

Supplementary Information

Homeostatic Dendritic Neuron Based on Co-Integrated Volatile and Non-Volatile Memristors for Neuromorphic Processing

Licheng Zhang^{1,2}, Teng Zhang³, Pek Jun Tiw³, Xulei Wu¹, Yuqi Su^{1,2} and Yuchao Yang^{1,2,3,4*}

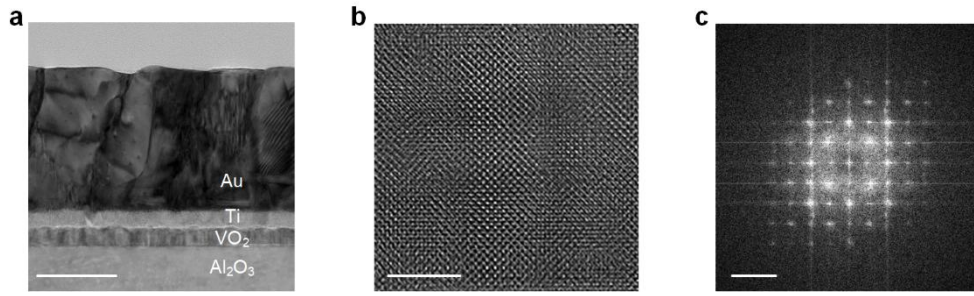
¹New Cornerstone Science Laboratory, Guangdong Provincial Key Laboratory of In-Memory Computing Chips, School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, Shenzhen, China.

²Peng Cheng Laboratory, Shenzhen, China.

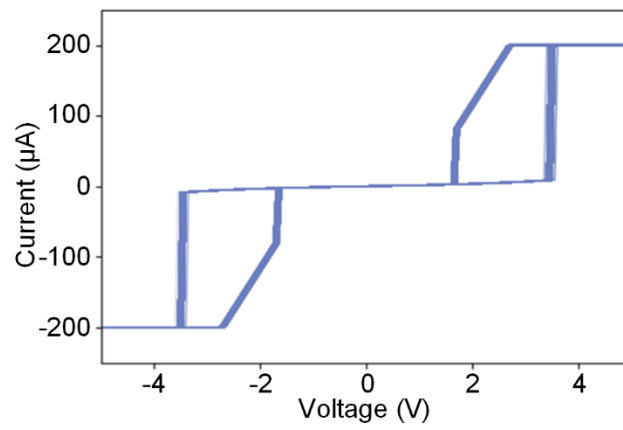
³New Cornerstone Science Laboratory, Beijing Advanced Innovation Center for Integrated Circuits, School of Integrated Circuits, Peking University, Beijing, China.

⁴Center for Brain Inspired Intelligence, Chinese Institute for Brain Research (CIBR), Beijing, China.

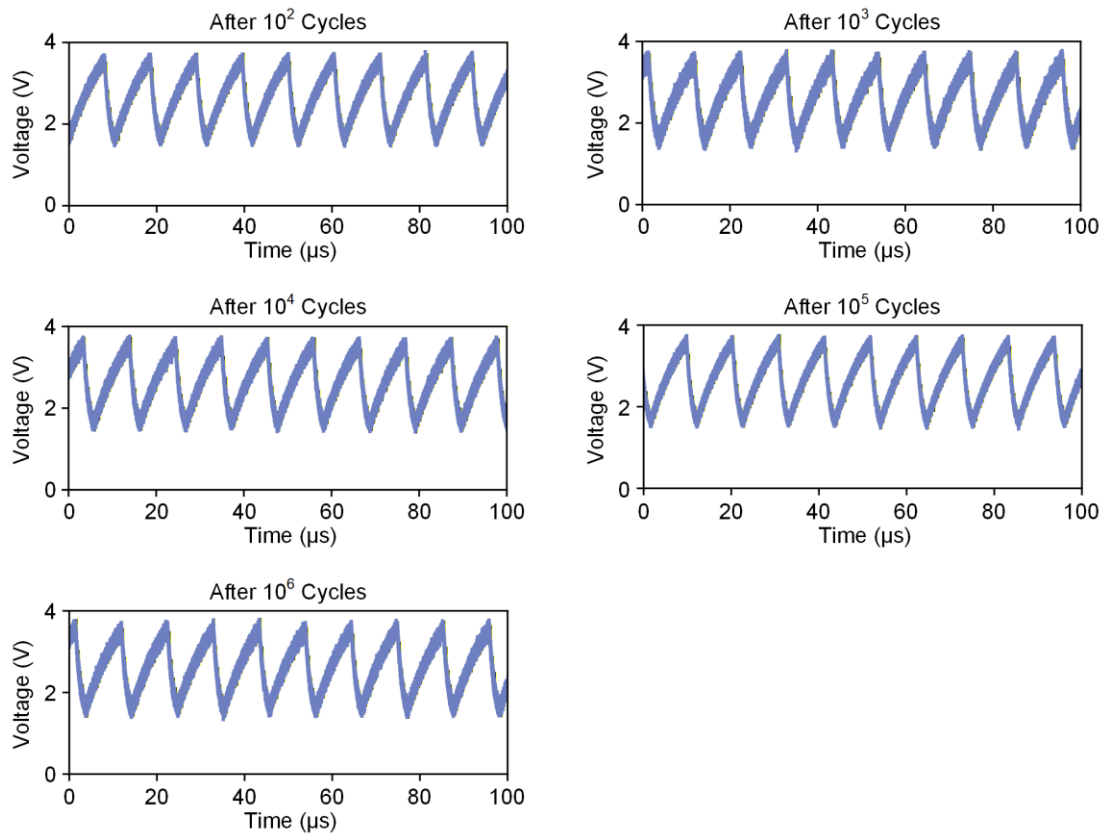
*Corresponding author: yuchaoyang@pku.edu.cn



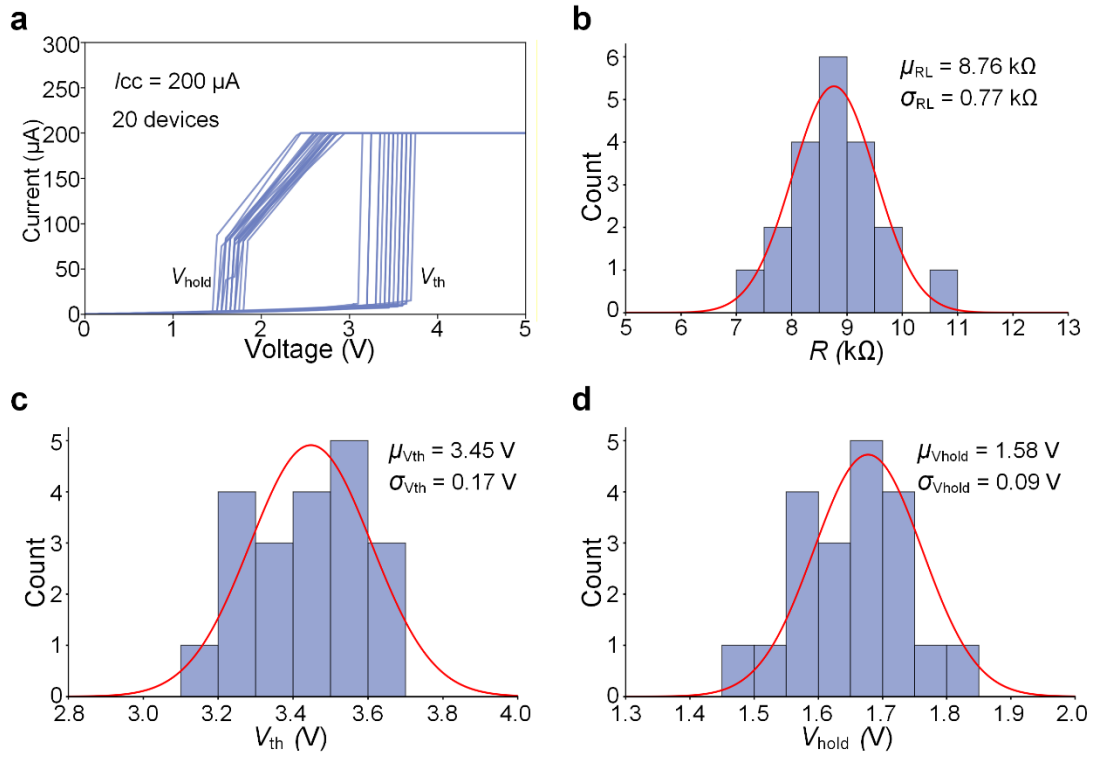
Supplementary Fig. 1 Characterization of the VO₂ device material stack. (a) Cross-sectional transmission electron microscopy (TEM) image of the device and the scale bar represents 50 nm. (b) Zoomed-in high-resolution TEM (HRTEM) image of the VO₂ thin film and the scale bar represents 40 nm. (c) Corresponding fast Fourier transform diffraction pattern of the VO₂ thin film and the scale bar represents 5 nm⁻¹.



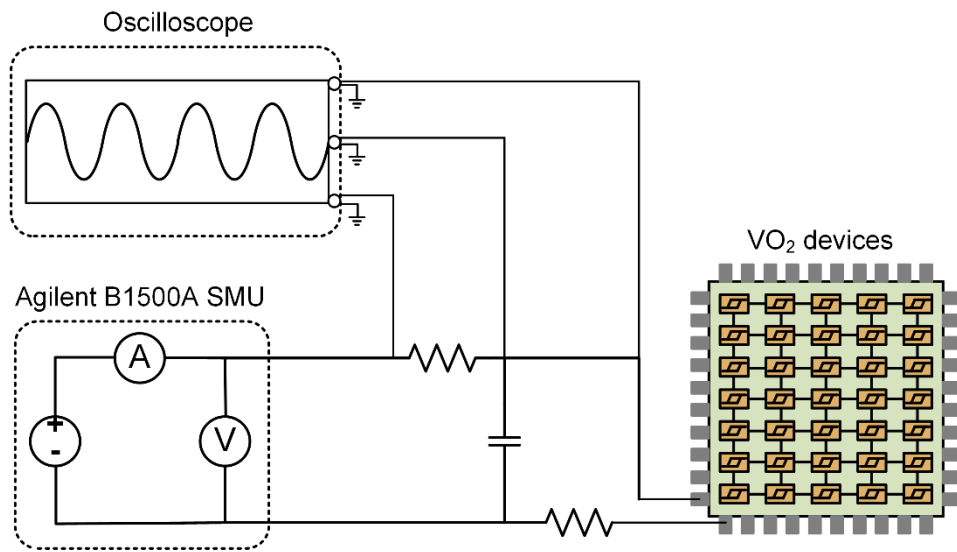
Supplementary Fig. 2 The I - V curves over 500 cycles under a compliance current of $\pm 200 \mu\text{A}$ demonstrate that the device exhibits consistent threshold and holding voltages under both positive and negative bias.



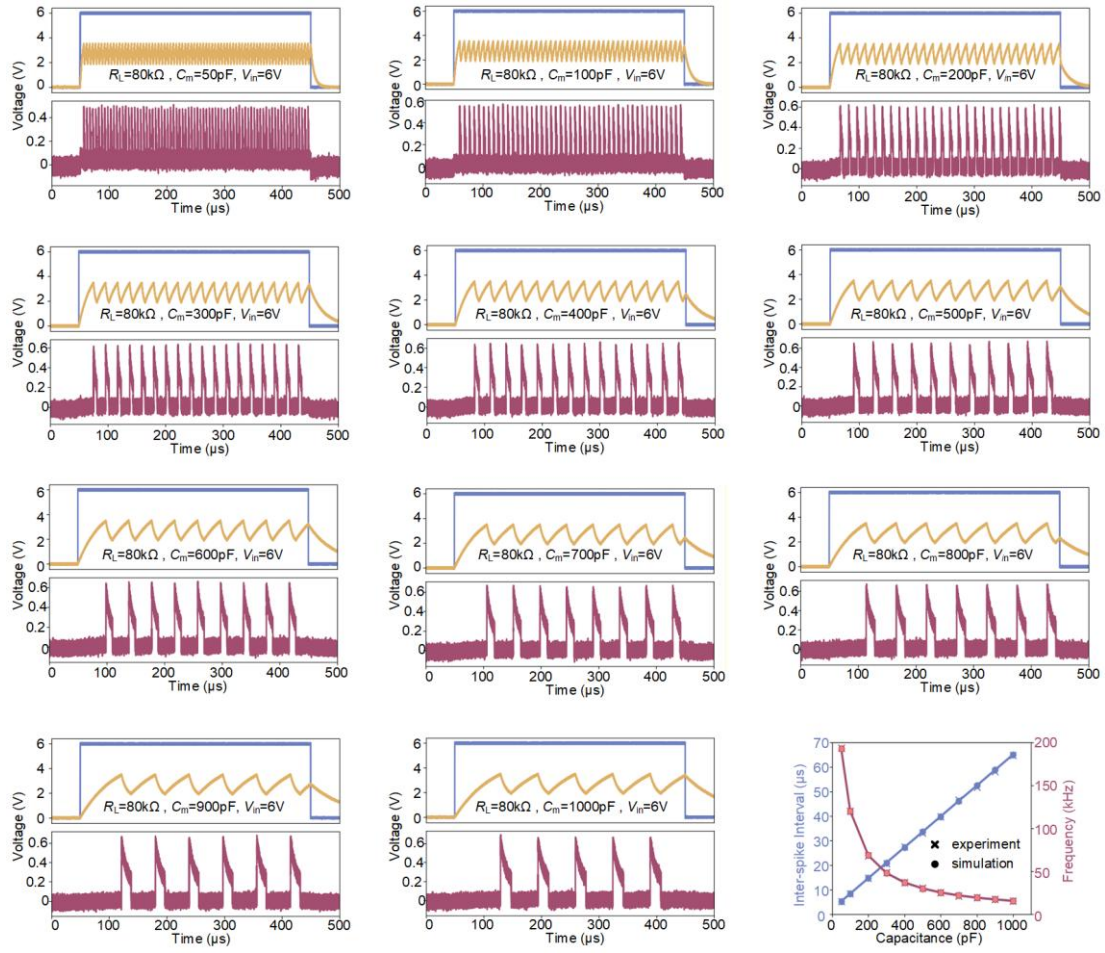
Supplementary Fig. 3 Endurance of the VO₂. Showing stable oscillating behavior of the VO₂ memristor circuit after 10², 10³, 10⁴, 10⁵ and 10⁶ switching cycles.



