MiST: Memory-inspired Token-Efficient Spiking Transformer

Anonymous authors
Paper under double-blind review

Abstract

Transformer-based Spiking Neural Networks (SNNs) combine the high performance of Transformers with the energy efficiency of SNNs, offering a promising paradigm for intelligent edge computing. However, they remain bottlenecked by the token-intensive processing inherited from conventional vision transformers, which hinders their deployment on resource-constrained edge devices. Inspired by the key-value memory mechanism of the human brain, we propose MiST, a memory-inspired token-efficient spiking transformer that exploits the inherent efficiency of transformer-based SNNs while maintaining competitive performance. Specifically, MiST decouples the full feature map into a memory repository and a small set of propagated query tokens, enabling forward inference to operate on only a fraction of the original tokens. The core component in MiST is the bidirectional spike-driven cross-attention (Bi-SDCA) module, which comprises two pathways: a forward pathway that retrieves semantic content from the repository into query tokens, and an inverse pathway that propagates enriched token representations back to update the repository. Furthermore, we observe that the memory repository in MiST exhibits high semantic disorder during early training, which slows convergence. To address this, we propose a semantic-guided self-distillation (SGSD) strategy that drives the repository to form compact semantic anchors. Extensive experiments on image classification and semantic segmentation show that MiST achieves accuracy comparable to token-intensive baselines using only 20$\sim$40% of tokens, establishing a new efficiency frontier for spiking transformer architectures.

1 Introduction

Brain-inspired spiking neural networks (SNNs) have emerged as an efficient alternative to next-generation neural networks due to their high biological plausibility and sparse event-driven computation Gerstner & Kistler (2002); Maass (1997); Li et al. (2024). Meanwhile, Transformer architectures in artificial neural networks (ANNs) have flourished in various fields, including computer vision Liu et al. (2021); Dosovitskiy (2020), natural language processing Kenton & Toutanova (2019); Vaswani et al. (2017), etc. Motivated by these advances, researchers have naturally integrated Transformers with SNNs, leading to high-performance and energy-efficient models such as Spikformer Zhou et al., Spike-Driven Transformer V1-V3 Yao et al. (2023; 2024; 2025), QKFormer Zhou et al. (2024), and SNN-ViT Wang et al. (2025). These efforts provide a promising pathway toward edge intelligent systems that require both high accuracy and computational efficiency.

While existing transformer-based SNNs have achieved notable progress, they largely inherit the traditional token-intensive computational paradigm of transformer architectures in ANNs Liu et al. (2021); Dosovitskiy (2020). Specifically, they retain a full feature map after the downsampling stage and apply spiking self-attention and feed-forward network operations to all tokens Yao et al. (2023); Zhou et al.; Shi et al. (2024); Zhou et al. (2024). This incurs substantial computational overhead, weakening the efficiency benefits of the models Zhuge et al. (2024); Kang et al. (2024); Lu et al. (2025). Meanwhile, recent research in the SNN community has largely focused on scaling up SNNs' model size, aimed at achieving overwhelming performance on various tasks Zheng et al. (2025); Yao et al. (2025); Lee et al. (2025). Under this trend, the drawbacks of the token-intensive paradigm become more severe, making deployment on edge devices increasingly difficult. Therefore, it is essential to develop a token-efficient design for transformer-based SNNs.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

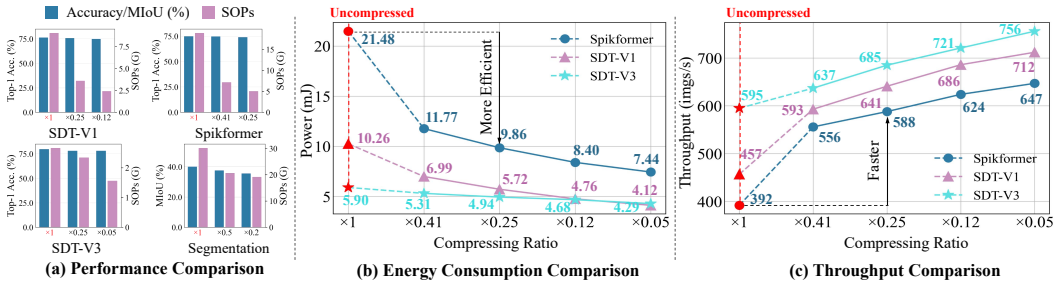


Figure 1: Comparison of MiST with three classical token-intensive spiking transformers. MiST attains comparable accuracy with far fewer tokens, while offering improved efficiency and throughput.

Images contain substantial redundancy, and vision models often rely on only a small subset of visual information for prediction Hosseini et al. (2010); Wang et al. (2021). This motivates us to reduce the number of tokens, improving efficiency without sacrificing accuracy. Recent studies have pursued this goal by pruning or merging less important tokens within each transformer block Zhuge et al. (2024); Wei et al. (2026). However, these methods rely on heuristic local scoring mechanisms that distinguish important tokens from unimportant ones across the entire feature map. Once tokens are identified as unimportant, they are irreversibly pruned or merged, and their visual cues and contextual information are discarded. Such irreversible processing may introduce semantic loss and increase the risk of discarding important regions. As a result, when designing token-efficient models, leveraging global context to guide token selection helps preserve semantic integrity and important regions.

With sophisticated memory mechanisms, the human brain can integrate global semantic information from large-scale inputs and rapidly extract key features, enabling efficient decision-making Attwell & Laughlin (2001). This provides a promising direction for designing effective token-efficient methods. Recent research in neuroscience shows that the human brain's memory mechanism can be viewed as a key-value memory system Gershman et al. (2025), which stores two types of information, i.e., indices used for retrieval (i.e., keys) and content that encodes detailed information (i.e., values). When the brain receives a sensory cue, it activates specific memory indices in the hippocampus and retrieves only the most important memory content. Then, the brain analyzes this retrieved content to make decisions in complex scenarios. Inspired by this Whittington et al. (2022), we employ a memory mechanism that leverages global context to guide token-efficient design in transformer-based SNNs.

In this paper, we introduce a token-efficient paradigm for transformer-based SNNs, leveraging their inherent computational efficiency while attaining competitive accuracy. Specifically, we propose a key-value memory-inspired spiking transformer, MiST, which treats the full feature map as a memory repository and propagates only small query tokens for inference. The core of MiST is the bidirectional spike-driven cross-attention (Bi-SDCA) module, which enables efficient two-way interaction between sparse query tokens and the memory repository. Furthermore, since the memory repository plays a central role in guiding token selection, we introduce a semantic-guided self-distillation (SGSD) to drive it rapidly from an initially disordered state into compact semantic anchors, thereby accelerating convergence and improving accuracy. As shown in Figure 1, MiST attains comparable accuracy with superior efficiency and higher throughput. Our contributions are summarized as follows:

- We take inspiration from the key-value memory theory and design a token-efficient spiking transformer MiST, with bidirectional spike-driven cross-attention as its core module. MiST leverages a memory repository to extract critical cues from input queries and enables forward propagation using only a small set of tokens. This design significantly reduces computational overhead, exploiting the inherent efficiency advantages of transformer-based SNNs.
- To ensure stable and precise semantic guidance from the memory repository, we propose a semantic-guided self-distillation strategy for MiST. This strategy allows the memory repository to quickly evolve from an initial disordered state into compact semantic anchors, accelerating convergence and improving performance.
- Experiments on various tasks show that MiST achieves SOTA performance and efficiency compared to existing token-efficient SNNs. Even with only 20∼40% of tokens, it maintains comparable accuracy to token-intensive architectures, offering valuable insights for designing efficient spiking transformer computational paradigms.

