SELECTIVE ACTIVATION VIA NEURONAL DYNAMICS FOR CONTINUAL LEARNING IN SPIKING NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual learning with Spiking Neural Networks (SNNs) offers a promising solution for low-power, rehearsal-free learning on resource-constrained devices. Brain-inspired selective activation, arising from neural competition and lateral inhibition, is crucial for learning sparse, task-specific representations. However, existing k-Winner-Take-All (k-WTA) SNNs allocate neurons statically, leaving many neurons inactive while task-relevant neurons remain vulnerable to subsequent interference. We propose STAR, a Spike-driven Task-Adaptive Resource allocation framework that manages the neuron lifecycle under a fixed capacity. During task training, Generate-and-Test (GnT) uses maturity-gated spike utility to replace low-utility neurons with mutated clones of high-utility neurons, thereby renewing plasticity. At task boundaries, Sleep Consolidation protects statistically selected experts and reinitializes redundant neurons to preserve knowledge and release capacity. Together, they form a generate–select–consolidate–release cycle without network expansion or data replay. Experiments on Split-MNIST, CIFAR-100, and TinyImageNet show that STAR consistently improves accuracy, memory retention, and neuron utilization. On Split-MNIST, STAR improves average accuracy by 6.03% percentage points and increases neuron utilization from 9.4% to 61.6%, while reducing forgetting across all three benchmarks. Ablation and neurodynamic analyses further validate the complementarity of GnT and Sleep Consolidation in addressing the stability–plasticity dilemma.

1 INTRODUCTION

Continual learning Parisi et al. (2019); De Lange et al. (2021) aims to enable neural networks to efficiently acquire new knowledge (plasticity) while retaining previously learned knowledge (stability) — a tension known as the stability-plasticity dilemma Grossberg (1988). When a model updates its parameters on new tasks, it tends to overwrite weights that are critical for old tasks, leading to catastrophic forgetting McCloskey & Cohen (1989); French (1999). This challenge is particularly pronounced in resource-constrained edge computing scenarios, where edge devices typically cannot store large amounts of historical data for replay-based training.

Spiking Neural Networks (SNNs) Maass (1997); Roy et al. (2019) have demonstrated unique potential in continual learning due to their inherent sparsity, event-driven computation Gerstner & Kistler (2002), and low-power nature Davies et al. (2018). While recent efforts have introduced the k-Winner-Take-All (k-WTA) mechanism into SNN-based continual learning Mozafari et al. (2018); Safa et al. (2021) to enforce sparsity and mitigate forgetting, they suffer from a static resource allocation deficiency. Over half of the neurons become "dead neurons" that contribute nothing yet persistently occupy the network's capacity. Conversely, highly active neurons lack dedicated protection mechanisms and remain susceptible to being overwritten during multi-task sequence learning, causing severe forgetting on earlier tasks.

Inspired by neurogenesis Eriksson et al. (1998) and sleep-dependent memory consolidation Walker & Stickgold (2004); Diekelmann & Born (2010) in the biological brain O'Keefe & Nadel (1978); Buzsáki (2010), we propose **Spike-driven Task-Adaptive Resource allocation (STAR)**, a dynamic neuron resource management framework for SNN-based continual learning. STAR achieves "sur-

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.

vival of the fittest, on-demand allocation" of neural resources through two complementary mechanisms: 1) **Generate-and-Test (GnT):** An intra-task strategy that dynamically identifies underperforming "dead neurons" and replaces them using the current best-performing "star neurons" via Gaussian mutation. This clears redundant capacity, ensures sufficient plasticity, and incorporates a task annealing mechanism to balance plasticity and stability throughout the sequence. 2) **Sleep Consolidation:** An inter-task mechanism mimicking biological sleep. It evaluates spike utility distributions to identify and permanently freeze the input weights of "cognitively critical" neurons, turning them into dedicated task experts. Simultaneously, redundant neurons are randomly reinitialized to free up capacity for subsequent tasks. We evaluate STAR on Split-MNIST, CIFAR-100, and TinyImageNet to examine its effects on continual learning performance and neuron utilization. The main contributions are as follows:

- We propose STAR, a spike-driven neuron resource allocation framework for rehearsal-free continual learning in SNNs. Under a fixed neuron budget, STAR combines intra-task neuron replacement with inter-task consolidation to balance plasticity renewal and knowledge retention.
- We develop two complementary mechanisms: Generate-and-Test replaces mature, low-utility neurons with perturbed copies of high-utility neurons, while consolidation freezes the input weights of selected experts and reinitializes the remaining plastic neurons. These mechanisms jointly guide the renewal and preservation of neuronal resources across tasks.
- Experiments demonstrate that STAR improves average accuracy by 6.03% on Split-MNIST and 2.05% on CIFAR-100, while reducing average forgetting across all three benchmarks. On Split-MNIST, neuron utilization increases from 9.4% to 61.6%. Component ablations further support the complementary roles of neuron replacement and consolidation.

2 Related Work

This section reviews related work from three perspectives, clarifying the key differences between STAR and existing approaches.

