Not All Timesteps Matter Equally: Selective Alignment Knowledge Distillation for Spiking Neural Networks

Kai Sun[1] , **Peibo Duan**[1*] , **Yongsheng Huang**[2] , **Guowei Zhang**[1] , **Benjamin Smith**[1] , **Nanxu Gong**[3] and **Levin Kuhlmann**[1*]
[1]Faculty of Information Technolody, Monash University, Australia
[2]School of Software, Northeastern University, China
[3]Department of Medicine, National University of Singapore, Singapore
{kai.sun1, peibo.duan, levin.kuhlmann}@monash.edu

Abstract

Spiking neural networks (SNNs), which are brain-inspired and spike-driven, achieve high energy efficiency. However, a performance gap between SNNs and artificial neural networks (ANNs) still remains. Knowledge distillation (KD) is commonly adopted to improve SNN performance, but existing methods typically enforce uniform alignment across all timesteps, either from a teacher network or through inter-temporal self-distillation, implicitly assuming that per-timestep predictions should be treated equally. In practice, SNN predictions vary and evolve over time, and intermediate timesteps need not all be individually correct even when the final aggregated output is correct. Under such conditions, effective distillation should not force every timestep toward the same supervision target, but instead provide corrective guidance to erroneous timesteps while preserving useful temporal dynamics. To address this issue, we propose **S**elective **A**lignment **K**nowledge **D**istillation **(SeAl-KD)**, which selectively aligns class-level and temporal knowledge by equalizing competing logits at erroneous timesteps and reweighting temporal alignment based on confidence and inter-timestep similarity. Extensive experiments on static image and neuromorphic event-based datasets demonstrate consistent improvements over existing distillation methods. The code is available at https://github.com/KaiSUN1/SeAl.

1 Introduction

Spiking Neural Networks (SNNs), often regarded as the third generation of neural networks, are biologically inspired models that communicate through discrete spikes rather than continuous-valued activations ([Maass, 1997]). By processing information in an event-driven manner with sparse spikes over time, SNNs offer the potential for high energy efficiency, especially when deployed on neuromorphic hardware ([Guo *et al.*, 2023]). Nevertheless, training SNNs to achieve accuracy comparable to Artificial Neural Networks (ANNs) remains challenging, because useful evidence is accumulated progressively through spikes, making different timesteps unevenly informative during learning ([Bellec *et al.*, 2018; Neftci *et al.*, 2019]).

To narrow the performance gap between ANNs and SNNs, knowledge distillation (KD) has become a widely adopted strategy ([Xu *et al.*, 2023; Qiu *et al.*, 2024]). Early distillation methods typically supervise only the final aggregated output by matching averaged logits, which may mix inconsistent temporal information during optimization ([Zhao *et al.*, 2025; Deng *et al.*, 2022]). More recent approaches move toward timestep-wise distillation by injecting teacher supervision at each timestep and aligning logits throughout the temporal dimension ([Yu *et al.*, 2025b; Yu *et al.*, 2025a]). In parallel, several methods further exploit the temporal structure of SNNs through self-distillation ([Ding *et al.*, 2025b]) or temporal consistency regularization ([Zhao *et al.*, 2025; Ding *et al.*, 2025a]), encouraging predictions or features from different timesteps to be consistent in order to stabilize optimization and improve performance ([Ding *et al.*, 2025b; Yu *et al.*, 2025b]).

The consistency assumption is partially misaligned with the intrinsic properties of SNNs and their prediction mechanism. Due to membrane potential integration and reset, spike firing may induce abrupt changes in membrane states, making it difficult for SNNs to maintain identical predictions across all timesteps, as detailed in **Appendix A**. Meanwhile, since the final decision is determined by temporal accumulation rather than any single timestep, an incorrect intermediate prediction does not necessarily imply an incorrect final outcome. As illustrated in the toy case in Figure 1(a), some intermediate timesteps are misclassified while the final temporally aggregated prediction remains correct. Figure 1(b) further shows that per-timestep accuracy is consistently lower than temporally aggregated accuracy. Moreover, Figure 1(c) reveals that more than 12% of correctly classified samples are misclassified at some intermediate timesteps, and some samples are never correctly classified at any individual timestep but still become correct after temporal aggregation. These observations suggest that a timestep should not be judged solely by whether it is already correct, but by whether it contributes useful evidence to the final temporal accumulation.

*Corresponding author.

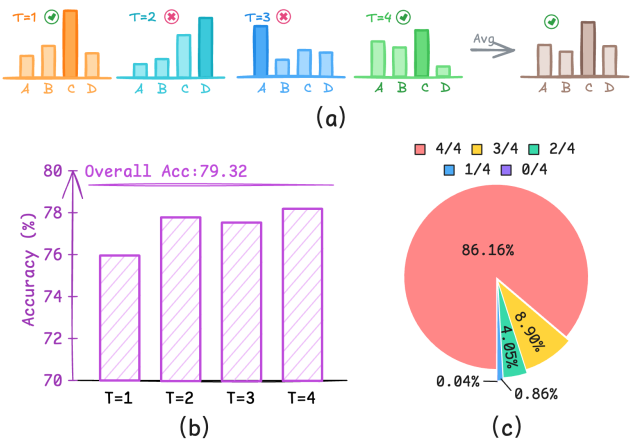

Figure 1: Mismatch between intermediate and final predictions in SNNs. (a) A toy example where intermediate predictions are wrong, but the final aggregated prediction is correct. (b) Per-timestep accuracy and final aggregated accuracy on CIFAR100. (c) Distribution of CIFAR100 samples with correct final predictions by the number of correctly predicted timesteps out of T.

