CLASSIFIER EVIDENCE GUIDED TOKEN SELECTION FOR SPIKING TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spiking transformers achieve strong performance in neuromorphic vision, but multi-step token processing still introduces substantial redundancy and inference cost. Existing token reduction methods mainly rely on response-driven cues and do not explicitly exploit the temporal class evidence encoded by the frozen classifier. Since token utility is ultimately reflected in downstream prediction, we investigate the classifier as an additional source of ranking information. We observe that different tokens exhibit distinct uncertainty trajectories across spiking steps, suggesting that temporally aggregated classifier evidence can provide a complementary signal for token selection. Based on this observation, we propose CETS (Classifier-Evidence Token Selection), a training-free and plug-and-play framework that models token-wise class evidence with a Dirichlet distribution and aggregates its temporal uncertainty statistics into an inference-time ranking score. CETS requires no auxiliary predictor or parameter update and can be directly applied to token reduction. Experiments on static and neuromorphic benchmarks demonstrate favorable accuracy-efficiency trade-offs under diverse retention ratios. Further analyses show that the discriminative strength of classifier-derived uncertainty varies across representations, providing empirical insight into when classifier-side evidence is informative for token selection.

1 INTRODUCTION

The human brain achieves remarkable visual efficiency under strict energy constraints, partly through selective neural activation, where only a subset of neurons responds strongly to relevant stimuli while less useful responses are suppressed Vinje & Gallant (2000); Yoshida & Ohki (2020). This selective processing mechanism suggests an appealing principle for efficient artificial perception: computational resources should be preferentially allocated to informative signals.

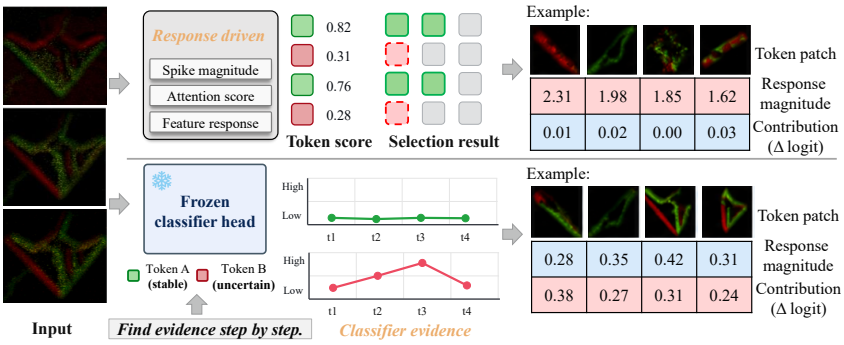

Figure 1: **Motivation of CETS.**

Inspired by such selective processing in biological vision, neuromorphic vision systems Zhou et al. (2023; 2025) exploit event-driven sensing Maass (1997) and spiking neural networks to achieve sparse and temporally precise perception. Building on this paradigm, spiking transformers further introduce self-attention to model long-range spatio-temporal dependencies. However, the sparsity of spikes does not automatically translate into efficient token processing: once transformed into token sequences across multiple simulation steps, spiking representations can still contain substantial redundancy, leading to nontrivial inference cost Hao et al. (2025); Huang et al. (2024; 2025).

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.

A natural way to improve efficiency is token reduction Bolya et al. (2023). Existing methods typically estimate token importance from response-driven cues, such as activation magnitude, firing statistics, feature similarity, or lightweight learned predictors, as illustrated in Figure 1. These signals are readily available, but they primarily characterize tokens from their intermediate responses and do not explicitly exploit how the frozen classifier maps token representations to task predictions. This motivates us to ask whether classifier-side information can provide an additional cue for token ranking, without introducing a separately trained selector Zeng et al. (2022); Rao et al. (2021). Specifically, a complementary question has been raised: *can classifier-derived evidence provide a useful token-ranking signal for training-free token selection?*

To probe classifier-side information, we apply the frozen classifier head to token representations at each spiking step, obtaining token-wise class logits and the corresponding uncertainty. Our key observation is that this uncertainty evolves differently across time for different tokens, as illustrated in Figure 1. Such heterogeneity motivates us to characterize classifier-derived evidence over the temporal dimension rather than at a single step.

Motivated by this observation, we propose **CETS**, a training-free and plug-and-play token ranking framework for spiking transformers. Specifically, CETS models token-wise class evidence with a Dirichlet distribution and summarizes its temporal statistics across spiking steps, yielding an evidence-derived ranking score for inference-time token selection. The resulting score can be directly applied to token pruning, without introducing auxiliary predictors or additional parameter updates. Experiments on both static and neuromorphic benchmarks demonstrate that CETS provides useful token-ranking signals under diverse retention ratios, while further analyses reveal that its effectiveness is closely related to the interaction between classifier-derived evidence and intermediate token representations.

Our contributions are summarized as follows:

- We study classifier-derived evidence as a training-free token-ranking signal in spiking transformers and analyze its temporal evolution across simulation steps.
- We propose **CETS**, which reuses the frozen classifier head to derive temporally aggregated Dirichlet evidence scores for token selection without additional training.
- We provide the analyses of ranking utility, classifier projection, score parameterization, and backbone dependence, revealing that the usefulness of classifier evidence is representation dependent and that temporal variation provides dataset-dependent refinement.

2 Preliminaries and Related Works

Spiking neural networks (SNNs), often considered the third generation of neural networks, have attracted growing interest due to their event-driven computation and remarkable energy efficiency. In this work, we employ the widely adopted Leaky Integrate-and-Fire (LIF) neuron Izhikevich (2004); Gerstner & Kistler (2002) model to validate the feasibility of the proposed CETS. At each time step t, the membrane potential integrates incoming spikes from the previous layer while gradually decaying over time:

$$U_i^l[t] = \tau U_i^l[t-1] + W_i^l S^{l-1}[t], \tag{1}$$

where $0 < \tau < 1$ denotes the membrane decay factor that models the leakage of the membrane potential. When the membrane potential exceeds the firing threshold V_{th}, the neuron emits a spike:

$$S_i^l[t] = H(U_i^l[t] - V_{th}), \tag{2}$$

where $H(\cdot)$ denotes the Heaviside step function. After spike emission, the membrane potential is reset as

$$U_i^l[t] = U_i^l[t](1 - S_i^l[t]). \tag{3}$$

The temporal integration and leakage process introduces variability across time steps. Early SNNs are primarily built upon convolutional architectures, which mainly focus on local feature extraction Zhong et al. (2024).