2 Related Work

Spiking Transformer Many researchers have integrated high-performance Transformer structures with energy-efficient SNNs to develop spiking transformers Fang et al.; Cao et al. (2025); Wang et al. (2025); Shi et al. (2024). For example, Spikformer proposes a spiking self-attention (SSA), establishing the first directly trained spiking transformer Zhou et al.. Spikingformer merges the MS Residual connection with self-attention in a biologically plausible way; it demonstrates strong effectiveness and generality across 13 datasets, establishing an important benchmark for SNNs Zhou et al. (2026). QKFormer implements a hierarchical architecture using query-key attention and spiking patch embeddings with deformable shortcuts, achieving milestone results on multiple datasets Zhou et al. (2024). In the SNN field, the spike-driven transformer series is widely recognized for leveraging a spike-driven architecture. Spike-Driven transformer (SDT-V1) develops the SSA with Hadamard product, achieving a fully spike-driven architecture Yao et al. (2023). SDT-V2 extends SDT-V1 into a meta architecture, exploring structure design, spike-driven attention, and skip connection to enhance performance Yao et al. (2024). SDT-V3 improves spiking neuron firing patterns with a quantization mechanism and achieves efficient training Yao et al. (2025). Recently, SpiLiFormer proposes a differential spiking transformer, achieving more effective attention allocation and enhanced robustness Zheng et al. (2025). Despite efficiency gains, these studies inherit the token-intensive computational paradigm of ANN-based transformers, incurring substantial computational complexity and overhead. This drawback becomes more pronounced as the model scales.

Token Compression in Spiking Transformer Existing token-efficient designs for spiking transformers mainly fall into two categories: token pruning and token merging. Most of these methods rely on heuristic local scoring functions to divide tokens into important and unimportant groups. For instance, SparseSpikformer adopts the firing rate of spiking neurons as a local scoring criterion, pruning tokens with low firing rates Liu et al. (2024). AT-SNN employs pairwise token similarity as its scoring basis, integrating adaptive token computation time to perform similarity-driven token merging Kang et al. (2024). TP-Spikformer, inspired by the human visual system, proposes a dedicated information-preserving scorer to identify and retain salient tokens Wei et al. (2026). A step toward semantic awareness is taken by STATA, which introduces an anchor token that consolidates classification-relevant information; tokens are then scored by their similarity to this anchor, with similar tokens retained and dissimilar ones discarded Zhuge et al. (2024). While this brings in some semantic awareness, STATA's scoring remains heuristic and still involves irreversible token discarding. In short, all existing methods share a fundamental limitation: once a token is deemed unimportant, it is irreversibly pruned or merged and excluded from subsequent forward propagation. The irreversibility makes the visual information encoded in those tokens inaccessible, which may cause partial semantic information loss and increase the risk of removing important regions.

3 Preliminaries

3.1 Spiking Transformer

Given a 2D image sequence $\mathbf{I} \in \mathbb{R}^{T\times C\times H\times W}$, the spiking patch splitting (SPS) module projects it into D-dimensional spiking features and partitions the result into a 2D grid of spike-form patches x. These patches are combined with spike-based positional encodings (PE) to form the initial feature map $\mathbf{X}_0$. The features are then processed by L spiking transformer blocks, each consisting of a spike-driven self-attention (SSA) module and a feed-forward network (FFN). After the final block, global average pooling (GAP) aggregates the output features into a D-dimensional vector, which is fed into a classification head (CH) to produce the final prediction $\mathbf{Y}$. This process is written as:

$$x = \mathrm{SPS}(\mathbf{I}), \qquad x \in \mathbb{R}^{T\times H\times W\times D}, \tag{1}$$

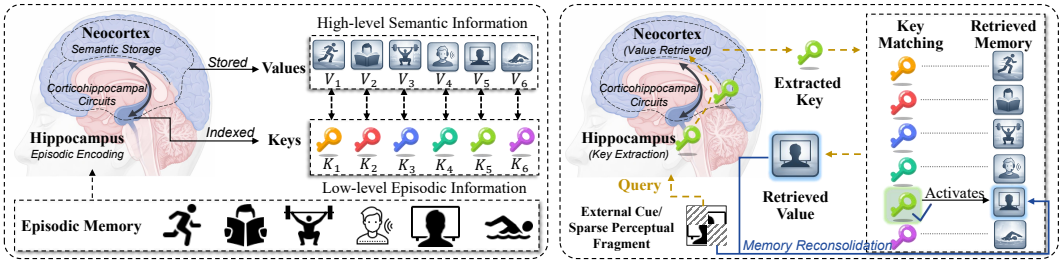
$$\mathbf{X}_0 = x + \mathrm{PE}(x), \qquad \mathbf{X}_0 \in \mathbb{R}^{T\times H\times W\times D}, \tag{2}$$

$$\mathbf{X}'_l = \mathrm{SSA}(\mathbf{X}_{l-1}) + \mathbf{X}_{l-1}, \qquad \mathbf{X}'_l \in \mathbb{R}^{T\times H\times W\times D}, \tag{3}$$

$$\mathbf{X}_l = \mathrm{FFN}(\mathbf{X}'_l) + \mathbf{X}'_l, \qquad \mathbf{X}_l \in \mathbb{R}^{T\times H\times W\times D}, \tag{4}$$

$$\mathbf{Y} = \mathrm{CH}(\mathrm{GAP}(\mathbf{X}_L)). \tag{5}$$

As the core of spiking transformer blocks, the SSA provides an efficient approach to model local-to-global information in images using spike-based queries ($\mathbf{Q}$), keys ($\mathbf{K}$), and values ($\mathbf{V}$) without



(a) Memory Formation Process (b) Memory Retrieval and Reconsolidation Process

Figure 2: Key-value memory in the brain, including memory formation, retrieval and reconsolidation.

relying on softmax. It is defined as:

$$\mathrm{SSA}(\mathbf{X}_{l-1}) = \mathbb{SN}(\mathbf{Q}\mathbf{K}^\top \mathbf{V} * s), \qquad \mathbf{x} = \mathbb{SN}(W_{\mathbf{x}} \cdot \mathbb{SN}(\mathbf{X}_{l-1})),\ \mathbf{x} \in \{\mathbf{Q}, \mathbf{K}, \mathbf{V}\}, \tag{6}$$

where $\mathbb{SN}(\cdot)$ denotes the spiking neuron that enables the sparse spike-driven computation. The computationally efficient leaky integrate-and-fire (LIF) neuron model is widely used, given by:

$$\tilde{\mathbf{U}}^l[t] = \tau \mathbf{U}^l[t-1] + \mathbf{X}^l[t], \qquad \mathbf{S}^l[t] = \mathrm{Heaviside}(\tilde{\mathbf{U}}^l[t]), \qquad \mathbf{U}^l[t] = \tilde{\mathbf{U}}^l[t]\left(1 - \mathbf{S}^l[t]\right). \tag{7}$$

where τ is the leaky factor, and $\mathrm{Heaviside}(\cdot)$ output a spike $\mathbf{S}^l[t]$ when $\tilde{\mathbf{U}}^l[t]$ surpasses a threshold θ.

3.2 Key-Value Memory Theory

The brain's memory system has long been studied for its remarkable capacity to store and retrieve information efficiently. The recent key-value memory theory offers a new perspective on memory mechanisms, including memory formation, retrieval, and reconsolidation in the brain, and has shown strong explanatory power across empirical studies Gershman et al. (2025); Tyulmankov et al. (2021).

Memory Formation Process The key-value memory theory suggests that memory formation depends on two complementary neural representations, namely 'keys' and 'values', which are processed by specialized neural systems to support efficient and flexible recall McClelland et al. (1995); O'Reilly & Norman (2002). Specifically, the medial temporal lobe system, especially the hippocampus, is responsible for forming 'key' representations that uniquely index individual experiences and enable the precise retrieval and reactivation of past events Teyler & DiScenna (1986); Goode et al. (2020). In parallel, the neocortex, with its stronger representational capacity and better cross-context generalization, mainly stores 'value' representations that encode perceptual details and semantic content Danker et al. (2017); Pacheco Estefan et al. (2019). Together, hippocampal keys and neocortical values are associated to form a jointly addressable memory structure.