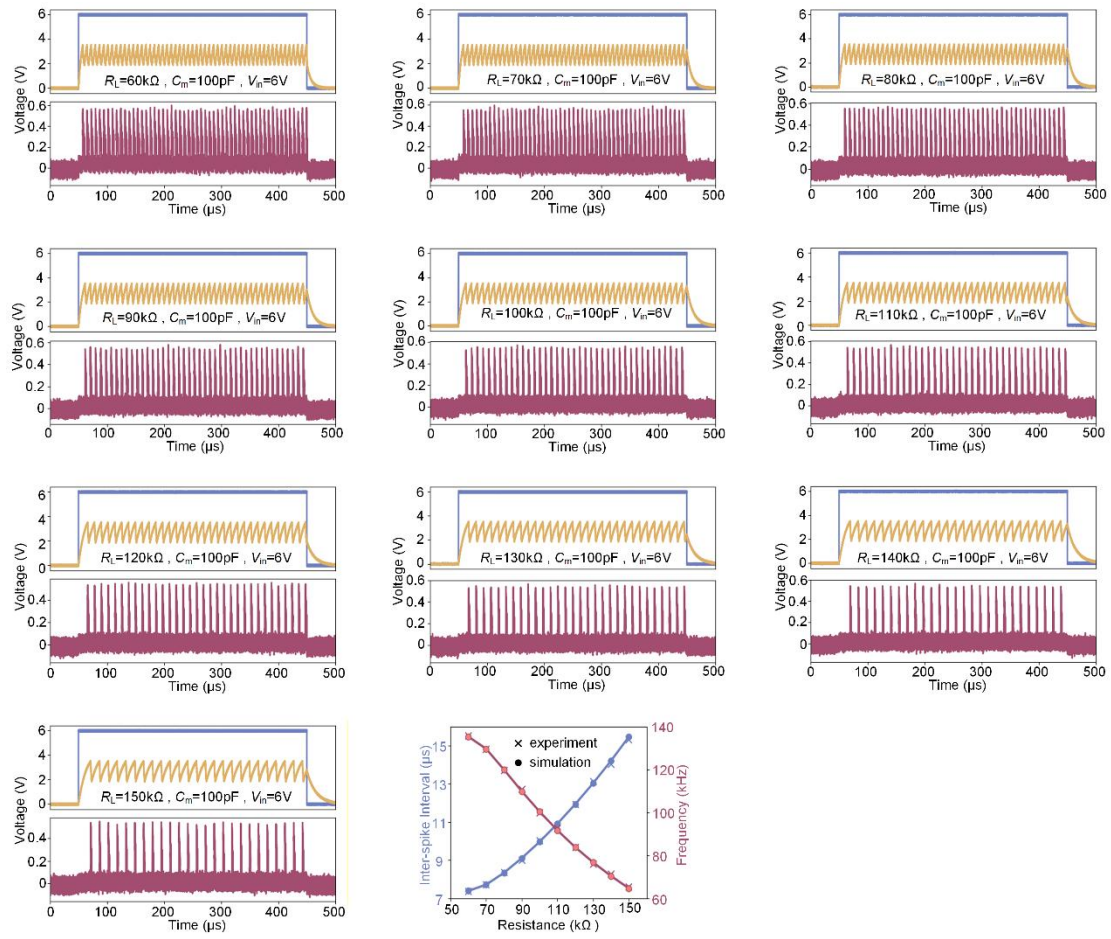
Supplementary Fig. 4. Device-to-device variation of V_{th} , V_{hold} and the low resistance state (R_{VO2}) of the VO₂ memristors, measured across 20 devices. (a) I - V sweeps of 20 devices superposed to illustrate the decent uniformity. (b) Histogram of R_{VO2} . The C_v is 8.79%. (c) Histogram of V_{th} . The C_v is 4.93%. (d) Histogram of V_{hold} . The C_v is 5.70%.



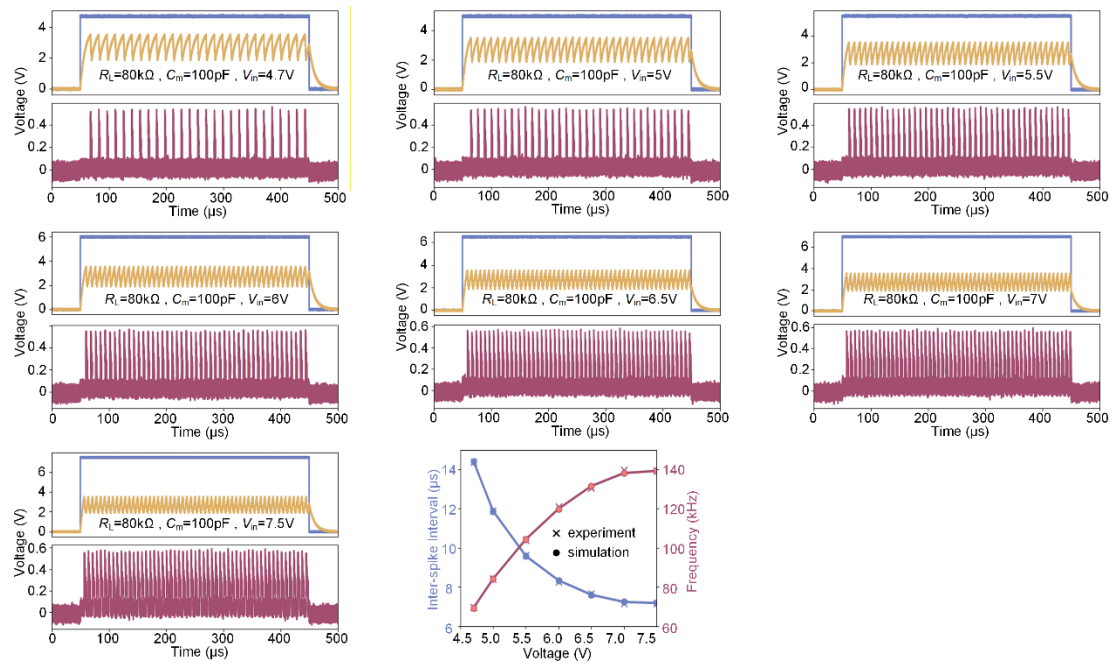
Supplementary Fig. 5 Experimental setup to measure the behavior of the LIF neuron circuit. The VO₂ device is first packaged in a QFP80 and soldered onto a custom-designed PCB, which integrates peripheral circuit components such as resistors and capacitors. The assembled system is then tested using an Agilent B1500A source measurement unit and an oscilloscope.



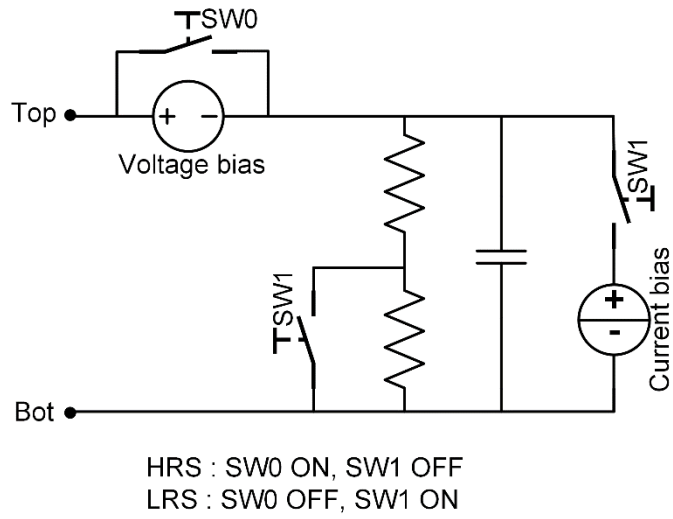
Supplementary Fig. 6 The response of the artificial neuron with different external capacitances (50 pF, 100 pF, 200 pF, 300 pF, 400 pF, 500 pF, 600 pF, 700 pF, 800 pF, 900 pF and 1000 pF) while maintaining a constant input voltage of 6 V and a load resistance of 80 k Ω . As the capacitance increases, the spiking frequency decreases. The results obtained from SPICE simulations using the constructed model show good agreement with the measured data.



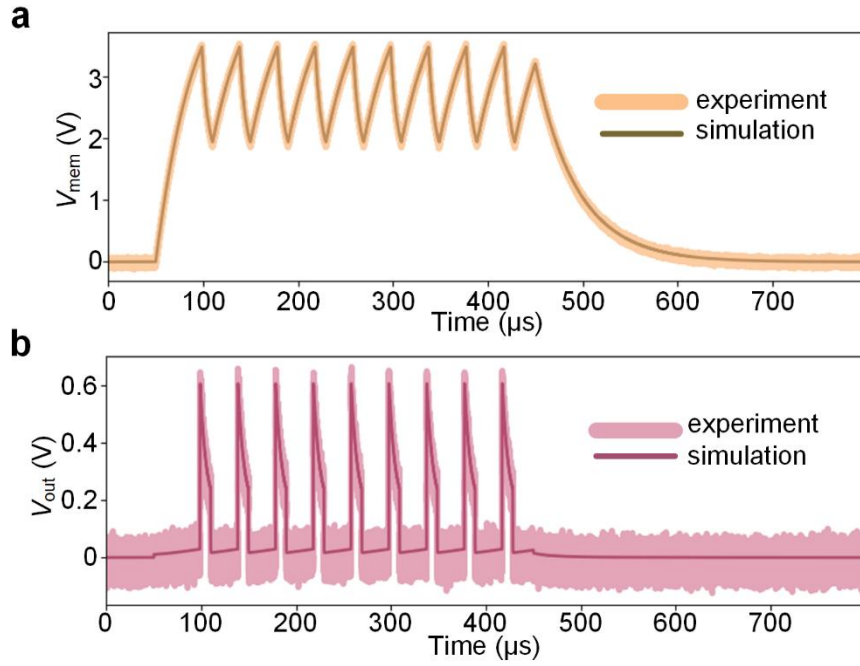
Supplementary Fig. 7 The response of the artificial neuron with different load resistances (60 kΩ, 70 kΩ, 80 kΩ, 90 kΩ, 100 kΩ, 110 kΩ, 120 kΩ, 130 kΩ, 140 kΩ and 150 kΩ) while maintaining a constant input voltage of 6 V and an external capacitance of 100 pF. As the load resistance increases, the charging process slows down and the spiking frequency decreases. The results obtained from SPICE simulations using the constructed model show good agreement with the measured data.



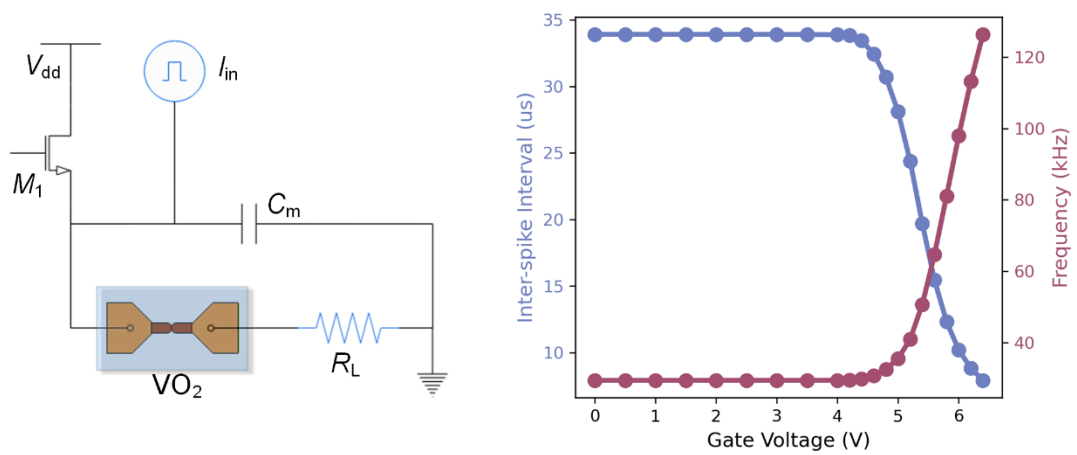
Supplementary Fig. 8 The response of the artificial neuron with voltage (4.7 V, 5 V, 5.5 V, 6 V, 6.5 V, 7 V and 7.5 V) while maintaining an external capacitance of 100 pF and a load resistance of 80 kΩ. As the voltage decreases, the charging process slows down and the spiking frequency decreases. The results obtained from SPICE simulations using the constructed model show good agreement with the measured data.



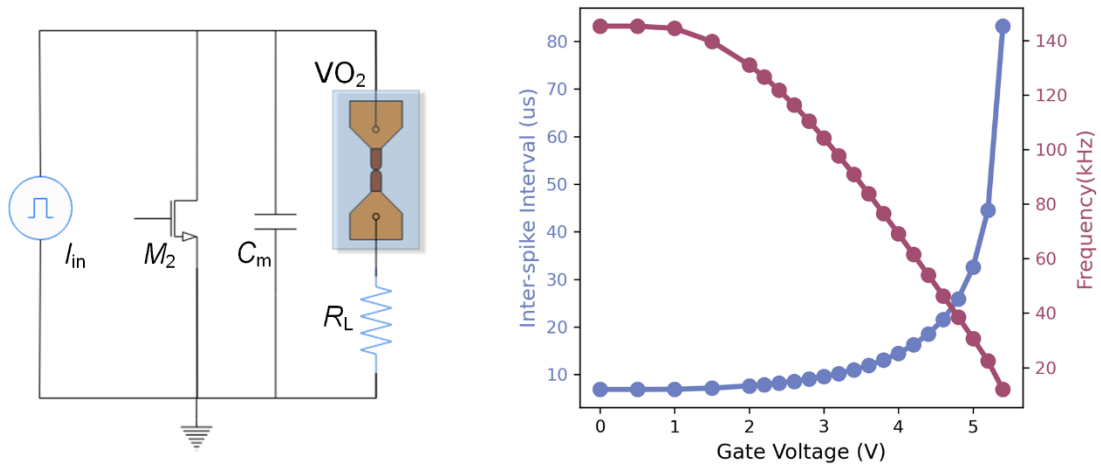
Supplementary Fig. 9 The SPICE model of VO₂ memristor. Initially, SW0 is closed, and SW1 is open, resulting in a series connection of a small resistor and a large voltage-dependent resistor, which is then connected in parallel with a capacitor. Under these conditions, the memristor remains in a high-resistance state. When the voltage across the memristor exceeds the threshold voltage, SW0 opens and SW1 closes, activating a bias voltage source and a bias current source in the model. It also effectively short-circuits the large resistor, causing the memristor to transition to a low-resistance state. As the voltage across the memristor decreases gradually while it remains in the low-resistance state, the voltage eventually falls below V_{hold} . At this point, SW0 closes, SW1 opens, and the memristor reverts to its initial high-resistance state.



