RISK-AWARE GRADIENT ALLOCATION FOR SPIKING NEURAL NETWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Surrogate-gradient learning enables the direct training of spiking neural networks (SNNs), but conventional threshold-centered surrogates concentrate backward gradient energy on the membrane states most susceptible to spike flipping. We recast surrogate design as a risk-aware gradient allocation problem and introduce Gradient Risk Exposure (GRE) to quantify this mismatch between gradient concentration and state instability. By minimizing GRE subject to a Kullback–Leibler penalty on deviations from the baseline allocation, we derive the Risk-Calibrated Surrogate Gradient (RCSG), a closed-form exponential calibration applicable to arbitrary baseline surrogates. RCSG redistributes gradient energy toward more stable membrane states without altering the forward neuronal dynamics or requiring an auxiliary robustness loss. On three static-image classification benchmarks, RCSG improves adversarial robustness while maintaining or improving clean accuracy, and its gains hold consistently across surrogate families, network architectures, and attack gradient estimators.

1 INTRODUCTION

Spiking neural networks (SNNs) communicate through discrete spikes while retaining temporal information in their neuronal states, making them a computational model for event-driven neuromorphic systems (Maass, 1997; Neftci et al., 2019; Tavanaei et al., 2019). Their discontinuous firing dynamics, however, pose a fundamental challenge for gradient-based optimization: the derivative of the spike generation function is zero almost everywhere and undefined at the firing threshold (Eshraghian et al., 2023; Kaiser et al., 2020). Surrogate-gradient (SG) learning circumvents this issue by replacing the non-differentiable firing derivative with a smooth proxy during backpropagation (Wu et al., 2018; Neftci et al., 2019). This strategy has enabled the direct training of increasingly deep and expressive SNN architectures, including residual networks and Transformer-based models (Fang et al., 2021; Zhou et al., 2022). Nevertheless, training performance remains highly sensitive to the choice and shape of the surrogate function.

Most existing SG methods aim to improve gradient approximation or optimization stability. Early approaches employ fixed smooth, triangular, or piecewise surrogate functions to provide non-zero gradients near the firing threshold (Shrestha & Orchard, 2018; Zenke & Vogels, 2021), while more recent methods refine this mechanism by adapting the surrogate function itself or its associated optimization process (Wang et al., 2023; Lian et al., 2023). Dspike improves the approximation of the discrete firing function through finite-difference-inspired learning (Li et al., 2021), and Temporal Efficient Training improves temporal credit assignment via time-dependent optimization objectives (Deng et al., 2022). MPD-aware approaches further exploit the membrane-potential distribution to adapt surrogate support, while A^2SG adjusts both surrogate width and asymmetry to improve temporal gradient consistency (Jiang et al., 2025; Kang et al., 2026). Despite this progress, existing SG designs overlook how gradient contributions should be distributed across unstable membrane states.

This issue bears directly on robustness. A spike output changes only when a perturbation pushes the membrane potential across the firing threshold, so threshold-neighboring states are inherently more prone to spike flipping (Shrestha & Orchard, 2018). Recent studies have linked this phenomenon to SNN robustness. TGO identifies threshold neighboring membrane states as a major source of adversarial vulnerability and improves robustness by driving membrane potentials away from the

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.

threshold (Wang et al., 2026b). MPD-SGR similarly examines the interaction between membrane-potential distributions and gradient-active regions, regularizing their overlap to reduce adversarial sensitivity (Jiang et al., 2026b). Other approaches improve robustness through adversarial training (Ding et al., 2022; Kundu et al., 2021), modified encoding or neuronal dynamics (Ding et al., 2024; Xu et al., 2024), gradient sparsity regularization (Liu et al., 2024), or certified defenses (Liang et al., 2022). Collectively, these methods establish the importance of membrane-state geometry for robust SNN learning, but they intervene primarily in forward neuronal states or training objectives, leaving the role of the backward gradient distribution itself comparatively underexplored.

In this work, we view surrogate-gradient learning as the allocation of backward gradient energy across membrane states, where excessive concentration on unstable states may expose learning signals to spike-instability risk. We quantify this exposure through Gradient Risk Exposure (GRE), defined as the expected spike-flip risk under the normalized gradient-energy distribution. To reduce GRE while retaining fidelity to a baseline allocation, we formulate a risk–distortion objective with a Kullback–Leibler penalty. Its optimal solution yields an exponential risk reweighting, from which we derive the Risk-Calibrated Surrogate Gradient (RCSG). Applicable to arbitrary baseline surrogates, RCSG modifies only the backward pass, preserving forward neuronal dynamics without requiring an auxiliary robustness objective.

Our contributions are threefold:

- We reinterpret surrogate-gradient learning as the allocation of backward gradient energy across membrane states and introduce Gradient Risk Exposure (GRE) to quantify how much of this energy is assigned to spike-unstable regions.
- We formulate gradient allocation as a risk-minimization problem under a KL-based fidelity constraint, derive its unique optimal solution, and map it analytically to RCSG.
- Through experiments, we validate that reallocating gradient energy toward more stable membrane states reduces risk exposure and improves adversarial robustness.