Memory Retrieval and Reconsolidation Process Memory retrieval can be viewed as a key–value matching process Tanaka et al. (2014); Gershman et al. (2025). When an external cue or a sparse and incomplete perceptual fragment is presented, the hippocampus first processes the input and extracts task-relevant information from redundant signals to form a compact 'key' representation. This key is then projected to the neocortex through corticohippocampal pathways, where it reinstates the corresponding cortical activity pattern and recovers the associated 'value' representation. This two-stage process, i.e., sparse indexing followed by content reinstatement, enables efficient and selective retrieval without exhaustive search over the full memory store. Moreover, neuroscientific studies also suggest that memory retrieval is not simply a readout process. Once retrieved, a memory becomes temporarily unstable and can be updated by new information, a process known as memory reconsolidation. Inspired by these mechanisms, we propose a token-efficient spiking transformer that mimics similar key-value dynamics to achieve a balance between performance and efficiency.

4 Method

In this section, we first present our memory-inspired token-efficient spiking transformer architecture, in which the core component is a bidirectional spike-driven cross-attention module. Second, we present a semantic-guided self-distillation training strategy, which stabilizes and accelerates the convergence of MiST without requiring an external teacher model. Finally, we theoretically analyze the computational cost of MiST and compare it with token-intensive baselines.

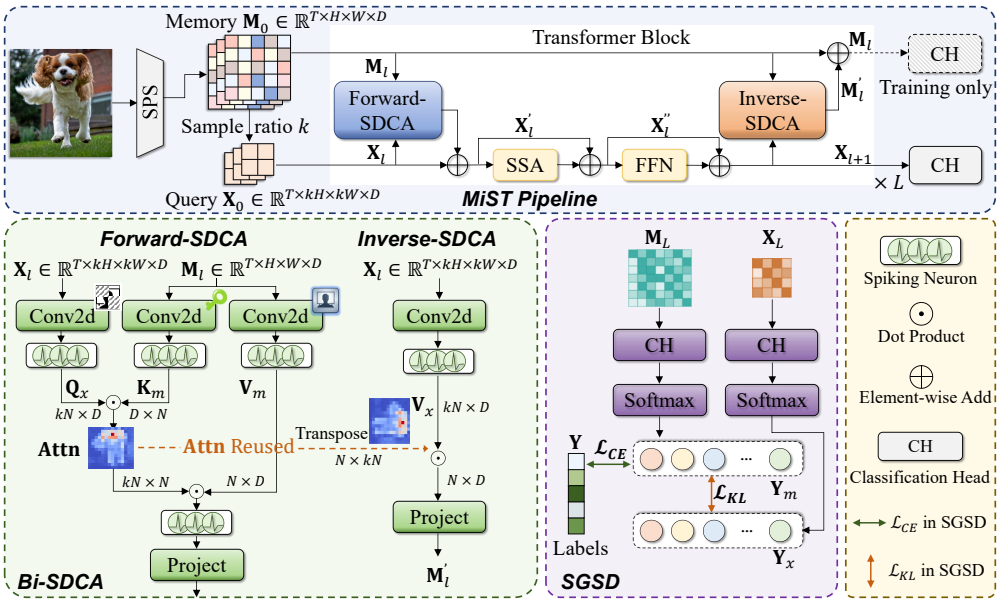


Figure 3: Workflow of MiST and its core components, including Bi-SDCA and the SGSD strategy.

4.1 Memory-inspired Token-Efficient Spiking Transformer

To enable token-efficient inference in spiking transformers, we seek to propagate only a small subset of tokens, which raises a critical question: how to select these tokens and ensure that they carry sufficient semantic information from the full feature map throughout forward propagation? Inspired by the key-value memory theory in Sec. 3.2, we propose MiST, a token-efficient spiking transformer consisting of a memory repository construction module and L transformer blocks, each comprising bidirectional spike-driven cross-attention (Bi-SDCA), SSA, and FFN.

Memory Repository Construction Given an input image $\mathbf{I}$, MiST first uses a SPS module to generate spiking features $\mathbf{M}_0$, which serve as the memory repository encoding the full scene context:

$$\mathbf{M}_0 = \text{SPS}(\mathbf{I}), \quad \mathbf{M}_0 \in \mathbb{R}^{T\times H\times W\times D}. \tag{8}$$

A small set of query tokens is then sampled from the memory repository $\mathbf{M}_0$ for forward propagation:

$$\mathbf{X}_0 = \text{Sample}(\mathbf{M}_0, k), \quad \mathbf{X} \in \mathbb{R}^{T\times kH\times kW\times D}, \tag{9}$$

where k is the compression ratio and $\mathbf{X}_0$ is the sampled sparse query tokens. The sparse tokens $\mathbf{X}_0$ are analogous to the incomplete perceptual fragments that trigger memory retrieval in the brain. They are then propagated through L transformer blocks to extract key information for decision-making.

Bidirectional Spike-Driven Cross-Attention The core component of each Transformer block is the Bi-SDCA module, which enables bidirectional interaction between sparse query tokens and the memory repository through two complementary pathways: Forward-SDCA and Inverse-SDCA.

Forward-SDCA (F-SDCA) is designed as a spike-driven cross-attention module that mimics the memory retrieval process, enabling the sparse query tokens to retrieve global semantic content from the memory repository. Within this cross-attention module, the sparse query tokens $\mathbf{X}_l$ are projected to queries $\mathbf{Q}_x$, while the memory repository $\mathbf{M}_l$ is used to generate keys $\mathbf{K}_m$ and values $\mathbf{V}_m$, mimicking how the hippocampus extracts key representations and the neocortex reinstates the associated value representations. Mathematically, the F-SDCA pathway is defined as follows:

$$\mathbf{X}'_l = \mathbf{X}_l + \text{F-SDCA}(\mathbf{X}_l, \mathbf{M}_l) = \mathbf{X}_l + \mathbb{SN}(\mathbf{Attn}\mathbf{V}_m * s), \tag{10}$$

$$\mathbf{Attn} = \mathbf{Q}_x\mathbf{K}_m^\top, \qquad \mathbf{Attn} \in \mathbb{R}^{T\times kN\times N}, \tag{11}$$

$$\mathbf{Q}_x = \mathbb{SN}(W_\mathbf{Q} \cdot \mathbb{SN}(\mathbf{X}_l)), \qquad \mathbf{Q}_x \in \mathbb{R}^{T\times kN\times D}, \tag{12}$$

$$\mathbf{K}_m = \mathbb{SN}(W^m_\mathbf{K} \cdot \mathbb{SN}(\mathbf{M}_l)), \qquad \mathbf{K}_m \in \mathbb{R}^{T\times N\times D}, \tag{13}$$

$$\mathbf{V}_m = \mathbb{SN}(W^m_\mathbf{V} \cdot \mathbb{SN}(\mathbf{M}_l)), \qquad \mathbf{V}_m \in \mathbb{R}^{T\times N\times D}. \tag{14}$$

Through F-SDCA, the sparse query tokens are enriched with the most relevant semantic content from the memory repository, compressing global context into a small number of representative tokens $\mathbf{X}'_l$.

Inverse-SDCA (I-SDCA) mimics the memory reconsolidation process, in which retrieved memories enter a labile state and are updated by new incoming signals. Specifically, I-SDCA projects the enriched query tokens $\mathbf{X}'_l$ into value representations $\mathbf{V}_x$, and then reuses the attention map $\mathbf{Attn}$ from the F-SDCA to propagate this updated information back to the memory repository $\mathbf{M}_l$:

$$\mathbf{M}_{l+1} = \mathbb{SN}(\mathbf{M}_l + \text{I-SDCA}(\mathbf{Attn}, \mathbf{X}'_l)), \tag{15}$$

$$\text{I-SDCA}(\mathbf{Attn}, \mathbf{X}'_l) = \mathbb{SN}(\mathbf{Attn}^\top \mathbf{V}_x * s), \tag{16}$$

$$\mathbf{V}_x = \mathbb{SN}(W^x_{\mathbf{V}} \cdot \mathbb{SN}(\mathbf{X}'_l)), \tag{17}$$

This feedback pathway ensures that the memory repository evolves smoothly across transformer blocks, avoiding the feature mismatch caused by a depth-invariant static memory store. Notably, by reusing rather than recomputing attention scores, I-SDCA introduces minimal additional cost.