However, existing timestep-wise distillation strategies do not explicitly account for this distinction. By encouraging each timestep to align with the teacher signal, they may overlook what corrective evidence is actually needed for improving temporal accumulation and which temporal sources can provide reliable support. This perspective leads to two key questions for timestep-wise distillation: **what corrective evidence should be injected into a currently erroneous timestep**, and **from which temporal sources should it be drawn**? For a currently erroneous timestep dominated by a confusing wrong class, directly transferring teacher preference over the full class distribution may dilute the needed correction and interfere with the adjustment of the most critical class relation, namely that between the ground-truth class and the wrongly favored class. Moreover, in temporal self-distillation, not all source timesteps are equally reliable: some provide confident and compatible evidence, while others may introduce noisy or misleading temporal signals. To address these issues, we propose **SeAl-KD**, a selective distillation framework with two components. **Error-aware Logit Alignment (ELA)** refines the class evidence received by erroneous timesteps, while **Selective Temporal Alignment (STA)** emphasizes reliable and compatible source timesteps during alignment.

Our main contributions are summarized as follows: (1) We reveal that erroneous timesteps are prevalent in SNNs and show that intermediate misclassifications do not necessarily impair the final prediction under temporal aggregation, exposing the mismatch between consistency and temporal evidence accumulation. (2) We propose SeAl-KD, a selective KD framework that improves corrective supervision for erroneous timesteps through ELA and STA. We further provide theoretical analysis to justify the proposed selective alignment. (3) We conduct extensive experiments on both static and neuromorphic image datasets. The results show that SeAl-KD preserves richer temporal distributions across timesteps and consistently improves performance.

2 Related Works

2.1 Knowledge Distillation for SNNs

KD improves SNN training by transferring final outputs, logits, or features from ANNs to narrow the gap with real-valued ANNs ([Xu *et al.*, 2023; Hong *et al.*, 2025]). Along this line, distillation strategies have evolved from global supervision on temporally aggregated outputs ([Zhang *et al.*, 2025]) to timestep-wise supervision that aligns logits or representations at each timestep ([Yu *et al.*, 2025b; Yu *et al.*, 2025a]). In addition, self-distillation and temporal consistency regularization further exploit the student's temporal structure to align outputs across timesteps ([Qiu *et al.*, 2024; Zuo *et al.*, 2024]). Despite these advances, most existing methods impose supervision on different timesteps in a largely uniform manner, without distinguishing whether a timestep is currently erroneous, what corrective information it actually needs, or whether the temporal source used for supervision is reliable.

2.2 Temporal Discrepancy in SNNs

Temporal discrepancy is intrinsic to SNNs, as membrane integration and spike-triggered resets cause neuronal states and predictions to evolve across timesteps ([Bellec *et al.*, 2018]). Prior work has improved temporal modeling and training stability through learnable membrane dynamics ([Fang *et al.*, 2021]), temporal normalization ([Zheng *et al.*, 2021; Duan *et al.*, 2022]), and surrogate-gradient design ([Li *et al.*, 2021a; Wang *et al.*, 2023]). Some studies further exploit temporal structure via timestep-dependent weighting ([Deng *et al.*, 2022]) or temporal consistency regularization across timesteps ([Zhao *et al.*, 2025; Ding *et al.*, 2025a]). However, these methods mostly regulate temporal discrepancy uniformly, without explicitly considering the role of each timestep in the final temporally accumulated prediction. Consequently, they do not distinguish what supervision a timestep needs or which temporal sources are reliable for providing it.

3 Preliminaries

Timestep-wise distillation. Consider an SNN unrolled over T timesteps. At timestep t, the student SNN model, denoted by the superscript S, produces a logit vector $\mathbf{z}_t^S \in \mathbb{R}^C$, where C denotes the number of classes. The final prediction is typically obtained by temporally averaging the logits over all timesteps, while each $\mathbf{z}_t^S$ can also be treated as an intermediate prediction. For the same input, the teacher ANN model, denoted by the superscript A, produces a time-invariant logit vector $\mathbf{z}^A$. The corresponding categorical distributions are given by the temperature-scaled softmax functions $\mathbf{p}_t^S = \mathrm{softmax}(\mathbf{z}_t^S/\tau)$ and $\mathbf{p}^A = \mathrm{softmax}(\mathbf{z}^A/\tau)$, where $\tau > 0$ is the temperature. The probability of class i under $\mathbf{p}_t^S$ and $\mathbf{p}^A$ is denoted by $p_{t,i}^S$ and p_i^A, respectively.

For each timestep t, the student is optimized with two objectives. The first objective $\mathcal{L}_{\mathrm{CLS}}$ applies cross-entropy (CE)