Spiking Transformers. Transformer architectures have recently demonstrated strong capability in capturing global contextual dependencies through self-attention Li et al. (2022); Liu et al. (2025). Inspired by the success, spiking transformers have been proposed to integrate the spike-driven computation of SNNs with the global modeling ability Yao et al. (2024a). Spikformer Zhou et al. (2023)

pioneers the introduction of self-attention and enables direct Transformer training in SNNs by employing spike-based Q, K, and V representations, eliminating floating-point multiplications in attention computation. Spikformer V2 further explores masked image modeling in spiking transformers Zhou et al. (2024b). SpikingResformer Shi et al. (2024) integrates dual-spike self-attention with a ResNet-style architecture, improving performance while reducing parameter complexity. QKFormer Zhou et al. (2024a) adopts spike-based Q and K for attention computation and introduces spiking patch embeddings with deformable shortcuts. Meanwhile, Spike-driven Transformer (SDT) series Yao et al. (2023; 2024b; 2025), which introduces spike-driven attention mechanisms to enable efficient Transformer computation in SNNs. Besides, MaxFormer restores high-frequency signals through frequency-enhancing operators Fang et al. (2025).

3 Empirical Observation

Spiking transformers process visual information through tokens that evolve across multiple discrete time steps. Let z_i^t denote the representation of token i at time step t, where $t = 1, \ldots, T$. Unlike conventional transformers with static token representations, spiking transformers exhibit temporally evolving token states induced by sparse neural firing, motivating us to examine whether classifier-derived scores show corresponding dynamics across simulation steps.

3.1 Temporal Dynamics and Selection Utility of Classifier Evidence

To better understand the temporal variation of tokens in spiking transformers, we analyze token-wise uncertainty trajectories across time steps. For each token i at time step t, we estimate an uncertainty score U_i^t from the frozen classifier responses using evidential learning Shafer (1976), forming a temporal trajectory $\{U_i^t\}_{t=1}^T$. We visualize representative tokens with consistently low uncertainty, pronounced temporal fluctuation, and persistently high uncertainty. As shown in Figure 2(a), different tokens exhibit distinct temporal behaviors, including consistently low, strongly fluctuating, and persistently high uncertainty. These patterns indicate that classifier-derived uncertainty is not fully characterized by a single simulation step, motivating us to summarize its temporal statistics for token ranking.

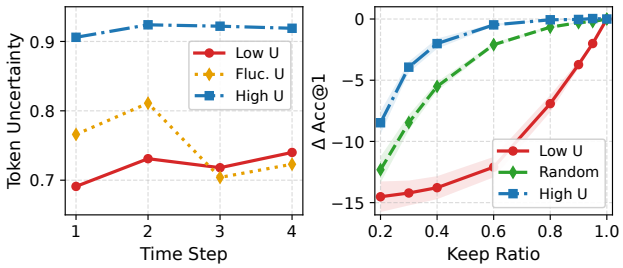

Figure 2: **Temporal evidence and selection utility on CIFAR-100.** (a) Representative low-, fluctuating-, and high-uncertainty trajectories across simulation steps. (b) Acc@1 change relative to the dense model when retaining low-U, random, or high-U token subsets at different retention ratios.

We examine whether the resulting ranking contains useful token-selection utility information. Under the same pretrained checkpoint and token retention ratio, we compare retaining high-U tokens, randomly selected tokens, and low-U tokens. As shown in Figure 2(b), high-U retention consistently preserves higher accuracy than both alternatives, with a larger separation under stronger token reduction. This comparison shows that the induced ordering is not arbitrary and motivates using temporally aggregated classifier evidence as an inference-time ranking signal.

4 Method

In this section, we introduce **Classifier-Evidence Token Selection (CETS)**, a training-free framework for token ranking in pretrained spiking transformers, as illustrated in Figure 3. CETS derives token-wise evidence from the frozen classifier head, summarizes its temporal behavior across simulation steps, and uses the resulting score to select spatial tokens for the target spiking self-attention module.

4.1 Spiking Token Representation

We adopt QKFormer Zhou et al. (2024a) as the primary spiking transformer backbone. Let $\mathcal{I} \in \mathbb{R}^{T_0 \times B \times C_{\text{in}} \times H \times W}$ denote a batched input containing T_0 frames. For a static RGB image, $T_0 = 1$

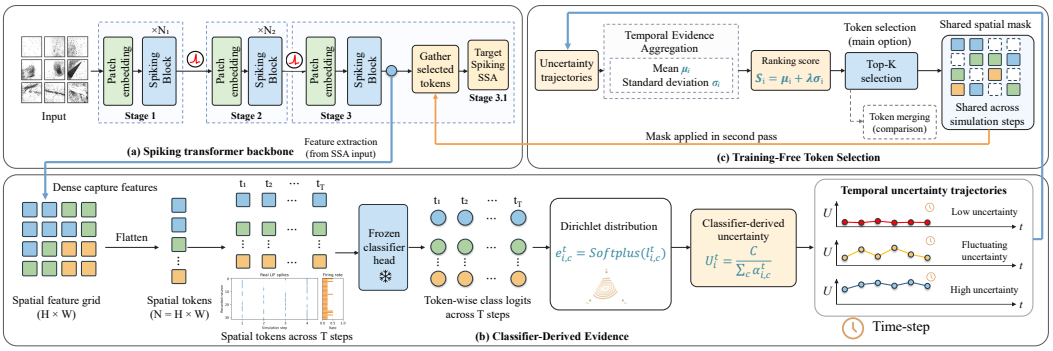

Figure 3: **Overall framework of the proposed CETS.** CETS includes the basic spiking transformer backbone (top) and the spiking token uncertainty estimation (bottom). (a) shows the spiking transformer backbone, instantiated with QKFormer Zhou et al. (2024a) or MaxFormer Fang et al. (2025). (b) presents the proposed spiking-temporal uncertainty modeling. (c) shows uncertainty-aware token selection for dropping and merging.

and $C_{\text{in}} = 3$; the image is repeated internally over T simulation steps to drive the spiking dynamics. For frame-based neuromorphic inputs, we use $T_0 = T$ event frames with two polarity channels, *i.e.,* $C_{\text{in}} = 2$.

The backbone contains three hierarchical stages. We denote the feature map at stage s as

$$X^{(s)} \in \mathbb{R}^{T\times B\times C_s\times H_s\times W_s}, \qquad s \in \{1, 2, 3\}, \tag{4}$$

where C_s and (H_s, W_s) denote its channel dimension and spatial resolution. Flattening the spatial dimensions yields the equivalent token representation

$$Z^{(s)} \in \mathbb{R}^{T\times B\times N_s\times D_s}, \qquad N_s = H_s W_s, \quad D_s = C_s. \tag{5}$$

Stage 1 applies the initial spiking patch embedding followed by a Token-QK Transformer module. Stage 2 performs spatial downsampling and channel expansion through another patch-embedding module, followed by a second Token-QK Transformer module. Stage 3 further downsamples the representation and processes it using multiple spiking transformer modules equipped with spike-driven self-attention. Consequently, spatial token count decreases while channel dimension increases with network depth, producing a hierarchical spatio-temporal token representation.

CETS operates on these frozen intermediate representations and reuses the pretrained classification head to construct its token-ranking signal. It does not update or otherwise modify the backbone parameters.