2 RELATED WORK

2.1 SURROGATE GRADIENT LEARNING IN SNNS

Surrogate-gradient (SG) learning approximates the non-differentiable firing derivative to enable gradient-based training of SNNs (Wu et al., 2018; Zenke & Ganguli, 2018; Neftci et al., 2019). SLAYER relates surrogate derivatives to the probability of spike-state changes under membrane perturbations (Shrestha & Orchard, 2018), while subsequent work analyzes the robustness of SG learning to surrogate design choices and its connections to stochastic differentiation (Zenke & Vogels, 2021; Gygax & Zenke, 2025). Beyond fixed surrogates, Dspike refines the approximation through finite differences (Li et al., 2021), while Temporal Efficient Training improves optimization through temporal loss design (Deng et al., 2022). Adaptive smoothing, learnable surrogates, and MPD-aware methods adjust gradient support to evolving neuronal states (Wang et al., 2023; Lian et al., 2023; Jiang et al., 2025; Nair et al., 2026), and A2SG adapts surrogate width and asymmetry to improve temporal gradient consistency (Kang et al., 2026). However, these methods do not explicitly account for spike-instability risk when allocating gradient energy, potentially concentrating learning signals on highly sensitive membrane states.

2.2 MEMBRANE POTENTIAL AND ROBUSTNESS IN SNNS

Membrane-potential distributions (MPDs) influence both SNN optimization and robustness. RecDis-SNN regularizes MPDs to mitigate membrane-potential degeneration and gradient mismatch (Guo et al., 2022), and RMP-Loss adjusts these distributions to reduce quantization-induced information loss (Guo et al., 2023). DS-ATGO jointly adjusts firing thresholds and surrogate gradients (Jiang et al., 2026a). For robustness, MPD-SGR regularizes the overlap between MPDs and gradient-active regions (Jiang et al., 2026b), whereas TGO moves membrane states away from vulnerable threshold neighborhoods (Wang et al., 2026b). Other approaches employ adversarial training or SNN-specific certification (Ding et al., 2022; Liang et al., 2022). These studies establish the importance of membrane-state sensitivity, but primarily address robustness through forward-state regulation or training objectives, leaving backward gradient allocation comparatively underexplored.

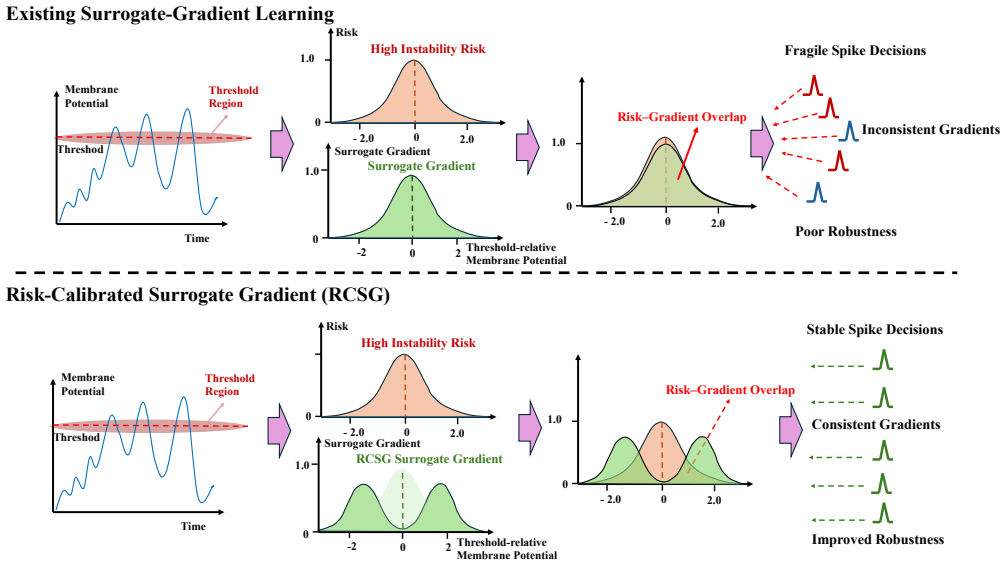


Figure 1: **Overview of RCSG. Top:** A conventional surrogate gradient concentrates backward learning signals near the firing threshold, where spike-instability risk is highest, so that gradient allocation substantially overlaps with risk. **Bottom:** RCSG down-weights this unstable, threshold-centered region and reallocates gradient energy toward more stable neighboring states. The resulting off-threshold gradient structure stabilizes backward learning while leaving the forward spiking dynamics unchanged.

3 METHOD

Figure 1 illustrates the motivation for and mechanism of RCSG. Conventional surrogate gradients produce their strongest backward responses near the firing threshold, precisely where spike decisions are most sensitive to small perturbations. RCSG corrects this mismatch by calibrating the backward gradient according to state-dependent instability risk, suppressing gradient emphasis at the most vulnerable threshold-neighboring states and redistributing it toward more stable states farther from the threshold.