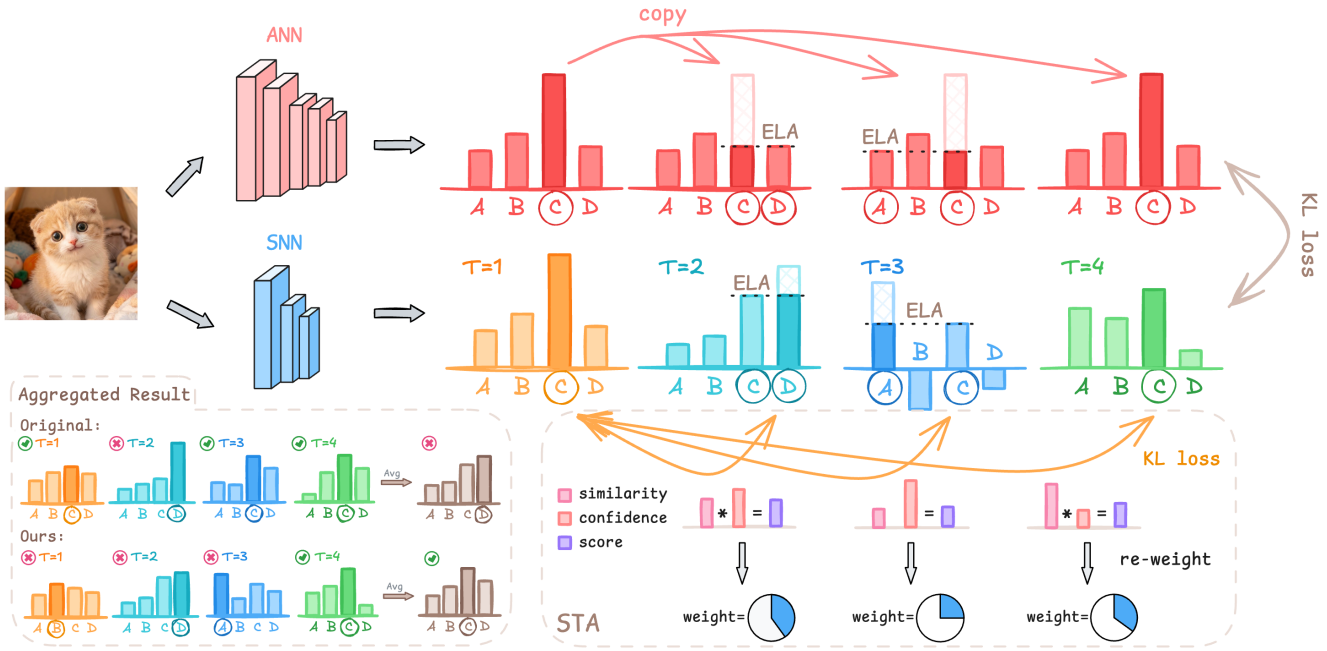

Figure 2: SeAl-KD framework. SNNs learn from the same copied ANN output across timesteps. ELA equalizes the true and predicted-false logits at erroneous timesteps before teacher–student KL. STA reweights temporal KL using confidence and inter-timestep similarity. C denotes the ground-truth class. Left-bottom illustration shows that our method aims to follow temporal diversity to obtain correct predictions, even when some intermediate predictions are incorrect.

loss for classification supervision at each timestep:

$$\mathcal{L}_{\mathrm{CLS}} = \frac{1}{T}\sum_{t=1}^{T} \ell_{\mathrm{CE}}(\mathbf{z}_t^S, \mathbf{y}) = \frac{1}{T}\sum_{t=1}^{T}\Big(-\sum_{i=1}^{C} \mathbf{y}_i \log p_{t,i}^S\Big), \tag{1}$$

where $\mathbf{y} \in \{0,1\}^C$ is the one-hot ground-truth label.

The second objective $\mathcal{L}_{\mathrm{KD}}$ aligns each temporal output with the same teacher distribution using the Kullback–Leibler (KL) divergence:

$$\mathcal{L}_{\mathrm{KD}} = \frac{1}{T}\sum_{t=1}^{T} \mathrm{KL}\big(\mathbf{p}^A \,\|\, \mathbf{p}_t^S\big) = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{C} p_i^A \log \frac{p_i^A}{p_{t,i}^S}. \tag{2}$$

The overall training objective $\mathcal{L}$ is defined as

$$\mathcal{L} = \mathcal{L}_{\mathrm{CLS}} + \lambda \mathcal{L}_{\mathrm{KD}}, \tag{3}$$

where λ controls the contribution of the distillation term.

4 Methodology

This section presents SeAl-KD, which consists of ELA and STA. As illustrated in Figure 2, ELA refines class-level correction at erroneous timesteps, while STA selects confident and compatible temporal sources for alignment.

4.1 Error-aware Logits Alignment Distillation

ELA performs timestep-wise distillation while accounting for prediction errors across timesteps in SNNs. Let $y^* \in \{1,\dots,C\}$ denote the ground-truth class index corresponding to the one-hot label $\mathbf{y}$, and let $z_{t,c}^S$ and $z_{t,c}^A$ denote the student and teacher logits for class c at timestep t, respectively. The student prediction at timestep t is $c_t^{\mathrm{pred}} = \arg\max_c z_{t,c}^S$. When an intermediate timestep is erroneous, ELA relaxes distillation on its dominant confusion instead of directly enforcing the correct class ordering, avoiding misleading correction on the confusing class pair while preserving supervision on the remaining classes.

Logit modification. If $c_t^{\mathrm{pred}} \neq y^*$, we focus on the class pair formed by the ground-truth class y^* and the predicted false class $c_t^{\mathrm{false}} = c_t^{\mathrm{pred}}$. We equalize only the logits of this pair and keep the remaining classes unchanged. Specifically, the logits of y^* and c_t^{false} are both set to the minimum of their original values, which avoids introducing unnecessary absolute shifts. The modified student and teacher logits, denoted by $\tilde{z}_{t,c}^S$ and $\tilde{z}_{t,c}^A$, are defined as

$$\tilde{z}_{t,c}^S = \begin{cases} \min\Big(z_{t,y^*}^S,\, z_{t,c_t^{\mathrm{false}}}^S\Big), & c \in \{y^*,\, c_t^{\mathrm{false}}\}, \\ z_{t,c}^S, & \text{otherwise.} \end{cases} \tag{4}$$

$$\tilde{z}_{t,c}^A = \begin{cases} \min\Big(z_{t,y^*}^A,\, z_{t,c_t^{\mathrm{false}}}^A\Big), & c \in \{y^*,\, c_t^{\mathrm{false}}\}, \\ z_{t,c}^A, & \text{otherwise.} \end{cases} \tag{5}$$

ELA objective. The ELA loss is then defined as

$$\mathcal{L}_{\mathrm{ELA}} = \frac{1}{T}\sum_{t=1}^{T} \mathrm{KL}\big(p(\tilde{\mathbf{z}}_t^A) \,\|\, p(\tilde{\mathbf{z}}_t^S)\big). \tag{6}$$

By removing the forced preference between the ground-truth class and the dominant false class at erroneous timesteps,

ELA prevents distillation from reinforcing misleading intermediate ordering while retaining teacher guidance on the remaining classes.