Spike-driven Self-Attention and Feed-Forward Network Following Bi-SDCA, the SSA and FFN modules further process the query tokens $\mathbf{X}'_l$ to enrich their representations:

$$\mathbf{X}''_l = \text{SSA}(\mathbf{X}'_l) + \mathbf{X}'_l, \tag{18}$$

$$\mathbf{X}_{l+1} = \text{FFN}(\mathbf{X}''_l) + \mathbf{X}''_l. \tag{19}$$

Since only k fraction of tokens are propagated through SSA and FFN, their computational cost is significantly reduced compared to token-intensive baselines.

In summary, MiST mirrors the brain's memory mechanism through three processes. First, the full feature map serves as a memory repository encoding global scene semantics, analogous to memory formation. Second, sparse query tokens retrieve global semantic content via F-SDCA, in which the memory repository dynamically generates keys and values that mimic hippocampal and neocortical representations, analogous to memory formation and retrieval. Third, I-SDCA feeds the enriched queries back to update the repository, analogous to memory reconsolidation. This design offers two key advantages. First, propagating only k fraction of tokens substantially reduces computational cost even with the extra Bi-SDCA module (see Sec. 4.3). Second, the memory repository provides global semantic guidance, which helps preserve semantic integrity under aggressive token compression.

4.2 Semantic-guided Self-distillation Strategy

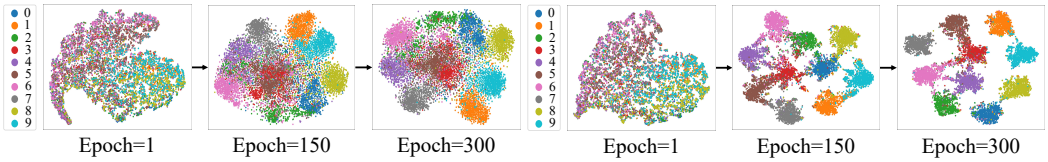
The memory repository plays a central role in MiST: it provides global semantic guidance that enables sparse query tokens to retrieve task-relevant information throughout forward propagation. Its semantic organization therefore directly affects both convergence speed and model performance.

Despite its central role, the memory repository does not inherently organize itself into semantically meaningful structures during early training. To investigate this, we visualize the evolution of the memory repository using t-distributed stochastic neighbor embedding (t-SNE) on CIFAR-10 with Spikformer-4-384. As shown in Figure 4(a), the memory repository fails to form distinguishable semantic clusters even after 300 epochs of training. Since the sparse query tokens rely on the memory repository for semantic guidance at every block, this disorder propagates through the entire forward pass, substantially weakening the quality of retrieved features and slowing model convergence. Therefore, the task loss alone is insufficient to drive the memory repository toward well-structured semantic representations. This motivates the need for an additional supervision signal that directly regularizes the semantic structure of the memory repository during training.

We propose the SGSD strategy, which imposes explicit supervision on the memory repository to drive it toward compact and stable semantic anchors. As shown in Figure 3(d), SGSD adopts a composite loss with two objectives. First, an auxiliary classification head is attached to the memory repository during training to provide class-discriminative supervision via cross-entropy:

$$\mathcal{L}_{CE}(\mathbf{Y}_m, \mathbf{Y}) = -\sum_{i=1}^{C} y_i \log(p_{m,i}), \tag{20}$$

where $\mathbf{Y}_m$ and $\mathbf{Y}$ is the memory repository predictions and ground-truth labels, C is the number of classes, and $p_{m,i}$ is the predicted probability of class i. While cross-entropy aligns the memory



(a) T-SNE visualization without SGSD (b) T-SNE visualization with SGSD

Figure 4: T-SNE visualization of the memory repository on CIFAR-10 with Spikformer-4-384. (a) MiST without SGSD, where semantic anchors remain diffuse after 300 epochs. (b) MiST with SGSD, where the memory repository rapidly evolves into compact and stable semantic anchors at 150 epochs.

repository to class boundaries, it does not ensure alignment with the compressed representations produced by the query pathway after L blocks. To bridge this gap, the second objective aligns the memory repository's output distribution to that of the final query tokens via KL divergence:

$$\mathcal{L}_{KL}(\mathbf{Y}_x \| \mathbf{Y}_m) = \sum_{i=1}^{C} p_{x,i} \log\left(\frac{p_{x,i}}{p_{m,i}}\right), \tag{21}$$

where $\mathbf{Y}_x$ is the prediction from query tokens. The two terms are combined into the final SGSD loss:

$$\mathcal{L}_{SGSD} = \mathcal{L}_{CE}(\mathbf{Y}_m, \mathbf{Y}) + \mathcal{L}_{KL}(\mathbf{Y}_x \| \mathbf{Y}_m). \tag{22}$$

Taken together, $\mathcal{L}_{CE}$ anchors the repository to ground-truth class boundaries, while $\mathcal{L}_{KL}$ further aligns it with query-distilled semantics. As shown in Figure 4(b), SGSD drives the memory repository to form compact and well-separated semantic clusters as early as epoch 150, ensuring stable and precise guidance to query tokens from early in training and thereby improving convergence speed.

4.3 Computational Cost Analysis

We analyze the per-block computational cost in terms of floating-point operations (FLOPs) and compare a vanilla spiking transformer (VST) with MiST. Specifically, in each VST block, SSA and FFN operate on all N tokens. The SSA incurs $3ND^2$ for Q/K/V projections, N^2D for attention, and ND^2 for output projection, totaling $4ND^2 + N^2D$ FLOPs. The FFN costs $4ND_{\text{in}}D_{\text{hid}}$ FLOPs. In contrast, each MiST block introduces Bi-SDCA while restricting SSA and FFN to kN tokens. The Bi-SDCA consists of F-SDCA and I-SDCA; crucially, I-SDCA reuses the attention map from F-SDCA at no extra cost. The combined Bi-SDCA cost is $(2k{+}2)ND^2$ for Q/K/V projections, $2kN^2D$ for cross-attention, and $(k{+}1)ND^2$ for output projection. SSA and FFN then cost $4kND^2 + k^2N^2D$ and $4kND_{\text{in}}D_{\text{hid}}$, respectively. Table 1 summarizes the per-block FLOPs for each module. Despite the additional Bi-SDCA overhead, MiST remains highly efficient overall, as the token reduction applied to SSA and FFN yields substantial savings that more than compensate for the added cost. For instance, at $k{=}0.25$, $N{=}196$, and $D{=}768$, MiST reduces per-block FLOPs from 2.34 to 1.01 GFLOPs, yielding a $2.32\times$ speedup.

Table 1: FLOPs comparison between VST and MiST. $k \in [0.12, 0.41]$ is the token compression ratio; $D_{\text{in}}/D_{\text{hid}}$ is the input/hidden dimensions.

Module	Operation	VST	MiST
Bi-SDCA	Q, K, V	—	$(2k{+}2)ND^2$
	Attention	—	$2kN^2D$
	Projection	—	$(k{+}1)ND^2$
SSA	Q, K, V	$3ND^2$	$3kND^2$
	Attention	N^2D	k^2N^2D
	Projection	ND^2	kND^2
FFN	Linear 1	$2ND_{\text{in}}D_{\text{hid}}$	$2kND_{\text{in}}D_{\text{hid}}$
	Linear 2	$2ND_{\text{in}}D_{\text{hid}}$	$2kND_{\text{in}}D_{\text{hid}}$

5 Experiment

In this section, we evaluate MiST on multiple benchmarks, including image classification and semantic segmentation, to demonstrate its effectiveness and efficiency. We then provide ablation studies, visualize the key components in the Bi-SDCA module, and analyze the impact of different compression ratios. Experimental details are provided in the Appendix A.