4.2 Classifier-Derived Evidence

Let $z_i^t \in \mathbb{R}^D$ denote the representation of token i at time step t, where $i = 1, \ldots, N$ indexes the spatial tokens and $t = 1, \ldots, T$ denotes the simulation time steps. To characterize token-level class evidence, we estimate temporal uncertainty from classifier responses using a Dirichlet distribution.

Specifically, given token representation z_i^t, the classifier head first produces class logits

$$l_{i,c}^t = g_c(z_i^t), \tag{6}$$

which are converted into non-negative evidence values using the Softplus activation,

$$e_{i,c}^t = \text{Softplus}(l_{i,c}^t) = \log(1 + \exp(l_{i,c}^t)), \tag{7}$$

where Softplus is adopted to ensure that the evidence values are non-negative. The Dirichlet parameters are then defined as

$$\alpha_{i,c}^t = e_{i,c}^t + 1, S_i^t = \sum_{c=1}^{C} \alpha_{i,c}^t \tag{8}$$

S_i^t denotes the total evidence. Following evidential uncertainty modeling, the uncertainty of token i at time step t is defined as

$$U_i^t = \frac{C}{S_i^t}. \tag{9}$$

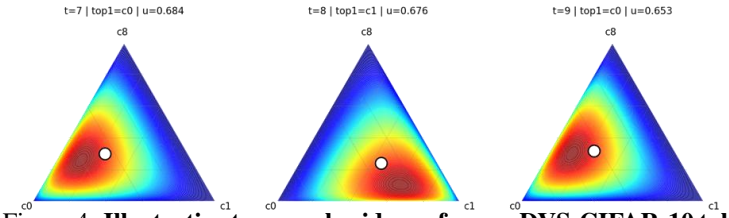

Figure 4: **Illustrative temporal evidence for one DVS-CIFAR-10 token.** Three consecutive steps ($t = 7, 8, 9$) of the same token are visualized over fixed class vertices. The marker represents the relative evidence assigned to the three displayed classes, illustrating how the class-evidence distribution evolves across time. The U_i^t is computed from the full class evidence vector.

This formulation yields a temporal uncertainty trajectory $\{U_i^t\}_{t=1}^T$ for each token, across simulation steps. Figure 4 visualizes the classifier-derived evidence of one token over consecutive simulation steps. Although only three class dimensions are shown for illustration, U_i^t is computed from the full class-evidence vector. The changing marker position illustrates the temporal evolution of the token's class-evidence distribution.

Why inverse evidence strength? Softmax probabilities describe relative class preference after normalization and are invariant to a common shift of all logits. They therefore do not provide an explicit measure of the overall evidence strength across classes. The Dirichlet parameterization instead maps the frozen classifier responses to non-negative evidence and summarizes their total concentration through $\sum_c \alpha_{i,c}^t$. The inverse Dirichlet strength yields a scalar U_i^t for each token and simulation step, which can be directly aggregated over time. We use this construction as a convenient training-free evidence parameterization, rather than as a claim that it is calibrated or universally superior to softmax-based scores.

4.3 Temporal Evidence Aggregation

Given the trajectory $\{U_i^t\}_{t=1}^T$, we summarize its temporal level and variation as

$$\mu_i = \frac{1}{T}\sum_{t=1}^{T} U_i^t, \qquad \sigma_i = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(U_i^t - \mu_i)^2}, \tag{10}$$

where μ_i captures the average inverse evidence strength across simulation steps, while σ_i measures its temporal fluctuation. We combine the two statistics to construct the ranking score

$$S_i = \mu_i + \lambda\sigma_i, \tag{11}$$

where λ controls the contribution of temporal variation. We use $\lambda = 0.9$ in the reported frozen protocol unless otherwise specified, and rank tokens in descending order of S_i (Sensitive ablation analysis can be found in Appendix A.6).

4.4 Training-Free Token Selection

Given the scores $\{S_i\}_{i=1}^N$, we retain the top-K spatial tokens under a prescribed retention ratio r:

$$\mathcal{I}_{\text{keep}} = \text{TopK}(S, N_{\text{keep}}), \qquad N_{\text{keep}} = \lfloor rN \rfloor. \tag{12}$$

The same spatial index set is shared across simulation steps, preserving the temporal correspondence of selected tokens. During the SSA computation, only the selected tokens $X_{t,\mathcal{I}_{\text{keep}}}^{\ell-1}$ are used to compute the query, key, and value representations:

$$\boldsymbol{Q} = f_q(X_{t,\mathcal{I}_{\text{keep}}}^{\ell-1}), \quad \boldsymbol{K} = f_k(X_{t,\mathcal{I}_{\text{keep}}}^{\ell-1}), \quad \boldsymbol{V} = f_v(X_{t,\mathcal{I}_{\text{keep}}}^{\ell-1}), \tag{13}$$

$$X_t^\ell = (\boldsymbol{Q}\boldsymbol{K}^\top)\boldsymbol{V}. \tag{14}$$

After attention, the updated selected tokens are scattered back to their original spatial locations to restore the dense layout required by subsequent layers. All model parameters remain frozen throughout the procedure. Besides, we have compared the merging operator and complete accuracy results in Appendix A.5.

Table 1: **Comparison of training-free token-ranking methods on CIFAR-100 and DVS-CIFAR-10.** All methods use the same frozen checkpoint, pruning site, and token retention ratio within each dataset. We report Acc@1 (%). Notably, the four retention ratios correspond to 60/64, 51/64, 38/64, and 25/64 retained tokens. Random reports mean ± standard deviation over five masks. IRToP is adapted from TP-Spikformer Wei et al. (2026).

Dataset	Token Selection	95%	80%	60%	40%
CIFAR-100	Dense (no pruning)	80.98	80.98	80.98	80.98
	Random	80.67±0.12	80.21±0.10	78.81±0.11	75.36±0.22
	Temporal variance	80.86	80.90	80.46	78.18
	IRToP (adapted)†	80.86	80.75	79.57	76.10
	Ours	**80.95**	**81.06**	**80.55**	**79.06**
DVS-CIFAR-10	Dense (no pruning)	82.30	82.30	82.30	82.30
	Random	81.88±0.39	79.92±0.29	74.22±0.43	**60.60**±0.66
	Temporal variance	**82.70**	80.30	**74.50**	54.60
	IRToP (adapted)†	82.40	80.00	72.70	50.00
	Ours	82.50	**81.20**	74.10	55.40

5 EXPERIMENT

To systematically assess the proposed method, we investigate the following four research questions from the perspectives of effectiveness, mechanism analysis, component contribution, and cross-architecture generalization.

RQ1: Can classifier-derived evidence provide a useful training-free token-ranking signal for spiking transformers? ***RQ2:*** How do the learned classifier projection and its evidence parameterization affect the resulting token ranking? ***RQ3:*** How does the ranking behavior vary across different spiking-transformer representations and backbones? ***RQ4:*** Does the temporal evolution of classifier evidence provide additional information beyond its time-averaged level?