4.2 Similarity-aware Temporal Alignment Distillation

Based on uniform temporal alignment (**UTA**) [Zhao *et al.*, 2025], which enforces unweighted pairwise alignment across timesteps, STA adopts a weighted temporal distillation scheme that allows each timestep t to selectively learn from confident and compatiable timesteps. This guides weak timesteps toward better temporal states that support effective temporal evidence accumulation.

Timestep confidence. The reliability of timestep t is quantified by an entropy-based confidence score. The entropy H_t measures the uncertainty of the class distribution and is normalized by the maximum entropy $\log C$:

$$\mathrm{Conf}_t = 1 - \frac{H_t}{\log C}, \qquad H_t = -\sum_{c=1}^{C} p(\mathbf{z}_t^S)_c \log p(\mathbf{z}_t^S)_c. \tag{7}$$

A lower entropy leads to a larger Conf_t, indicating that the timestep provides a more reliable source.

Timestep compatibility. The compatibility between a target timestep t and a source timestep t' is measured by cosine similarity in the student logit space. This similarity reflects the consistency of their class-level preferences:

$$\mathrm{Sim}(t, t') = \frac{\mathbf{z}_t^S \cdot \mathbf{z}_{t'}^S}{\|\mathbf{z}_t^S\| \, \|\mathbf{z}_{t'}^S\|}. \tag{8}$$

A larger $\mathrm{Sim}(t, t')$ indicates that the two timesteps exhibit more compatible class preferences.

Source weighting. For each target timestep t, an unnormalized source score $s_{t,t'}$ is computed for source timestep t' by combining its confidence with its compatibility to the target:

$$s_{t,t'} = \mathrm{Conf}_{t'} \cdot \mathrm{Sim}(t, t'). \tag{9}$$

The final weights are obtained by applying a softmax over all source timesteps satisfying $t' \neq t$, with $w_{t,t} = 0$:

$$w_{t,t'} = \frac{\exp(s_{t,t'})}{\sum_{j \neq t} \exp(s_{t,j})}, \quad t' \neq t. \tag{10}$$

Thus, each target timestep primarily learns from confident and compatible source timesteps.

STA objective. Based on the obtained weights, we encourage each target timestep to align with other source timesteps through a weighted KL divergence. The STA loss is defined as

$$\mathcal{L}_{\mathrm{STA}} = \frac{1}{T} \sum_{t=1}^{T} \sum_{t' \neq t} w_{t,t'} \, \mathrm{KL}\big(p(\mathbf{z}_{t'}^S) \,\|\, p(\mathbf{z}_t^S)\big). \tag{11}$$

This objective lets each timestep absorb selectively weighted temporal guidance, instead of being uniformly aligned to all other timesteps, thereby providing more effective support for weak timesteps and reducing interference on already reliable ones.

4.3 Selective Alignment Distillation

SeAl-KD combines ELA in Eq. (6) and STA in Eq. (11) to realize selective temporal supervision. Specifically, ELA regulates what class-level knowledge is injected at erroneous timesteps, while STA guides weak timesteps toward weighted better temporal states, so that temporal correction better supports the final evidence accumulation process. Together with the classification loss, the overall training objective is

$$\mathcal{L}_{\text{SeAl-KD}} = \mathcal{L}_{\mathrm{CLS}} + \alpha \, \mathcal{L}_{\mathrm{ELA}} + \beta \, \mathcal{L}_{\mathrm{STA}}, \tag{12}$$

where α and β control the contributions of ELA and STA, respectively.

4.4 Theoretical and Statistical Analysis

We analyze SeAl-KD from the perspective of the additional update introduced by distillation beyond the standard cross-entropy objective. The analysis focuses on three representative timestep conditions: erroneous timesteps, weak timesteps, and already-correct timesteps. For space efficiency, in the main paper we select one timestep for each condition. The full results over additional sampled timesteps are provided in **Appendix B**. These diagnostics show that SeAl-KD intervenes selectively: it localizes correction when the prediction is wrong, transfers temporal guidance from more reliable temporal states, and remains restrained when the prediction is already reliable.

Proposition 1 (Localized correction)**.** *At an erroneous timestep, ELA encourages the distillation update to concentrate on the error-relevant class subspace, rather than spreading the correction uniformly over all classes.*

Consider an erroneous timestep t with ground-truth class index y^* and currently predicted wrong class c_t^{false}. Let $\bar{z}_{t,\mathrm{rest}}^S$ denote the mean student logit over the remaining classes outside $\{y^*, c_t^{\mathrm{false}}\}$, and let ∇_{θ_l} denote the gradient with respect to the parameters of layer l. To examine whether the ELA update is concentrated on the dominant confusion pair, we define three directional effects:

$$\begin{aligned} D_{\mathrm{true}}^l &= \left|\left\langle \nabla_{\theta_l} \mathcal{L}_{\mathrm{ELA}}, \nabla_{\theta_l} z_{t,y^*}^S \right\rangle\right|, \\ D_{\mathrm{false}}^l &= \left|\left\langle \nabla_{\theta_l} \mathcal{L}_{\mathrm{ELA}}, \nabla_{\theta_l} z_{t,c_t^{\mathrm{false}}}^S \right\rangle\right|, \\ D_{\mathrm{rest}}^l &= \left|\left\langle \nabla_{\theta_l} \mathcal{L}_{\mathrm{ELA}}, \nabla_{\theta_l} \bar{z}_{t,\mathrm{rest}}^S \right\rangle\right|. \end{aligned} \tag{13}$$