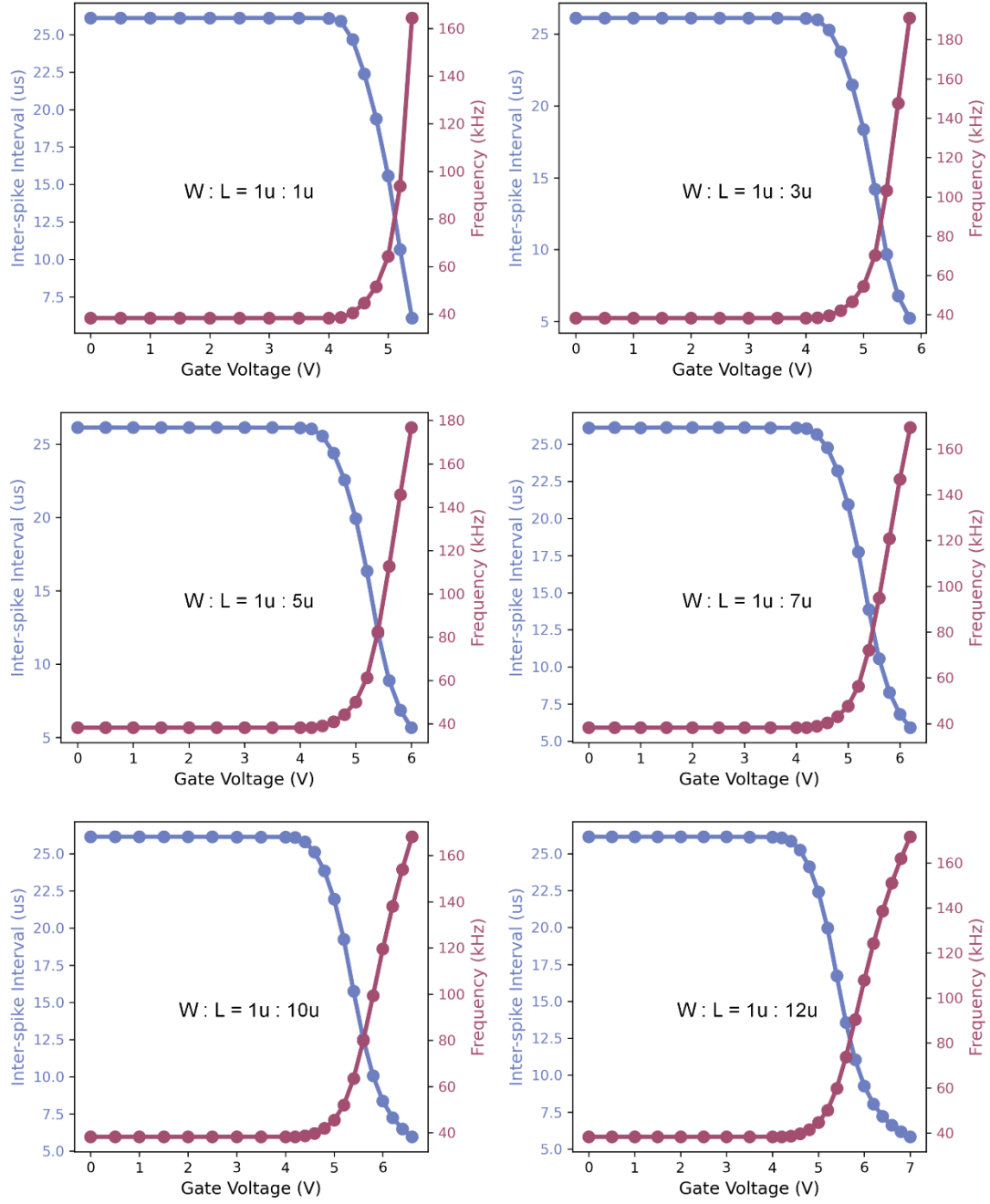
Supplementary Fig. 10 For the LIF neuron with a membrane capacitance of $C_m=600$ pF and load resistance $R_L=80$ k Ω , a voltage input of 6 V with a pulse width of 400 μs is applied to test the membrane potential and spike output. In addition, simulations are performed using the Verilog-A model on the HSPICE platform. The simulation results demonstrate good agreement with the experimental data, further confirming the model's validity. (a) Membrane potential dynamics of the neuron under both simulation and experimental measurement conditions. (b) Spike firing results of the neuron obtained from both simulation and experimental measurements.



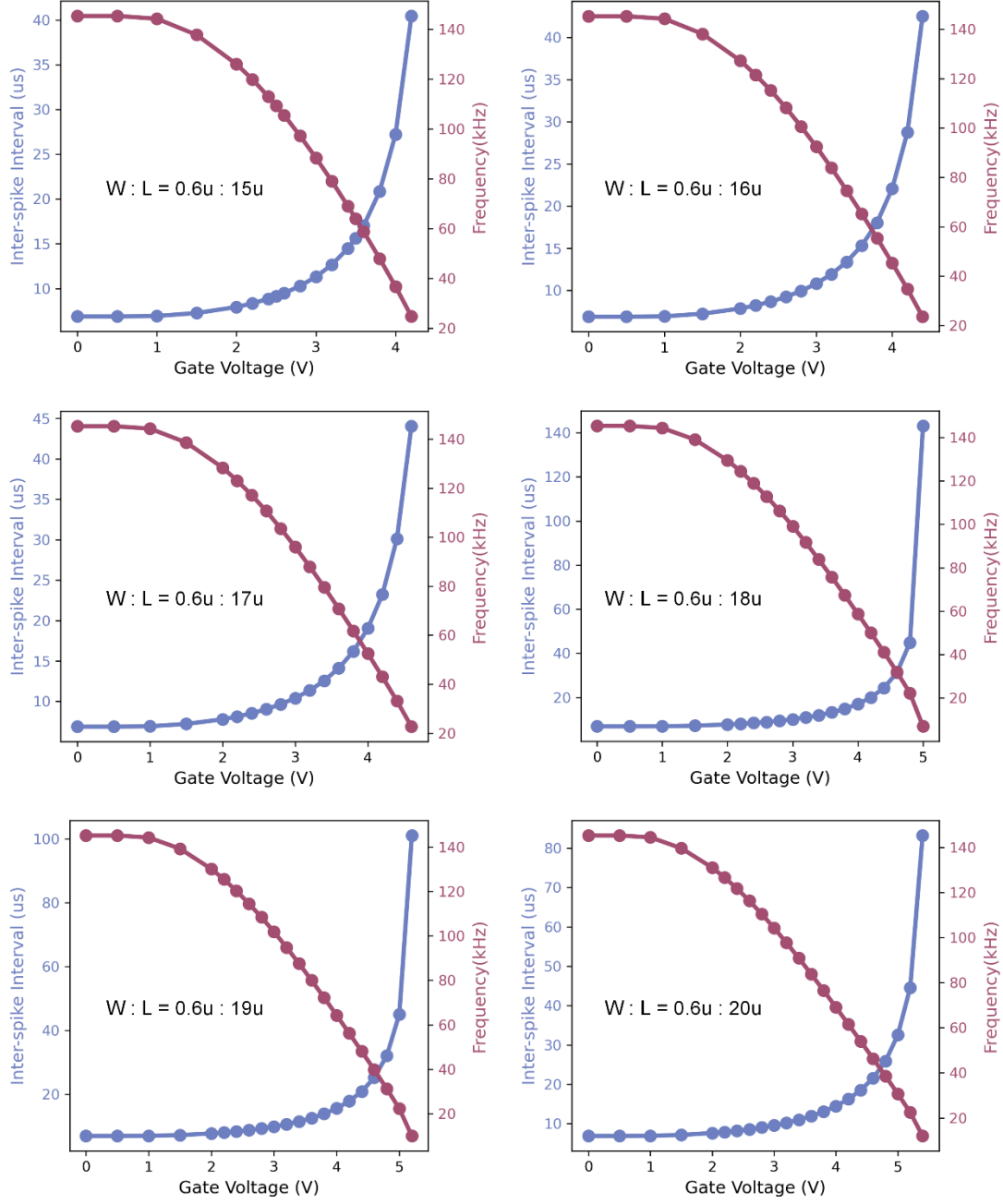
Supplementary Fig. 11 We introduced transistor M_1 , connected to the voltage source V_{dd} , into the existing current-input neuron model with a small input current of 13 μA . By varying the gate voltage applied to M_1 , we tested the neuron's spike output frequency under different gate voltages to analyze the role of the M_1 transistor. When the gate voltage exceeds 4 V, M_1 becomes active, supplementing the input current and thereby increasing the spike output frequency. As the gate voltage continues to rise, the neuron's spike frequency increases and the inter-spike interval decreases. This behavior is primarily due to the enhanced gate voltage boosting the supplementary current provided by V_{dd} through M_1 , which effectively raises the neuron's total input current.



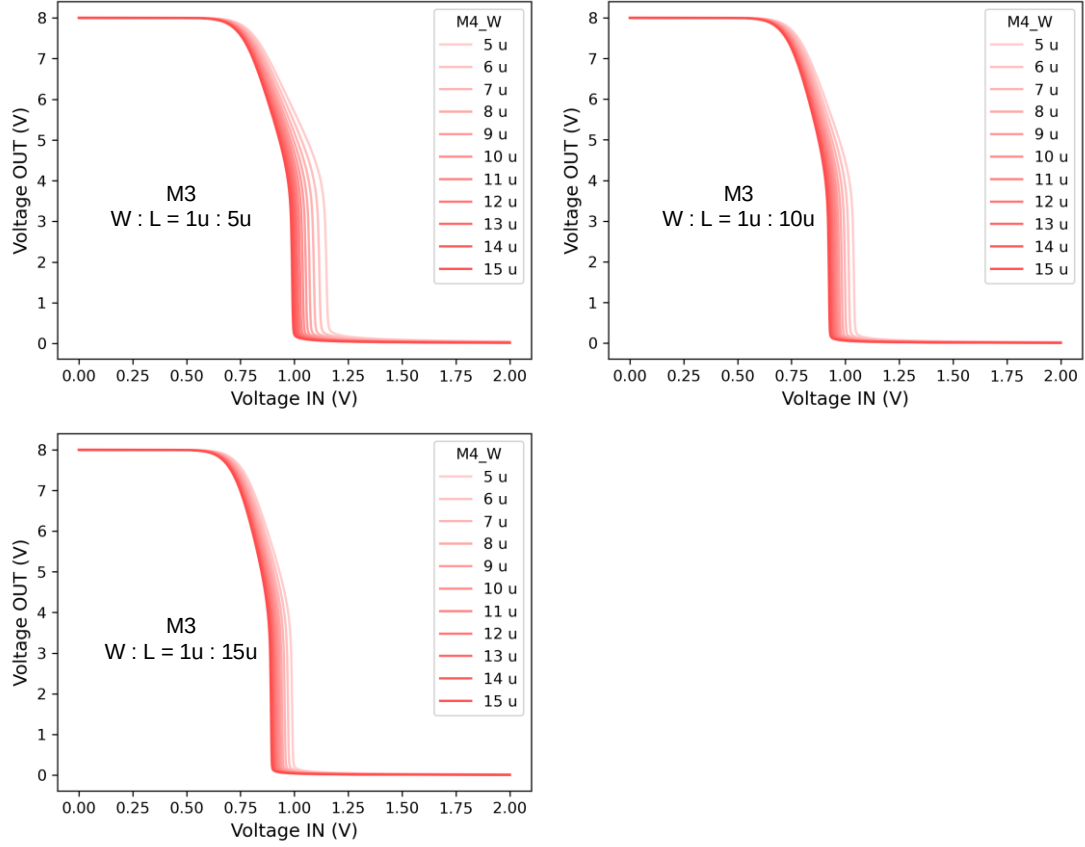
Supplementary Fig. 12 We integrated transistor M_2 into the existing current-input neuron model with a large input current of $40 \mu\text{A}$. By varying the gate voltage applied to M_2 , we tested the neuron's spike output frequency under different gate voltages to analyze the role of the M_2 transistor. When the gate voltage exceeds 1 V, M_2 becomes active, diverting part of the input current and thereby reducing the neuron's spike output frequency. As the gate voltage continues to rise, the spike frequency decreases and the inter-spike interval lengthens. This behavior arises because the elevated gate voltage increases the current flowing through M_2 , resulting in greater leakage of the input current and effectively lowering the neuron's total input current. To achieve oscillation frequency modulation and restore stability, the width-to-length ratio of M_2 must be smaller than that of M_1 .



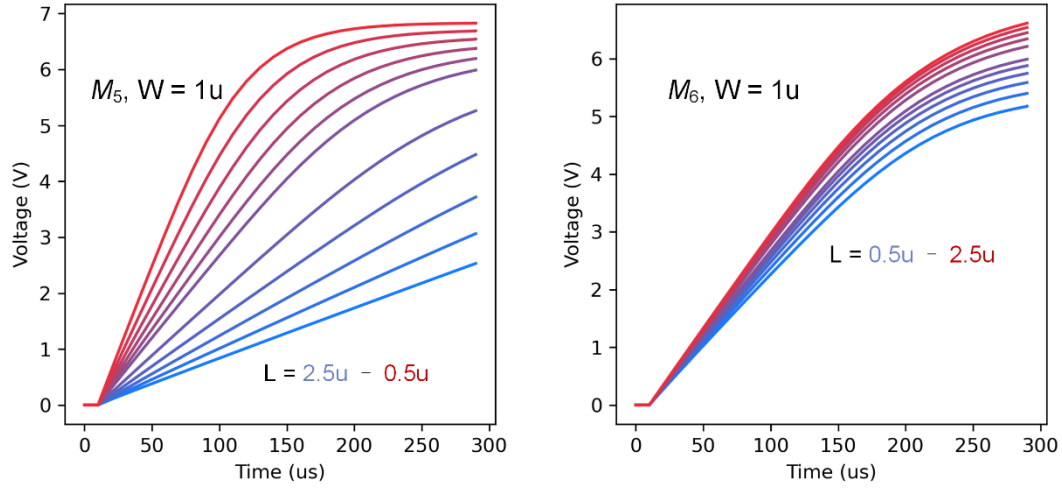
Supplementary Fig. 13 We conducted a more in-depth analysis of the modulation effect of the M_1 transistor's gate voltage on neuronal spike frequency under varying width-to-length (W/L) ratios. The observed variation in the maximum spike frequencies across different configurations is primarily attributed to the finite resolution of gate voltage increments, which constrained our ability to precisely capture the peak spiking frequency. Comparative results demonstrate that, for neurons incorporating M_1 transistors with larger W/L ratios, a lower gate voltage is sufficient to elicit a given spike frequency. This suggests that increasing the W/L ratio enhances the transistor's current-driving capability, thereby enabling more efficient modulation of the effective input current and neuronal firing dynamics.