Table 2: Performance comparison on image classification using Spikformer-4-384. ***#N*** denotes the token compression ratio. The top three results are highlighted in first, second, and third.

Method	CIFAR-10			CIFAR-100			DVS-CIFAR10		
	#N	***T***	Acc (%)	***#N***	***T***	Acc (%)	***#N***	***T***	Acc (%)
Spikformer Zhou et al.	×1	4	95.19	×1	4	78.21	×1	16	80.9
SparseSpikformer Liu et al. (2024)	×0.85	4	95.18 (-0.01)	×0.85	4	77.70 (-0.51)	×0.85	16	79.3 (-1.6)
	×0.70	4	95.03 (-0.16)	×0.70	4	77.07 (-1.14)	×0.70	16	78.4 (-2.5)
	×0.63	4	94.77 (-0.42)	×0.63	4	76.78 (-1.43)	×0.63	16	79.1 (-1.8)
AT-SNN Kang et al. (2024)	×0.28	4	95.06 (-0.13)	×0.75	4	78.14 (-0.07)	-	-	-
	×0.21	4	94.88 (-0.31)	×0.58	4	77.27 (-0.94)	-	-	-
STATA Zhuge et al. (2024)	×0.50	4	95.00 (-0.19)	×0.50	4	77.70 (-0.51)	×0.50	16	80.7 (-0.2)
TP-Spikformer Wei et al. (2026)	×0.25	4	95.16 (-0.03)	×0.60	4	78.48 (+0.27)	×0.78	16	81.0 (+0.1)
	×0.20	4	95.12 (-0.07)	×0.55	4	77.83 (-0.38)	×0.50	16	80.7 (-0.2)
Proposed MiST	×0.25	4	95.46 (+0.27)	×0.25	4	78.75 (+0.54)	×0.25	16	80.5 (-0.4)
	×0.12	4	95.33 (+0.14)	×0.12	4	78.69 (+0.48)	×0.12	16	80.2 (-0.7)

Table 3: Segmentation result on ADE20K.

Method	Param (M)	*T*	MIoU (%)	***#N***	Power (mJ)
Yu et al. (2022)	15.5	1	32.9	×1	147.1
Zhang et al. (2022)	68.1	1	42.7	×1	1240.6
Yao et al. (2024)	16.5	4	33.6	×1	88.1
	59.8	4	35.3	×1	183.6
Yao et al. (2025)	11.5	2	31.9	×1	18.4
	11.5	4	40.1	×1	27.2
	11.5	8	41.4	×1	33.6
MiST	12.4	4	37.6	×**0.5**	**19.5**
	12.4	4	35.6	×**0.2**	**18.2**

Table 4: Ablation study. "w/o" and "r/w" denote without and replaced with, respectively. "C., D., E." replace SDCA with several local scorers.

	Method	Acc.(%)	Compare.	Reduce.
A.	MiST	78.09	–	–
B.	w/o SGSD	77.58	*A.*	0.51 ↓
C.	r/w Random	74.00	*B.*	3.58 ↓
D.	r/w L1-norm	75.19	*B.*	2.39 ↓
E.	r/w Attention	76.57	*B.*	1.01 ↓

5.1 Performance Comparison

Image Classification We first evaluate MiST on CIFAR-10/100 and DVS-CIFAR10 with the Spikformer-4-384 backbone. As shown in Table 2, MiST achieves SOTA performance with very few tokens, outperforming the latest TP-Spikformer by 0.30% on CIFAR-10 and 0.27% on CIFAR-100 under the same ×0.25 compression ratio. We then evaluate MiST on large-scale ImageNet against token-intensive baselines and TP-Spikformer. As shown in Table 5, MiST surpasses TP-Spikformer on both SDT-V1 and SDT-V3 backbones under tighter token budgets, e.g., achieving 78.09% on SDT-V3 at ×0.25 vs. 77.55% at ×0.56. Notably, MiST achieves 78.06% accuracy using only 5% of the tokens, demonstrating clear advantages in both accuracy and efficiency.

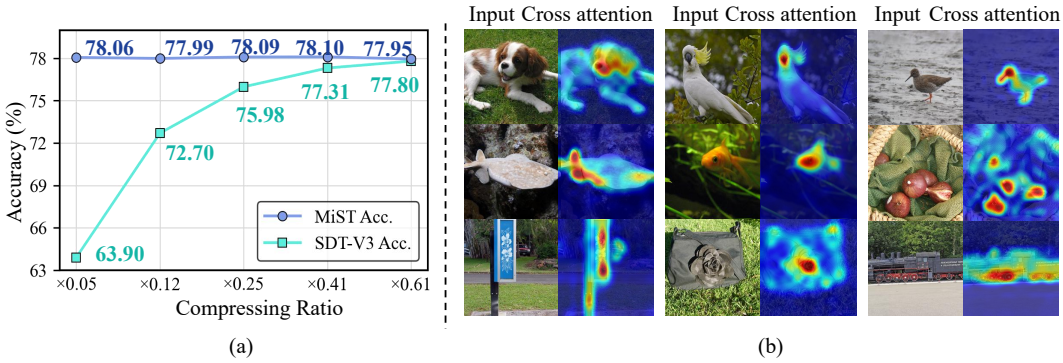
Semantic Segmentation We further evaluate MiST on semantic segmentation following prior work, with k=0.5 and k=0.2. To preserve full-resolution feature maps for the task head, we downsample the memory repository at the start of each block for efficient inference, and restore the tokens to their original spatial locations before the task head. As shown in Table 3, with only 20% of tokens, MiST attains higher mIoU (35.6% vs. 31.9%) and lower energy consumption (18.2 mJ vs. 18.4 mJ) than SDT-V3, confirming its effectiveness for dense prediction tasks under aggressive token compression.

5.2 Ablation and Validation Study

Ablation Study We ablate the two key components of MiST, i.e., Bi-SDCA and SGSD, on ImageNet with SDT-V3 at k=0.25. As shown in Table 4, removing SGSD causes a 0.51% accuracy drop (*A* vs. *B*), indicating its role in stabilizing memory convergence. In addition, replacing Bi-SDCA with heuristic local scorers (random, L1-norm, attention-based) leads to a further 1.01–3.58% drop (*B* vs. *C–E*), showing the clear advantage of global memory-guided token selection over local scoring. More detailed metrics comparing MiST with the uncompressed baseline are shown in Figure 1.

Table 5: Comparison on ImageNet. '†' is the synaptic operation of all transformer blocks. 'Thr.' is the throughput of each method. All results are evaluated on an H100 GPU under the same batch size.

Method	Architecture	T	Acc. (%)	#N	SOPs† (G)	Power (mJ)	Thr. (imgs/s)
SEW-ResNet Fang et al. (2021)	SEW-ResNet-152	4	69.26	×1	–	12.89	–
SpikingResformer Shi et al. (2024)	SpikingResformer	4	77.24	×1	–	5.46	–
SNN-ViT Wang et al. (2025)	SNN-ViT-8-512	4	80.23	×1	–	35.75	–
SDT-V2 Yao et al. (2024)	SDT-V2-8-512	4	77.20	×1	–	32.8	–
Spikformer Zhou et al.	Spikformer-8-768	4	74.81	×1	18.91	21.48	392
STATA Zhuge et al. (2024)	Spikformer-8-768	4	74.03	×0.50	–	11.16	–
Proposed MiST	Spikformer-8-768	4	74.66	×0.41	7.17	11.77	556
	Spikformer-8-768	4	74.02	×0.25	5.05	9.86	588
SDT-V1 Yao et al. (2023)	SDT-V1-8-768	4	76.32	×1	9.04	10.26	457
TP-Spikformer Wei et al. (2026)	SDT-V1-8-768	4	74.79	×0.51	4.71	6.36	–
Proposed MiST	SDT-V1-8-768	4	75.79	×0.25	3.60	5.72	641
	SDT-V1-8-768	4	74.86	×0.12	2.42	4.76	686
SDT-V3 Yao et al. (2025)	E-SpikeFormer-L	4	79.78	×1	2.66	5.90	595
TP-Spikformer Wei et al. (2026)	E-SpikeFormer-L	4	77.55	×0.56	1.55	4.25	–
Proposed MiST	E-SpikeFormer-L	4	78.09	×0.25	2.34	4.94	685
	E-SpikeFormer-L	4	78.06	×0.05	1.56	4.29	756



(a) (b)

Figure 5: (a) Performance comparison between MiST and SDT-V3 with direct token pruning. (b) Visualization of the cross-attention maps in MiST.