The pair-concentration score is then defined as

$$\mathrm{PairShare}_t^l = \frac{D_{\mathrm{true}}^l + D_{\mathrm{false}}^l}{D_{\mathrm{true}}^l + D_{\mathrm{false}}^l + D_{\mathrm{rest}}^l}. \tag{14}$$

The inner products measure how strongly an update along the ELA direction acts on each logit. A larger $\mathrm{PairShare}_t^l$ means that the update is more concentrated on the confusing pair $\{y^*, c_t^{\mathrm{false}}\}$ than on the remaining classes. As shown in Figure 3(a), SeAl-KD gives a higher pair-share score than timestep-wise KD, suggesting that ELA yields more localized correction at erroneous timesteps.

Proposition 2 (Reference-guided correction)**.** *For a weak timestep, STA encourages the update to reduce its margin discrepancy to reliability-weighted temporal references, thereby providing guidance from confident and compatible states.*

Figure 3: Layer-wise statistics over all timesteps for the three propositions: (a) the fraction of the ELA update assigned to the ground-truth class and the dominant false class at erroneous timesteps; (b) the cosine similarity between the STA update and the direction that reduces the gap to reliability-weighted temporal references at weak timesteps; (c) the ratio between the distillation-gradient norm and the classification-gradient norm at already-correct timesteps. Statistics are computed from five randomly selected samples and reported as mean ± std.

	Method	Architecture	CIFAR-10		CIFAR-100	
			T=4	T=6	T=4	T=6
w/o KD	Dspike [Li *et al.*, 2021b]	ResNet-18	93.66	94.25	73.35	74.24
	GLIF [Yao *et al.*, 2022]		94.67	94.88	76.42	77.28
	RateBP [Yu *et al.*, 2024]		95.61	95.90	78.26	79.02
	STBP-tdBN [Zheng *et al.*, 2021]	ResNet-19	92.92	93.16	–	–
	TET [Deng *et al.*, 2022]		94.44	94.50	74.47	74.72
	GLIF [Yao *et al.*, 2022]		94.85	95.03	77.05	77.35
	LSG [Lian *et al.*, 2023]		95.17	95.52	76.85	77.13
	RateBP [Yu *et al.*, 2024]		96.26	96.36	80.71	80.83
w/ KD	KDSNN [Xu *et al.*, 2023]	ResNet-18	93.41	–	–	–
	Joint A-SNN [Guo *et al.*, 2023]		95.45	–	77.39	–
	Rate-based KD [Yang *et al.*, 2025]		95.92	96.14	78.85	79.40
	Logit-SNN [Yu *et al.*, 2025a]		95.57	95.96	79.10	79.80
	SeAl-KD (Ours)		**95.88**$_{\pm 0.12}$	**96.18**$_{\pm 0.06}$	**79.88**$_{\pm 0.13}$	**80.25**$_{\pm 0.12}$
	SAKD [Qiu *et al.*, 2024]	ResNet-19	96.06	–	80.10	–
	HTA-KL [Zhang *et al.*, 2025]		96.76	–	81.03	–
	Logit-SNN [Yu *et al.*, 2025a]		96.97	97.00	82.47	82.56
	SeAl-KD (Ours)		**97.14**$_{\pm 0.06}$	**97.23**$_{\pm 0.05}$	**83.04**$_{\pm 0.05}$	**83.37**$_{\pm 0.08}$

Table 1: Comparison of different direct-training and distillation methods on CIFAR-10 and CIFAR-100.

Let $m_t = z^S_{t,y^*} - z^S_{t,c_t^{\text{false}}}$ denote the student margin at timestep t, where the margin measures the logit gap between the ground-truth class y^* and the dominant false class c_t^{false}. A larger margin indicates that the student assigns stronger relative preference to the ground-truth class over the confusing false class. Let $m_{\text{ref},t}$ be the reference margin for timestep t, obtained by aggregating source-timestep margins according to the STA weights. We measure whether the STA update is aligned with the direction that reduces the discrepancy between m_t and $m_{\text{ref},t}$:

$$\text{RefAlign}_t^l = \cos\Big(\nabla_{\theta_l}\mathcal{L}_{\text{STA}}, \nabla_{\theta_l}(m_t - m_{\text{ref},t})^2\Big). \quad (15)$$

A larger RefAlign_t^l indicates stronger alignment with the discrepancy-reduction direction, suggesting that STA guides weak timesteps toward reliability-weighted temporal references. Rather than forcing each intermediate timestep to be independently correct, STA promotes more consistent evidence accumulation across timesteps. As shown in Figure 3(b), SeAl-KD produces positive alignment, whereas timestep-wise KD shows weaker or negative alignment.