3.1 RISK-AWARE GRADIENT FORMULATION

Consider a spiking neuron with membrane potential U, firing threshold ϑ, threshold-relative membrane potential $Z = U - \vartheta$, and spike output $S = H(Z)$, where $H(\cdot)$ is the Heaviside step function. During backpropagation, the derivative of H is replaced by a surrogate gradient $\phi(Z)$. Collecting the remaining local Jacobian factors into a single upstream signal A, we write the resulting local gradient contribution as

$$G = A\phi(Z). \tag{1}$$

Let $q(z)$ denote the density of membrane states and $m(z) = \mathbb{E}[A^2 \mid Z = z]$ the conditional second moment, or energy, of the upstream signal. The surrogate ϕ then induces the normalized distribution of squared gradient contributions

$$p_\phi(z) = \frac{q(z)m(z)\phi(z)^2}{\int q(s)m(s)\phi(s)^2\,\mathrm{d}s}. \tag{2}$$

This distribution describes how the effective backward gradient energy is spread across membrane states, jointly reflecting state occupancy, upstream gradient strength, and the surrogate response.

We further associate each membrane state with a spike-instability risk. Given an additive perturbation Ξ, we define

$$r(z) = \mathbb{P}[H(z + \Xi) \neq H(z)] \tag{3}$$

as the probability that perturbing the membrane state z flips the resulting spike decision. For instance, under Gaussian membrane perturbations $\Xi \sim \mathcal{N}(0, \sigma^2)$,

$$r_\sigma(z) = \Phi\left(-\frac{|z|}{\sigma}\right), \tag{4}$$

where σ is the perturbation scale and $\Phi(\cdot)$ is the standard normal cumulative distribution function. As shown in the upper part of Figure 1, this risk peaks at the firing threshold and decays as the membrane state moves away from it. Conventional threshold-centered surrogates, however, typically place their strongest responses in exactly this high-risk region, creating a potential mismatch.

To quantify the risk carried by the backward learning signal, we define the Gradient Risk Exposure (GRE) of a surrogate ϕ as the expected state risk under its induced gradient-energy distribution:

$$\mathcal{R}(\phi) = \mathbb{E}_{Z\sim p_\phi}[r(Z)] = \int r(z)\, p_\phi(z)\, \mathrm{d}z. \tag{5}$$

Unlike a simple count of near-threshold neurons, $\mathcal{R}(\phi)$ measures how much gradient energy is assigned to perturbation-sensitive states. It therefore directly captures the mismatch described above, in which a threshold-centered surrogate may place a disproportionate share of its effective learning signal where spike-instability risk is highest. This observation motivates treating surrogate-gradient design as a risk-aware redistribution of backward gradient energy.

Let ϕ_0 denote a baseline surrogate and $p_0 = p_{\phi_0}$ its induced gradient-energy distribution. We seek an alternative allocation p that reduces exposure to unstable states while remaining close to this baseline learning geometry, which leads to the regularized optimization problem

$$\min_{p \ll p_0} \Big\{ \mathbb{E}_p[r(Z)] + \tau\, \mathrm{KL}(p \,\|\, p_0) \Big\}. \tag{6}$$

Here, $\mathrm{KL}(p \,\|\, p_0)$ measures the distortion of p relative to the baseline allocation, the absolute-continuity condition $p \ll p_0$ restricts p to the support of p_0, and $\tau > 0$ controls the trade-off between reducing risk exposure and preserving fidelity to the baseline surrogate.

3.2 Optimal Risk-Aware Allocation

We now derive the optimal gradient-energy allocation for (6). Define the normalizing constant

$$\mathcal{Z}_\tau = \mathbb{E}_{Z\sim p_0}\left[\exp\left(-\frac{r(Z)}{\tau}\right)\right], \tag{7}$$

which depends on the baseline distribution p_0, the state-dependent risk r, and the trade-off parameter τ. Because $0 \le r(z) \le 1$, $\mathcal{Z}_\tau$ is finite and strictly positive.

The objective in (6) admits the closed-form minimizer

$$p^*_\tau(z) = \frac{p_0(z)\exp[-r(z)/\tau]}{\mathcal{Z}_\tau}. \tag{8}$$

To verify this, we expand the KL divergence between any admissible $p \ll p_0$ and p^*_τ:

$$\begin{aligned} \tau\, \mathrm{KL}(p\|p^*_\tau) &= \tau \int p(z) \log \frac{p(z)}{p_0(z)\exp[-r(z)/\tau]/\mathcal{Z}_\tau}\, \mathrm{d}z \\ &= \tau\, \mathrm{KL}(p\|p_0) + \mathbb{E}_p[r(Z)] + \tau \log \mathcal{Z}_\tau. \end{aligned} \tag{9}$$

Rearranging yields

$$\mathbb{E}_p[r(Z)] + \tau\, \mathrm{KL}(p\|p_0) = \tau\, \mathrm{KL}(p\|p^*_\tau) - \tau \log \mathcal{Z}_\tau. \tag{10}$$

The second term on the right-hand side does not depend on p, and the KL divergence is non-negative, vanishing only when $p = p^*_\tau$ almost everywhere. Hence, (8) is the unique minimizer up to p_0-null sets, with minimum value $-\tau \log \mathcal{Z}_\tau$.