Study of Compressing Ratio We further analyze the effect of different token ratios in MiST and compare it with SDT-V3 under direct token pruning. As shown in Figure 5(a), direct token pruning on SDT-V3 causes clear performance degradation, which indicates that simply reducing the token count is insufficient for achieving an efficient and reliable design in token-intensive architectures. In contrast, MiST maintains stable accuracy as k decreases, demonstrating its ability to preserves global semantic information under aggressive token compression. This further verifies the efficacy of MiST.

Effectiveness of Cross-Attention To qualitatively validate the Bi-SDCA module, we visualize its cross-attention maps. As shown in Figure 5(b), the attention maps consistently highlight semantically meaningful regions of the target object. This selective focus enables the query tokens to capture the most task-relevant regions, effectively compressing global contextual information into a small number of tokens. This highlights the role of Bi-SDCA in semantic-preserving token compression.

6 Conclusion

This paper presents a token-efficient spiking transformer MiST. By introducing a memory repository with Bi-SDCA and a SGSD strategy, MiST enables sparse query tokens to efficiently retrieve global semantic information while maintaining stable convergence. Unlike previous token-efficient SNNs that rely on irreversible local scoring, MiST uses global context to guide token selection and better preserves semantic information under aggressive compression. Extensive experiments show that MiST achieves accuracy comparable to token-intensive baselines using only 20∼40% of the tokens, highlighting its potential for deploying spiking transformers on resource-limited edge devices.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript writing, including language editing, grammar correction, and improving the clarity and conciseness of the presentation. Generative AI was not used to generate experimental results or make scientific conclusions. All AI-assisted content was carefully reviewed and revised by the authors to ensure technical accuracy and consistency with the underlying research. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

David Attwell and Simon Laughlin. An energy budget for signaling in the grey matter of the brain. *Journal of cerebral blood flow and metabolism : official journal of the International Society of Cerebral Blood Flow and Metabolism*, 21:1133–45, 10 2001. doi: 10.1097/00004647-200110000-00001.

Honglin Cao, Zijian Zhou, Wenjie Wei, Yu Liang, Ammar Belatreche, Dehao Zhang, Malu Zhang, Yang Yang, and Haizhou Li. Binary event-driven spiking transformer. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 4110–4118, 2025.

Jared F Danker, Alexa Tompary, and Lila Davachi. Trial-by-trial hippocampal encoding activation predicts the fidelity of cortical reinstatement during subsequent retrieval. *Cerebral Cortex*, 27(7): 3515–3524, 2017.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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

Yuetong Fang, Deming Zhou, Ziqing Wang, Hongwei Ren, ZeCui Zeng, Lusong Li, Renjing Xu, et al. Spiking neural networks need high-frequency information. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Samuel J. Gershman, Ila Fiete, and Kazuki Irie. Key-value memory in the brain. *Neuron*, 113(11):1694–1707.e1, 2025. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2025.02.029. URL https://www.sciencedirect.com/science/article/pii/S0896627325001722.

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

Travis D Goode, Kazumasa Z Tanaka, Amar Sahay, and Thomas J McHugh. An integrated index: engrams, place cells, and hippocampal memory. *Neuron*, 107(5):805–820, 2020.

Mark Horowitz. 1.1 computing's energy problem (and what we can do about it). In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, pp. 10–14, 2014a. doi: 10.1109/ISSCC.2014.6757323.

Mark Horowitz. 1.1 computing's energy problem (and what we can do about it). In *2014 IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 10–14. IEEE, 2014b.

Reshad Hosseini, Fabian Sinz, and Matthias Bethge. Lower bounds on the redundancy of natural images. *Vision Research*, 50(22):2213–2222, 2010. ISSN 0042-6989. doi: https://doi.org/10.1016/j.visres.2010.07.025. URL https://www.sciencedirect.com/science/article/pii/S004269891000372X. Mathematical Models of Visual Coding.

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.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, pp. 2. Minneapolis, Minnesota, 2019.

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

Guoqi Li, Lei Deng, Huajin Tang, Gang Pan, Yonghong Tian, Kaushik Roy, and Wolfgang Maass. Brain-inspired computing: A systematic survey and future trends. *Proceedings of the IEEE*, 2024.

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

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Chengzhuo Lu, Huilin Du, Wenjie Wei, Qian Sun, Yuchen Wang, Dingyi Zeng, Wenyu Chen, Malu Zhang, and Yang Yang. Estsformer: Efficient spatio-temporal spiking transformer. *Neural Networks*, pp. 107786, 2025.

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

James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995.

Randall C O'Reilly and Kenneth A Norman. Hippocampal and neocortical contributions to memory: Advances in the complementary learning systems framework. *Trends in cognitive sciences*, 6(12): 505–510, 2002.

D Pacheco Estefan, Martí Sánchez-Fibla, Armin Duff, Alessandro Principe, Rodrigo Rocamora, Hui Zhang, Nikolai Axmacher, and Paul FMJ Verschure. Coordinated representational reinstatement in the human hippocampus and lateral temporal cortex during episodic memory retrieval. *Nature communications*, 10(1):2255, 2019.

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

Kazumasa Z Tanaka, Aleksandr Pevzner, Anahita B Hamidi, Yuki Nakazawa, Jalina Graham, and Brian J Wiltgen. Cortical representations are reinstated by the hippocampus during memory retrieval. *Neuron*, 84(2):347–354, 2014.

Timothy J Teyler and Pascal DiScenna. The hippocampal memory indexing theory. *Behavioral neuroscience*, 100(2):147, 1986.

Danil Tyulmankov, Ching Fang, Annapurna Vadaparty, and Guangyu Robert Yang. Biological learning in key-value memory networks, 2021. URL https://arxiv.org/abs/2110.13976.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

Longguang Wang, Xiaoyu Dong, Yingqian Wang, Xinyi Ying, Zaiping Lin, Wei An, and Yulan Guo. Exploring sparsity in image super-resolution for efficient inference. In *CVPR*, 2021.

Shuai Wang, Malu Zhang, Dehao Zhang, Ammar Belatreche, Yichen Xiao, Yu Liang, Yimeng Shan, Qian Sun, Enqi Zhang, and Yang Yang. Spiking vision transformer with saccadic attention. *arXiv preprint arXiv:2502.12677*, 2025.

Wenjie Wei, Yu Liang, Ammar Belatreche, Yichen Xiao, Honglin Cao, Zhenbang Ren, Guoqing Wang, Malu Zhang, and Yang Yang. Q-snns: Quantized spiking neural networks. *arXiv preprint arXiv:2406.13672*, 2024a.

Wenjie Wei, Malu Zhang, Jilin Zhang, Ammar Belatreche, Jibin Wu, Zijing Xu, Xuerui Qiu, Hong Chen, Yang Yang, and Haizhou Li. Event-driven learning for spiking neural networks. *arXiv preprint arXiv:2403.00270*, 2024b.

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 *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=L5llQD0nMf.

James C. R. Whittington, Joseph Warren, and Timothy E. J. Behrens. Relating transformers to models and neural representations of the hippocampal formation, 2022. URL https://arxiv.org/abs/2112.04035.

Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, and Luping Shi. Spatio-temporal backpropagation for training high-performance spiking neural networks. *Frontiers in neuroscience*, 12:331, 2018.

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

Man Yao, JiaKui Hu, Tianxiang Hu, Yifan Xu, Zhaokun Zhou, Yonghong Tian, Bo Xu, and Guoqi Li. Spike-driven transformer v2: Meta spiking neural network architecture inspiring the design of next-generation neuromorphic chips. *arXiv preprint arXiv:2404.03663*, 2024.

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*, 2025.

Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10819–10829, 2022.

Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2736–2746, 2022.

Zeqi Zheng, Yanchen Huang, Yingchao Yu, Zizheng Zhu, Junfeng Tang, Zhaofei Yu, and Yaochu Jin. Spiliformer: Enhancing spiking transformers with lateral inhibition, 2025. URL https://arxiv.org/abs/2503.15986.

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 q-k attention. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=AVd7DpiooC.

Chenlin Zhou, Liutao Yu, Zhaokun Zhou, Han Zhang, Jiaqi Wang, Huihui Zhou, Zhengyu Ma, and Yonghong Tian. Spikingformer: A key foundation model for spiking neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2236–2244, 2026.

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

Zhengyang Zhuge, Peisong Wang, Xingting Yao, and Jian Cheng. Towards efficient spiking transformer: a token sparsification framework for training and inference acceleration. In *Forty-first International Conference on Machine Learning*, 2024.

A Experimental Details

A.1 Image Classification

Dataset We assess the performance of MiST across both static image datasets and event-based dynamic data. For static benchmarks, we adopt CIFAR-10, CIFAR-100, and ImageNet-1K. CIFAR-10 consists of 60,000 low-resolution (32×32) images evenly distributed over 10 classes, serving as a standard testbed for general visual recognition. CIFAR-100 follows the same image format but increases the task complexity by introducing 100 fine-grained categories grouped into 20 higher-level classes. In contrast, ImageNet-1K provides a much larger and more diverse evaluation setting, containing over one million training images and 1,000 object categories, making it a widely recognized benchmark for large-scale classification. To further examine temporal processing capability, we include the event-driven dataset DVS-CIFAR10, which is generated from CIFAR-10 using a dynamic vision sensor. Unlike frame-based datasets, DVS-CIFAR10 captures asynchronous spike events with microsecond-level temporal precision and a spatial resolution of 128×128. It comprises 9,000 training samples and 1,000 test samples, enabling a comprehensive evaluation of the model's performance on neuromorphic, time-dependent visual data.

Experimental Setup For experiments on CIFAR-10, CIFAR-100, and DVS-CIFAR10, we evaluate our model under two unified token compressing ratio to study its behavior across varying sparsity levels. Specifically, the average token retention ratio $\#N$ is set to 0.25 and 0.12, using a Spikformer-4-384 backbone with a learning rate of 5e-4 and 4 time steps. For DVS-CIFAR10, we adopt a lighter Spikformer-2-384 architecture with same token compression ratio and learning rate, while keeping other settings consistent with prior work. For large-scale ImageNet-1K, we further evaluate 3 different architectures, with 2 compression ratios each architecture. During training, we set up 300 epochs for QKFormer and SDT-V1, and 200 epochs for SDT-V3 with a learning rate of 6e-4, using per-GPU batch sizes of 64 and 512 for different model scales. The token preservation ratio is set to 0.41 and 0.25 for spikformer, 0.25 and 0.12 for SDT-V1, 0.25 and 0.05 for SDT-V3.

A.2 Semantic Segmentatio

Dataset ADE20K is a large-scale benchmark widely adopted for evaluating semantic segmentation models. The dataset contains around 25K images, which are split into over 20K training samples, 2K validation samples, and 3K test samples. Each image is annotated at the pixel level with 150 semantic classes, covering both object categories (e.g., humans, vehicles, and animals) and diverse scene components such as sky, buildings, and vegetation. The rich contextual variations and fine-grained annotations introduce significant challenges for dense prediction tasks. In our work, ADE20K provides a rigorous testbed to assess the effectiveness of the proposed memory-driven token-efficient framework, particularly its ability to preserve global semantic information under aggressive token compression while maintaining competitive segmentation performance.

Experimental Setup In this work, we first extend the *mmsegmentation* framework into a spike-based paradigm, following the design principles of SDT-V3 Yao et al. (2025). To adapt it to our memory-inspired token-efficient architecture, we adopt SDT-V3-19M as the backbone and integrate it with a spike-driven FPN for dense prediction. The backbone is initialized with ImageNet pretrained weights to provide robust feature representations. All training configurations are aligned with SDT-V3 to ensure fair comparison. During fine-tuning, we evaluate two compression ratios consistent with the object detection setting, using 4×H100 GPUs with a batch size of 16 per GPU. As reported in Table 3, our method achieves comparable segmentation performance to SDT-V3 while significantly reducing energy consumption. We emphasize that the goal of this comparison is not to pursue absolute accuracy gains, but to demonstrate that the proposed memory-driven token-efficient design can effectively preserve performance under substantial token reduction, highlighting its advantage in efficiency-oriented scenarios.

B Learning Algorithm of MiST

MiST is trained end-to-end using spatiotemporal backpropagation (STBP) Wu et al. (2018), which unrolls gradients across both the spatial (layer-wise) and temporal (time-step) dimensions of SNNs.

Concretely, the gradient of the loss $\mathcal{L}$ with respect to a synaptic weight W^l at layer l is decomposed via the chain rule and accumulated over all T time steps as follows:

$$\frac{\partial \mathcal{L}}{\partial W^l} = \sum_{t=1}^{T} \left(\frac{\partial \mathcal{L}}{\partial \mathbf{S}^l[t]} \frac{\partial \mathbf{S}^l[t]}{\partial \tilde{\mathbf{U}}^l[t]} \frac{\partial \tilde{\mathbf{U}}^l[t]}{\partial W^l} + \frac{\partial \mathcal{L}}{\partial \mathbf{U}^l[t+1]} \frac{\partial \mathbf{U}^l[t+1]}{\partial \tilde{\mathbf{U}}^l[t]} \frac{\partial \tilde{\mathbf{U}}^l[t]}{\partial W^l} \right), \tag{23}$$

where the error signals $\partial \mathcal{L}/\partial \mathbf{S}^l[t]$ and $\partial \mathcal{L}/\partial \mathbf{U}^l[t+1]$ are propagated iteratively, while the remaining partial derivatives with respect to the pre-synaptic membrane potential $\tilde{\mathbf{U}}^l[t]$ can be derived analytically from the LIF neuron dynamics defined in Eq. 7.

A fundamental challenge inherent to this training procedure arises from the non-differentiable nature of the spike generation mechanism. Specifically, the Heaviside step function governing spike emission in Eq. 7 possesses a zero-valued derivative almost everywhere and remains undefined at the firing threshold θ. Consequently, the term $\partial \mathbf{S}^l[t]/\partial \tilde{\mathbf{U}}^l[t]$ is analytically intractable, precluding the direct application of gradient-based optimization. To circumvent this limitation, we follow the widely adopted surrogate gradient approach Wu et al. (2018), which replaces the true but ill-defined derivative with a smooth approximation during the backward pass while leaving the forward dynamics unchanged. Several surrogate functions have been proposed in the literature, including rectangular Wei et al. (2024a) and linear Wei et al. (2024b) variants. In this work, we adopt the triangular surrogate gradient, defined as:

$$\frac{\partial \mathbf{S}^l[t]}{\partial \tilde{\mathbf{U}}^l[t]} = \max\Big(0,\ \beta - \big|\tilde{\mathbf{U}}^l[t] - \theta\big|\Big), \tag{24}$$

where $\beta > 0$ governs the effective support of the gradient region, and θ denotes the neuronal firing threshold consistent with Eq. 7. This formulation provides a computationally efficient and empirically effective approximation. With this surrogate substitution in place, gradients can flow through the spiking layers without interruption, rendering the entire MiST network, including all LIF neuron layers, fully compatible with standard error backpropagation.