Proposition 3 (Restrained correction)**.** *When a timestep is already correctly classified, the distillation update remains small relative to the task gradient, reducing unnecessary interference with reliable predictions.*

For already-correct timesteps, a large distillation update may perturb a reliable prediction. We therefore measure the relative strength of the distillation gradient against the task gradient:

$$\text{KDRatio}_t^l = \frac{\|\nabla_{\theta_l}\mathcal{L}_{\text{KD}}\|}{\|\nabla_{\theta_l}\mathcal{L}_{\text{CLS}}\|}. \quad (16)$$

A smaller KDRatio_t^l indicates weaker interference from the distillation objective. As shown in Figure 3(c), SeAl-

	Method	Acc(%)
w/o KD	TET [Deng *et al.*, 2022]	68.00
	Dspike [Li *et al.*, 2021b]	68.19
	GLIF [Yao *et al.*, 2022]	67.52
	RecDis-SNN [Guo *et al.*, 2022]	67.33
	RateBP [Yu *et al.*, 2024]	70.01
w/ KD	KDSNN [Xu *et al.*, 2023]	67.18
	LaSNN [Hong *et al.*, 2025]	66.94
	BKDSNN [Xu *et al.*, 2024]	67.21
	Logit-SNN [Yu *et al.*, 2025a]	71.04
	SeAl-KD (Ours)	**72.11**

Table 2: Comparison of different direct-training and distillation methods on ImageNet with ResNet-34.

KD yields a much smaller KD-to-CLS gradient ratio than timestep-wise KD at the representative correct timestep, indicating that it remains restrained when the prediction is already reliable.

5 Experiments

5.1 Implementation Details

Datasets. We evaluate the proposed method on four benchmark datasets, including three static image datasets, namely CIFAR-10, CIFAR-100, and ImageNet, as well as one neuromorphic dataset, DVS-CIFAR10. This selection allows us to assess the effectiveness of the proposed approach on both frame-based vision tasks and event-based neuromorphic data. Dataset details are provided in the **Appendix C**.

Training Settings. All experiments are implemented in PyTorch and trained on NVIDIA A100 GPUs. The SNNs are built with leaky integrate-and-fire neurons and trained using surrogate gradient learning [Neftci *et al.*, 2019]. All experiments are conducted three times and the average results are reported, except for ImageNet. For neuromorphic datasets, ANN training uses the average value of the event data as input. Training details are provided in the **Appendix D**.

5.2 Performance Comparison

The results on CIFAR-10, CIFAR-100, ImageNet, and DVS-CIFAR10 are summarized in Tables 1, 2, and 3. Across all evaluated benchmarks, SeAl-KD consistently outperforms directly trained SNNs and achieves superior or competitive performance compared with representative logit-based SNN distillation methods and other distillation baselines under comparable architectures and inference timesteps. The improvements are observed on both static frame-based datasets and event-based neuromorphic datasets, indicating that the proposed method generalizes well across different data modalities. In addition, SeAl-KD only introduces lightweight training-time objectives, and a detailed energy analysis is provided in **Appendix E**.

5.3 Ablation Study

Component Ablation of SeAl-KD

We first evaluate the individual and joint contributions of ELA and STA in Table 4. Removing either module degrades performance, and removing both causes a larger drop. This indicates that SeAl-KD benefits from making alignment selective in both class and temporal dimensions: ELA reduces misleading supervision on erroneous classes, while STA provides temporal guidance from more reliable states.

	Method	Architecture	T	Acc(%)
w/o KD	STBP-tdBN [Zheng *et al.*, 2021]	ResNet-19	10	67.80
	Dspike [Li *et al.*, 2021b]	ResNet-18	10	75.40
	TET [Deng *et al.*, 2022]	VGG-11	10	83.17
	SLTT [Meng *et al.*, 2023]	VGG-11	10	82.20
	Spikformer [Zhou *et al.*, 2023]		16	80.90
	Spike-driven Transformer [Yao *et al.*, 2023]		16	80.00
w/ KD	SAKD [Qiu *et al.*, 2024]	VGG-11	4	81.50
		ResNet-19	4	80.30
	TSSD [Zuo *et al.*, 2024]	ResNet-18	8	72.90
			16	81.60
	LogitSNN [Yu *et al.*, 2025a]	ResNet-18	4	83.50
			10	86.40
	SeAl-KD (Ours)	ResNet-18	4	**84.00**$_{\pm 0.20}$
			10	**86.70**$_{\pm 0.20}$

Table 3: Comparison of different direct-training and distillation methods on DVS-CIFAR10.

Method	CIFAR10	CIFAR100	DVS-C10
Timestep-wise KD	95.04	78.11	80.90
w/ ELA	95.95	79.86	**84.20**
w/ STA	95.60	78.76	83.80
SeAl-KD	**96.00**	**80.01**	**84.20**

Table 4: Component ablation of SeAl-KD using ResNet-18 across datasets under $T = 4$.

Ablation Study of ELA Variants

We analyze several ELA variants, with results reported in Table 5. **ELA-S** applies error-aware modification only to the student logits, **ELA-A** only to the teacher logits, **ELA-AS** introduces error awareness on the ANN teacher side and aligns the student accordingly, and **ELA-Both** extends the alignment to three classes when the teacher prediction is also incorrect. All variants outperform the baseline without ELA, showing that class-level selective correction is beneficial once alignment is no longer imposed uniformly. Our ELA achieves the best overall performance, and student-driven variants are consistently more effective than teacher-driven ones, because student errors directly identify the timesteps where naive alignment is most misleading. This supports that ELA should be selective, student-driven, and focused on correcting the dominant confusing class relation rather than enforcing broader per-timestep matching.