Equation (8) gives a direct interpretation of the proposed risk–distortion principle. The optimal allocation multiplicatively tilts the baseline allocation $p_0(z)$ by the risk weight

$$w_\tau(z) = \exp\left[-\frac{r(z)}{\tau}\right], \tag{11}$$

so that states with higher spike-instability risk receive less gradient energy. Because $w_\tau(z) > 0$, this tilt preserves the support of p_0 and thus retains the structure inherited from the baseline surrogate.

The parameter τ offers a single continuous control over the risk–fidelity trade-off. As $\tau \to \infty$, $w_\tau(z) \to 1$ and $p^*_\tau \to p_0$, recovering the baseline allocation. As τ decreases, the risk-dependent tilt strengthens and gradient energy is progressively reallocated toward more stable membrane states.

3.3 Risk-Calibrated Surrogate Gradient

The optimal distribution in (8) specifies where gradient energy should be allocated, but training requires a local surrogate function that realizes this allocation. For a fixed backward context, the baseline surrogate induces

$$p_0(z) = \frac{q(z)m(z)\phi_0(z)^2}{\int q(s)m(s)\phi_0(s)^2\,\mathrm{d}s}, \tag{12}$$

whereas a candidate surrogate ϕ induces

$$p_\phi(z) = \frac{q(z)m(z)\phi(z)^2}{\int q(s)m(s)\phi(s)^2\,\mathrm{d}s}. \tag{13}$$

Requiring $p_\phi = p^*_\tau$ and combining (8), (12), and (13), we obtain

$$\frac{p_\phi(z)}{p_0(z)} \;\propto\; \frac{\phi(z)^2}{\phi_0(z)^2} \;\propto\; \exp\left[-\frac{r(z)}{\tau}\right], \tag{14}$$

where the proportionality constants are state-independent normalizing integrals. The state density $q(z)$ and the upstream energy $m(z)$ appear in both (12) and (13) and therefore cancel.

Taking the non-negative square root of (14) yields the proposed Risk-Calibrated Surrogate Gradient (RCSG):

$$\phi_{\mathrm{RCSG}}(z) = C_\tau\,\phi_0(z)\exp\left[-\frac{r(z)}{2\tau}\right], \tag{15}$$

where $\phi_0(z)$ is an arbitrary baseline surrogate, $r(z)$ is the spike-instability risk at state z, τ controls the calibration strength, and $C_\tau > 0$ is a state-independent normalization factor. The factor $\exp[-r(z)/(2\tau)]$ suppresses the backward response in unstable regions while preserving the support and structural prior of ϕ_0. As shown in the lower part of Figure 1, this calibration reduces gradient concentration at the high-risk threshold and reallocates the learning signal toward nearby, more stable states.

The normalization factor C_τ does not affect the normalized distribution $p_{\phi_{\mathrm{RCSG}}}$ but determines the overall gradient magnitude. We set it by matching the gradient energy of the baseline surrogate,

$$\mathbb{E}\left[A^2\phi_{\mathrm{RCSG}}(Z)^2\right] = \mathbb{E}\left[A^2\phi_0(Z)^2\right], \tag{16}$$

which gives

$$C_\tau = \left(\frac{\int q(z)m(z)\phi_0(z)^2\,\mathrm{d}z}{\int q(z)m(z)\phi_0(z)^2\exp[-r(z)/\tau]\,\mathrm{d}z}\right)^{1/2} = \mathcal{Z}_\tau^{-1/2}. \tag{17}$$

RCSG therefore reshapes the distribution of gradient energy rather than its global scale. The normalization statistics are treated as stop-gradient quantities, so the calibration introduces no additional optimization path.

RCSG can also induce off-threshold gradient peaks without explicitly prescribing their locations. Consider an even, positive baseline surrogate ϕ_0 and a symmetric risk profile r, both differentiable for $z > 0$. Then

$$\frac{\mathrm{d}}{\mathrm{d}z}\log\phi_{\mathrm{RCSG}}(z) = \frac{\phi_0'(z)}{\phi_0(z)} - \frac{r'(z)}{2\tau}, \tag{18}$$

so any interior local maximum $z^*_\tau > 0$ must satisfy

$$-\frac{\phi_0'(z^*_\tau)}{\phi_0(z^*_\tau)} = -\frac{r'(z^*_\tau)}{2\tau}. \tag{19}$$

This condition balances the relative decay of the baseline surrogate against the decrease in risk away from the threshold. By symmetry, any such peak is accompanied by a mirror peak at $-z^*_\tau$.

RCSG modifies only the backward pass and leaves the forward membrane dynamics, firing threshold, and spike generation rule unchanged:

$$S = H(U - \vartheta), \qquad \frac{\partial S}{\partial U} \approx \phi_{\mathrm{RCSG}}(U - \vartheta). \tag{20}$$

The multiplicative calibration in (15) therefore integrates into standard surrogate-gradient training without additional neuronal states or an auxiliary robustness loss.