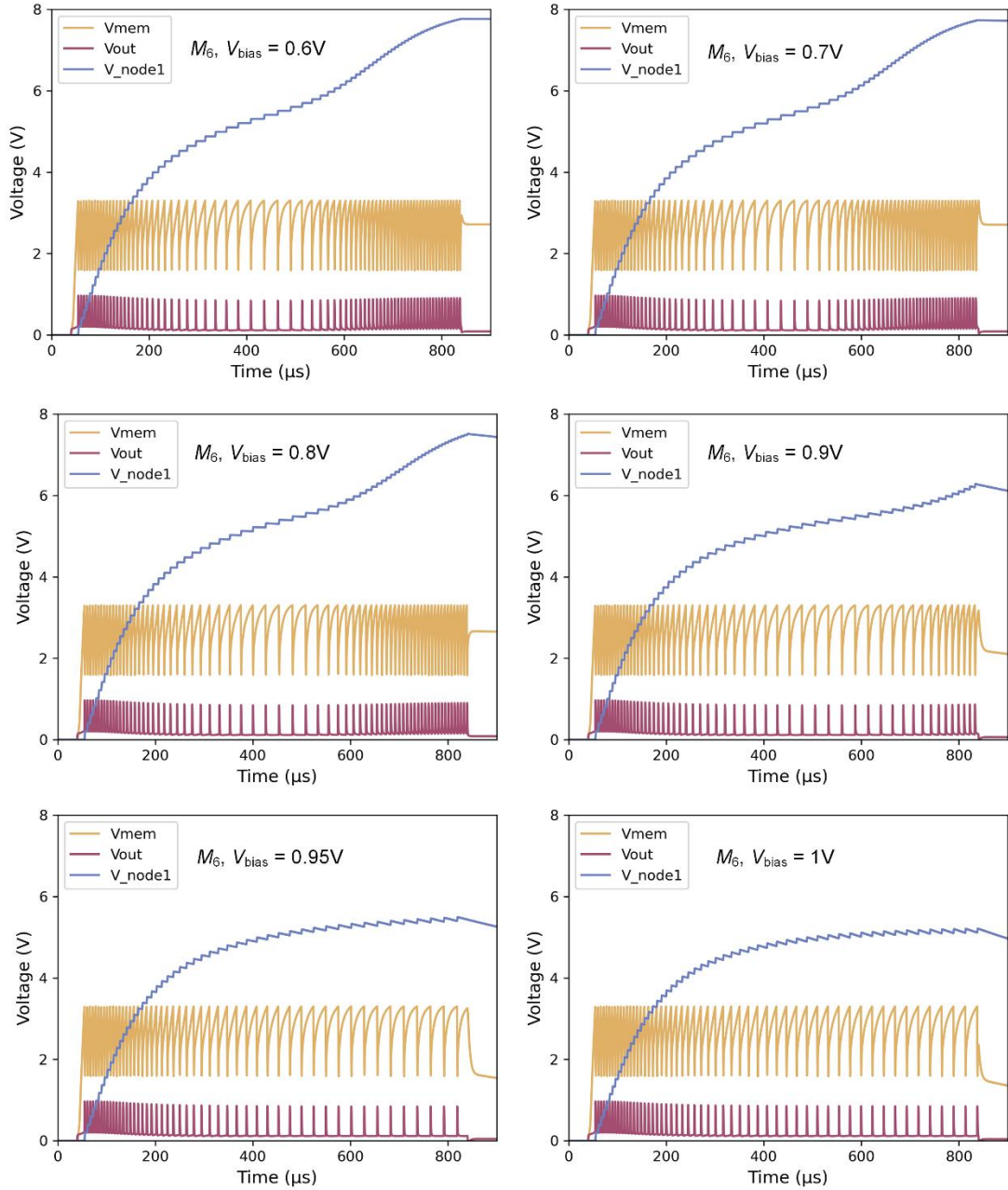
Supplementary Fig. 14 We further investigated the regulatory effect of the M_2 transistor's gate voltage on neuronal spike frequency under different width-to-length (W/L) ratios. Under a fixed W/L ratio, the spike output frequency of the neuron consistently decreases with increasing gate voltage, indicating enhanced current leakage through M_2 . Moreover, comparative analysis across various W/L ratios reveals that neurons incorporating M_2 transistors with larger W/L ratios require a lower gate voltage to achieve a given level of spike frequency suppression. These findings suggest that increasing the W/L ratio enhances the current leakage capability of M_2 , thereby enabling more efficient modulation of the neuron's effective input current and firing behavior.



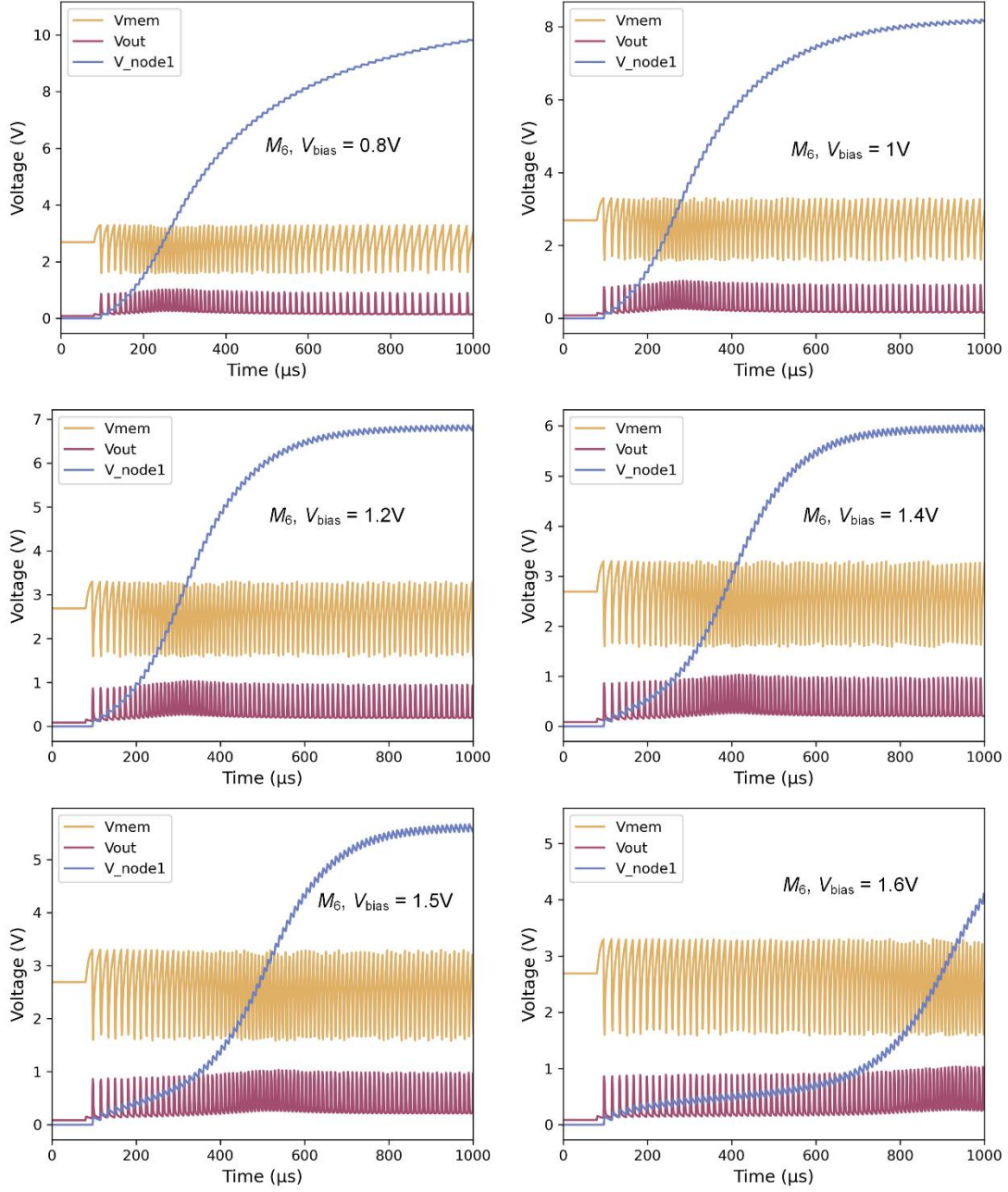
Supplementary Fig. 15 We further investigated the impact of varying the width-to-length (W/L) ratios of transistors M_3 and M_4 on neuronal circuit performance. Specifically, under three fixed W/L configurations for M_3 (1:5, 1:10, and 1:15), we systematically compared different W/L ratios of M_4 (with length fixed at 1 μ and width ranging from 5 μ to 15 μ). The analysis reveals that when M_3 has a W/L ratio of 1:10 and M_4 is configured at 10:1, the circuit effectively amplifies the output spikes. This configuration fulfills the signal enhancement requirement of the feedback module, indicating its suitability for feedback-driven neuromorphic circuit applications.



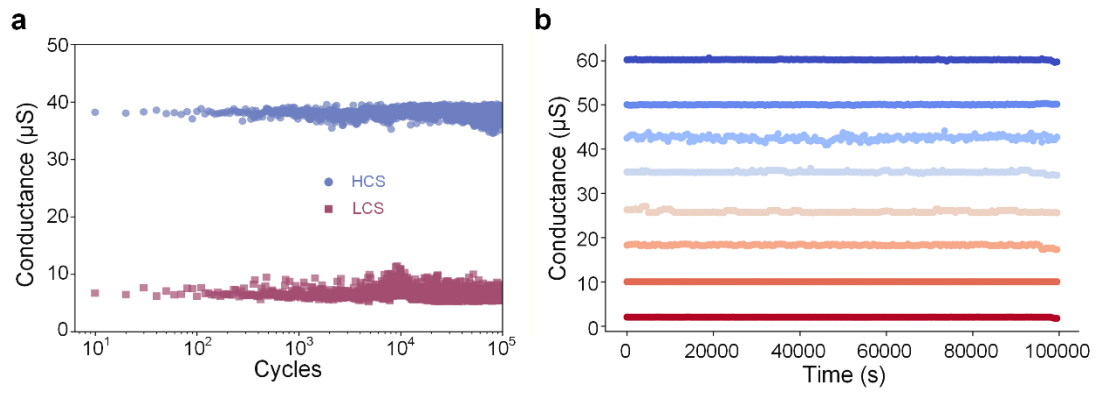
Supplementary Fig. 16 To evaluate the design parameters of the feedback module, a spiking input pattern with a period of 10 μ s and a pulse width of 500 ns is employed. A systematic one variable at a time approach was adopted, beginning with an initial configuration comprising a 500 pF feedback capacitor and identical width-to-length (W/L) ratios of 1:1 for both transistors M_5 and M_6 . Subsequently, the transistor lengths were varied from 0.5 μ m to 2.5 μ m in the following steps: 0.5 μ m, 0.6 μ m, 0.7 μ m, 0.8 μ m, 0.9 μ m, 1.0 μ m, 1.3 μ m, 1.6 μ m, 1.9 μ m, 2.2 μ m, and 2.5 μ m, while the transistor widths remained constant. The results indicate that the W/L ratio of transistor M_5 exerts a more substantial influence on the feedback behavior than that of M_6 . This effect is mainly due to the significantly lower gate voltage applied to M_6 during operation, which constrains its ability to modulate current flow compared to M_5 .



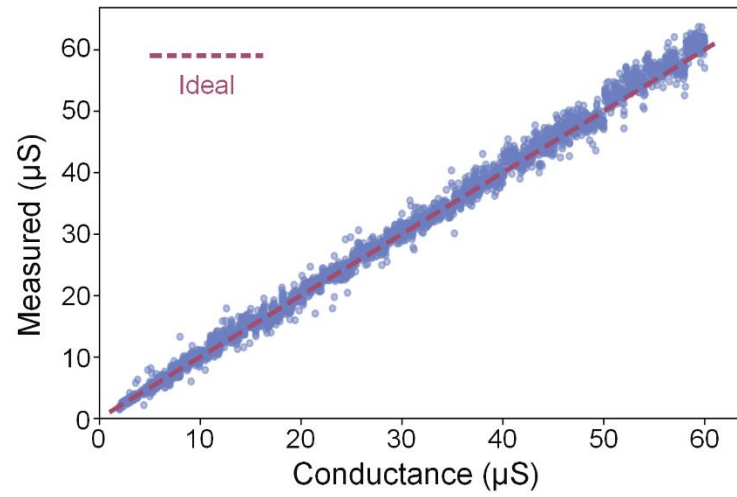
Supplementary Fig. 17 For the constructed self-tuning neuron, the spiking behavior was analyzed under different gate bias voltages applied to M_6 . At lower bias voltages of 0.6 V and 0.7 V, the feedback loop fails to discharge the node₁ voltage after the removal of the input signal, leading to persistent inhibitory behavior and a non-reproducible return to the resting state. At higher bias voltages of 0.95 V and 1 V, the voltage at node₁ cannot exceed 5 V due to the constraints of the feedback circuit, rendering transistor M_1 inactive and preventing the neuron from returning to its resting state. In contrast, moderate bias voltages of 0.8 V and 0.9 V strike a balance: they ensure the neuron can reliably return to a stable state while maintaining repeatability of the spiking-reset process.



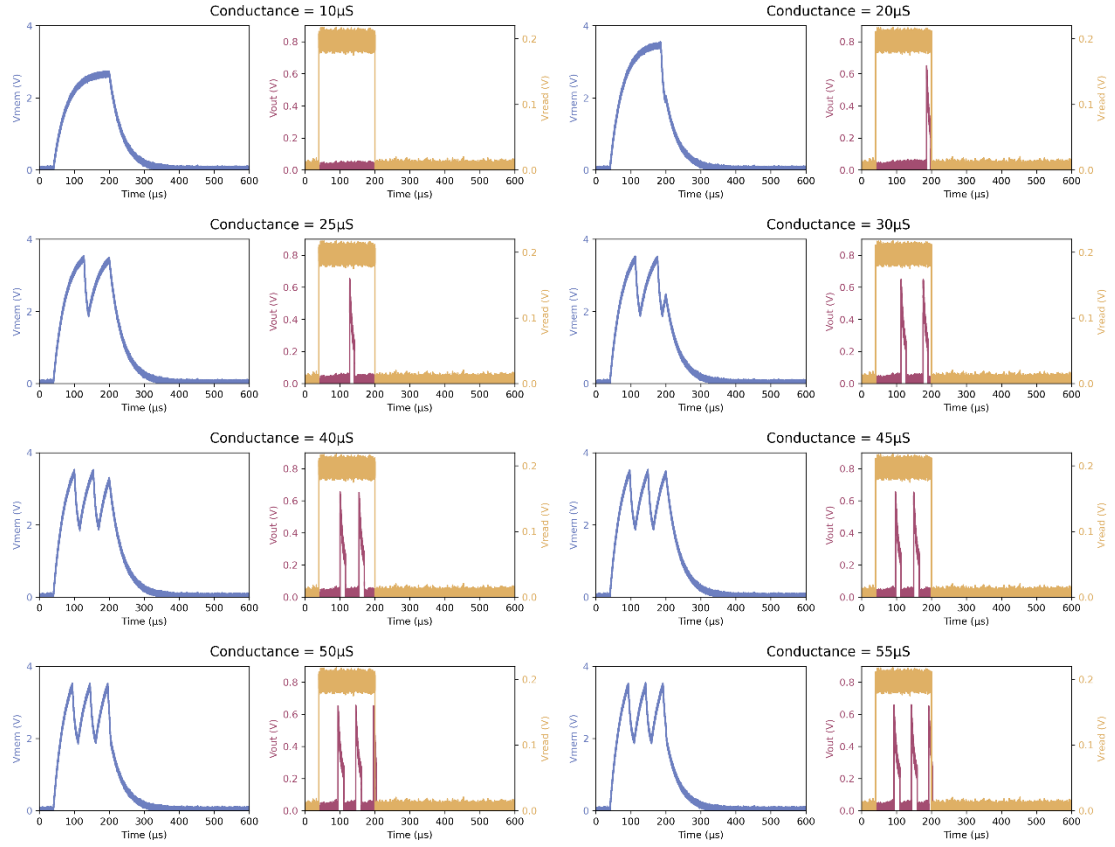
Supplementary Fig. 18 The spiking behavior of the constructed excitatory self-tuning neuron was analyzed under varying gate bias voltages applied to transistor M_6 . At a low bias voltage (e.g., 0.8 V), M_6 discharges the capacitor slowly, allowing the feedback-induced voltage at node₁ to rise rapidly. This enables the neuron to quickly reach the excitatory threshold and promptly return to a stable resting state after firing. As the gate voltage increases, the discharge rate of M_6 accelerates, which in turn slows down the rise of the node₁ voltage. Consequently, a longer time is required to trigger the excitatory response and to restore stability. When the gate voltage exceeds 1.4 V, the discharge rate of M_6 matches the charging rate of M_5 before the node₁ voltage can reach its effective operating point. This prevents further increase in node₁ voltage, thereby hindering the system's ability to return to a stable resting state. As the bias voltage continues to rise, the excitatory spiking behavior of the neuron is further suppressed.