5.1 SETUP

Models and Datasets. We use QKFormer Zhou et al. (2024a) as the primary backbone for experiments on CIFAR-10 Krizhevsky (2009), CIFAR-100 Krizhevsky (2009), DVS-CIFAR-10 Li et al. (2017), and DVS-Gesture Amir et al. (2017), while MaxFormer Fang et al. (2025) is used only for the cross-backbone analysis in Section 5.4. For QKFormer, CETS is applied before the stage-3.1 spike-driven self-attention module, whose input contains $N = 64$ spatial tokens. CIFAR-10 and CIFAR-100 images are evaluated at 32×32 resolution with $T = 4$ simulation steps. DVS-CIFAR-10 uses $T = 10$ at 128×128 resolution, while DVS-Gesture follows its standard event-frame preprocessing with $T = 16$ steps. Details can be found in Appendix A.2.

Experimental Details We conduct all the experiments on one NVIDIA RTX 4090 GPU with 24 GB of memory. All token-selection experiments use frozen pretrained checkpoints, with no parameter updates. For each input sample, a single spatial mask is shared across all simulation steps. The results in the main paper use the CIFAR-100 checkpoint trained with seed 42. Random token selection is averaged over five independently generated masks. To evaluate checkpoint-level robustness, we additionally repeat the 80% and 60% retention experiments using three independently trained checkpoints with seeds 41, 42, and 43, which can be found in the Appendix A.7.

5.2 TOKEN-SELECTION UTILITY OF CLASSIFIER EVIDENCE (RQ1)

Analysis of token-selection utility. Table 1 compares CETS with representative training-free token-ranking criteria. On CIFAR-100, CETS outperforms random selection at all evaluated retention ratios and maintains accuracy close to the dense model under moderate pruning. Specifically, it achieves 81.06% Acc@1 at 80% retention, slightly above the dense reference, and 80.55% at 60% retention. On DVS-CIFAR-10, CETS outperforms random selection at the nominal 95%, 80%, and 60% retention ratios, but underperforms it at 40%. It achieves the highest accuracy among the compared ranking methods at 80% retention, whereas feature temporal variance performs best at 95%

Table 2: **Comparison with representative spiking-transformer models on static and neuromorphic benchmarks.** We report the number of parameters (Param. (M)), simulation steps (T), and Acc@1 (%) on CIFAR-10, CIFAR-100, DVS-Gesture, and DVS-CIFAR-10.

Method	CIFAR-10			CIFAR-100			DVS-Gesture			DVS-CIFAR-10		
	Param. (M)	T	Acc@1 (%)	Param. (M)	T	Acc@1 (%)	Param. (M)	T	Acc@1 (%)	Param. (M)	T	Acc@1 (%)
Spikformer Zhou et al. (2023)	9.32	4	95.51	9.32	4	78.21	2.57	16	98.30	2.57	16	80.90
S-Transformer Yao et al. (2023)	10.28	4	95.60	10.28	4	78.40	2.57	16	99.30	2.57	16	80.00
SWformer Fang et al. (2024)	7.51	4	96.10	7.51	4	79.30	-	-	-	2.05	16	83.90
STAtten Lee et al. (2025)	-	4	96.03	-	4	80.20	-	-	-	-	16	83.90
QKFormer Zhou et al. (2024a)	6.74	4	96.31	6.74	4	80.98	1.99	16	98.61	1.50	10	82.30
QKFormer *w/* Ours	6.74	4	96.30	6.74	4	81.06	1.99	16	96.88	1.50	10	82.50

and 60%, and random selection performs best at 40%. These results demonstrate the token-selection utility of CETS, while showing that its advantage over other ranking criteria depends on the dataset and retention ratio. More analysis can be found in Appendix A.3.

Analysis of existing spiking-transformer models. Table 2 shows that CETS can be applied to pretrained QKFormer models while preserving competitive classification accuracy across both static and neuromorphic benchmarks at moderate retention ratios. On CIFAR-100 and DVS-CIFAR-10, the selected-token models achieve accuracy comparable to or slightly above the corresponding dense checkpoints, whereas the behavior is less consistent on DVS-Gesture. These results support the practical applicability of CETS across different input modalities, while the ranking comparisons in Table 1 provide the primary evidence for its token-selection utility.

5.3 Understanding Classifier-Derived Evidence (RQ2)

In Table 3 (a), we report the Acc@1 difference between high-score and low-score retention, where a positive value indicates that retaining high-score tokens is more effective. The pretrained classifier produces a clear positive separation, whereas replacing it with a row-norm-matched random projection reverses the ordering. This indicates that the evidence transformation alone does not automatically yield a useful token ranking.

Table 3: **Comparison of ranking sources and classifier-side uncertainty signals on CIFAR-100.** Panel (a) reports the high-score minus low-score Acc@1 gap for different ranking sources. Panel (b) reports Acc@1 (%) for different classifier-derived uncertainty signals under the same temporal aggregation and token retention settings.

(a) Ranking source

Source	Perturbation	80%	60%
Trained head	Pretrained	+7.23	+12.12
Channel-perm.	Input channels	$+7.53\pm0.21$	$+12.15\pm0.11$
Random head	Norm matched	-3.37 ± 0.23	-6.40 ± 0.79
Feature norm	Temporal RMS	+8.03	+12.52

(b) Classifier-side signals

Signal	Step-wise score q_i^t	80%	60%
Max prob.	$1-\max_c p_{i,c}^t$	76.87	73.50
Entropy	$-\frac{\sum_c p_{i,c}^t \log p_{i,c}^t}{\log C}$	75.82	71.02
Prob. margin	$1-(p_{i,(1)}^t - p_{i,(2)}^t)$	78.44	77.00
Inv. Dirichlet	$\frac{C}{\sum_c \alpha_{i,c}^t}$	**81.06**	**80.55**

Interestingly, permuting the input channels of the trained classifier largely preserves the high–low gap. Thus, the observed ranking does not require an exact channel-wise correspondence between the intermediate representation and classifier weights in this setting. Feature norm also exhibits a strong high–low gap, suggesting that substantial selection structure is already present in the intermediate representation.

We further study how the classifier response should be parameterized. We compare inverse Dirichlet strength with three uncertainty signals derived from the same frozen classifier: one minus the maximum softmax probability, normalized predictive entropy, and one minus the top-two probability margin. As shown in Table 3 (b), inverse Dirichlet strength outperforms the strongest softmax-based alternative by 2.62 and 3.55 Acc@1 points at 80% and 60% retention, respectively. These results support the use of an explicit evidence-strength parameterization for token ranking in the evaluated QKFormer setting.

Visualization of uncertainty token-selection. Figure 5 visualizes the token-wise uncertainty maps and the resulting shared spatial mask on DVS-CIFAR-10. The uncertainty pattern changes across simulation steps, while the temporally aggregated statistics produce a single spatial ranking shared

across time. The visualization provides an example of how CETS converts temporally varying classifier evidence into a shared token-selection mask.