2.1 Continual Learning Methods

The core challenge of continual learning is mitigating catastrophic forgetting without historical data. Existing approaches are broadly categorized into regularization-based methods (e.g., EWC Kirkpatrick et al. (2017), SI Zenke et al. (2017), MAS Aljundi et al. (2018)), which penalize updates to critical parameters but often suffer from plasticity decay due to accumulated constraints; dynamic architecture-based methods (e.g., PNN Rusu et al. (2016), DEN Yoon et al. (2017)), which allocate dedicated parameters for new tasks yet incur unsustainable linear growth; and replay-based methods (e.g., Experience Replay Rolnick et al. (2019); Chaudhry et al. (2019), Generative Replay Shin et al. (2017)), which rely on stored or generated data and thus face high storage overhead and privacy risks. In contrast, STAR addresses these limitations architecturally rather than via loss penalties or memory buffers: it achieves dynamic resource scheduling under a strictly fixed neuron budget via dead neuron replacement and targeted expert freezing, remaining entirely rehearsal-free.

2.2 SNN-based Continual Learning

The temporal dynamics and sparse firing of SNNs naturally enable soft representational separation Bellec et al. (2018). While early works directly ported ANN regularization to SNNs Draelos et al. (2017); Soures et al. (2024) and recent studies introduced k-WTA mechanisms to enforce extreme sparsity Ororbia & Kifer (2022), existing SNN continual learning frameworks still face severe limitations. Although our base architecture Shen et al. (2024) systematically validated k-WTA-SNNs for continual learning, its static resource management leads to widespread dead neurons that waste capacity, the overwriting of neurons encoding old knowledge, and extremely low overall neuron utilization.

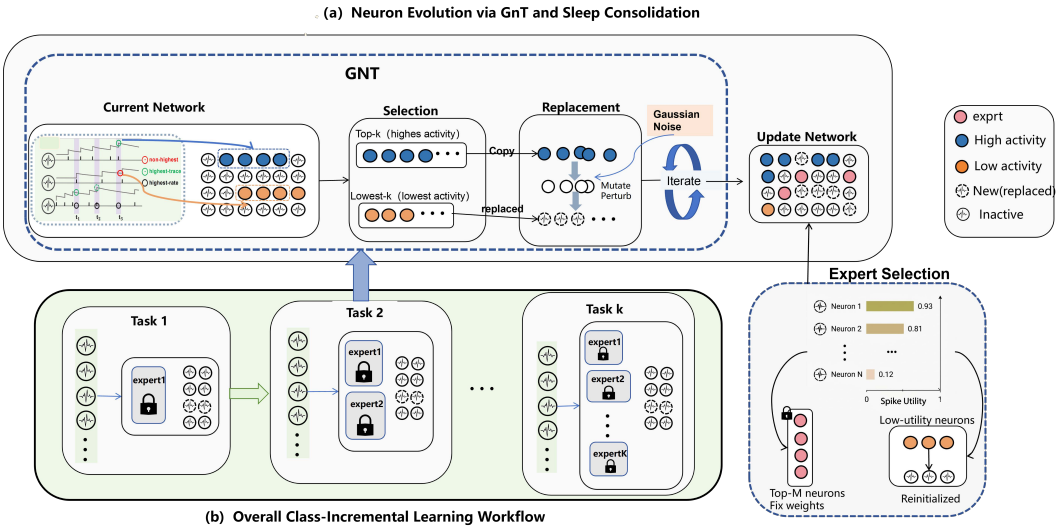

Figure 1: Overview of STAR. (a) Intra-task: consolidated experts are preserved while non-expert neurons are ranked by spike activity, and low-activity ones are replaced by mutated clones of top-performing ones via GnT. (b) Inter-task: the highest-utility neurons are frozen as task experts and sub-threshold ones are reinitialized, so that experts $(\text{expert}_1, \ldots, \text{expert}_K)$ accumulate across the task sequence.

2.3 Biologically Inspired Resource Management Mechanisms

The design of STAR draws direct inspiration from two classical neuroscience findings: neurogenesis and sleep consolidation. In mammalian brains, the hippocampus continuously generates new neurons Eriksson et al. (1998) for pattern separation Clelland et al. (2009). While prior computational models introduced neurogenesis by randomly adding neurons, leading to unbounded capacity growth Draelos et al. (2017), STAR's GnT improves this by replacing dead neurons to maintain a constant capacity, utilizing the best-performing star neuron as a mutation template, and introducing maturity gating. Furthermore, regarding memory consolidation during sleep Walker & Stickgold (2004); Diekelmann & Born (2010), the Synaptic Homeostasis Hypothesis (SHY) Tononi & Cirelli (2014) posits that sleep restores plasticity via global downscaling while preserving critical connections. While previous machine learning models utilize sleep replay Golden et al. (2022) or periodic resetting Masana et al. (2021), STAR mimics SHY by statistically freezing high-utility neurons as task experts and resetting inefficient ones based on data-driven spike utility distributions. In summary, STAR is the first to jointly model spike-driven, neuron-granularity resource scheduling and sleep consolidation in SNNs, achieving a complete lifecycle under a fixed network capacity.

3 Method

This section details the STAR framework, which consists of the foundational SNN architecture, the intra-task GnT mechanism, and the inter-task Sleep Consolidation mechanism. Fig. 1 provides an overview of the framework.

3.1 Overview of the STAR Framework

As illustrated in Fig. 1, STAR operates under a strictly fixed neuron budget through two temporally separated mechanisms. During task training, previously consolidated experts (red circles) are preserved while non-expert neurons are ranked by spike activity into a Top-k and a Lowest-k pool (Fig. 1(a)); the latter are eliminated and replaced by Gaussian-perturbed clones of the former, so that wasted capacity is continuously reclaimed and plasticity renewed. At each task boundary, neurons with the highest spike utility are promoted to experts (pink circles) with their input weights perma-

nently frozen, whereas sub-threshold neurons are reinitialized to release clean capacity for future tasks; as tasks arrive sequentially (Fig. 1(b)), experts ($\text{expert}_1, \ldots, \text{expert}_K$) thus accumulate and remain locked while the remaining plastic neurons keep interacting with GnT.