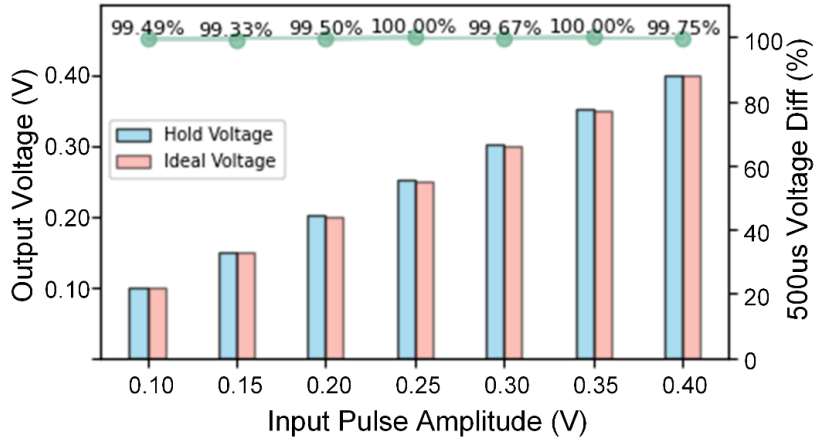
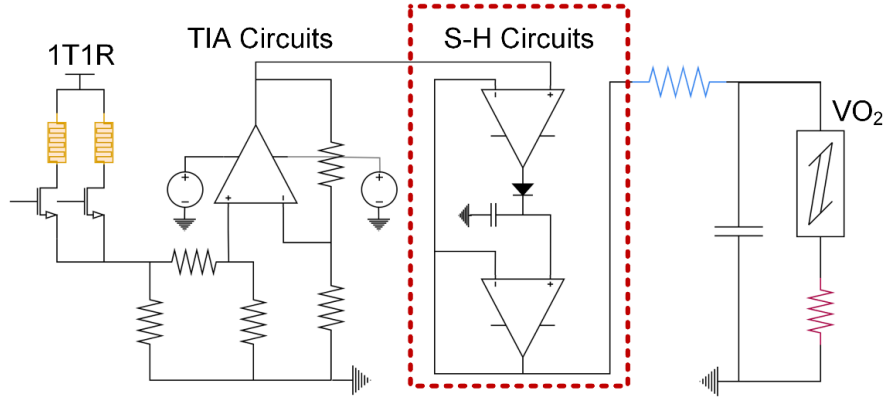
Supplementary Fig. 19 The endurance and retention performance of the non-volatile memristor. (a) Endurance of 10^5 cycles. (b) 8 conductance states can be retained for 10^5 seconds. From the lowest to the highest conductance state, the observed drift errors were 1.5%, 0.01%, 1.18%, 1.03%, 0.6%, 1.2%, 0.15%, and 0.14%.



Supplementary Fig. 20 Continuous analog values programmed into the array, and the measured conductance closely follow the intended target values.

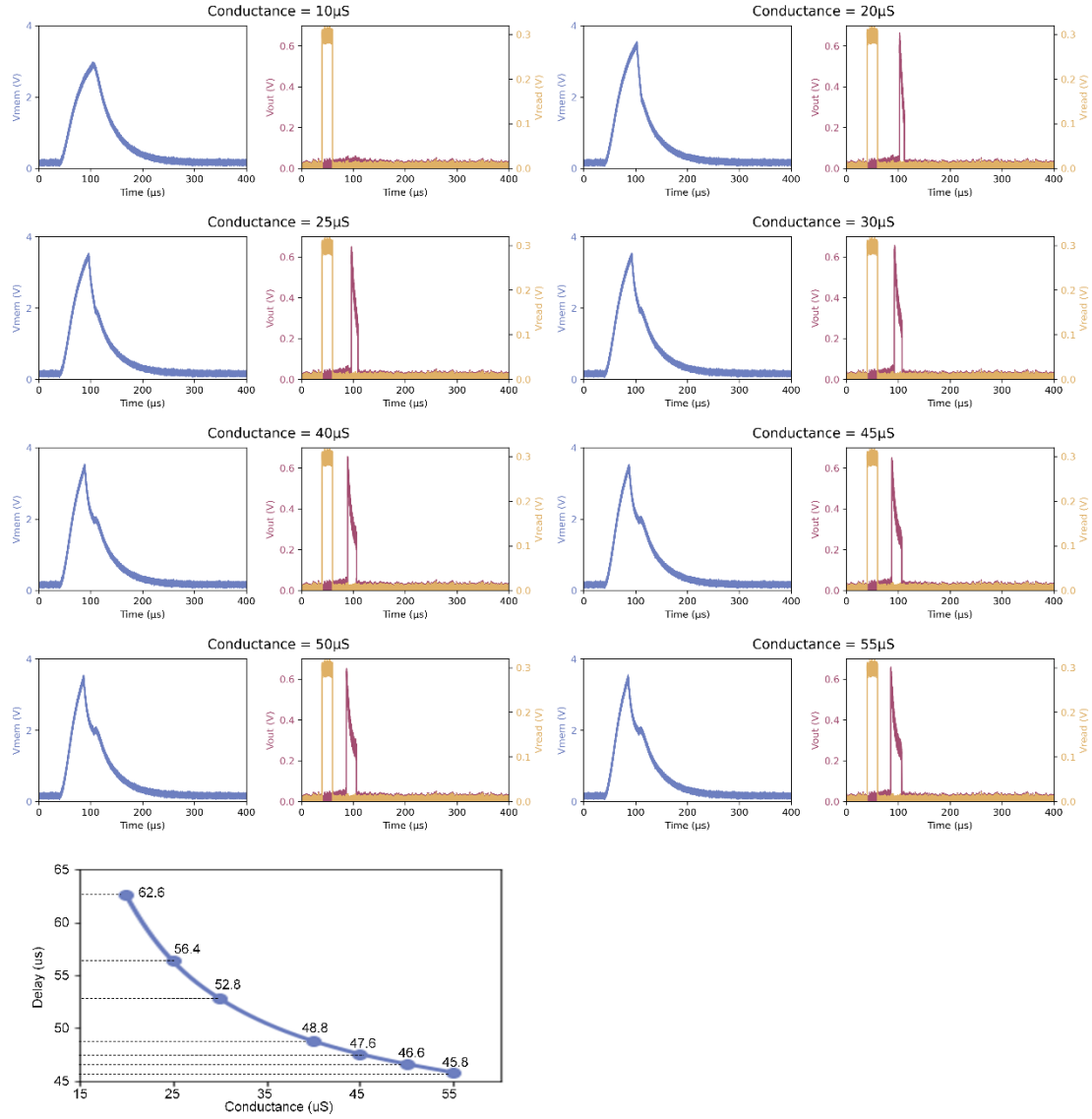


Supplementary Fig. 21 RRAM devices were programmed to distinct conductance states of 20 μS , 25 μS , 30 μS , 40 μS , 45 μS , 50 μS , and 55 μS . A read pulse with an amplitude of 0.2 V and a duration of 160 μs was applied to the dendritic structure, and the timing of the resulting output spike was monitored. The experimental results demonstrate that the latency of output spike generation increases monotonically with decreasing conductance. Notably, when the conductance was reduced to 10 μS , the 160 μs read pulse failed to induce an output spike within the pulse duration, indicating that the spiking latency exceeded 160 μs under this low-conductance condition.

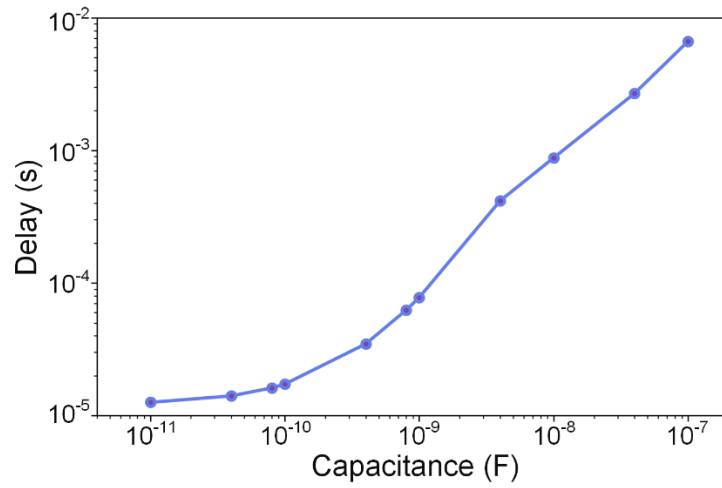


Supplementary Fig.22 We introduced a peak-hold circuit between the TIA circuit and the neuron to widen narrow pulses. The overall operation process is as follows: The pulse voltage applied to the RRAM generates a current, which is converted to a high-voltage output by the TIA circuit. This high-voltage output is then broadened by the TIA circuit and applied to the neuron, resulting in spike generation.

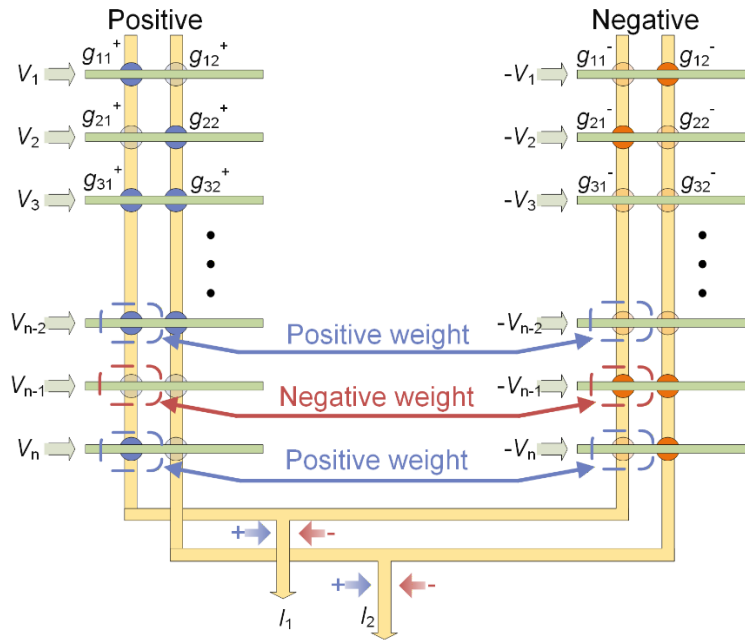
For the peak-hold circuit, the input voltage is fed into the first-stage comparator. When the input voltage exceeds the output voltage, the positive voltage of the diode increases, causing the diode to conduct and charge the capacitor C . At the same time, the voltage at this node is fed back to the other input of the op-amp via a voltage buffer, reducing the voltage difference across the op-amp. This feedback limits the further increase in the output node voltage, ensuring that the output voltage is sufficient to trigger a pulse in the LIF neuron. Disregarding the reverse bias current of the diode and the leakage current at the transistor's gate, when the input voltage decreases, the voltage at the C_{hold} node remains constant and the output voltage from the voltage buffer remains the peak value of the input voltage. We tested the voltage holding capability of the peak-hold circuit under an input voltage range of 0.1–0.4 V. The results indicate that the error between the output peak value and the input peak value remains within 1%, satisfying our design requirements. Additionally, the output voltage decay after 500 μs is also within a 1% range.



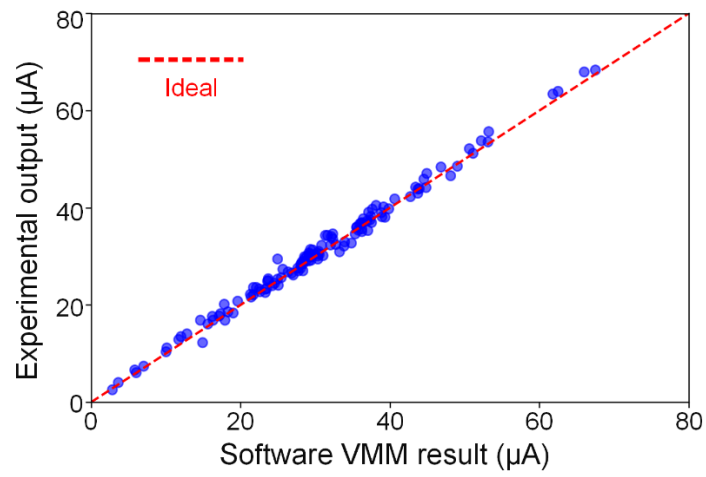
Supplementary Fig. 23 RRAM devices were programmed to distinct conductance states of 10 μS , 20 μS , 25 μS , 30 μS , 40 μS , 45 μS , 50 μS , and 55 μS . A read pulse with an amplitude of 0.3 V and a duration of 20 μs was applied to the modified dendritic structure, resulting in output spikes generated after varying temporal delays. Experimental results reveal that the spike generation delay increases as the conductance decreases, indicating an inverse relationship between conductance and the temporal response of the system.



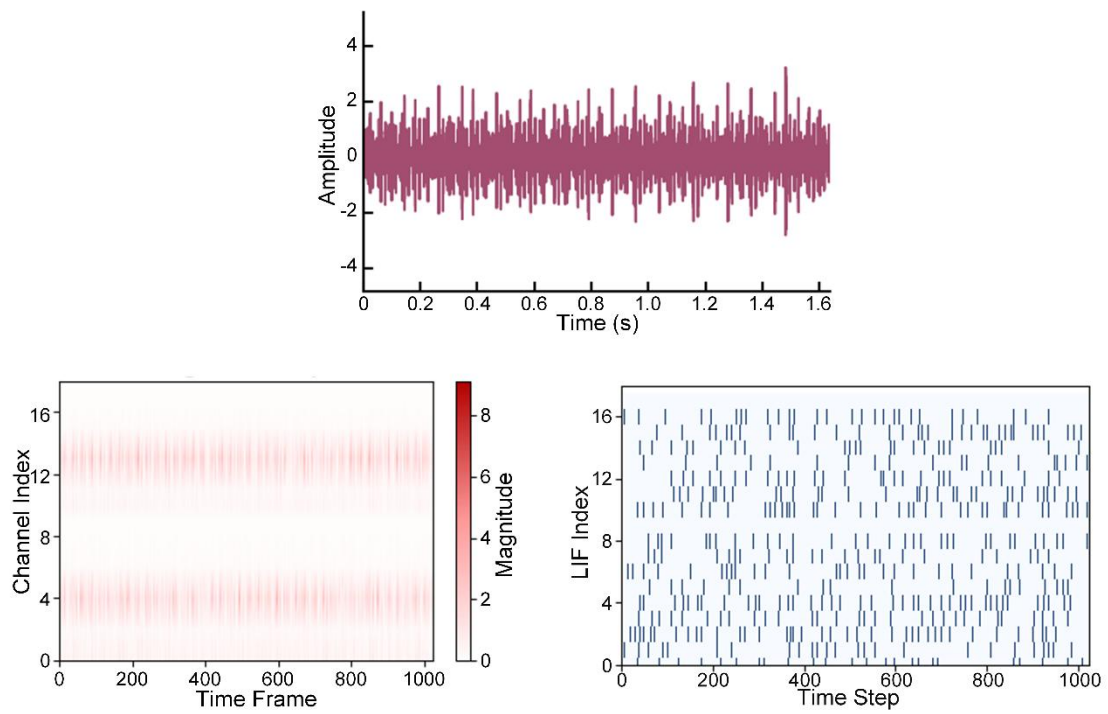
Supplementary Fig. 24 The RRAM device was programmed to a conductance state of $20 \mu\text{S}$. By adjusting the capacitance connected in parallel with the VO_2 device, the output pulse delay can be effectively tuned. Experimental results indicate that as the parallel capacitance increases, the delay in pulse generation becomes longer.