5.4 Classifier-Evidence Ranking Across Spiking-Transformer Backbones (RQ3)

Table 4 shows how the ranking signal changes across spiking-transformer backbones. We compare QKFormer and MaxFormer on the full CIFAR-100 test set using their respective frozen checkpoints and target attention sites. Both models use $T = 4$ simulation steps and $N = 64$ tokens at the selection site. For each token, we compare the CETS evidence score $R_i^{\mathrm{E}} = \mu_i + 0.9\sigma_i$ with the temporal RMS feature norm:

$$R_i^{\mathrm{FN}} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} \|z_i^t\|_2^2}. \tag{15}$$

For either ranking signal R, the high–low accuracy gap is defined as the Acc@1 difference between retaining its top-K and bottom-K tokens under the same retention ratio. At 80% and 60% retention, $K = 51$ and 38, respectively.

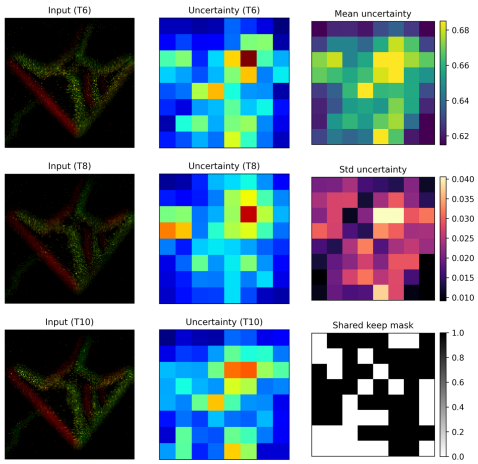

Figure 5: **Visualization of CETS on DVS-CIFAR-10 at 60% token retention.** Warmer colors indicate higher uncertainty, and white regions denote retained tokens.

We further compute the sample-wise Spearman correlation (details can be found in Appendix A.3) between R^{E} and R^{FN}, together with their Top-20% and Top-40% overlap, and average these statistics over the test set. As shown in Table 4, QKFormer exhibits pronounced high–low accuracy gap for both classifier evidence and feature norm. At 80%/60% retention, the evidence-based gaps reach 7.23/12.12 Acc@1 points, whereas they decrease to only 0.17/0.11 points on MaxFormer. Feature norm shows a similar reduction, from 8.03/12.52 to 1.06/1.74 points. The evidence–feature relationship also changes markedly across backbones. QKFormer shows a strong positive correlation ($\rho = 0.914$) with 75.54%/84.82% Top-20%/40% overlap, whereas MaxFormer shows a negative correlation ($\rho = -0.267$) and substantially lower overlaps of 7.43%/26.61%. These results show that evidence-ranking utility is strongly backbone dependent and co-varies with its relation to the intermediate representation.

5.5 Contribution of Temporal Variation (RQ4)

Table 5 examines whether temporal variation provides additional information beyond the mean uncertainty level. Δ denotes the Acc@1 difference between mean-plus-standard-deviation and mean-only scoring, in percentage points. Adding the standard-deviation term gives modest improvements on CIFAR-100 and DVS-CIFAR-10, with the largest gain of 0.80 Acc@1 points on DVS-CIFAR-10. In contrast, it reduces accuracy on DVS-Gesture. These results indicate that temporal fluctuation acts as a dataset- and retention-dependent refinement, rather than a universally beneficial source of ranking information. More analysis can be found in Appendix A.4.

Table 4: **Comparison of evidence-ranking behavior across spiking-transformer backbones.** We report high–low Acc@1 gaps and evidence–feature ranking relations for QKFormer and MaxFormer.

Backbone	Evidence gap		Feature-norm gap	
	80%	**60%**	**80%**	**60%**
QKFormer	+7.23	+12.12	+8.03	+12.52
MaxFormer	+0.17	+0.11	+1.06	+1.74
Evidence–feature ranking relation				
QKFormer	$\rho = +0.914$	Top-20%: 75.54%	Top-40%: 84.82%	
MaxFormer	$\rho = -0.267$	Top-20%: 7.43%	Top-40: 26.61%	

5.6 Efficiency Analysis

Target-module cost. Table 6 reports the theoretical operations per image for the selected SSA forward pass under different token retention ratios. The requested retention ratios correspond to 64,

Table 5: **Comparison of temporal aggregation strategies across datasets.** We compare mean-only scoring with mean-plus-variation scoring and report Acc@1 (%).

Dataset	Tokens	μ	$\mu + 0.9\sigma$	Δ
CIFAR-100	51/64	80.88	**81.06**	+0.18
CIFAR-100	38/64	80.44	**80.55**	+0.11
DVS-CIFAR-10	60/64	82.40	**82.50**	+0.10
DVS-CIFAR-10	51/64	80.40	**81.20**	+0.80
DVS-Gesture	51/64	**97.22**	96.88	−0.34
DVS-Gesture	25/64	**88.89**	87.50	−1.39

Table 6: **Comparison of target-SSA operation counts on CIFAR-100.** We report Acc@1 together with theoretical SOP and MAC counts under different token retention ratios.

Keep	Acc@1	SOP (M)	MAC (M)
100%	80.98	6.889	113.246
95%	81.03	6.348	106.168
90%	81.02	5.940	100.860
80%	**81.06**	5.124	90.243
60%	80.55	3.389	67.240
40%	78.91	1.819	44.237

60, 57, 51, 38, and 25 retained tokens out of 64 after integer truncation. As the number of retained tokens decreases, both spike-driven SOPs and non-binary projection MACs decrease substantially, while the accuracy remains relatively stable under moderate pruning. For example, retaining 38 tokens reduces SOPs from 6.889 M to 3.389 M and MACs from 113.246 M to 67.240 M, while maintaining an Acc@1 of 80.55%. These operation counts characterize the pruned SSA forward pass only.

To isolate the token-aggregation choice, we compare dropping low-ranked tokens with merging them into the retained tokens under the same frozen checkpoint, uncertainty ranking, target SSA, and token retention ratios.

Figure 6(a) shows that merging is comparable to dropping at moderate retention, but is less robust under aggressive pruning. In panel (b), the observed mean Merge-minus-Drop latency is positive across the tested ratios, although several paired confidence intervals overlap zero.

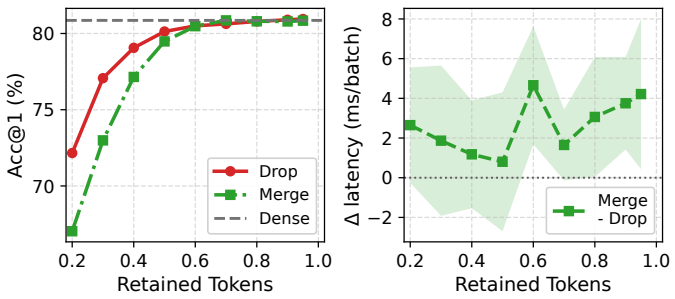

Figure 6: **Comparison of token dropping and merging on CIFAR-100.** Both strategies use the same frozen checkpoint, uncertainty ranking, target SSA, and token retention setting. Merging assigns discarded tokens to their cosine-nearest retained tokens and averages their features. We report the merge-minus-drop latency difference with approximate 95% confidence intervals.