3.2 Preliminaries and Network Architecture

We adopt the classic Integrate-and-Fire (IF) neuron model. For neuron i in layer l, the dynamics of its membrane potential $v_i^{(l)}(t)$ are governed by:

$$v_i^{(l)}(t) = v_i^{(l)}(t-1) + x_i^{(l)}(t) - s_i^{(l)}(t-1) \cdot \theta_i^{(l)} \tag{1}$$

where $x_i^{(l)}(t)$ is the synaptic input and $\theta_i^{(l)}$ is the firing threshold. When the membrane potential exceeds the threshold, the neuron emits a spike:

$$s_i^{(l)}(t) = H\left(v_i^{(l)}(t) - \theta_i^{(l)}\right) \tag{2}$$

where $H(\cdot)$ denotes the Heaviside step function using a surrogate gradient Neftci et al. (2019). The accumulated spike count over time window T is defined as:

$$u_i = \sum_{t=1}^{T} s_i^{(l)}(t) \tag{3}$$

which serves as the core metric for neuronal utility. To enforce sparsity, we apply the k-WTA mechanism ($k = 100$), allowing only the top-k neurons with the highest spike counts to fire while others are inhibited Ororbia & Kifer (2022).

Consistent with prior work Shen et al. (2024), our network comprises a Convolutional Encoder $\text{ConvE}(\cdot)$, a k-WTA sparse intermediate layer with N IF neurons, and a Linear Classifier. The network adheres to Dale's Principle ($\mathbf{W} \geq 0$) Dayan & Abbott (2005) and input normalization to stabilize sparse representations Shen et al. (2024).

3.3 GnT: Generate-and-Test for Intra-Task Neuron Screening

GnT operates periodically during task training to identify and replace dead (silent) neurons while maintaining a fixed capacity. It establishes a maturity constraint via age counters $a_i \in \mathbb{N}_0$, including a neuron in the elimination candidate pool only after its age reaches $M = 100$ batches:

$$\mathcal{M} = \{i \mid a_i \geq M\} \tag{4}$$

Neurons maintain a global utility score defined as:

$$u_i^{\text{total}} = u_i^{\text{curr}} + u_i^{\text{prev}} \tag{5}$$

To adaptively determine the pruning scale, we introduce a task annealing factor:

$$\gamma_{\text{anneal}}(t) = \max\left(0, 1 - \frac{t-1}{T}\right) \tag{6}$$

yielding the replacement count $N_{\text{replace}} = \lfloor N_{\text{base}} \times \gamma_{\text{anneal}}(t) \rfloor$. Excluding immature and frozen expert neurons ($\mathcal{F}$) as non-eliminable, the adaptive pruning threshold is defined as:

$$\theta_{\text{prune}} = \max\left(\theta_{\min}, \text{median}\left(\{u_i^{\text{total}} \mid i \in \mathcal{E}\}\right) \times r_{\text{med}}\right) \tag{7}$$

GnT then performs two-stage screening to locate dead neurons $\mathcal{I}_{\text{dead}}$ whose utilities rank in the lowest N_{replace} and fall below θ_{prune}.

Once $\mathcal{I}_{\text{dead}}$ is non-empty, GnT executes generation and replacement: (1) It selects the highest-utility non-expert neuron $i^* = \arg\max_{i \notin \mathcal{F}} u_i^{\text{total}}$ as the cloning template. (2) For each dead neuron $j \in \mathcal{I}_{\text{dead}}$, input weights are mutated via:

$$\mathbf{W}_1[j,:] \leftarrow \mathbf{W}_1[i^*,:] + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}\left(0, \frac{0.5}{\sqrt{d_{\text{in}}}}\right) \tag{8}$$

and output weights are reinitialized via Xavier initialization. (3) State cleanup is executed to reset age, spike counters, Adam optimizer momentum terms, and clear EWC Fisher information matrix entries, preventing spurious regularization penalties on newborn neurons.

Algorithm 1 STAR Training Pipeline

Input: Task sequence $\mathcal{T}_1, \ldots, \mathcal{T}_T$, base replacement count N_{base}, maturity threshold M, median ratio r_{med}
Initialize: Network parameters $\mathbf{W}_1, \mathbf{W}_2$, initialize all neuron states, $\mathcal{F} \leftarrow \emptyset$
Output: Trained network parameters $\mathbf{W}_1, \mathbf{W}_2$ and expert set $\mathcal{F}$