4 Experiments

4.1 Experimental Setup

We evaluate models on CIFAR-10, CIFAR-100 Krizhevsky et al. (2009), and Tiny-ImageNet Le et al. (2015). The default backbone is SEW-ResNet-18 with $T = 4$ simulation steps, Spikformer is also used to assess transfer across architectures (Fang et al., 2021; Zhou et al., 2022). We consider three training regimes: standard training, FGSM adversarial training, and PGD adversarial training. Within each regime, paired methods use identical initializations, data orderings, optimizers, training schedules, and hyperparameter search budgets. All models are trained for 300 epochs using AdamW with an initial learning rate of 5×10^{-4}, a weight decay of 0.05, and a batch size of 64. All experiments are implemented in PyTorch 2.8.0 with CUDA 12.8 and run on NVIDIA V100 GPU.

Given N sampled threshold-relative membrane states z_i, corresponding local gradient factors a_i, and a surrogate ϕ, we estimate GRE directly from the samples:

$$\widehat{\mathcal{R}}(\phi) = \frac{\sum_{i=1}^{N} a_i^2 \phi(z_i)^2 r_\sigma(z_i)}{\sum_{i=1}^{N} a_i^2 \phi(z_i)^2}, \qquad r_\sigma(z) = \Phi\left(-\frac{|z|}{\sigma}\right). \tag{21}$$

To distinguish state occupancy from gradient allocation near the firing threshold, we measure the fraction of states within $|z| < \delta$,

$$Q_\delta = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(|z_i| < \delta), \tag{22}$$

and the fraction of local gradient energy assigned to this region,

$$M_\delta(\phi) = \frac{\sum_{i=1}^{N} a_i^2 \phi(z_i)^2 \mathbf{1}(|z_i| < \delta)}{\sum_{i=1}^{N} a_i^2 \phi(z_i)^2}. \tag{23}$$

Comparing $M_\delta(\phi)$ with Q_δ reveals whether near-threshold states receive a disproportionate share of gradient energy relative to their frequency.

We evaluate untargeted ℓ_∞ attacks with $\epsilon = 8/255$, using attack-time surrogate gradients selected independently of the training surrogate. The suite includes multi-restart PGD with many iterations, SNN-adaptive gradient estimators, and query-based attacks (Croce & Hein, 2020; Bu et al., 2023; Wang et al., 2026a).

4.2 Gradient–Risk Mismatch

We examine the proposed mechanism in a controlled synthetic spiking system. We sample threshold-relative membrane states $Z \sim \mathcal{N}(0, 0.45^2)$ and generate spikes as $S = H(Z)$. Spike instability is measured under independent Gaussian perturbations $\Xi \sim \mathcal{N}(0, \sigma^2)$ with $\sigma \in \{0.10, 0.15\}$. Gradient allocation is evaluated using an arctangent baseline surrogate ϕ_0 and independent local gradient factors $A \sim \mathcal{N}(0, 1)$.

Figure 2(a) compares Monte Carlo estimates of spike-flip probability (markers) with the analytical risk $r_\sigma(z)$ (solid curves). The estimates closely match the prediction, confirming that spike instability is highest near the threshold and decreases rapidly with $|Z|$. Figure 2(b) compares state occupancy Q_δ with gradient-energy mass $M_\delta(\phi_0)$ within $|Z| < \delta$, with error bars indicating variability across independent runs. For small δ, $M_\delta(\phi_0)$ substantially exceeds Q_δ, showing that the baseline surrogate assigns disproportionate gradient energy to states most susceptible to spike flipping.

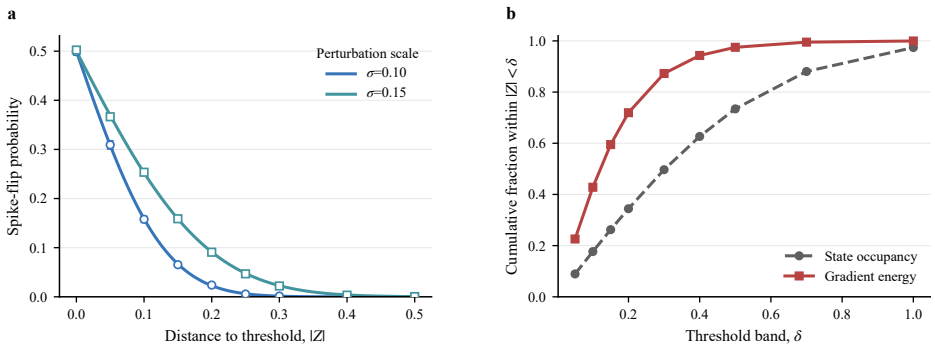


Figure 2: **Gradient–risk mismatch in a synthetic spiking system.** (a) Spike-flip risk versus threshold distance. (b) Near-threshold state occupancy and gradient-energy allocation.

