaluation on Spikingformer Backbone

We show that the proposed BASA and Pre-TS are also effective on another spiking Transformer backbone. Specifically, we evaluate them on Spikingformer (Zhou et al., 2023a), which follows a ViT-like single-stage architecture comprising a tokenizer and stacked Transformer blocks while keeping a constant feature resolution (Dosovitskiy et al., 2021).

As shown in Table 5, the original Spikingformer employs conventional LIF neurons, whereas all others are trained using integer-valued spiking neurons (Luo et al., 2024; Yao et al., 2025). Replacing LIF neurons with integer-valued spiking neurons substantially improves performance on IN-1k, IN-C, and IN-A. Increasing the number of timesteps from 4 to 8 further improves most metrics. BASA and Pre-TS consistently improve the performance of the corresponding Spikingformer baseline, demonstrating that the proposed method is not limited to a specific backbone, but can generalize to different spiking Transformer architectures.

Table 5: Accuracy (%) comparison of VSSA, BASA, and Pre-TS on the Spikingformer backbone.

Model	Neuron	Param (M)	Timestep	IN-1k	IN-C	IN-A	IN-R
Spikingformer (VSSA)	LIF	17	4	72.45	30.19	7.01	37.36
Spikingformer (VSSA)	Integer Spike	17	4	74.37	31.38	10.37	36.90
Spikingformer (VSSA)	Integer Spike	17	8	75.67	33.54	13.27	36.54
Spikingformer (BASA)	Integer Spike	17	8	76.17	34.93	13.88	37.06
Spikingformer (BASA+Pre-TS)	Integer Spike	17	8	**76.57**	**35.15**	**14.69**	**37.70**

C Comparison under Same Data Augmentation Settings

The original implementation of SDT-V3 does not employ several common data augmentation strategies, such as Mixup and color jitter (Yao et al., 2025). For a fair comparison, we retrain SDT-V3 using the same data augmentation recipe as BARS-Former, denoted as SDT-V3† in Table 6. Applying the same data augmentation improves the robustness of SDT-V3, but its accuracy on clean ImageNet-1K slightly decreases. Nevertheless, under the same model sizes, timesteps, and training settings, BARS-Former consistently outperforms SDT-V3† across all benchmarks. These results demonstrate that the proposed architecture provides performance gains beyond those introduced by training settings, particularly in terms of robustness.

Table 6: Accuracy (%) comparison between BARS-Former and SDT-V3 under the same training recipe. † denotes models trained using the same data augmentation recipe as BARS-Former. Power is the average theoretical energy consumption for predicting a sample from IN-1k test set.

Model	Param(M)	Timestep	Power(mJ)	IN-1k	IN-C	IN-A	IN-R
SDT-V3	5	4	1.7	75.30	41.90	7.39	36.19
SDT-V3†	5	8	3.2	74.74	44.27	8.51	38.66
BARS-Former	5	8	3.0	**76.16**	**46.82**	**11.20**	**40.07**
SDT-V3	10	4	3.0	78.50	45.95	12.11	39.49
SDT-V3	19	4	5.9	79.80	47.78	16.01	40.85
SDT-V3†	19	4	6.0	79.30	50.90	16.48	43.89
BARS-Former	19	4	6.0	**80.59**	**53.97**	**21.59**	**44.75**
SDT-V3†	19	8	9.6	80.19	53.03	20.85	45.02
BARS-Former	19	8	11.3	**81.22**	**55.36**	**24.04**	**46.36**

Table 7: Comparison of mIoU (%) on Cityscapes (City) and Cityscapes-C (City-C) using 19M-parameter models with 8 timesteps. † denotes model trained using the same data augmentation recipe as BARS-Former.

Backbone	City	City-C
SDT-V3	70.26	35.18
SDTV3†	67.72	38.12
BARS-Former	**72.43**	**41.53**

D More Implementation Details

Here we provide implementation details for the experiments. Detailed training hyperparameters, data augmentation configurations, and model configurations are summarized in Table 8, Table 9 and Table 10, respectively.

Training configuration. Models are trained from scratch for 200 epochs using distributed training on four GPUs. The batch size is set to 256 per GPU with one gradient accumulation step, resulting in an effective batch size of 1024. We employ the LAMB optimizer with a weight decay of 0.05. The base learning rate is set to 6×10^{-4} and scaled linearly according to the effective batch size. The learning rate is adjusted at each training iteration, with 5 warm-up epochs and a minimum learning rate of 1×10^{-6}. We adopt a rectangular surrogate gradient for spiking neuron, and gradient clipping is not applied. Automatic mixed-precision is used during training.