5.7 Discussion

Classifier-derived uncertainty is a criterion for token selection, not a direct measure of intrinsic token importance: a positive high–low accuracy gap supports a selection strategy but does not establish individual causal contributions. Strong ranking agreement with feature norm and a preserved gap under classifier-channel permutation suggest that aggregate projection and representation properties may matter more than precise class-semantic alignment. Moreover, reduced target-SSA operations do not establish end-to-end acceleration, given the overhead of the two-pass procedure. Future work should assess the additional information provided by classifier-derived evidence beyond inexpensive feature statistics and explore single-pass implementations.

6 Conclusion

We present CETS, a training-free token-selection framework that reuses the frozen classifier of a spiking transformer to construct ranking scores from temporal statistics of classifier-derived uncertainty. Experiments on static and neuromorphic benchmarks demonstrate the effectiveness of these scores for token ranking under diverse token retention ratios. Extensive analyses highlight the role of the pretrained classifier projection and the dependence of ranking effectiveness on backbone representations. Inverse Dirichlet strength provides a more effective ranking signal than the tested softmax-based alternatives, while temporal variation offers a dataset-dependent refinement. These findings support classifier-derived evidence as a complementary, representation-dependent cue for token selection.

AI USE STATEMENT

In this work, we used generative AI tools exclusively for grammatical error correction and language polishing. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content of this work.

REFERENCES

Arnon Amir, Brian Taba, David J. Berg, Timothy Melano, Jeffrey L. McKinstry, Carmelo di Nolfo, Tapan K. Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, Jeff Kusnitz, Michael DeBole, Steven K. Esser, Tobi Delbrück, Myron Flickner, and Dharmendra S. Modha. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7388–7397, 2017.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *Proceedings of the International Conference on Learning Representations*, 2023.

Tong Bu, Xinyu Shi, and Zhaofei Yu. Activity pruning for efficient spiking neural networks. In *Advances in Neural Information Processing Systems*, volume 38, pp. 189601–189620, 2025.

Yuetong Fang, Ziqing Wang, Lingfeng Zhang, Jiahang Cao, Honglei Chen, and Renjing Xu. Spiking wavelet transformer. In *Proceedings of the European Conference on Computer Vision*, pp. 19–37. Springer, 2024.

Yuetong Fang, Deming Zhou, Ziqing Wang, Hongwei Ren, Zecui Zeng, Lusong Li, Shibo Zhou, and Renjing Xu. Spiking neural networks need high-frequency information. In *Advances in Neural Information Processing Systems*, volume 38, pp. 114506–114527, 2025.

Wulfram Gerstner and Werner M Kistler. *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge university press, 2002.

Zecheng Hao, Qichao Ma, Kang Chen, Yi Zhang, Zhaofei Yu, and Tiejun Huang. Faster and stronger: When ANN-SNN conversion meets parallel spiking calculation. In *Proceedings of the International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22103–22115, 2025.

Zihan Huang, Xinyu Shi, Zecheng Hao, Tong Bu, Jianhao Ding, Zhaofei Yu, and Tiejun Huang. Towards high-performance spiking transformers from ANN to SNN conversion. In *Proceedings of the ACM International Conference on Multimedia*, pp. 10688–10697, 2024.

Zihan Huang, Wei Fang, Tong Bu, Peng Xue, Zecheng Hao, Wenxuan Liu, Yuanhong Tang, Zhaofei Yu, and Tiejun Huang. Differential coding for training-free ANN-to-SNN conversion. In *Proceedings of the International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25235–25264, 2025.

Eugene M Izhikevich. Which model to use for cortical spiking neurons? *IEEE Transactions on Neural Networks*, 15(5):1063–1070, 2004.

Donghwa Kang, Youngmoon Lee, Eun-Kyu Lee, Brent Kang, Jinkyu Lee, and Hyeongboo Baek. At-SNN: Adaptive tokens for vision transformer on spiking neural network. *arXiv preprint arXiv:2408.12293*, 2024.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`.

Donghyun Lee, Yuhang Li, Youngeun Kim, Shiting Xiao, and Priyadarshini Panda. Spiking transformer with spatial-temporal attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13948–13958, 2025.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontañón. FNet: Mixing tokens with fourier transforms. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4296–4313, 2022.

Hongmin Li, Hanchao Liu, Xiangyang Ji, Guoqi Li, and Luping Shi. Cifar10-dvs: an event-stream dataset for object classification. *Frontiers in Neuroscience*, 11:309, 2017.

Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. MViTv2: improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4794–4804, 2022.

Wenxuan Liu, Xuemei Jia, Xian Zhong, Kui Jiang, Xiaohan Yu, and Mang Ye. Dynamic and static mutual fitting for action recognition. *Pattern Recognition*, 157:110948, 2025.

Yue Liu, Shanlin Xiao, Bo Li, and Zhiyi Yu. SparseSpikformer: A co-design framework for token and weight pruning in spiking transformer. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6410–6414. IEEE, 2024.

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

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. DynamicViT: efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13937–13949, 2021.

G Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

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*, pp. 5610–5619, 2024.

William E Vinje and Jack L Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–1276, 2000.

Wenjie Wei, Xiaolong Zhou, Malu Zhang, Ammar Belatreche, Qian Sun, Yimeng Shan, Dehao Zhang, Zijian Zhou, Zeyu Ma, Yang Yang, and Haizhou Li. TP-Spikformer: Token pruned spiking transformer. In *Proceedings of the International Conference on Learning Representations*, 2026.

Man Yao, Jiakui Hu, Zhaokun Zhou, Li Yuan, Yonghong Tian, Bo Xu, and Guoqi Li. Spike-driven transformer. In *Advances in Neural Information Processing Systems*, volume 36, pp. 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 *Proceedings of the International Conference on Learning Representations*, 2024a.

Man Yao, Ole Richter, Guangshe Zhao, Ning Qiao, Yannan Xing, Dingheng Wang, Tianxiang Hu, Wei Fang, Tugba Demirci, Michele De Marchi, et al. Spike-based dynamic computing with asynchronous sensing-computing neuromorphic chip. *Nature Communications*, 15(1):4464, 2024b. doi: 10.1038/s41467-024-47811-6.

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.

Takashi Yoshida and Kenichi Ohki. Natural images are reliably represented by sparse and variable populations of neurons in visual cortex. *Nature Communications*, 11(1):872, 2020.

Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11091–11101, 2022.

Xian Zhong, Shengwang Hu, Wenxuan Liu, Wenxin Huang, Jianhao Ding, Zhaofei Yu, and Tiejun Huang. Towards low-latency event-based visual recognition with hybrid step-wise distillation spiking neural networks. In *Proceedings of the ACM International Conference on Multimedia*, MM '24, pp. 9828–9836, 2024.

