nature materials

Supplementary information

https://doi.org/10.1038/s41563-026-02815-z

Tunable diffusive memristor arrays for neuromorphic temporal computing

In the format provided by the authors and unedited

Supplementary Information

Contents

Supplementary Figures

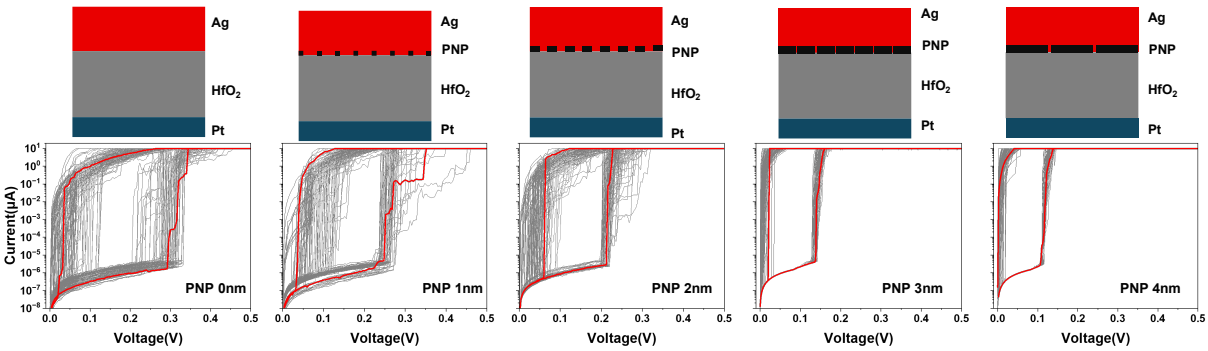
Fig. 1 I–V uniformity with varying PNP thickness (Ag devices) 2
Fig. 2 I–V uniformity with varying PNP thickness (Cu devices) 3
Fig. 3 Threshold-voltage distributions with varying PNP thickness 4
Fig. 4 I–V uniformity with varying HfO_2 deposition temperature 5
Fig. 5 Switching probability enhancement 6
Fig. 6 Endurance and relaxation stability 7
Fig. 7 Device-area dependence of threshold switching 8
Fig. 8 Two types of relaxation behaviour 9
Fig. 9 Compliance-current-dependent relaxation time 10
Fig. 10 TEM analysis of PNP morphology 11
Fig. 11 Lamella preparation for in situ TEM biasing 12
Fig. 12 Electron energy loss spectroscopy mapping 13
Fig. 13 Charge transfer and interfacial built-in electric field 14
Fig. 14 Lattice mismatch at Ag(111)/Pt(111) interfaces 15
Fig. 15 In situ STEM observation of Ag filament formation 16
Fig. 16 Time surface generation 17
Fig. 17 Array packaging and acquisition-board assembly 18
Fig. 18 Hardware workflow of the HOTS implementation 19
Fig. 19 Two relaxation windows for time surface generation 20
Fig. 20 HOTS classification with optimized and control devices 21
Fig. 21 HOTS robustness against relaxation time variability 22
Fig. 22 Voltage- and current-controlled sweeps 23

Supplementary Notes

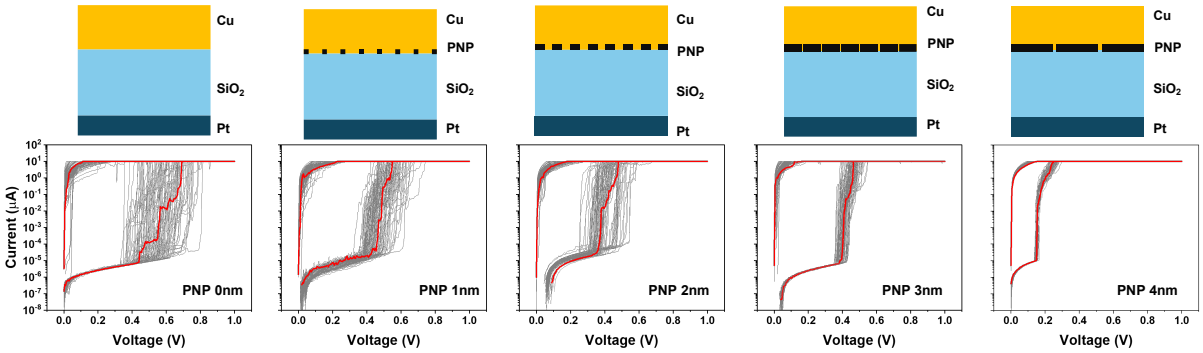
Note 1 Relaxation kinetics and compliance-current regulation 24
Note 2 Lattice mismatch at Ag(111)/Pt(111) interfaces 26
Note 3 Energy comparison for HOTS hardware inference 27