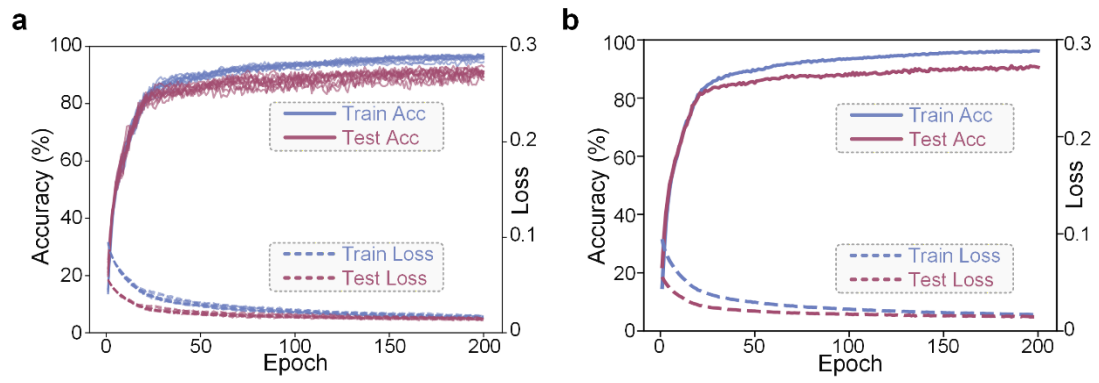
Supplementary Fig. 25 Schematic illustration of the differential-pair conductance mapping used to realize positive and negative weights in an RRAM crossbar array. Each input V_i is connected to two parallel devices storing the positive and negative conductance components, respectively. The left devices receive the original input voltages ($+V_i$), while the right devices receive their inverted counterparts ($-V_i$). The resulting branch currents are summed through separate routing lines and combined at the output to produce a signed weighted current via differential subtraction. This structure enables the hardware representation of signed weights using only positive-conductance memory devices.



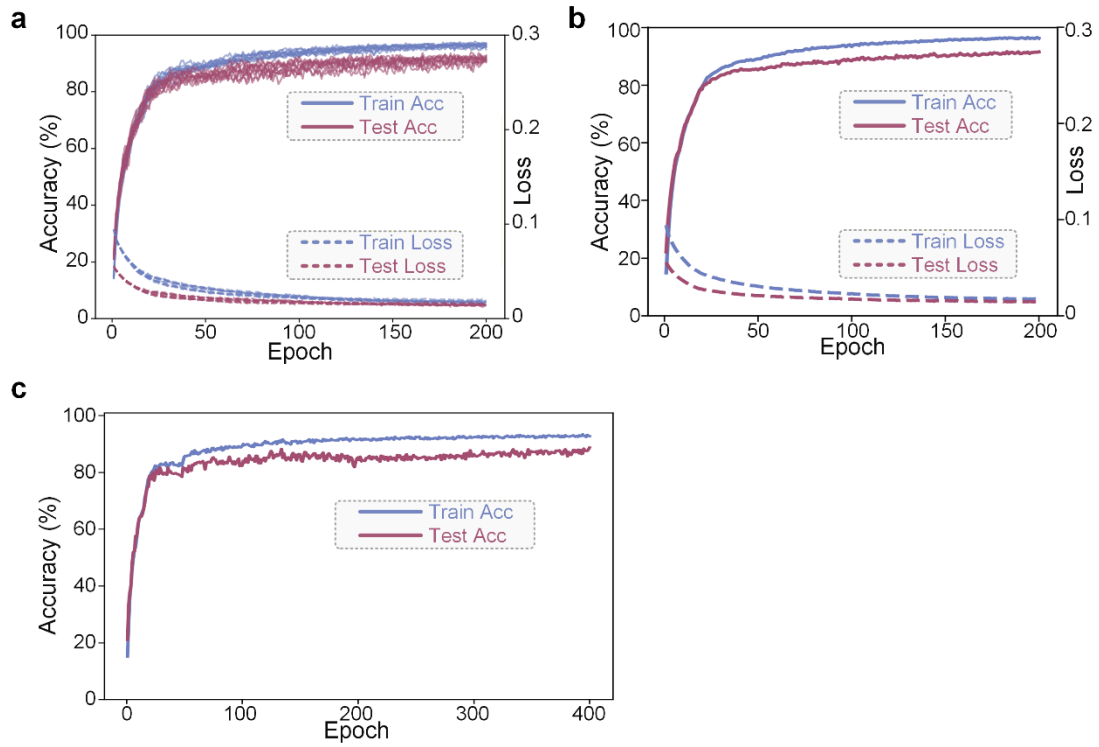
Supplementary Fig. 26 Matrix–vector multiplication test on the RRAM array. A subset of the trained weights from the CWRU task was mapped onto the array. The outputs obtained from the array show strong agreement with the expected results, demonstrating the capability of the RRAM array to perform analog in-memory computing.



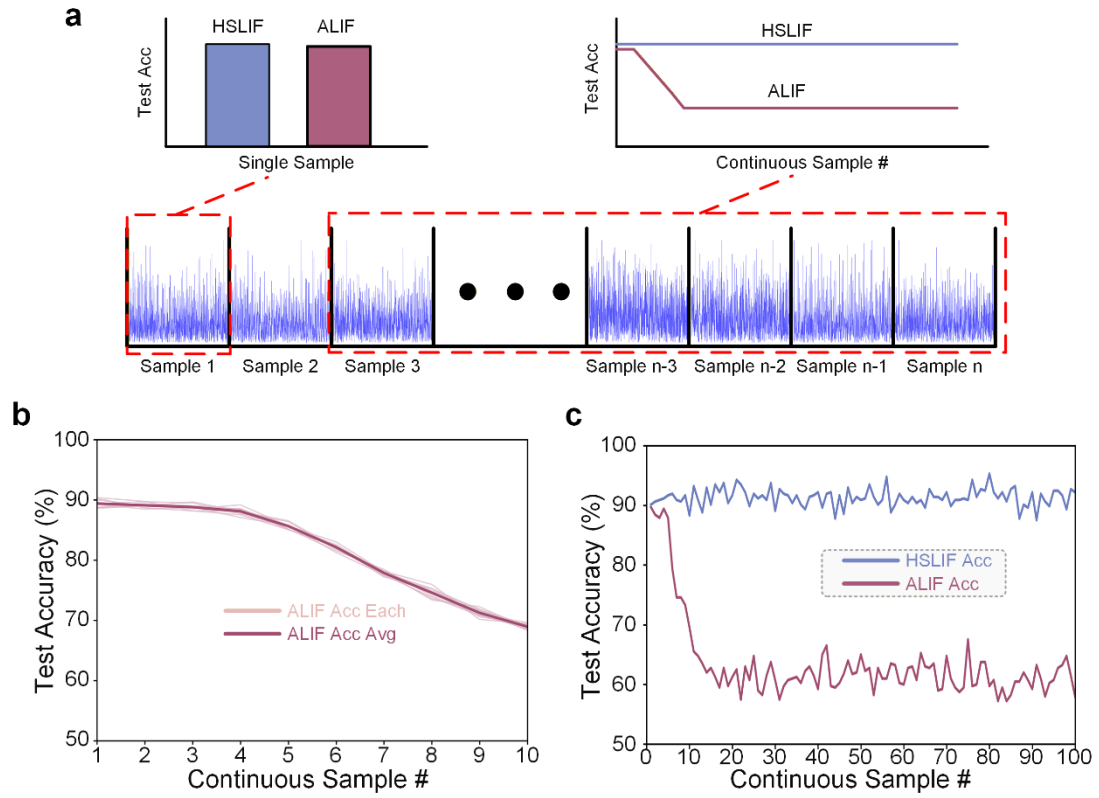
Supplementary Fig. 27. The original defect signal is processed using Short-Time Fourier Transform (STFT) with a Hamming window. The window length, overlap length, and FFT length are all set to 16. The processed results are shown in the figure. The resulting data is sequentially fed into individual LIF neurons across different channels, generating a series of spike outputs.



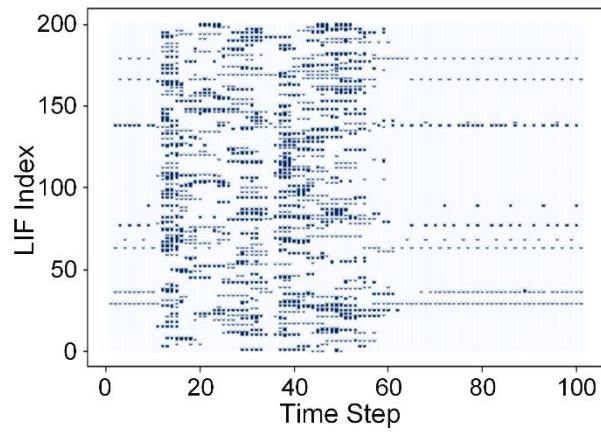
Supplementary Fig. 28. Accuracy and loss curves during the training of the SNN with HSLIF. (a) Accuracy and loss curves over 200 epochs for ten independent training runs of the SNN constructed with HSLIF neurons, where each curve corresponds to one training trial. (b) Averaged training and testing accuracy and loss across the ten runs, where for each epoch, the accuracy and loss values were averaged to illustrate the overall training trend.



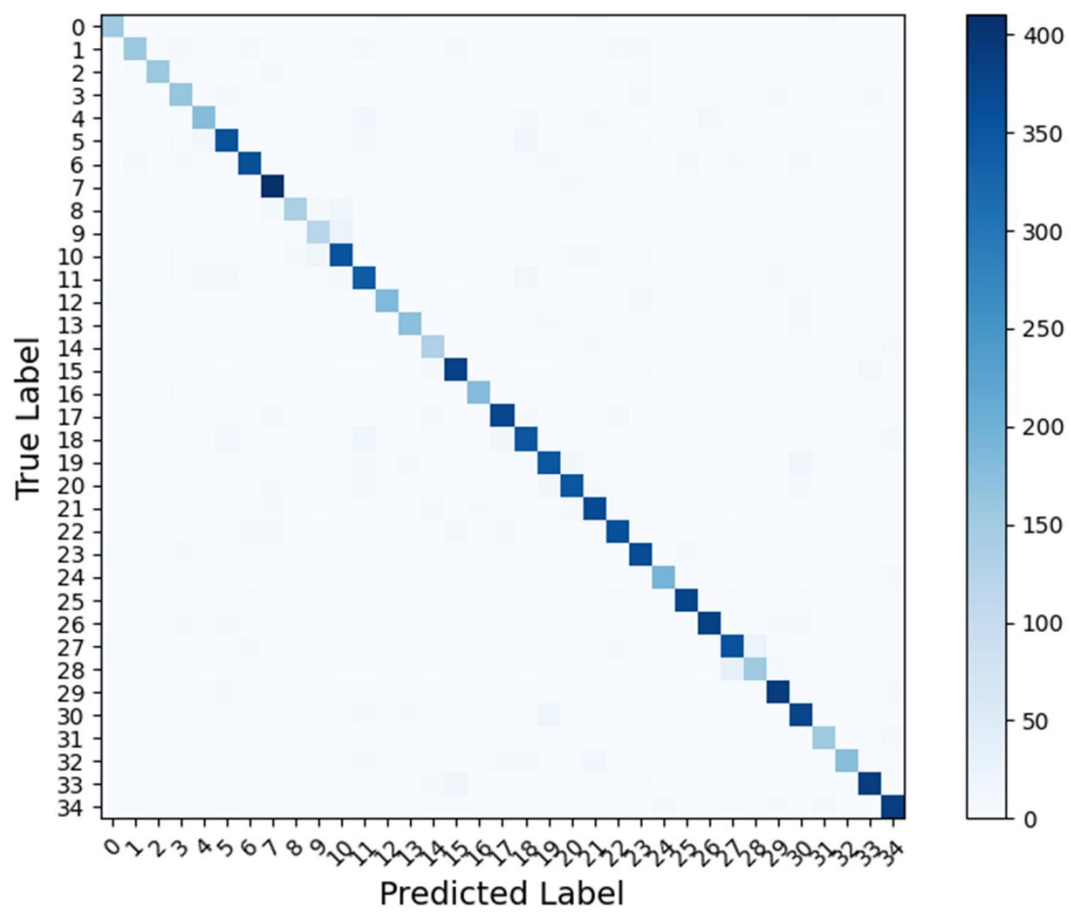
Supplementary Fig. 29. Accuracy and loss curves during the training of the SNN with LIF. (a) Accuracy and loss curves over 200 epochs for ten independent training runs of the SNN constructed with basic LIF neurons, where each curve corresponds to one training trial. (b) Averaged training and testing accuracy and loss across the ten runs, where for each epoch, the accuracy and loss values were averaged to illustrate the overall training trend. (c) Evolution of training accuracy for the same-sized SNN employing LIF neurons, as the number of training epochs was increased from 200 to 400. The results indicate that the network had already converged within 200 epochs, and extending the training to 400 epochs did not yield any further improvement in accuracy.