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. In *Advances in Neural Information Processing Systems*, volume 37, pp. 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 *Proceedings of the International Conference on Learning Representations*, 2023.

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.

Zhaokun Zhou, Kaiwei Che, Jun Niu, Man Yao, Guoqi Li, Li Yuan, Guibo Luo, and Yuesheng Zhu. Spatial-temporal spiking feature pruning in spiking transformer. *IEEE Transactions on Cognitive and Developmental Systems*, 17(3):644–658, 2025.

Zhengyang Zhuge, Peisong Wang, Xingting Yao, and Jian Cheng. Towards efficient spiking transformer: A token sparsification framework for training and inference acceleration. In *Proceedings of the International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62768–62778, 2024.

A Appendix

This appendix provides additional details and comprehensive experimental results to support the main manuscript.

- Additional related work on token selection in spiking transformers.
- Detailed experimental protocols.
- Supplementary Ranking Definition.
- Additional experimental results, including temporal scoring variants, drop-versus-merge comparisons, sensitivity to λ, and variability across training seeds.

A.1 Token Selection in Spiking Transformers

Given the heavy computational cost of Transformers, several works have attempted to develop more efficient Transformer architectures for both NLP and vision tasks Bolya et al. (2023); Lee-Thorp et al. (2022). Since transformers can operate on a variable number of tokens, several recent works have attempted to prune tokens to improve the efficiency of transformers, including spiking transformers Liu et al. (2024); Kang et al. (2024); Zhuge et al. (2024). SparseSpikformer proposes a co-design framework that jointly performs token pruning and weight pruning to improve the efficiency of spiking transformers Liu et al. (2024). AT-SNN introduces an adaptive token mechanism that dynamically selects informative tokens in spiking vision transformers to reduce redundant computation Kang et al. (2024). STATA proposes a token sparsification framework with a learnable importance predictor to remove redundant tokens during training and inference Zhuge et al. (2024). TP-SpikFormer prunes tokens based on spike firing statistics, leveraging the intrinsic sparsity of spiking activations in a training-free manner Wei et al. (2026). Meanwhile, Bu et al. (2025) has explored suppressing neuronal activity to improve SNN efficiency by introducing adaptive neuron AT-LIF, which dynamically adjusts firing thresholds to reduce spike rates.

A.2 Detailed Experimental Protocol

Random token selection is evaluated over five mask realizations and reported as mean ± standard deviation. The channel-permuted and matched-random classifier-head controls are likewise reported over five repeats. These repeats evaluate randomized selection or scoring under a frozen checkpoint; they do not represent independently trained backbones. **Two-pass evaluation.** The evaluator first performs an unpruned capture pass to obtain the tokens used for scoring, followed by a selected-token evaluation pass that applies the resulting mask. The target-SSA SOP and MAC counts in the main paper describe only the selected SSA computation. They exclude the dense capture pass, computation outside the target SSA, and scoring and Top-K overhead. Consequently, these counts do not establish end-to-end latency or energy reductions for the two-pass evaluator.

A.3 Ranking Scores

Classifier-side scores. Let $z_i^t \in \mathbb{R}^D$ denote token i at step t, and $\ell_{i,c}^t$ its frozen-classifier logit for class c. We use $\alpha_{i,c}^t = 1 + \mathrm{softplus}(\ell_{i,c}^t)$ and $p_{i,c}^t = \exp(\ell_{i,c}^t)/\sum_{j=1}^{C}\exp(\ell_{i,j}^t)$. The classifier-side comparison in the main paper uses inverse strength $U_i^t = C/\sum_c \alpha_{i,c}^t$, maximum-probability uncertainty $1 - \max_c p_{i,c}^t$, normalized entropy $-\sum_c p_{i,c}^t \log p_{i,c}^t / \log C$, and margin uncertainty $1 - (p_{i,(1)}^t - p_{i,(2)}^t)$, where $p_{i,(1)}^t$ and $p_{i,(2)}^t$ are the largest and second-largest class probabilities. For each step-wise signal q_i^t, the final ranking is

$$R_i^q = \mathrm{mean}_t(q_i^t) + 0.9\,\mathrm{std}_t(q_i^t). \tag{16}$$

The same temporal aggregation and selection operator are used for all four signals. The highest-scoring K tokens are retained. The mean-only and uncertainty-std-only variants use $\mathrm{mean}_t(U_i^t)$ and $\mathrm{std}_t(U_i^t)$, respectively.

Classifier-head controls. The channel-permuted control permutes the input channels of the trained classifier weights. The matched-random control uses random linear heads with the same input/output dimensions and matched row norms. Permuting class outputs alone would leave total Dirichlet strength unchanged and therefore would not alter the ranking. These controls do not establish that class-semantic alignment is the source of the observed selection utility.

Representation-side baselines. The feature-norm reference uses the temporal RMS norm, $R_i^{\mathrm{FN}} = \sqrt{T^{-1}\sum_t \|z_i^t\|_2^2}$. A token-independent scaling factor, such as $1/\sqrt{D}$, leaves its ranking unchanged. Feature temporal variance is computed from intermediate features, whereas uncertainty std is computed from classifier-derived inverse strength; these are distinct ranking signals. The adapted IRToP baseline transfers its scoring criterion to the same shared-time selection protocol; it does not reproduce the full TP-Spikformer pipeline or its IR-Arc early-stopping operator.

High–low accuracy gap. For ranking R and $K = \lfloor rN \rfloor$, we compute

$$\Delta_R(r) = \mathrm{Acc@1}\big(\mathrm{TopK}(R,K)\big) - \mathrm{Acc@1}\big(\mathrm{BottomK}(R,K)\big). \tag{17}$$

The two accuracies use the same checkpoint, retention ratio, and operator. A positive gap favors high-score retention; a negative gap indicates the reverse ordering, rather than the absence of ranking information.

Evidence–feature agreement. For each sample, we compute Spearman correlation as the Pearson correlation between the token ranks induced by R^{E} and R^{FN}. Top-set overlap is defined as

$$\mathrm{Overlap@}\,p = \frac{|\,\mathrm{TopK}(R^{\mathrm{E}},K_p) \cap \mathrm{TopK}(R^{\mathrm{FN}},K_p)|}{K_p}, \qquad K_p = \lfloor pN \rfloor. \tag{18}$$

For $N = 64$, $p = 0.2$ and 0.4 correspond to 12 and 25 tokens. We average the sample-wise correlations and overlaps over the test set, rather than computing a single correlation across pooled tokens. The numerical results are reported in the main paper.

A.4 Additional Temporal Ranking Analysis

Table 7 extends the main paper's mean-versus- mean-plus-variation comparison with uncertainty-std-only scoring and feature temporal variance. All methods use the same frozen checkpoint, target site, retention ratio, and drop operator within each dataset.

Feature temporal variance performs best among these signals on DVS-Gesture at every listed retention ratio. On DVS-CIFAR-10, the ordering changes with retention: feature variance performs best at 95% and 60%, while uncertainty std alone performs best at 40%. These results further illustrate the dataset- and retention ratio-dependent behavior of temporal ranking signals.