for $t = 1$ to T **do**
  for each training batch $(\mathbf{x}, y) \in \mathcal{T}_t$ **do**
    Perform forward pass with k-WTA sparse activation
    Compute joint loss $\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{reg}}$
    Compute gradients $\nabla\mathbf{W}_1, \nabla\mathbf{W}_2$
    $\nabla\mathbf{W}_1[\mathcal{F}, :] \leftarrow 0$ {Block gradient updates for expert neurons}
    Update parameters $\mathbf{W}_1, \mathbf{W}_2$ using optimizer with constraints
    if should_trigger_GnT() **then**
      $\mathcal{I}_{\text{dead}} \leftarrow \text{GnT_screen}(\mathbf{u}, \mathbf{a}, \mathcal{F}, M, r_{\text{med}}, N_{\text{base}}, t, T)$
      GnT_replace($\mathcal{I}_{\text{dead}}, \mathcal{F}$)
    end if
  end for
  $\mathcal{E}^*_{\text{new}} \leftarrow$ Sleep_select_experts($\mathbf{u}, \mathcal{F}$)
  $\mathcal{F} \leftarrow \mathcal{F} \cup \mathcal{E}^*_{\text{new}}$ {Freeze new experts}
  Sleep_reset_non_experts($\mathcal{P} \setminus \mathcal{E}^*_{\text{new}}$) {Reset redundant neurons}
end for
return $\mathbf{W}_1, \mathbf{W}_2, \mathcal{F}$

3.4 Sleep Consolidation: Inter-Task Memory Consolidation

Triggered after each task is completed, Sleep Consolidation mimics the Synaptic Homeostasis Hypothesis (SHY) Tononi & Cirelli (2014); Walker & Stickgold (2004); Diekelmann & Born (2010) by partitioning neurons into frozen experts $\mathcal{F}$ and plastic neurons $\mathcal{P} = \{1, \ldots, N\} \setminus \mathcal{F}$.

The expert selection threshold combines statistical distribution and safety bounds:

$$\theta_{\text{expert}} = \max\left(\mu + \sigma,\ 0.20 \times u_{\max}\right) \tag{9}$$

To balance allocation, we enforce capacity controls where new experts $\mathcal{E}^*_{\text{new}}$ are capped at 50% of $\mathcal{P}$ and lower-bounded by $\min(10, |\mathcal{P}|)$. During consolidation, new experts are added to $\mathcal{F}$ and their input weight gradients are explicitly zeroed out during subsequent task training:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}_1[\mathcal{F}, :]} \leftarrow 0 \tag{10}$$

while output weights remain plastic. Meanwhile, non-expert plastic neurons undergo complete redundancy reset and reinitialization, releasing clean capacity for future tasks.

3.5 Training Pipeline and Integration

Algorithm 1 summarizes the STAR training pipeline, uniting GnT (intra-task dynamic replacement) and Sleep Consolidation (inter-task memory consolidation). STAR is fully orthogonal and composable with parameter-granularity regularization methods like EWC, providing hard resource scheduling at the neuron granularity.

4 Experiments

To validate the effectiveness and internal mechanisms of STAR, this section provides a thorough empirical evaluation. We first outline the experimental setup and present STAR's performance advantages across various standard continual learning benchmarks. Furthermore, we conduct extensive ablation studies to isolate and verify the contributions of STAR's core components. Beyond these performance and component-level evaluations, we further dive into an in-depth mechanistic analysis to reveal why and how STAR achieves its remarkable results, accompanied by a rigorous hyperparameter sensitivity analysis to demonstrate its robustness and stability.

4.1 Experimental Setup

We evaluate on three continual learning benchmarks. Split-MNIST partitions the 10 MNIST classes into 5 tasks of 2 classes each in a class-incremental setting, where the model must sequentially learn to distinguish 0–1, 2–3, …, 8–9 and ultimately discriminate all 10 digits without access to task identity. CIFAR-100 contains 100 classes split into 10 tasks of 10 classes; since training SNNs directly from raw pixels converges poorly, we follow prior work and feed a frozen non-spiking CNN feature extractor, conducting continual learning on the resulting embeddings. TinyImageNet serves as a harder large-scale benchmark, with 200 classes of 64×64 images split into 10 tasks of 20 classes.

The base architecture is a k-WTA-SNN that employs the kWTASpiktr module for trace-based hard-mask spike selection, combined with EWC regularization and the Adam optimizer; this baseline already achieves competitive rehearsal-free performance on several continual learning benchmarks. All compared methods share the same architecture, optimizer, and hyperparameters, differing only in whether the GnT and Sleep Consolidation components of STAR are enabled. We report two metrics. Average accuracy (AT) is the mean test accuracy over all tasks once the full sequence has been learned, $\text{AT} = \frac{1}{T}\sum_{i=1}^{T} a_{T,i}$, where $a_{T,i}$ is the accuracy on task i after learning the T-th task. Forgetting measures how much old-task knowledge is overwritten, $f_i = \max_{j\in\{1,\ldots,T-1\}} a_{j,i} - a_{T,i}$, i.e., the gap between the best historical accuracy on task i and its final accuracy.

The network comprises a VGG-SNN convolutional encoder, a 2000-dimensional k-WTA sparse intermediate layer ($k = 100$), and a Softmax classifier. The IF time window is $T = 8$, and inputs are normalized to the unit hypersphere. Training uses the Adam optimizer with a learning rate of 5×10^{-2} and a batch size of 64. GnT is triggered after each training epoch, with maturity threshold $M = 100$, median ratio $r_{\text{med}} = 0.1$, and base replacement count $N_{\text{base}} = 20$. All experiments run on NVIDIA RTX GPUs, with EWC ($\lambda = 10^5$) as the base regularization method across all runs.

4.2 Main Results

4.2.1 Overall Performance

Table 1 and Table 3 present the comprehensive performance comparison of STAR against other continual learning methods across benchmarks. In terms of average accuracy (Table 1), STAR achieves competitive performance, reaching 89.35% on Split-MNIST and 38.52% on CIFAR-100, outperforming the baseline and standard SNN continual learning variants (such as EWC-SNN and MAS-SNN).