Ablation Study of STA Variants

We analyze several STA variants, with results reported in Table 6. **STA-NoConf** removes confidence weighting, **STA-NoSim** removes similarity filtering, **STA-Dist** replaces cosine similarity with cosine distance, and **UTA** enforces unweighted pairwise alignment across all timesteps. All variants outperform the baseline without STA, confirming the

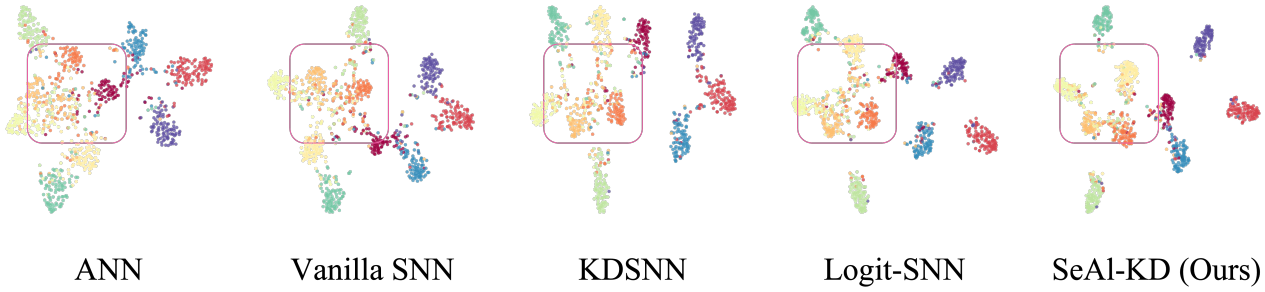

Figure 5: t-SNE visualization of learned feature representations on DVS-CIFAR10 under different direct-training and distillation methods.

Method	CIFAR-10	CIFAR-100	DVS-CIFAR10
w/o ELA	95.04	78.11	80.9
ELA-S	95.85	79.56	**84.2**
ELA-A	95.78	79.34	83.6
ELA-AS	95.76	79.29	82.9
ELA-Both	95.82	79.36	83.9
ELA (Ours)	**95.95**	**79.86**	**84.2**

Table 5: Ablation study of ELA variants using ResNet-18 across different datasets under $T = 4$.

Method	CIFAR-10	CIFAR-100	DVS-CIFAR10
w/o STA	95.04	78.11	80.9
UTA	95.55	78.58	82.4
STA-NoConf	95.32	78.57	**84.1**
STA-NoSim	95.58	78.60	83.6
STA-Dist	95.37	78.60	83.7
STA (Ours)	**95.60**	**78.76**	83.8

Table 6: Ablation study of STA variants using ResNet-18 across different datasets under $T = 4$.

benefit of temporal guidance, while the full STA performs best overall. This suggests that temporal correction should not be propagated uniformly across timesteps, but should instead guide erroneous timesteps using temporal states that are both reliable and compatible. Confidence weighting suppresses noisy sources, similarity filtering avoids mismatched guidance, and their combination yields the most effective selective temporal correction.

5.4 Visualization

Temporal Discrepancy

To better understand how selective alignment affects temporal prediction dynamics, we visualize per-timestep class logits on DVS-CIFAR10 in Figure 6. Compared with Logit-SNN, which exhibits a largely time-invariant pattern where a subset of classes remains dominant across most timesteps, our method produces more structured temporal trajectories. The dominant class can shift over time, and certain classes show larger logit variation across timesteps. For example, given that the ground-truth label is $C3$, the logit of $C0$ transitions from negative to strongly positive. Overall, the visualization suggests that SeAI-KD redistributes evidence over time and

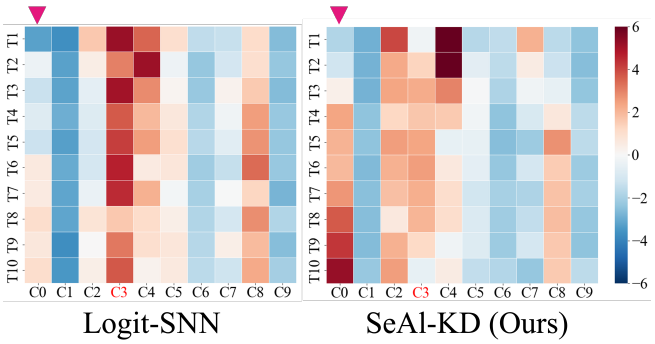

Figure 6: Heatmap of per-timestep class logits on DVS-CIFAR10 for Logit-SNN and SeAI-KD.

encourages temporally differentiated class responses, rather than maintaining a fixed class preference across timesteps.

t-SNE

To better understand the effect of selective temporal alignment, we visualize the learned representations using t-SNE, as shown in Figure 5. Compared with direct training and previous distillation methods, our approach produces more compact class clusters with clearer separation. Notably, the circled regions exhibit less class overlap and fewer scattered samples. Overall, the visualization indicates that selective alignment leads to more discriminative and well-structured representations over time.

6 Conclusion

This work examines the limitations of uniform timestep-wise KD in SNNs and shows that treating all temporal states equally can impose overly rigid supervision, even though intermediate timesteps need not all be individually correct. Instead, erroneous timesteps should receive guidance that moves them toward the correct final outcome. We propose SeAl-KD, which selectively aligns class-level and temporal knowledge via ELA and STA. Our analysis shows that ELA and STA reduce the influence of erroneous, low-confidence, or incompatible timesteps during alignment, enabling more effective correction where it is needed. Experiments on both static and neuromorphic datasets further confirm that this selective alignment strategy consistently improves accuracy with only minimal computational overhead.

Ethical Statement

There are no ethical issues.

Acknowledgements

This work was supported by start-up funds with No. MSRI8001004 and No. MSRI9002005 and Monash eResearch capabilities, including M3.

References

[Bellec *et al.*, 2018] Guillaume Bellec, Darjan Salaj, Anand Subramoney, Robert Legenstein, and Wolfgang Maass. Long short-term memory and learning-to-learn in networks of spiking neurons. *Advances in neural information processing systems*, 31, 2018.