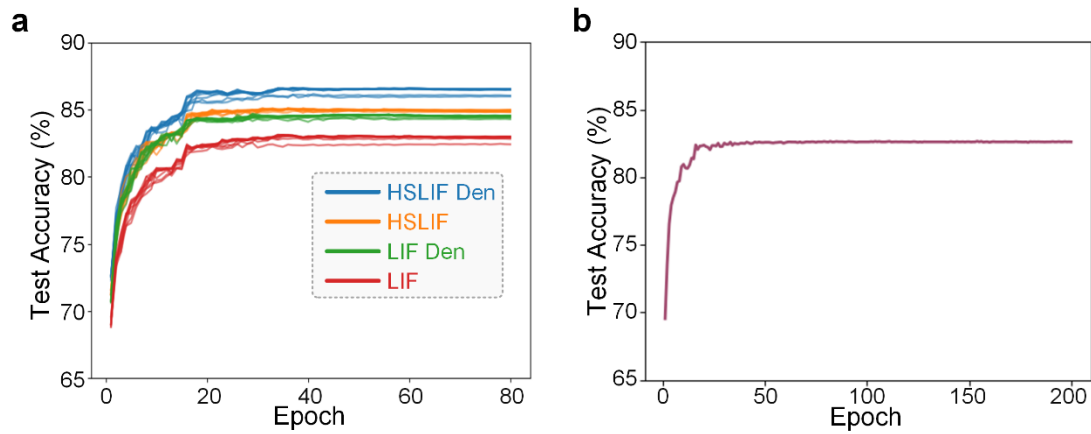
Supplementary Fig. 30 Performance of HSLIF and ALIF neurons in long-term classification tasks. (a) Schematic of our testing methodology. We first trained and evaluated the network model using independent samples. Under this setting, the performance achieved by the HSLIF neurons shows a certain improvement compared with the ALIF neurons, although the enhancement is not very pronounced. We then conducted a continuous testing experiment by concatenating multiple segments to form a long-duration input sequence. (b) Ten randomly selected data segments were concatenated to form a new continuous sequence, and the previously trained ALIF-based network was used for testing. The results indicate that the classification accuracy gradually decreases as inference progresses. (c) One hundred randomly selected data segments were concatenated to form a longer continuous sequence, and testing was performed using both the previously trained ALIF-based and HSLIF-based networks. The results show that the classification accuracy of the ALIF-based network decreases significantly over time, whereas the HSLIF-based network maintains a consistently high accuracy throughout long-term inference.



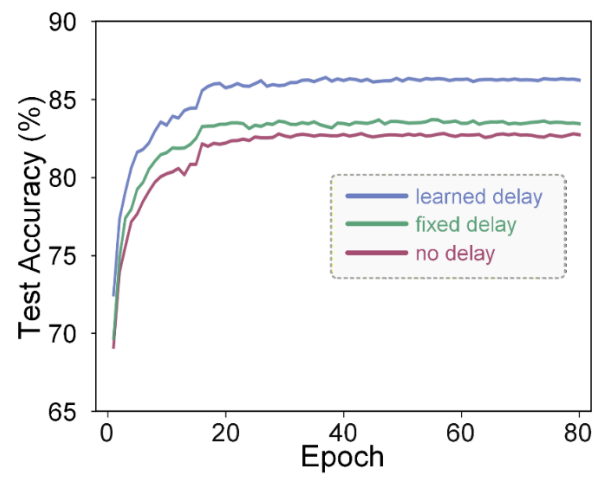
Supplementary Fig. 31. Spike emission behavior of 200 neurons over 101 time steps for the trained optimal model, and demonstrating a sparsity of 89.78%.



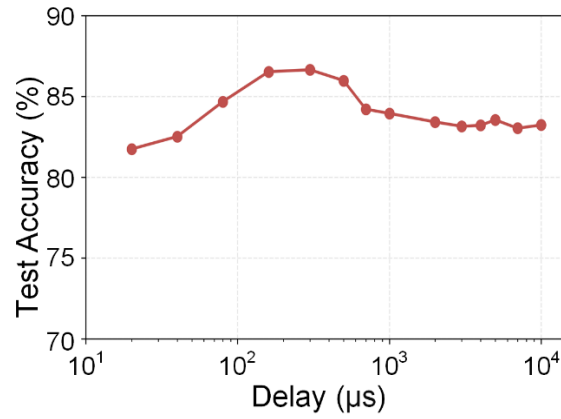
Supplementary Fig. 32. The trained model is evaluated using the test data from the dataset. The resulting confusion matrix, as shown above, demonstrates high testing accuracy.



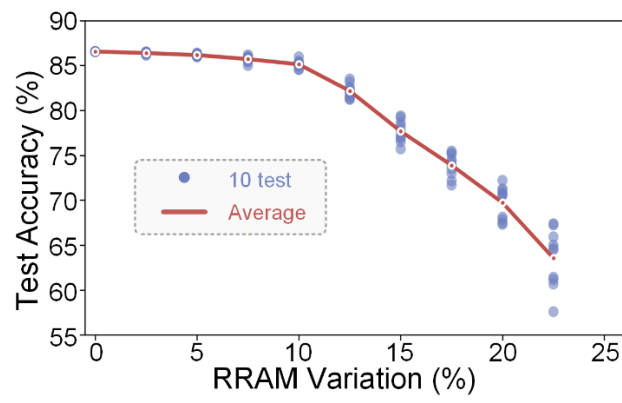
Supplementary Fig. 33. Accuracy of SNNs with different neuron structures. (a) Test accuracy evolution of SNNs constructed with four different neuron configurations from ten independent training runs: (1) regular LIF neurons only, (2) homeostatic neurons only, (3) dendritic structures with regular LIF neurons, and (4) a combination of homeostatic neurons and dendritic structures. Each model was trained independently ten times for 80 epochs. Each curve corresponds to one training run, showing the accuracy variation during the training process. (b) Testing accuracy of the same-sized SNN using regular LIF neurons when the number of training epochs is extended from 80 to 200. The results demonstrate that the network had already converged within 80 epochs, and further increasing the number of epochs did not result in a significant improvement in accuracy.



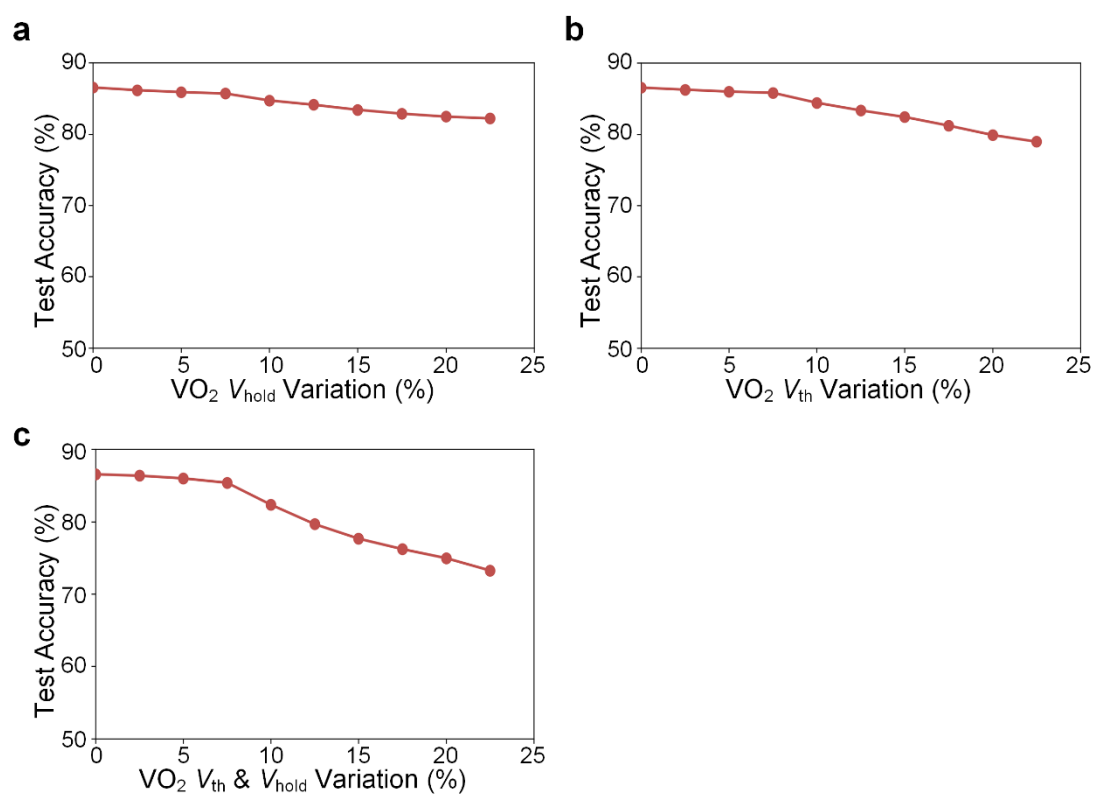
Supplementary Fig. 34. Performance comparison between trainable delays, fixed (non-trainable) delays, and the baseline SNN without delays.



Supplementary Fig. 35. Impact of dendritic delay on system-level performance. In the analysis, the temporal dimension of the input is represented by 101 discrete feature frames obtained from log-Mel preprocessing. For quantitative evaluation, each frame is rescaled to an effective duration of 10 μs in the system-level simulation. As the dendritic delay increases from 20 μs to 160 μs , the classification accuracy improves due to enhanced spatiotemporal integration across multiple frames. The accuracy remains relatively stable for larger delays but slightly degrades when the delay exceeds $\sim 500 \mu\text{s}$, which is attributed to excessive temporal overlap of the input features.



Supplementary Fig. 36. Accuracy of the proposed SNN under RRAM conductance variations.



Supplementary Fig. 37. SNN Accuracy under VO₂ Device Variations. (a) Impact of V_{hold} variation on SNN accuracy. (b) Impact of V_{th} variation on SNN accuracy. (c) Combined impact of V_{th} and V_{hold} variations on SNN accuracy.

Supplementary Note. 1

The width-to-length ratio of M_1 and M_2 , as well as the V_{dd1} voltage. Based on the test results of LIF neurons, it is observed that oscillation can only occur when the input current is within the range of 10-60 μA . Therefore, the current applied to the neuron, which results from the combined effects of the current supplied by M_1 , the leakage current from M_2 , and the input current, must lie within this range. Additionally, during operation, the voltage at node₂ oscillates within the range of 1.8-3.5 V at a high frequency. By statistically analyzing the oscillation amplitude at different frequencies and referencing the simulated PWM output signal, it is concluded that the voltage at this point is approximately 2.5 V. The gate voltages of M_1 and M_2 gradually increase as the capacitor charges, thus M_2 turning on earlier than M_1 . Initially, current leakage occurs through M_2 , which results in a decrease in spike frequency. To restore the oscillation frequency to its initial level after the frequency decrease, the current supplied by M_1 , as the gate voltage increases further, compensates for the leakage current of M_2 . To ensure that both transistors can turn on and the current meets the requirement for the neuron to generate oscillatory output, we designed V_{dd1} to be 8 V, with the width-to-length ratio of M_1 as $1\mu:9\mu$ and the width-to-length ratio of M_2 as $0.6\mu:20\mu$. Since the output pulse amplitude is around 1.1 V, the operational amplifier's bias point needs to be within 1.1 V. Therefore, we designed the width-to-length ratio of M_3 as $1\mu:10\mu$ and M_4 as $10\mu:1\mu$. To ensure that the time constant of the charging and discharging structure formed by capacitor C_1 and the M_5 , M_6 transistors is at least one order of magnitude greater than that of the time constant of the LIF neuron, we designed the capacitor to be 500 pF, with both M_5 and M_6 having a width-to-length ratio of $1\mu:1\mu$, and the gate voltage of M_6 set to 0.8 V.

Supplementary Note. 2

Initially, the VO₂ memristor is in the HRS, and the voltage across all capacitors is zero. When the current input is applied, the voltage across capacitor C_m increases. Once the voltage exceeds the threshold voltage V_{th} of the VO₂ device, the device rapidly switches to the LRS, causing the voltage across the device to drop. Since the capacitor voltage cannot change instantaneously, a sharp voltage spike is generated at the output node due to the threshold switching behavior. As the voltage across the VO₂ device continues to drop, the device reverts to HRS until the V_{hold} . This cycle repeats, generating a series of spike signals. Once a spike is generated, it is amplified by transistors M_3 and M_4 . The amplified signal then controls transistor M_5 , which charges capacitor C_1 .

Meanwhile, the gate voltage of transistors M_1 and M_2 rises with the increasing voltage across C_2 . When the voltage at node₁ is below the transistor turn-on threshold, neither additional current is supplied nor leakage occurs. When the voltage at node₁ surpasses the turn-on threshold but remains below a critical level, the leakage current of M_1 dominates over the supplementary current from M_2 , reducing the neuron's input current and thereby decreasing the spike frequency. Conversely, once the voltage at node₁ exceeds this threshold, the supplementary current from M_2 surpasses the leakage current of M_1 , increasing the neuron's input current and restoring the spike frequency to its stable state.

Supplementary Note. 3

Each 1T1R device is connected as follows: the top electrode of the memristor is connected to the bit line (BL), the bottom electrode of the memristor is connected to the drain of the transistor, the source of the transistor is connected to the source line (SL), and the gate of the transistor is connected to the word line (WL). The three bit lines allow selective access to the devices, and by applying different voltages on these lines, programming or read operations can be carried out. To ensure accurate programming, the write-verify operation is repeatedly performed during the programming process until the desired resistance value is achieved. During the set process, the WL selector is used to choose the column for operation, connecting the gates of all transistors in that column to a 3.3 V voltage, while the gates of other columns are grounded. All SL lines are also grounded. The BL selector is then used to select the row for operation, applying a V_{set} voltage to that row while grounding the other rows. In the reset process, all BL lines are grounded. The SL selector selects the corresponding column, and a V_{reset} voltage is applied to the selected port. In the read process, the WL selector chooses the column and applies a 5 V gate voltage. The BL selector selects the row and applies a V_{read} voltage, while the SL port is connected to the TIA circuit.

Supplementary Note. 4

To support both positive and negative synaptic weights in the proposed hardware neural network, we implemented a differential-pair conductance mapping scheme based on a dual-column RRAM crossbar structure (Supplementary Fig. 25). For each input voltage V_i , two parallel RRAM devices (the positive branch g_{i+} and the negative branch g_{i-}) are allocated symmetrically in the crossbar. The left devices store the positive component of the weight, whereas the right devices store the negative component. A scalar synaptic weight w_i is decomposed into two non-negative conductance values: $w_i = G_{i+} - G_{i-}$, where G_{i+} is programmed into the positive RRAM device g_{i+} , G_{i-} is programmed into the negative RRAM device g_{i-} . To realize weight subtraction physically, the two halves of the array are biased using opposite input polarities. The source lines on both sides are routed and merged at the output summing node using a pair of differential current-collecting paths. This differential configuration naturally enables both positive and negative weight representations without relying on devices that provide negative conductance. In addition, device variations and non-idealities are alleviated through differential cancellation, as the two branches experience nearly identical voltage conditions, wire resistances, and temporal fluctuations.

Supplementary Table 1. Comparative benchmark of different networks on the CWRU dataset.¹⁻⁷

	Model	Accuracy	Params	Flops
Chen et al. ¹	ResNet50	99.92%	47.2M	718M
Yoo et al. ²	CNN0	96.64%	901K	1.8M
Rosa et al. ³	WDCNN	89.70%	40.7K	-
Wang et al. ⁴	CWT-2DCNN	83.10%	1.34M	840M
Wang et al. ⁵	MS-ER	98.80%	3.95M	2.8G
Lim et al. ⁶	SpikeCAN	99.86%	1.45M	-
Fang et al. ⁷	Spiking-ResNet	87.46%	3.85M	-
This work	SNN-HSLIF	92.14%	22K	140K

Supplementary Table 2. Comparison of Four Network Configurations for the GSC Task.