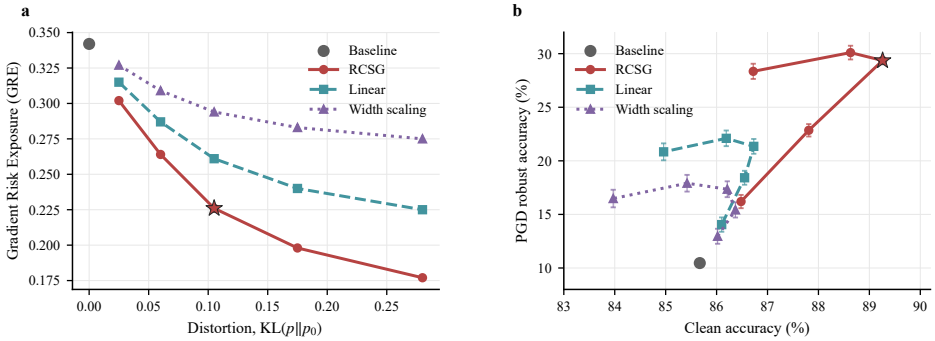


Figure 3: **Mechanistic validation of RCSG.** (a) Risk–distortion comparison in fixed backward contexts. (b) Clean–robust accuracy trade-off across calibration strengths.

4.3 Mechanistic Validation

We examine whether RCSG realizes the risk–distortion principle derived in Section 3. We extract membrane states and local gradient factors from trained CIFAR-10 baseline checkpoints and hold them fixed while varying the surrogate. Figure 3(a) compares GRE against $\mathrm{KL}(p_\phi \| p_0)$ for RCSG, linear risk weighting, and surrogate-width scaling. At comparable distortion levels, RCSG achieves the lowest GRE, consistent with the optimal exponential allocation in Eq. (8).

Figure 3(b) shows clean and PGD robust accuracy on CIFAR-10 across calibration strengths; the star marks the selected RCSG setting. Moderate calibration improves robustness while maintaining clean accuracy, whereas excessive calibration reduces clean accuracy. This pattern is consistent with the proposed risk–fidelity trade-off.

We use controlled ablations to assess the role of state-dependent risk allocation. Uniform attenuation rescales the surrogate by a constant, shuffled risk disrupts the correspondence between risk and membrane state, and anti-risk reverses the sign of the risk exponent in Eq. (15) at matched KL distortion. We also evaluate linear risk weighting and RCSG without energy normalization.

Table 1 reports GRE, near-threshold gradient-energy mass $M_{0.1}$ within $|Z| < 0.1$, and clean and PGD robust accuracy on CIFAR-10. PGD uses an ℓ_∞ perturbation budget of $\epsilon = 8/255$. Uniform attenuation leaves GRE and $M_{0.1}$ unchanged and yields little robustness improvement; shuffled risk also provides limited gains. Anti-risk increases both risk metrics and reduces robustness, supporting the importance of the allocation direction. Linear weighting lowers GRE but achieves lower robust accuracy than RCSG. Removing energy normalization retains the reported risk reduction but lowers clean and robust accuracy relative to full RCSG, suggesting that normalization supports effective

Table 1: **Mechanism ablation on CIFAR-10.** Clean and PGD robust accuracy are reported as percentages.

Backward intervention	GRE ↓	$M_{0.1}$ ↓	Clean ↑	PGD ↑
Baseline ϕ_0	0.342	0.428	85.67	10.46
Uniform attenuation	0.342	0.428	84.93	10.81
Shuffled risk	0.336	0.416	85.31	11.74
Anti-risk, matched KL	0.397	0.512	84.61	7.86
Linear risk weighting	0.268	0.301	87.06	21.35
RCSG w/o normalization	**0.226**	**0.238**	86.42	25.87
RCSG	**0.226**	**0.238**	**89.26**	**29.36**

Table 2: **Clean and robust accuracy (%).** All methods use the same architecture, optimization protocol, and training budget.

	CIFAR-10			CIFAR-100			Tiny-ImageNet		
Method	Clean	FGSM	PGD	Clean	FGSM	PGD	Clean	FGSM	PGD
Arctangent SG (Fang et al., 2021)	85.67	21.83	10.46	62.41	13.57	6.82	54.55	11.64	7.25
Dspike (Li et al., 2021)	86.90	23.41	12.18	63.76	14.62	7.91	55.48	12.83	8.16
LSG (Lian et al., 2023)	88.14	24.76	13.07	65.83	15.39	8.54	57.26	13.71	8.95
MPD-AGL (Jiang et al., 2025)	87.42	26.32	14.85	64.95	17.24	10.16	56.71	15.68	10.73
A^2SG (Kang et al., 2026)	**89.58**	27.80	16.12	67.12	18.73	11.28	59.14	16.91	11.86
MPD-SGR (Jiang et al., 2026b)	84.92	38.74	27.31	61.73	25.83	16.48	53.81	**23.17**	18.72
TGO (Wang et al., 2026b)	83.76	39.52	27.86	60.84	26.47	16.83	52.96	21.84	17.53
RCSG	89.26	**41.27**	**29.36**	**67.71**	**27.54**	**17.92**	**59.81**	22.69	**19.61**

optimization. Full RCSG achieves the highest clean and PGD robust accuracy among the evaluated controls.