A.5 Drop-versus-Merge Details

Drop-versus-merge protocol. Let $\mathcal{K}$ denote the indices of the K tokens with the highest CETS ranking scores, and let $\mathcal{D} = \{1, \dots, N\} \setminus \mathcal{K}$ denote the discarded-token indices. For cosine matching, we first average each token representation over simulation steps:

$$\bar{z}_i = \frac{1}{T} \sum_{t=1}^{T} z_i^t. \tag{19}$$

Each discarded token $j \in \mathcal{D}$ is assigned to the retained anchor with the highest cosine similarity:

$$a(j) = \arg\max_{i \in \mathcal{K}} \frac{\bar{z}_j^\top \bar{z}_i}{\|\bar{z}_j\|_2 \|\bar{z}_i\|_2}. \tag{20}$$

This assignment is shared across all simulation steps. Let $\mathcal{A}_i = \{j \in \mathcal{D} : a(j) = i\}$ denote the discarded tokens assigned to anchor i. At each simulation step, we uniformly average the anchor and its assigned tokens to obtain the merged SSA input:

$$\tilde{z}_i^t = \frac{z_i^t + \sum_{j \in \mathcal{A}_i} z_j^t}{1 + |\mathcal{A}_i|}, \qquad i \in \mathcal{K}. \tag{21}$$

Both variants therefore process exactly K tokens in the target SSA module: dropping uses $\{z_i^t\}_{i \in \mathcal{K}}$, whereas merging uses $\{\tilde{z}_i^t\}_{i \in \mathcal{K}}$. The remainder of the network, including the residual path, is unchanged. This comparison evaluates the specified uniform averaging operator rather than token-merging methods in general.

The main paper compares dropping with merging under the same CETS ranking, checkpoint, target site, and retained-token ratios. Table 8 provides the accuracy values underlying the comparison.

Merging remains close to direct dropping under moderate token reduction: between 60 and 38 retained tokens, the absolute Acc@1 difference is at most 0.23 percentage points. Merging slightly improves accuracy at 44 tokens by 0.10 points, but degrades increasingly under more aggressive reduction. At 32, 25, 19, and 12 retained tokens, merging underperforms dropping by 0.64, 1.67, 4.27, and 5.11 points, respectively. This suggests that uniformly averaging discarded features into their nearest retained anchors can preserve coarse information at moderate retention ratios but introduces harmful feature mixing when many tokens are compressed into a small number of anchors.

A.6 Impact of Temporal Uncertainty Parameter λ

We further evaluate the sensitivity of the weighting factor λ in the proposed token ranking score, where λ balances the contribution of temporal mean and temporal variation. As shown in Table 9, the performance remains highly consistent across different λ values under varying token retention ratios, indicating that the proposed scoring strategy is not sensitive to the specific choice of λ. Among the tested settings, $\lambda = 0.9$ achieves slightly better performance in the moderate pruning regime (*e.g.,* retention ratios of 0.8–0.9), and is therefore adopted as the default setting in our experiments.

A.7 Random Seed

To examine whether the selection behavior depends on a particular pretrained checkpoint, we repeat the high-score, random, and low-score retention experiments using three independently trained

Table 7: Acc@1 (%) for additional temporal ranking comparisons. Uncertainty std is computed from inverse Dirichlet strength; feature temporal variance is computed from intermediate token features.

Dataset	Ranking signal	95%	80%	60%	40%
DVS-Gesture	Mean $+0.9$ std	97.57	96.88	94.79	87.50
	Uncertainty std only	97.57	96.88	95.49	90.97
	Feature temporal variance	98.26	98.26	97.57	97.22
DVS-CIFAR-10	Mean $+0.9$ std	82.50	81.20	74.10	55.40
	Uncertainty std only	81.90	81.20	74.10	57.90
	Feature temporal variance	82.70	80.30	74.50	54.60

Table 8: **Drop-versus-merge Acc@1 (%) on CIFAR-100.** Both variants use the same frozen seed-42 checkpoint, CETS ranking, target SSA module, shared-time mask, and retained-token ratios. Results are evaluated on the full test set.

Retained tokens	60	57	51	44	38	32	25	19	12
Drop	**81.03**	**81.02**	**81.06**	80.66	**80.55**	**80.12**	**78.91**	**77.09**	**72.13**
Merge	80.95	80.90	80.89	**80.76**	80.32	79.48	77.24	72.82	67.02
Merge − Drop	-0.08	-0.12	-0.17	$+0.10$	-0.23	-0.64	-1.67	-4.27	-5.11

QKFormer checkpoints with training seeds 41, 42, and 43. All checkpoints are evaluated on the full CIFAR-100 test set using the same target module (`stage3[1].ssa`), $N = 64$ input tokens, shared-time mask, $\lambda = 0.9$, and drop operator. The 80% and 60% settings retain $K = 51$ and $K = 38$ tokens, respectively.

For the deterministic high-score and low-score strategies, each checkpoint produces one evaluation result. For Random, we conduct five independent random-selection runs for each checkpoint and first average these five results within the checkpoint. We then compute the mean and sample standard deviation over the three checkpoint-level results. Thus, the reported variation measures sensitivity to independently trained checkpoints rather than variation among random masks.

Table 9: Sensitivity analysis of the weighting factor λ in CETS. Each entry is reported as Acc@1 / Acc@5 (%).

Retention ratio	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.7$	$\lambda = 0.9$
1.00	80.98/95.13	80.98/95.13	80.98/95.13	80.98/95.13	80.98/95.13
0.90	80.94/95.30	80.96/95.34	80.90/95.24	80.91/95.25	81.02/95.25
0.80	80.91/95.44	80.89/95.33	80.89/95.38	80.92/95.39	81.06/95.45
0.60	80.48/95.33	80.50/95.31	80.57/95.21	80.57/95.22	80.55/95.23
0.40	79.21/94.82	79.06/94.68	79.16/94.75	79.11/94.76	78.91/94.77
0.30	77.19/94.19	77.18/94.10	77.09/94.15	77.08/94.06	77.09/94.05
0.20	72.13/92.82	72.23/92.84	72.30/92.86	72.19/92.77	72.13/92.74

Table 10: **Selection utility across independently trained checkpoints.** Values are Acc@1 (%) reported as mean $\pm$ sample standard deviation over three checkpoint seeds (41, 42, and 43). For Random, five mask realizations are first averaged within each checkpoint before computing the across-checkpoint statistics.

Retained subset	**80%**	**60%**
High-score (CETS)	$\mathbf{80.95} \pm 0.10$	$\mathbf{80.52} \pm 0.11$
Random	80.29 ± 0.05	78.87 ± 0.08
Low-score	74.01 ± 0.62	68.83 ± 0.78
high–low accuracy gap (pp)	$+6.94 \pm 0.68$	$+11.70 \pm 0.88$