Furthermore, Table 3 details the performance regarding forgetting and neuron utilization across all three benchmarks (Split-MNIST, CIFAR-100, and TinyImageNet). STAR consistently achieves lower forgetting rates compared to the baseline (k-WTA+EWC), reducing forgetting from 14.67% to 5.06% on Split-MNIST, from 19.35% to 9.12% on CIFAR-100, and from 9.73% to 7.73% on TinyImageNet. In terms of neuron utilization, STAR yields a substantial increase across all datasets—elevating utilization from 9.4% to 61.6% on Split-MNIST, from 33.4% to 57.8% on CIFAR-100, and from 52.3% to 62.5% on TinyImageNet. This demonstrates that the GnT dead neuron replacement strategy successfully activates and maintains network capacity that would otherwise remain idle.

4.2.2 Memory Retention Analysis

Fig. 2 presents a panoramic view of cross-task memory retention in the form of an accuracy heatmap. The rows of the heatmap denote the currently learned task (horizontal axis), and the columns denote the task being evaluated (vertical axis). Diagonal elements represent the accuracy immediately after learning a given task, while elements below the diagonal reflect the impact of learning subsequent tasks on earlier ones.

In the baseline (left panel), a pronounced "color fading" pattern is evident below the diagonal: when the model has progressed to Task 5 (bottom row), the accuracy of Task 1 has plummeted from nearly 100% on the diagonal to 80.7%, which is a classic manifestation of catastrophic forgetting. In contrast, in STAR (right panel), the region below the diagonal retains deep color tones: upon reaching Task 5, the accuracy of Task 1 remains at 94.0%. This comparison directly demonstrates

Table 1: Average accuracy comparison on continual learning benchmarks

Methods	Split-MNIST (5 tasks)	CIFAR-100 (10 tasks)
None (Fine-tuning)	19.96 ($\pm$ 0.01)	4.28 ($\pm$ 0.38)
Joint Training	97.72 ($\pm$ 0.04)	65.28 ($\pm$ 0.15)
EWC-SNN	19.89 ($\pm$ 0.00)	3.93 ($\pm$ 0.24)
MAS-SNN	19.91 ($\pm$ 0.01)	3.09 ($\pm$ 0.45)
SA-SNN(rate)	50.22 ($\pm$ 0.91)	21.37 ($\pm$ 0.77)
SA-SNN	60.06 ($\pm$ 2.16)	22.86 ($\pm$ 0.64)
Baseline (k-WTA+EWC)	83.32 ($\pm$ 1.12)	36.47 ($\pm$ 2.13)
STAR (Ours)	**89.35** ($\pm$ **0.54**)	**38.52** ($\pm$ **1.75**)

that STAR's Sleep Consolidation expert freezing mechanism effectively preserves the knowledge representations of old tasks, shielding them from interference caused by parameter updates during subsequent task training.

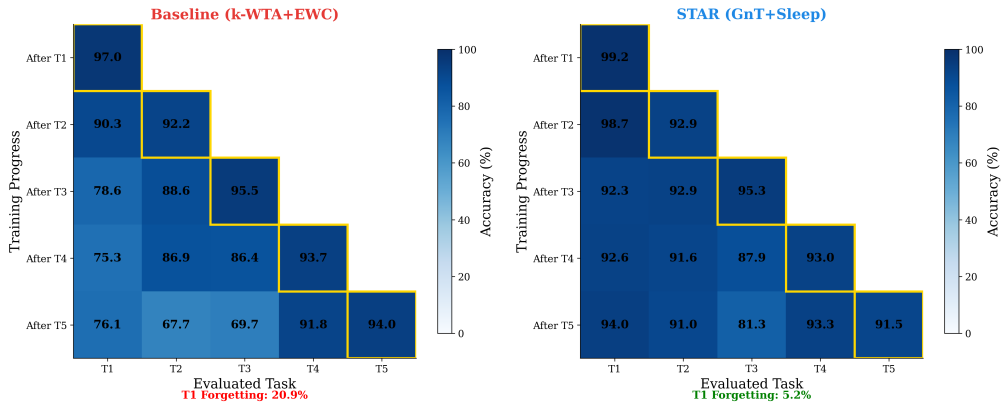

Figure 2: Task accuracy matrices and Task-1 forgetting across the training progress (After T1–T5). Left: baseline; right: STAR.

4.3 Ablation Study

To validate the independent contributions of GnT and Sleep Consolidation, we conduct a component ablation study (Table 2). The experiment compares three variants: (1) GnT only (without Sleep Consolidation); (2) Sleep Consolidation only (without GnT); and (3) full STAR (GnT + Sleep Consolidation).

The core contribution of GnT lies in providing a sustained source of plasticity — by identifying and replacing dead neurons, it ensures that the network always has sufficient active neurons to learn new knowledge. This is directly reflected in the leap in neuron utilization from below 40% to 61%. The core contribution of Sleep Consolidation lies in providing an anti-forgetting protection mechanism — through expert freezing and redundancy reset, old knowledge is physically locked in place, while new tasks are given clean, free capacity. The separation of the two fundamentally reduces inter-task interference.