C Energy Consumption Calculation

Following established practice in the SNN literature Zhou et al.; 2024), we estimate the theoretical energy consumption of MiST by assuming that all operations are executed on 45nm CMOS hardware Horowitz (2014a), where a multiply-accumulate (MAC) operation consumes $E_{\text{MAC}} = 4.6\,\text{pJ}$ and a spike-driven accumulate (AC) operation costs $E_{\text{AC}} = 0.9\,\text{pJ}$ Horowitz (2014b). Since dense floating-point multiplications only occur in the first convolutional layer of the patch embedding stage, while all subsequent computations are driven by binary spike activations, the total energy expenditure of MiST is formulated as:

$$E_{\text{total}} = E_{\text{MAC}} \cdot \text{FLOPs}^1_{\text{emb}} + E_{\text{AC}} \cdot \left(\sum_{n=2}^{N} \text{SOPs}^n_{\text{emb}} + \sum_{l=1}^{L} \text{SOPs}^l_{\text{block}} + \text{SOPs}_{\text{head}} \right), \tag{25}$$

where SOPs denotes the number of synaptic operations (i.e., spike-driven AC operations); $\text{FLOPs}^1_{\text{emb}}$ is the floating-point cost of the first convolutional layer in the patch embedding module; $\text{SOPs}^n_{\text{emb}}$ and $\text{SOPs}_{\text{head}}$ account for the spike-driven operations in the remaining embedding layers and the classification head, respectively; and $\text{SOPs}^l_{\text{block}}$ is the synaptic operation count of the l-th transformer block. For any spiking layer, the SOPs are related to its floating-point equivalent via the average neuronal firing rate:

$$\text{SOPs}^l = f_r^l \cdot T \cdot \text{FLOPs}^l, \tag{26}$$

where f_r^l denotes the average firing rate of layer l across all T time steps, and FLOPs^l is the floating-point operation count of the equivalent dense layer. The overall architecture of MiST consists of three stages: patch embedding, L transformer blocks, and a classification head. The energy costs of the patch embedding and classification head are unaffected by token compression and thus remain the same as in the uncompressed baseline. The primary efficiency gains come from the transformer blocks, where both SSA and FFN operate on only k fraction of the full token set, directly reducing $\text{SOPs}^l_{\text{block}}$ by a corresponding factor relative to token-intensive architectures.

D HARDWARE COST OF MEMORY REPOSITORY ACCESS

To assess whether the memory repository introduced in MiST imposes a non-trivial hardware burden, we conduct a quantitative analysis of its memory access energy overhead under realistic hardware assumptions. A key factor that inherently constrains this cost is the binary representation of the stored contents: unlike conventional ANN feature maps that require full-precision floating-point values, the memory repository holds 1-bit spike activations, which substantially reduces both its storage footprint and the associated read energy.

Following the standard 45nm CMOS process assumptions Horowitz (2014b), we adopt an SRAM read energy of $E_{\text{SRAM}} = 0.16\,\text{pJ/bit}$ and a DRAM read energy of $E_{\text{DRAM}} = 20\,\text{pJ/bit}$. Since the memory repository is accessed at every transformer block during the forward pass, the total read energy accumulated over all L blocks is formulated as:

$$E_{\text{read}} = \sum_{l=1}^{L} E_{\text{DRAM}} \cdot N \cdot D \cdot T \cdot n \cdot 1\,\text{bit}, \tag{27}$$

where N is the number of spatial tokens in the full feature map, D is the feature dimension, T is the simulation time steps, n denotes the number of memory accesses per data block within each transformer block, and the factor 1 bit reflects the binary precision of the stored spike representations. The corresponding results for MiST evaluated on ImageNet are summarized in Table 6.

Table 6: Memory access energy cost of MiST on ImageNet under the 45nm process assumption.

Spatial Tokens (N)	Feature Dim. (D)	Time Steps (T)	Accesses per Block (n)	Transformer Blocks (L)	Read Energy (E_{read})
196	768	4	2	8	0.193 mJ

As reported in Table 6, the total memory access energy for MiST on ImageNet amounts to only $0.193\,\text{mJ}$, which is negligible in comparison to the model's overall power consumption as reported in Table 5. This result confirms that the memory repository introduces no significant hardware overhead, and that the 1-bit spike representation is a practically effective strategy for keeping access costs minimal. These results show that MiST's memory mechanism is not only algorithmically efficient but also hardware-friendly, highlighting its suitability for deployment on resource-limited devices.

E ABLATION ON BIDIRECTIONAL VS. UNIDIRECTIONAL SDCA

To investigate the necessity of the bidirectional design in Bi-SDCA, we compare the full Bi-SDCA module against a unidirectional ablation in which the I-SDCA pathway is removed, leaving the memory repository static and unchanged throughout all L transformer blocks. In the full Bi-SDCA design, the I-SDCA pathway propagates enriched query representations back to update the memory repository after each block, enabling it to evolve alongside the query tokens during forward propagation. Removing this feedback pathway effectively decouples the repository from the query dynamics, reducing Bi-SDCA to a one-way retrieval mechanism analogous to standard cross-attention.

Table 7: Ablation study comparing bidirectional and unidirectional SDCA on CIFAR-10 with Spikformer-4-384 at $\times 0.12$ compression.

Architecture	$\#N$	Cross-Attention	Accuracy (%)
Spikformer-4-384	$\times 0.12$	Bidirectional (Ours)	**95.33**
Spikformer-4-384	$\times 0.12$	Unidirectional	94.95

As reported in Table 7, removing the I-SDCA pathway leads to a clear accuracy drop from 95.33% to 94.95% on CIFAR-10 under the $\times 0.12$ compression ratio with the Spikformer-4-384 backbone. We attribute this degradation to progressive feature misalignment: without feedback from the updated query tokens, the memory repository remains static and grows increasingly inconsistent with the

evolving query representations across deeper blocks. This misalignment weakens the quality of semantic guidance that the repository provides to subsequent blocks, ultimately impacting classification performance. In contrast, retaining I-SDCA allows the repository to continuously absorb task-relevant information from the compressed queries, sustaining coherent semantic guidance throughout all transformer blocks. This gap is likely to become more pronounced on larger and semantically richer benchmarks such as ImageNet, where cross-block feature alignment plays a more critical role.

F Robustness of Token Sampling Strategy

We conduct a ablation study to examine the sensitivity of MiST to the choice of initial token sampling strategy. Specifically, we replace the default L1-norm sampling with random sampling and evaluate the resulting performance on CIFAR-10. This analysis is motivated by the observation that token sampling is not a primary design contribution of MiST. The core performance drivers are Bi-SDCA, which retrieves global semantic content from the memory repository into sparse query tokens, and SGSD, which regularizes the semantic structure of the repository during training. The sampling step simply determines which subset of tokens from the memory repository is passed as the initial query set $\mathbf{X}_0$, and thus has limited influence on overall model accuracy.

Table 8: Ablation of token sampling on CIFAR-10 with Spikformer-4-384 at $\times 0.12$ compression.

Architecture	$\#N$	Sampling Method	Accuracy (%)
Spikformer-4-384	$\times 0.12$	L1-norm (Default)	**95.33**
Spikformer-4-384	$\times 0.12$	Random	95.23

As shown in Table 8, replacing L1-norm sampling with random sampling leads to a marginal accuracy drop of only 0.10% (95.33% vs. 95.23%) at a $\times 0.12$ compression ratio with the Spikformer-4-384 backbone. This negligible difference indicates that MiST's accuracy is largely independent of the specific criterion used to select the initial query tokens. Taken together, these results confirm that MiST is robust to the choice of sampling strategy and that L1-norm sampling serves as a simple default that requires no additional learned parameters or computational overhead.