Data augmentation We use Mixup with $\alpha = 0.8$, while CutMix is disabled. Label smoothing is set to 0.1 and is incorporated into the soft targets generated by Mixup. The model is optimized using the soft-target cross-entropy loss. The remaining augmentation settings follow the default training recipe (Yao et al., 2025), including RandAugment with the policy of `rand-m9-mstd0.5-inc1` and random erasing with a probability of 0.25. During training, we use random resized cropping to

224×224. For evaluation, images are resized such that the shorter side is 256 pixels and then center-cropped to 224×224. We use the standard ImageNet normalization with mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$.

Table 8: Training configuration.

Configuration	Value
Training epochs	200
Optimizer	LAMB
Weight decay	0.05
Base learning rate	6×10^{-4}
Learning rate schedule	Cosine decay
Warm-up epochs	5
Minimum learning rate	1×10^{-6}
Batch size (per GPU)	256
Number of GPUs	4
Gradient accumulation steps	1
Mixed precision training	Enabled
Label smoothing	0.1

Table 9: Data augmentation configuration.

Augmentation	Configuration
Random resized crop	224×224
RandAugment	`rand-m9-mstd0.5-inc1`
Mixup	$\alpha = 0.8$
CutMix	Disabled
Mixup probability	1.0
Label smoothing	0.1
Color jitter	0.4
Random erasing probability	0.25
Random erasing mode	Pixel-level
Number of erased regions	1

Table 10: Architectural configurations of BARS-Former variants.

Configuration	5M	19M	77M
Number of stage	4	4	3
Stage dimension	$[32, 64, 128, 192]$	$[64, 128, 256, 360]$	$[128, 256, 512]$
Attention head	8	8	8
MLP expansion ratio	4	4	4

E Detailed Results on Corruption Benchmarks

We further report detailed results for every corruption type on ImageNet-C and Cityscapes-C in Table 11 and Table 12.

Table 11: Comparison of accuracy (%) on ImageNet-C (IN-C). Each corruption column reports the average accuracy over severity levels 1–5. IN-C is the overall average over the 15 corruption types and five severity levels.