4.4 Main Robustness Results

Table 2 compares RCSG with surrogate-gradient and robustness-oriented methods under natural training. Adaptive surrogate methods improve clean accuracy but offer limited adversarial robustness, whereas MPD-SGR and TGO achieve stronger robustness with lower clean accuracy. RCSG achieves the highest PGD accuracy on all three datasets while maintaining competitive clean accuracy through backward gradient allocation alone. On Tiny-ImageNet, MPD-SGR performs slightly better under FGSM, but RCSG achieves higher clean and PGD accuracy.

Since RCSG modifies the training-time backward estimator, we assess potential surrogate-gradient masking using longer PGD runs, an independent attack-time surrogate, an SNN-adaptive gradient attack, and a query-based attack. As shown in Table 3, RCSG outperforms the baseline under every evaluation.

Increasing PGD from 20 to 100 steps reduces RCSG accuracy from 29.36% to 27.84%, while switching to an independent attack-time surrogate yields a further decrease to 26.91%. The SNN-adaptive and query-based attacks yield comparable accuracies of 25.88% and 26.34%, respectively. RCSG consistently outperforms the baseline across these evaluations, suggesting that its robustness gains are not solely attributable to the training-time gradient estimator.

4.5 Transfer Across Surrogates and Architectures

The preceding experiments examine whether risk-aware gradient allocation improves robustness and whether these gains persist across attack estimators. We next assess whether the benefits depend on the baseline surrogate or network architecture. Since RCSG calibrates an existing surrogate through multiplicative risk weighting, we apply a single validation-selected calibration setting across five surrogate families and four architectures without further retuning.

Table 4 shows consistent improvements in clean and FGSM accuracy across all evaluated settings. With SEW-ResNet-18, RCSG improves FGSM accuracy across the five surrogate families by 16.48–

Table 3: **Attack strength and estimator dependence on CIFAR-10.** Clean and robust accuracy (%) under attacks with varying iteration counts and gradient estimators, including a query-based attack.

Evaluation	Baseline	RCSG
Clean	85.67	**89.26**
FGSM	21.83	**41.27**
Training SG, PGD-20	10.46	**29.36**
Training SG, PGD-100	8.92	**27.84**
Independent SG, PGD-100	8.47	**26.91**
SNN-adaptive gradient attack	7.95	**25.88**
Query-based attack	8.21	**26.34**

Table 4: **Transfer across surrogate families and architectures on CIFAR-10.** Clean and FGSM accuracy (%) are reported for matched baselines and RCSG variants. Surrogate comparisons use SEW-ResNet-18; architecture comparisons use the arctangent surrogate. All variants share the same calibration setting without retuning.

Setting	**Variant**	Clean	Clean + RCSG	FGSM	FGSM + RCSG
Surrogate	Arctangent	85.67	89.26	21.83	**41.27**
	Triangular	84.91	87.63	20.47	38.56
	Sigmoid	86.12	88.54	22.36	39.74
	SuperSpike	86.48	88.91	24.15	40.63
	Dspike	86.90	89.11	23.41	40.32
Architecture	VGG-11	84.23	86.71	19.86	36.92
	WRN-16-4	87.35	89.10	24.73	42.16
	SEW-ResNet-18	85.67	89.26	21.83	41.27
	Spikformer	94.07	94.98	35.61	**44.84**

19.44 percentage points. With the arctangent surrogate, it improves FGSM accuracy across the four architectures by 9.23–19.44 percentage points, extending the gains from convolutional SNNs to Spikformer. Together with the preceding mechanism analyses, these results support risk-aware gradient allocation as a broadly applicable calibration strategy, with benefits that extend beyond a particular surrogate shape or architecture.

5 Conclusion

Conventional surrogate gradients concentrate backward gradient energy near the firing threshold, where membrane states are most susceptible to spike flipping. To characterize this mismatch between gradient allocation and state stability, we introduced Gradient Risk Exposure (GRE) and derived the Risk-Calibrated Surrogate Gradient (RCSG) by minimizing GRE under a Kullback–Leibler (KL) fidelity penalty. RCSG reallocates gradient energy toward more stable membrane states without modifying the forward neuronal dynamics or requiring an auxiliary robustness loss. Experiments on three static-image datasets show that RCSG improves adversarial robustness while maintaining competitive clean accuracy.

References

Tong Bu, Jianhao Ding, Zecheng Hao, and Zhaofei Yu. Rate gradient approximation attack threats deep spiking neural networks. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7896–7906. IEEE, 2023.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.

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

Jianhao Ding, Tong Bu, Zhaofei Yu, Tiejun Huang, and Jian Liu. Snn-rat: Robustness-enhanced spiking neural network through regularized adversarial training. *Advances in Neural Information Processing Systems*, 35:24780–24793, 2022.

Jianhao Ding, Zhiyu Pan, Yujia Liu, Zhaofei Yu, and Tiejun Huang. Robust stable spiking neural networks. *arXiv preprint arXiv:2405.20694*, 2024.

Jason K Eshraghian, Max Ward, Emre O Neftci, Xinxin Wang, Gregor Lenz, Girish Dwivedi, Mohammed Bennamoun, Doo Seok Jeong, and Wei D Lu. Training spiking neural networks using lessons from deep learning. *Proceedings of the IEEE*, 111(9):1016–1054, 2023.