Table 3: Comparison of Forgetting and Utilization on continual learning benchmarks

Methods	Forgetting (%) $\downarrow$			Utilization (%) $\uparrow$		
	Split-MNIST	CIFAR-100	TinyImageNet	Split-MNIST	CIFAR-100	TinyImageNet
Baseline (k-WTA+EWC)	14.67 ± 3.35	19.35 ± 2.41	9.73 ± 0.43	9.4 ± 0.80	33.4 ± 1.80	52.3 ± 0.40
STAR (Ours)	$\mathbf{5.06 \pm 1.57}$	$\mathbf{9.12 \pm 1.35}$	$\mathbf{7.73 \pm 1.13}$	$\mathbf{61.6 \pm 0.50}$	$\mathbf{57.8 \pm 2.10}$	$\mathbf{62.5 \pm 2.40}$

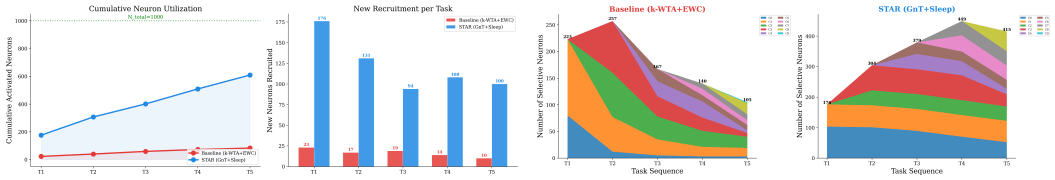

Figure 3: Plasticity dynamics on Split-MNIST. (a) Cumulative activated neurons and newly recruited neurons per task. (b) Task-selective neurons across the task sequence. In both panels, left is the baseline and right is STAR.

Due to space limits, we defer the complete neurodynamic analysis to Appendix A. Here we summarize the four key observations.

Table 2: Component ablation study on Split-MNIST

Variant	Average Accuracy (%)
Baseline (k-WTA+EWC)	83.32 ± 1.12
+ GnT only	84.64 ± 3.25
+ Sleep only	84.09 ± 0.60
+ Full STAR (GnT + Sleep)	$\mathbf{89.35 \pm 0.54}$

Neuron Pool Expansion (Fig. 3(a)): The left panel depicts the cumulative number of activated neurons as tasks progress, while the right panel shows the number of newly recruited neurons engaged by each new task. In the baseline, the cumulative number of activated neurons grows extremely slowly, where the majority of neurons remain in a permanently dead state, resulting in extremely low network capacity utilization. In STAR, by contrast, the cumulative number of activated neurons grows monotonically with the number of tasks, with each new task stably recruiting approximately 100 new neurons for representation learning. This phenomenon is directly attributable to the GnT dead neuron replacement mechanism: dead neurons from old tasks, after mutation cloning, participate in learning new tasks with a completely fresh initial state, thereby ensuring that the network always possesses ample plastic capacity.

Neuron Selectivity Growth (Fig. 3(b)): The stacked area plot illustrates how the number of task-specific selective neurons evolves over the course of training. In the baseline (left panel), neuron selectivity is disordered: training on new tasks causes the area bands of old-task-specific neurons to narrow or even be squeezed out, indicating that the representation space is encroached upon by new tasks. STAR (right panel), by contrast, exhibits a near-perfect monotonically increasing structure: each task acquires a stably growing and inviolable dedicated subset of neurons, with the area bands of different tasks neither overlapping nor interfering with one another. This validates, from the selectivity perspective, the expert freezing strategy of the Sleep Consolidation mechanism: once a neuron is confirmed as an expert for a given task, its weights are locked and cannot be modified by the training of subsequent tasks, thereby ensuring the independence of task-specific representations.

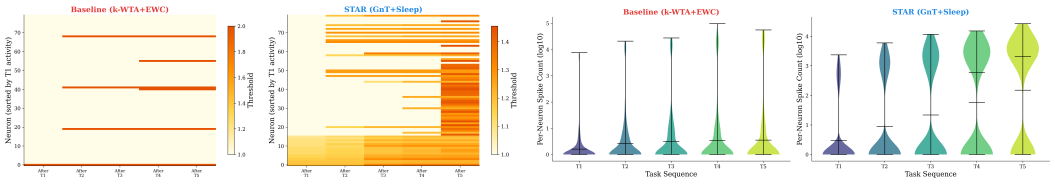

Figure 4: Neural dynamics on Split-MNIST. (a) Per-neuron firing-threshold heatmaps. (b) Per-neuron spike-count distributions ($\log_{10}$ scale). Left: baseline; right: STAR.

Adaptive Threshold Evolution (Fig. 4(a)): The heatmap shows how neuronal firing thresholds evolve across tasks, providing insight into how STAR preserves previously learned knowledge. In the baseline (left panel), the thresholds of nearly all neurons remain at the minimum value of 1.0, showing little adaptation to their activation history. In STAR (right panel), the thresholds of neurons identified as experts increase from 1.0 toward the upper bound of 2.0 after each task. These higher thresholds make expert neurons less likely to activate during subsequent tasks, helping to limit interference from new data. This behavior is consistent with the broader idea of homeostatic regulation associ-

ated with the Synaptic Homeostasis Hypothesis (SHY), although firing-threshold adaptation differs from synaptic downscaling. Meanwhile, non-expert neurons retain low thresholds, preserving their responsiveness and the network's plasticity for learning new tasks.