Dataset	Structure	Accuracy	Area (μm^2)	Basic circuit unit
GSC (35 classes)	LIF (VO ₂ based)	83.03%	78×10^3	1C 1R 1VO ₂
GSC (35 classes)	HSLIF (VO ₂ based)	85.09%	93×10^3	6T 2C 1R 1VO ₂
GSC (35 classes)	LIF + Dendrite (RRAM & VO ₂)	84.61%	86×10^3	2T 3C 7R 2AMP 1DIO 1RRAM 2VO ₂
GSC (35 classes)	HSLIF + Dendrite (RRAM & VO ₂)	86.53%	102×10^3	8T 4C 7R 2AMP 1DIO 1RRAM 2VO ₂

Supplementary Table 3. Comparative benchmark of different networks on the GSC dataset.^{8–17}

	Model	Num Classes	Accuracy	Params	Flops
Niizumi et al. ⁸	M2D	35	98.50%	86M	-
Chen et al. ⁹	EAT	35	98.30%	88M	-
Andrade et al. ¹⁰	Attention-RNN	20	94.50%	202K	44.6M
Kim et al. ¹¹	BC-ResNet-8	12	98.70%	321K	171.8M
Majumdar et al. ¹²	Matchbox-Net	35	97.40%	140K	34.2M
Bittar et al. ¹³	SNN-RadLIF	35	94.51%	1.2M	-
Hammouamri et al. ¹⁴	SNN-delays	35	95.35%	2.5M	-
Deckers et al. ¹⁵	SNN-d-cAdLIF	35	95.69%	610K	-
Sun et al. ¹⁶	SNN-delays	20(SHD)	92.45%	109K	-
D'agostino et al. ¹⁷	SNN-delays(fixed)	20(SHD)	90.88%	224K	-
Deckers et al. ¹⁵	SNN-LIF	35	86.21%	300k	-
This work	SNN-HSLIF-delays	35	86.53%	33K	3.13M

Supplementary Table 4. Comparison of hardware implementations for the GSC dataset task.^{18–26}

	Technology	Model	Num Classes	Accuracy	Frequency (Hz)	Energy for 1s speech (J)	Latency (s)	Area (um ²)
Zol. et al. ^{18,19}	14nm CMOS + IBM PCM	CNN	11	89.30%	400x10 ⁶	78x10 ⁻⁶	348x10 ⁻⁶	144x10 ⁶
Giraldo et al. ²⁰	65nm	LSTM	10	91%	250x10 ³	16.2x10 ⁻⁶	-	2.56x10 ⁶
Li et al. ²¹	65nm	GRU	2	98%	-	122.3x10 ⁻⁶	211x10 ⁻⁶	4.38x10 ⁶
Kim et al. ²²	65nm	RNN	12	86%	250x10 ³	23x10 ⁻⁶	12.4x10 ⁻³	2.03x10 ⁶
Shan et al. ²³	28nm	DSCNN	2	95%	40x10 ³	0.98x10 ⁻⁶	-	230x10 ³
Liu et al. ²⁴	65nm	FS-CNN + LIF	4	95%	100x10 ³	7.7x10 ⁻⁶	18.7x10 ⁻³	917x10 ³
Wang et al. ²⁵	65nm	SNN	5	90%	120x10 ³	0.57x10 ⁻⁶	100x10 ⁻³	2.71x10 ⁶
Shaban et al. ²⁶	50um RRAM + 0.5um CMOS	SNN	12	91%	1x10 ³	696x10 ⁻⁶	1.2	11x10 ³ (sim 65nm)
This work	180nmRRAM + 180nm CMOS	SNN	35	86.53%	180x10 ³	12.9x10 ⁻⁶	1.66x10 ⁻³	102x10 ³

Supplementary Table 5. Detailed comparison between neuron circuits.^{26–32}

	Neuron type	Area (μm^2)	Energy per Spike (J)	Frequency (Hz)	Circuit
Shaban et al. ²⁶	ALIF (frequency increase)	11435.7 (all CMOS 65nm)	158×10^{-12}	1×10^3	6T + 10xRAM
Ahma. et al. ²⁷	ALIF (frequency increase)	863 (CMOS 180nm)	140×10^{-12}	$\sim 1 \times 10^3$	24T + 3C + RRAM
Moradi et al. ²⁸	ALIF (frequency decrease)	8572 (CMOS 180nm)	260×10^{-12}	$\sim 100 \times 10^3$	30T + 3C
Chak. et al. ²⁹	ALIF (frequency decrease)	2177 (CMOS 65nm)	$\sim 28.6 \times 10^{-9}$	$\sim 6.7 \times 10^3$	T and C
Yang et al. ³⁰	HH	100 (5 μm)	1.06×10^{-12}	320×10^3	2R + 2C + 2NbO ₂ + 2 voltage source
Xiao et al. ³¹	LIF (frequency decrease)	no area info	$\sim 600 \times 10^{-12}$	160×10^3	2R + 2C + 2NbO ₂ + 2ECRAM
Qiao et al. ³²	LIF (frequency increase or decrease and return to stable)	1848 (CMOS 180nm)	$\sim 72 \times 10^{-12}$	$< 1 \times 10^3$	Purely CMOS
This work	LIF (frequency increase or decrease and return to stable)	~ 30197 (main cost by cap) / ~ 76.8 (only transistors)	19.29×10^{-12}	180×10^3	6T + 2C + 1R + 1VO ₂

Supplementary Table 6. Detailed comparison between dendritic circuits.^{17,33–37}

	Device	Dendritic Mechanism	Dendritic Type	Energy per Spike (J)	Time Constant (s)	Dendritic Circuit
D'Ago. et al. ¹⁷	RRAM + 180nm CMOS	temporal process & sequence selectivity	one synapse per dendrite	58.5×10^{-12}	8×10^{-3} $\sim 58 \times 10^{-3}$	40T + 1C + 2RRAM (analysis-based)
Sche. et al. ³³	65nm CMOS	non-linearity & multi-compartment	multiple synapses share a dendrite	$\sim 174 \times 10^{-12}$	N/A	N/A
Li et al. ³⁴	memristor (Pt/TaO _x /AlO _x /Al)	filter & integrate	multiple synapses share a dendrite	$\sim 600 \times 10^{-12}$	$\sim 10 \times 10^{-3}$	1 memristor
Baek et al. ³⁵	multigate silicon nanowire FET	non-linearity & direction selectivity	multiple synapses share a dendrite	$\sim 130 \times 10^{-12}$	250×10^{-3}	1FET + 1 sol-gel layer
Yang et al. ³⁶	anti-ambipolar transistor	meta-plasticity	multiple synapses share a dendrite	N/A	$\sim 5 \times 10^{-6}$	1AAB transistor
Moro et al. ³⁷	RRAM + 130nm CMOS	temporal process & coincidence detection	one synapse per dendrite	N/A	$\sim 20 \times 10^{-6}$	20T + 3C + 2RRAM + 1AMP
This work	VO ₂ + RRAM + 180nm CMOS	integrate & temporal process	multiple synapses share a dendrite	19.29×10^{-12}	15×10^{-6} $\sim 5 \times 10^{-3}$	2T + 6R + 2RRAM + 1AMP

Supplementary References

1. Chen, Z., Cen, J. & Xiong, J. Rolling Bearing Fault Diagnosis Using Time-Frequency Analysis and Deep Transfer Convolutional Neural Network. *IEEE Access* **8**, 150248–150261 (2020).
2. Yoo, Y., Jo, H. & Ban, S.-W. Lite and Efficient Deep Learning Model for Bearing Fault Diagnosis Using the CWRU Dataset. *Sensors* **23**, 3157 (2023).
3. Rosa, R. K., Braga, D. & Silva, D. Benchmarking deep learning models for bearing fault diagnosis using the CWRU dataset: A multi-label approach. Preprint at <https://doi.org/10.48550/arXiv.2407.14625> (2024).
4. Wang, X., Hua, T., Xu, S. & Zhao, X. A Novel Rolling Bearing Fault Diagnosis Method Based on BLS and CNN with Attention Mechanism. *Machines* **11**, 279 (2023).
5. Wang, H., Wang, H. & Xie, J. Fault Diagnosis of Rolling Bearings Based on HFMD and Dual-Branch Parallel Network Under Acoustic Signals. *Sensors (Basel)* **25**, 5338 (2025).
6. Lim, J. & Kim, S.-E. Spike-Driven Channel-Temporal Attention Network with Multi-Scale Convolution for Energy-Efficient Bearing Fault Detection. *Applied Sciences* **15**, 7622 (2025).
7. Fang, W. *et al.* Deep Residual Learning in Spiking Neural Networks. Preprint at <https://doi.org/10.48550/arXiv.2102.04159> (2022).
8. Niizumi, D., Takeuchi, D., Ohishi, Y., Harada, N. & Kashino, K. Masked Modeling Duo: Learning Representations by Encouraging Both Networks to Model the Input. in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 1–5 (2023).
9. Chen, W., Liang, Y., Ma, Z., Zheng, Z. & Chen, X. EAT: self-supervised pre-training with efficient audio transformer. in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence* 3807–3815 (Jeju, Korea, 2024).
10. Andrade, D. C. de, Leo, S., Viana, M. L. D. S. & Bernkopf, C. A neural attention

- model for speech command recognition. Preprint at <https://doi.org/10.48550/arXiv.1808.08929> (2018).
11. Kim, B., Chang, S., Lee, J. & Sung, D. Broadcasted Residual Learning for Efficient Keyword Spotting. Preprint at <https://doi.org/10.48550/arXiv.2106.04140> (2023).
 12. Majumdar, S. & Ginsburg, B. MatchboxNet: 1D Time-Channel Separable Convolutional Neural Network Architecture for Speech Commands Recognition. in *Interspeech 2020* 3356–3360 (ISCA, 2020).
 13. Bittar, A. & Garner, P. N. A surrogate gradient spiking baseline for speech command recognition. *Front. Neurosci.* **16**, (2022).
 14. Hammouamri, I., Khalfaoui-Hassani, I. & Masquelier, T. Learning Delays in Spiking Neural Networks using Dilated Convolutions with Learnable Spacings. Preprint at <https://doi.org/10.48550/arXiv.2306.17670> (2023).
 15. Deckers, L., Van Damme, L., Van Leekwijck, W., Tsang, I. J. & Latré, S. Co-learning synaptic delays, weights and adaptation in spiking neural networks. *Front. Neurosci.* **18**, (2024).
 16. Sun, P., Eqlimi, E., Chua, Y., Devos, P. & Botteldooren, D. Adaptive axonal delays in feedforward spiking neural networks for accurate spoken word recognition. in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2023).
 17. D’Agostino, S. *et al.* DenRAM: neuromorphic dendritic architecture with RRAM for efficient temporal processing with delays. *Nat Commun* **15**, 3446 (2024).
 18. Zolfagharinejad, M. *et al.* Analogue speech recognition based on physical computing. *Nature* **645**, 886–892 (2025).
 19. Le Gallo, M. *et al.* A 64-core mixed-signal in-memory compute chip based on phase-change memory for deep neural network inference. *Nat Electron* **6**, 680–693 (2023).
 20. Giraldo, J. S. P., Lauwereins, S., Badami, K., Van Hamme, H. & Verhelst, M. 18 μ W SoC for near-microphone Keyword Spotting and Speaker Verification. in *2019 Symposium on VLSI Circuits* C52–C53 (2019).

21. Li, Q. *et al.* NS-FDN: Near-Sensor Processing Architecture of Feature-Configurable Distributed Network for Beyond-Real-Time Always-on Keyword Spotting. *IEEE Transactions on Circuits and Systems I: Regular Papers* **68**, 1892–1905 (2021).
22. Kim, K. *et al.* A 23 μ W Solar-Powered Keyword-Spotting ASIC with Ring-Oscillator-Based Time-Domain Feature Extraction. in *2022 IEEE International Solid-State Circuits Conference (ISSCC)* vol. 65 1–3 (2022).
23. Shan, W. *et al.* A 510-nW Wake-Up Keyword-Spotting Chip Using Serial-FFT-Based MFCC and Binarized Depthwise Separable CNN in 28-nm CMOS. *IEEE Journal of Solid-State Circuits* **56**, 151–164 (2021).
24. Liu, Y. *et al.* A Keyword-Spotting(KWS) Chip Featuring a Bio-Inspired Neuron Model in 65-nm CMOS. in *2025 6th International Conference on Computing, Networks and Internet of Things (CNIOT)* 1–6 (2025).
25. Wang, D., Kim, S. J., Yang, M., Lazar, A. A. & Seok, M. 9.9 A Background-Noise and Process-Variation-Tolerant 109nW Acoustic Feature Extractor Based on Spike-Domain Divisive-Energy Normalization for an Always-On Keyword Spotting Device. in *2021 IEEE International Solid-State Circuits Conference (ISSCC)* vol. 64 160–162 (2021).
26. Shaban, A., Bezugam, S. S. & Suri, M. An adaptive threshold neuron for recurrent spiking neural networks with nanodevice hardware implementation. *Nat Commun* **12**, 4234 (2021).
27. Ahmadi-Farsani, J. *et al.* A CMOS-memristor hybrid system for implementing stochastic binary spike timing-dependent plasticity. *Philos Trans A Math Phys Eng Sci* **380**, 20210018 (2022).
28. Moradi, S., Qiao, N., Stefanini, F. & Indiveri, G. A Scalable Multicore Architecture With Heterogeneous Memory Structures for Dynamic Neuromorphic Asynchronous Processors (DYNAPs). *IEEE Transactions on Biomedical Circuits and Systems* **12**, 106–122 (2018).
29. Chakraborty, N. N., Das, H. & Rose, G. S. Homeostatic Plasticity in a Leaky Integrate and Fire Neuron Using Tunable Leak. in *2023 IEEE 66th International*

Midwest Symposium on Circuits and Systems (MWSCAS) 738–742 (2023).

30. Yang, Y. *et al.* Firing feature-driven neural circuits with scalable memristive neurons for robotic obstacle avoidance. *Nat Commun* **15**, 4318 (2024).
31. Xiao, Y. *et al.* Bio-plausible reconfigurable spiking neuron for neuromorphic computing. *Science Advances* **11**, eadr6733 (2025).
32. Qiao, N., Bartolozzi, C. & Indiveri, G. An Ultralow Leakage Synaptic Scaling Homeostatic Plasticity Circuit With Configurable Time Scales up to 100 ks. *IEEE Transactions on Biomedical Circuits and Systems* **11**, 1271–1277 (2017).
33. Schemmel, J., Kriener, L., Müller, P. & Meier, K. An accelerated analog neuromorphic hardware system emulating NMDA- and calcium-based non-linear dendrites. in *2017 International Joint Conference on Neural Networks (IJCNN)* 2217–2226 (2017).
34. Li, X. *et al.* Power-efficient neural network with artificial dendrites. *Nat. Nanotechnol.* **15**, 776–782 (2020).
35. Baek, E. *et al.* Neuromorphic dendritic network computation with silent synapses for visual motion perception. *Nat Electron* **7**, 454–465 (2024).
36. Yang, Y. *et al.* Bio-realistic and versatile artificial dendrites made of anti-ambipolar transistors. *Neuromorph. Comput. Eng.* **5**, 024020 (2025).
37. Moro, F. *et al.* Neuromorphic object localization using resistive memories and ultrasonic transducers. *Nat Commun* **13**, 3506 (2022).