	Noise			Blur				Weather				Digital				
Model	Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright.	Contrast	Elastic	Pixel	JPEG	IN-C
CAFormer (26M)	64.09	62.47	63.19	53.15	39.45	55.99	48.68	60.37	64.85	70.06	77.52	69.58	56.82	60.27	67.21	60.91
CAFormer (39M)	66.42	65.03	65.65	56.22	40.40	60.04	53.62	62.27	66.99	70.68	78.91	71.19	58.77	66.96	69.66	63.52
CAFormer (56M)	68.19	66.82	67.55	57.15	41.90	60.83	56.01	64.00	67.71	73.15	79.56	72.69	60.33	64.84	70.52	64.75
Max-Former (16M, $T = 4$)	36.13	34.97	35.57	33.14	25.41	36.41	34.33	35.08	36.66	50.12	67.52	47.21	44.15	40.25	53.13	40.67
Max-Former (29M, $T = 4$)	39.74	38.08	37.94	36.67	27.31	39.98	38.03	39.69	42.48	54.62	70.49	53.55	45.23	45.99	57.07	44.46
Max-Former (64M, $T = 4$)	48.12	45.82	47.55	38.51	29.43	43.51	41.38	44.87	47.92	59.26	73.25	58.65	48.46	49.88	59.36	49.06
QKFormer (16M, $T = 4$)	29.04	28.47	21.00	30.51	24.04	35.90	35.51	31.20	33.93	52.91	69.52	48.05	42.66	42.63	54.05	38.63
QKFormer (65M, $T = 4$)	42.61	41.59	35.52	42.19	30.97	45.20	46.18	40.82	43.66	61.31	76.89	60.96	48.64	50.45	63.32	48.69
SpikingResformer (11M, $T = 4$)	32.43	31.31	27.32	30.17	24.13	36.95	32.78	36.63	34.64	45.99	64.18	43.50	42.40	46.28	50.97	38.65
SpikingResformer (18M, $T = 4$)	36.56	35.31	31.42	33.72	25.20	38.79	35.90	37.72	37.19	49.64	66.63	47.92	44.59	46.58	53.04	41.35
SpikingResformer (36M, $T = 4$)	36.79	35.10	30.44	34.26	27.48	39.97	36.76	39.88	40.63	51.27	68.62	49.58	46.86	49.18	55.22	42.80
SpikingResformer (60M, $T = 4$)	42.04	41.25	37.87	36.92	29.17	42.63	39.10	43.23	43.45	53.16	70.46	57.20	46.97	51.66	57.21	46.16
Spikingformer (66M, $T = 4$)	29.80	28.79	23.30	25.48	28.06	33.62	31.37	38.12	43.42	54.49	67.11	45.92	45.54	37.89	49.88	38.85
SpiLiFormer (69M, $T = 4$)	41.51	40.53	33.21	41.47	32.92	48.76	47.53	42.15	44.31	62.75	78.34	58.94	50.94	54.36	66.13	49.59
SDT (36M, $T = 4$)	21.72	20.97	14.50	23.09	22.51	30.81	30.45	28.17	32.86	44.95	62.84	37.48	40.52	37.57	47.33	33.05
SDT (66M, $T = 4$)	25.04	23.69	15.96	28.11	24.17	33.78	34.28	30.45	35.37	48.94	65.28	42.03	42.52	41.54	50.54	36.11
SDT-V2 (55M, $T = 4$)	39.55	39.21	41.04	38.21	28.46	43.15	39.32	39.24	40.63	51.55	71.25	55.67	45.72	50.82	57.52	45.42
SDT-V3 (5M, $T = 4$)	36.45	35.77	37.01	35.41	27.24	40.51	37.26	35.22	36.61	47.53	67.00	50.78	44.10	44.16	53.36	41.90
SDT-V3 (10M, $T = 4$)	40.81	40.35	42.57	39.04	28.91	44.65	41.49	39.89	39.67	50.95	70.34	54.65	45.98	52.16	57.75	45.95
SDT-V3 (19M, $T = 4$)	44.67	43.74	46.45	40.80	28.83	45.89	42.83	41.65	42.10	52.40	71.65	56.18	47.68	53.40	58.50	47.78
SDT-V3 (83M, $T = 4$)	45.75	43.88	42.16	40.14	28.55	46.00	43.78	46.05	48.91	61.24	75.39	61.37	47.39	47.03	60.14	49.19
SDT-V3 (83M, $T = 8$)	40.26	38.69	34.43	37.34	26.69	42.34	39.77	42.72	44.21	59.43	74.85	59.23	46.02	41.22	56.64	45.59
BARS-Former (5M, $T = 8$)	47.83	45.62	48.02	34.50	25.95	42.41	38.00	47.50	50.25	56.55	68.04	57.73	42.63	42.99	54.30	46.82
BARS-Former (19M, $T = 4$)	56.14	53.17	55.61	43.06	30.62	51.68	45.81	53.46	56.08	63.21	72.81	63.15	48.52	56.07	60.10	53.97
BARS-Former (19M, $T = 8$)	56.46	54.43	56.31	45.44	31.94	51.88	47.39	55.17	57.12	65.15	73.40	64.46	49.45	59.57	62.21	55.36
BARS-Former (77M, $T = 8$)	59.17	57.31	58.45	45.07	32.20	51.40	46.44	53.83	60.08	64.55	76.72	67.66	49.09	54.88	63.76	**56.04**

Table 12: Comparison of mIoU (%) on Cityscapes (City) and Cityscapes-C (City-C) using 19M-parameter models with 8 timesteps. Each corruption column reports the average mIoU over severity levels 1–5. City-C reports the overall average over all 15 corruption types and five severity levels. † denotes the model trained using the same data augmentation recipe as BARS-Former.

		Noise			Blur				Weather				Digital				
Backbone	City	Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright.	Contrast	Elastic	Pixel	JPEG	City-C
SDT-V3	70.26	10.77	12.55	16.53	44.46	37.15	49.15	13.39	19.73	18.25	49.72	63.94	51.58	66.82	47.16	26.54	35.18
SDT-V3†	67.72	23.92	27.79	26.23	46.27	38.55	48.67	13.17	22.02	23.70	51.82	61.97	57.38	63.26	46.52	20.51	38.12
BARS-Former	**72.43**	19.56	22.21	23.11	51.16	43.21	54.07	18.22	23.80	28.46	57.81	68.73	60.50	68.87	54.04	29.25	**41.53**