Spike Distribution Shift (Fig. 4(b)): The violin plots on a logarithmic scale depict the distribution of per-neuron spike counts after each task is completed. In the baseline (left panel), spike firing is extremely impoverished: the spike counts of a large number of neurons are concentrated near zero ($\log_{10} \approx 0$), with only a very small number of neurons exhibiting weak firing activity. This "spike impoverishment" prevents effective feedforward propagation of information through the SNN and leads to severe attenuation of gradient signals. STAR (right panel), by contrast, exhibits a markedly healthier and more active spike distribution: the spike counts span a wider dynamic range, and the median is visibly shifted upward. This confirms that the GnT replacement strategy successfully cultivates more active neurons, safeguarding the flow of information within the SNN.

Weight Distribution Evolution (Fig. 5): The distribution histograms illustrate how the input weight distribution $\mathbf{W}_1$ evolves across different task stages. In the baseline (left panel), the weight distribution remains persistently concentrated within a narrow interval near zero — the network fails to develop meaningfully differentiated responses across feature dimensions. STAR (right panel) causes the weight distribution to gradually become broader and more diverse, exhibiting a more discriminative distributional profile. This is attributable to the combined effect of two factors: (1) GnT enriches the initialization patterns of weights by cloning high-performing neurons and injecting mutation noise; and (2) Sleep Consolidation preserves high-quality, already-differentiated weights through expert freezing, preventing them from being smoothed out by the averaging effect of gradient updates during subsequent training.

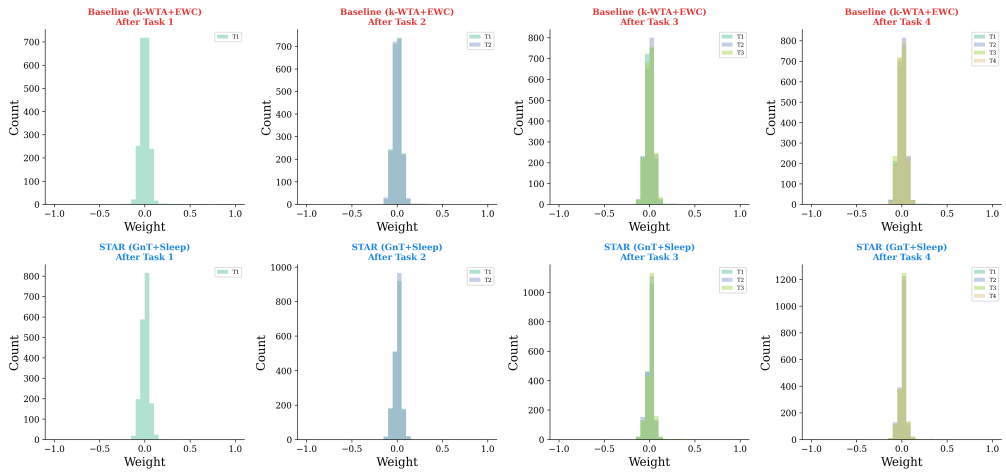

Figure 5: Evolution of the input weight distribution $\mathbf{W}_1$ across tasks. Top: baseline; bottom: STAR.

5 Conclusion

This paper introduced STAR, a spike-driven framework for dynamic neuron resource allocation in rehearsal-free continual SNN learning. Under a fixed network capacity, GnT renews plasticity by replacing mature low-utility neurons, while Sleep Consolidation preserves knowledge by protecting task-relevant experts and releasing redundant neurons. Together, they establish a generate–select–consolidate–release lifecycle that addresses the stability–plasticity dilemma. Experiments on Split-MNIST, CIFAR-100, and TinyImageNet demonstrate consistent improvements in accuracy, memory retention, and neuron utilization. In particular, STAR improves average accuracy from 83.32% to 89.35% and neuron utilization from 9.4% to 61.6% on Split-MNIST. These results demonstrate the effectiveness of spike-driven neuron lifecycle management for resource-efficient continual learning.

AI USE STATEMENT

The authors used generative AI tools to assist with manuscript drafting, paraphrasing, and language editing. These tools were not used to generate experimental data, formulate mathematical claims, develop proofs, or design the methodology or experiments. All AI-assisted text was reviewed and revised by the authors, who take responsibility for the final manuscript and its claims.

ETHICS STATEMENT

This work studies continual learning in spiking neural networks through selective neuron replacement and consolidation under a fixed network capacity. Experiments use the publicly available MNIST, CIFAR-100, and TinyImageNet datasets and do not involve human participants or newly collected personal data. Although the method does not retain historical samples for replay, this property alone does not guarantee privacy. Potential deployment in adaptive systems should account for classification errors, forgetting, and performance degradation under distribution shifts.

REPRODUCIBILITY STATEMENT

Section 3 describes the network architecture, neuron utility measures, Generate-and-Test mechanism, and task-boundary consolidation procedure. Algorithm 1 summarizes the training pipeline. Section 4.1 reports the datasets, task splits, network size, simulation time window, optimizer, learning rate, batch size, and principal hyperparameters for neuron replacement and EWC regularization. Section 4.3 presents component ablations, and Section 4.4 examines sensitivity to the replacement rate. Experiments are conducted on NVIDIA 4090 GPUs.

REFERENCES

R. Aljundi, F. Babiloni, M. Elhoseiny, et al. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 139–154, 2018.

G. Bellec, D. Salaj, A. Subramoney, et al. Long short-term memory and learning-to-learn in networks of spiking neurons. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

G. Buzsáki. Neural syntax: cell assemblies, synapsembles, and readers. *Neuron*, 68(3):362–385, 2010.

A. Chaudhry, M. Rohrbach, M. Elhoseiny, et al. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

C. D. Clelland, M. Choi, C. Romberg, et al. A functional role for adult hippocampal neurogenesis in spatial pattern separation. *Science*, 325(5937):210–213, 2009.