[Deng *et al.*, 2009] 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*, pages 248–255. Ieee, 2009.

[Deng *et al.*, 2022] Shikuang Deng, Yuhang Li, Shanghang Zhang, and Shi Gu. Temporal efficient training of spiking neural network via gradient re-weighting. In *International Conference on Learning Representations*, 2022.

[Ding *et al.*, 2025a] Yongqi Ding, Lin Zuo, Mengmeng Jing, Pei He, and Hanpu Deng. Rethinking spiking neural networks from an ensemble learning perspective. *International Conference on Learning Representations*, 2025.

[Ding *et al.*, 2025b] Yongqi Ding, Lin Zuo, Mengmeng Jing, Kunshan Yang, Pei He, and Tonglan Xie. Synergy between the strong and the weak: Spiking neural networks are inherently self-distillers. *Advances in neural information processing systems*, 2025.

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

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

[Guo *et al.*, 2022] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 326–335, 2022.

[Guo *et al.*, 2023] Yufei Guo, Weihang Peng, Yuanpei Chen, Liwen Zhang, Xiaode Liu, Xuhui Huang, and Zhe Ma. Joint a-snn: Joint training of artificial and spiking neural networks via self-distillation and weight factorization. *Pattern Recognition*, 142:109639, 2023.

[Han *et al.*, 2015] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015.

[Hong *et al.*, 2025] Di Hong, Yu Qi, and Yueming Wang. Lasnn: Layer-wise ann-to-snn distillation for effective and efficient training in deep spiking neural networks. *Neurocomputing*, page 131351, 2025.

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

[Li *et al.*, 2021a] 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.

[Li *et al.*, 2021b] 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.

[Lian *et al.*, 2023] Shuang Lian, Jiangrong Shen, Qianhui Liu, Ziming Wang, Rui Yan, and Huajin Tang. Learnable surrogate gradient for direct training spiking neural networks. In *IJCAI*, pages 3002–3010, 2023.

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

[Meng *et al.*, 2023] Qingyan Meng, Mingqing Xiao, Shen Yan, Yisen Wang, Zhouchen Lin, and Zhi-Quan Luo. Towards memory-and time-efficient backpropagation for training spiking neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6166–6176, 2023.

[Neftci *et al.*, 2019] 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.

[Orchard *et al.*, 2015] 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.

[Qiu *et al.*, 2024] Haonan Qiu, Munan Ning, Zeyin Song, Wei Fang, Yanqi Chen, Tao Sun, Zhengyu Ma, Li Yuan, and Yonghong Tian. Self-architectural knowledge distillation for spiking neural networks. *Neural Networks*, 178:106475, 2024.

[Wang *et al.*, 2023] 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*, pages 35798–35816. PMLR, 2023.

[Xu *et al.*, 2023] Qi Xu, Yaxin Li, Jiangrong Shen, Jian K Liu, Huajin Tang, and Gang Pan. Constructing deep spiking neural networks from artificial neural networks with knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7886–7895, 2023.

[Xu *et al.*, 2024] Zekai Xu, Kang You, Qinghai Guo, Xiang Wang, and Zhezhi He. Bkdsnn: Enhancing the performance of learning-based spiking neural networks training with blurred knowledge distillation. In *European Conference on Computer Vision*, pages 106–123. Springer, 2024.

[Yang *et al.*, 2025] Shu Yang, Chengting Yu, Lei Liu, Hanzhi Ma, Aili Wang, and Erping Li. Efficient ann-guided distillation: Aligning rate-based features of spiking neural networks through hybrid block-wise replacement. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10025–10035, 2025.

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

[Yao *et al.*, 2023] 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.

[Yu *et al.*, 2024] Chengting Yu, Lei Liu, Gaoang Wang, Erping Li, and Aili Wang. Advancing training efficiency of deep spiking neural networks through rate-based backpropagation. *Advances in Neural Information Processing Systems*, 37:115786–115815, 2024.

[Yu *et al.*, 2025a] Chengting Yu, Xiaochen Zhao, Lei Liu, Shu Yang, Gaoang Wang, Erping Li, and Aili Wang. Efficient logit-based knowledge distillation of deep spiking neural networks for full-range timestep deployment. *In Proceedings of the 42nd International Conference on Machine Learning*, 2025.

[Yu *et al.*, 2025b] Kairong Yu, Chengting Yu, Tianqing Zhang, Xiaochen Zhao, Shu Yang, Hongwei Wang, Qiang Zhang, and Qi Xu. Temporal separation with entropy regularization for knowledge distillation in spiking neural networks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8806–8816, 2025.

[Zhang *et al.*, 2025] Tianqing Zhang, Zixin Zhu, Kairong Yu, and Hongwei Wang. Head-tail-aware kl divergence in knowledge distillation for spiking neural networks. *arXiv preprint arXiv:2504.20445*, 2025.

[Zhao *et al.*, 2025] Dongcheng Zhao, Guobin Shen, Yiting Dong, Yang Li, and Yi Zeng. Improving stability and performance of spiking neural networks through enhancing temporal consistency. *Pattern Recognition*, 159:111094, 2025.

[Zheng *et al.*, 2021] Hanle Zheng, Yujie Wu, Lei Deng, Yifan Hu, and Guoqi Li. Going deeper with directly-trained larger spiking neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11062–11070, 2021.

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

[Zuo *et al.*, 2024] Lin Zuo, Yongqi Ding, Mengmeng Jing, Kunshan Yang, and Yunqian Yu. Self-distillation learning based on temporal-spatial consistency for spiking neural networks. *arXiv preprint arXiv:2406.07862*, 2024.