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

Yufei Guo, Xinyi Tong, Yuanpei Chen, Liwen Zhang, Xiaode Liu, Zhe Ma, and Xuhui Huang. Recdis-snn: Rectifying membrane potential distribution for directly training spiking neural networks. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 326–335. IEEE, 2022.

Yufei Guo, Xiaode Liu, Yuanpei Chen, Liwen Zhang, Weihang Peng, Yuhan Zhang, Xuhui Huang, and Zhe Ma. Rmp-loss: Regularizing membrane potential distribution for spiking neural networks. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17345–17355. IEEE, 2023.

Julia Gygax and Friedemann Zenke. Elucidating the theoretical underpinnings of surrogate gradient learning in spiking neural networks. *Neural Computation*, 37(5):886–925, 2025.

Jiaqiang Jiang, Lei Wang, Runhao Jiang, Jing Fan, and Rui Yan. Adaptive gradient learning for spiking neural networks by exploiting membrane potential dynamics. *arXiv preprint arXiv:2505.11863*, 2025.

Jiaqiang Jiang, Wenfeng Xu, Jing Fan, and Rui Yan. Ds-atgo: dual-stage synergistic learning via forward adaptive threshold and backward gradient optimization for spiking neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1855–1863, 2026a.

Runhao Jiang, Chengzhi Jiang, Rui Yan, and Huajin Tang. Mpd-sgr: Robust spiking neural networks with membrane potential distribution-driven surrogate gradient regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1864–1872, 2026b.

Jacques Kaiser, Hesham Mostafa, and Emre Neftci. Synaptic plasticity dynamics for deep continuous local learning (decolle). *Frontiers in Neuroscience*, 14:424, 2020.

Yechan Kang, Yongjin Kweon, Mingyeong Seo, Sohee Park, Yeonguk Jeon, Jongkil Park, Hyun Jae Jang, Jaewook Kim, YeonJoo Jeong, Suyoun Lee, et al. A2sg: Adaptive and asymmetric surrogate gradients for training deep spiking neural networks. *arXiv preprint arXiv:2606.11236*, 2026.

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

Souvik Kundu, Massoud Pedram, and Peter A Beerel. Hire-snn: Harnessing the inherent robustness of energy-efficient deep spiking neural networks by training with crafted input noise. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5189–5198. IEEE, 2021.

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

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

Yuhang Li, Yufei Guo, Shanghang Zhang, Shikuang Deng, Yongqing Hai, and Shi Gu. Differentiable spike: Rethinking gradient-descent for training spiking neural networks. *Advances in neural information processing systems*, 34:23426–23439, 2021.

Shuang Lian, Jiangrong Shen, Qianhui Liu, Ziming Wang, Rui Yan, and Huajin Tang. Learnable surrogate gradient for direct training spiking neural networks. In *Ijcai*, pp. 3002–3010, 2023.

Ling Liang, Kaidi Xu, Xing Hu, Lei Deng, and Yuan Xie. Toward robust spiking neural network against adversarial perturbation. *Advances in Neural Information Processing Systems*, 35:10244–10256, 2022.

Yujia Liu, Tong Bu, Jianhao Ding, Zecheng Hao, Tiejun Huang, and Zhaofei Yu. Enhancing adversarial robustness in snns with sparse gradients. *arXiv preprint arXiv:2405.20355*, 2024.

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

Kiran Nair, Rodrigue Rizk, and KC Santosh. Sage: Surrogate-gradient adaptation via attention-guided entropy for spiking transformers. *arXiv preprint arXiv:2608.13702*, 2026.

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.

Garrick Orchard, Ajinkya Jayawant, Gregory K Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9:437, 2015.

Sumit B Shrestha and Garrick Orchard. Slayer: Spike layer error reassignment in time. *Advances in neural information processing systems*, 31, 2018.

Amirhossein Tavanaei, Masoud Ghodrati, Saeed Reza Kheradpisheh, Timothée Masquelier, and Anthony Maida. Deep learning in spiking neural networks. *Neural networks*, 111:47–63, 2019.

Jihang Wang, Dongcheng Zhao, Ruolin Chen, Qian Zhang, and Yi Zeng. Towards reliable evaluation of adversarial robustness for spiking neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20691–20700, 2026a.

Shuai Wang, Malu Zhang, Yulin Jiang, Dehao Zhang, Ammar Belatreche, Yu Liang, Yimeng Shan, Zijian Zhou, Yang Yang, and Haizhou Li. Robust spiking neural networks against adversarial attacks. *arXiv preprint arXiv:2602.20548*, 2026b.

Ziming Wang, Runhao Jiang, Shuang Lian, Rui Yan, and Huajin Tang. Adaptive smoothing gradient learning for spiking neural networks. In *International conference on machine learning*, pp. 35798–35816. PMLR, 2023.

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.

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.

Friedemann Zenke and Surya Ganguli. Superspike: Supervised learning in multilayer spiking neural networks. *Neural computation*, 30(6):1514–1541, 2018.

Friedemann Zenke and Tim P Vogels. The remarkable robustness of surrogate gradient learning for instilling complex function in spiking neural networks. *Neural computation*, 33(4):899–925, 2021.

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