M. Davies, N. Srinivasa, T. H. Lin, et al. Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro*, 38(1):82–99, 2018.

P. Dayan and L. F. Abbott. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. MIT Press, 2005.

M. De Lange, R. Aljundi, M. Masana, et al. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7): 3366–3385, 2021.

S. Diekelmann and J. Born. The memory function of sleep. *Nature Reviews Neuroscience*, 11(2): 114–126, 2010.

T. J. Draelos, N. E. Miner, C. C. Lamb, et al. Neurogenesis deep learning: Extending deep networks to accommodate new classes. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 526–533. IEEE, 2017.

P. S. Eriksson, E. Perfilieva, T. Björk-Eriksson, et al. Neurogenesis in the adult human hippocampus. *Nature Medicine*, 4(11):1313–1317, 1998.

R. M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3 (4):128–135, 1999.

W. Gerstner and W. M. Kistler. *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge University Press, 2002.

R. Golden, J. E. Delanois, P. Sanda, et al. Sleep prevents catastrophic forgetting in spiking neural networks by forming a joint synaptic weight representation. *PLoS Computational Biology*, 18 (11):e1010628, 2022.

S. Grossberg. Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural Networks*, 1(1):17–61, 1988.

J. Kirkpatrick, R. Pascanu, N. Rabinowitz, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

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

M. Masana, T. Tuytelaars, and J. Van de Weijer. Ternary feature masks: zero-forgetting for task-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3570–3579, 2021.

M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pp. 109–165. Academic Press, 1989.

M. Mozafari, S. R. Kheradpisheh, T. Masquelier, et al. First-spike-based visual categorization using reward-modulated STDP. *IEEE Transactions on Neural Networks and Learning Systems*, 29(12): 6178–6190, 2018.

E. O. Neftci, H. Mostafa, and F. Zenke. Surrogate gradient learning in spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63, 2019.

J. O'Keefe and L. Nadel. *The Hippocampus as a Cognitive Map*. Oxford University Press, 1978.

A. Ororbia and D. Kifer. The neural coding framework for learning generative models. *Nature Communications*, 13(1):2064, 2022.

G. I. Parisi, R. Kemker, J. L. Part, et al. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

D. Rolnick, A. Ahuja, J. Schwarz, et al. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

K. Roy, A. Jaiswal, and P. Panda. Towards spike-based machine intelligence with neuromorphic computing. *Nature*, 575(7784):607–617, 2019.

A. A. Rusu, N. C. Rabinowitz, G. Desjardins, et al. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

A. Safa, F. Corradi, L. Keuninckx, et al. Improving the accuracy of spiking neural networks for radar gesture recognition through preprocessing. *IEEE Transactions on Neural Networks and Learning Systems*, 34(6):2869–2881, 2021.

J. Shen, W. Ni, Q. Xu, et al. Efficient spiking neural networks with sparse selective activation for continual learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 611–619, 2024.

H. Shin, J. K. Lee, J. Kim, et al. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

N. Soures, P. Helfer, A. Daram, et al. TACOS: Task agnostic continual learning in spiking neural networks. *arXiv preprint arXiv:2409.00021*, 2024.

G. Tononi and C. Cirelli. Sleep and the price of plasticity: from synaptic and cellular homeostasis to memory consolidation and integration. *Neuron*, 81(1):12–34, 2014.

M. P. Walker and R. Stickgold. Sleep-dependent learning and memory consolidation. *Neuron*, 44 (1):121–133, 2004.

J. Yoon, E. Yang, J. Lee, et al. Lifelong learning with dynamically expandable networks. *arXiv preprint arXiv:1708.01547*, 2017.

F. Zenke, B. Poole, and S. Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pp. 3987–3995. PMLR, 2017.

A Appendix

A.1 Sensitivity Analysis

Fig. 6 illustrates the effect of the base replacement rate R_{rep} on average accuracy, with the horizontal axis shown on a logarithmic scale. The highest accuracy is achieved at $R_{\text{rep}} = 0.01$, corresponding to a base replacement count of $N_{\text{base}} = 20$ out of 2000 neurons before task annealing and candidate screening. The choice of replacement rate balances plasticity renewal against representation preservation: too few replacements may limit the recruitment of new neurons for learning subsequent tasks, whereas too many may disrupt useful representations and increase forgetting. Task annealing reduces the replacement count in later tasks, but a large base rate can still result in excessive replacement. Within the tested range, accuracy changes only modestly between 0.001 and 0.01, while larger rates lead to a more pronounced decline. These results suggest that GnT tolerates moderate reductions from the selected rate but is more sensitive to excessive replacement. We therefore set $R_{\text{rep}} = 0.01$ in our experiments.

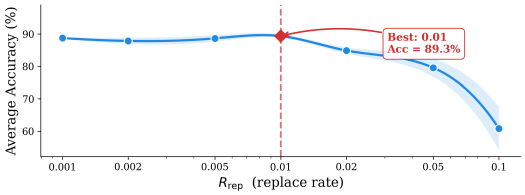

Figure 6: Average accuracy vs. the base replacement rate R_{rep} (log scale), peaking at $R_{\text{rep}} = 0.01$.