Supplementary Tables

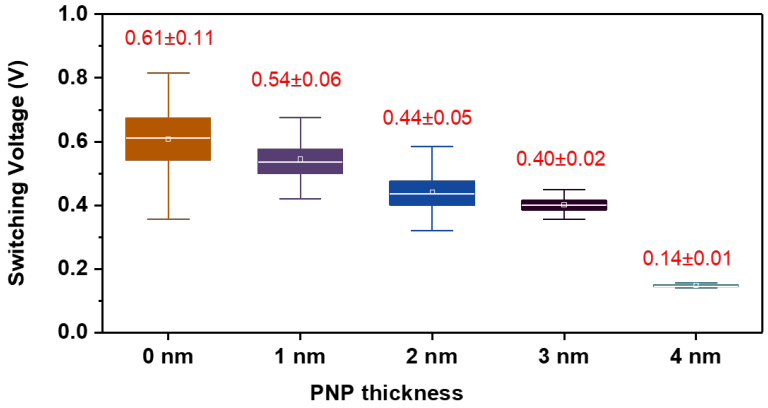
Table 1 Relaxation-time benchmark of volatile switching devices 34
Table 2 Simulated N-MNIST recognition accuracy 36
Supplementary References 37



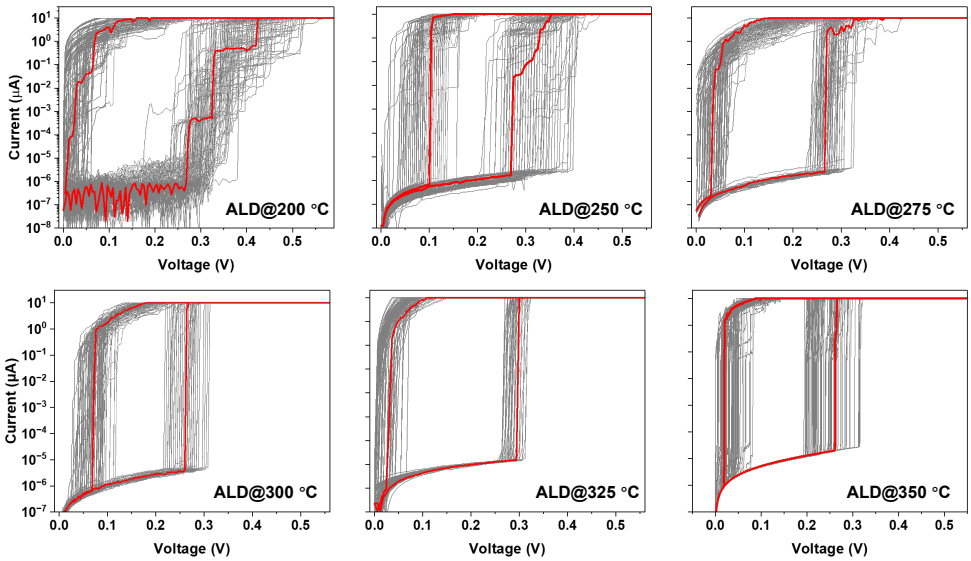
Supplementary Figure 1 | Uniform *I-V* curves of Ag(20 nm)/α-HfO_2(5 nm)/Pt(15 nm) memristors with varying PNP layer thickness. Overlaid 100 consecutive switching cycles of the Ag/PNP/α-HfO_2/Pt memristors with different thickness of the PNP layer, highlighted by the averaged curve.



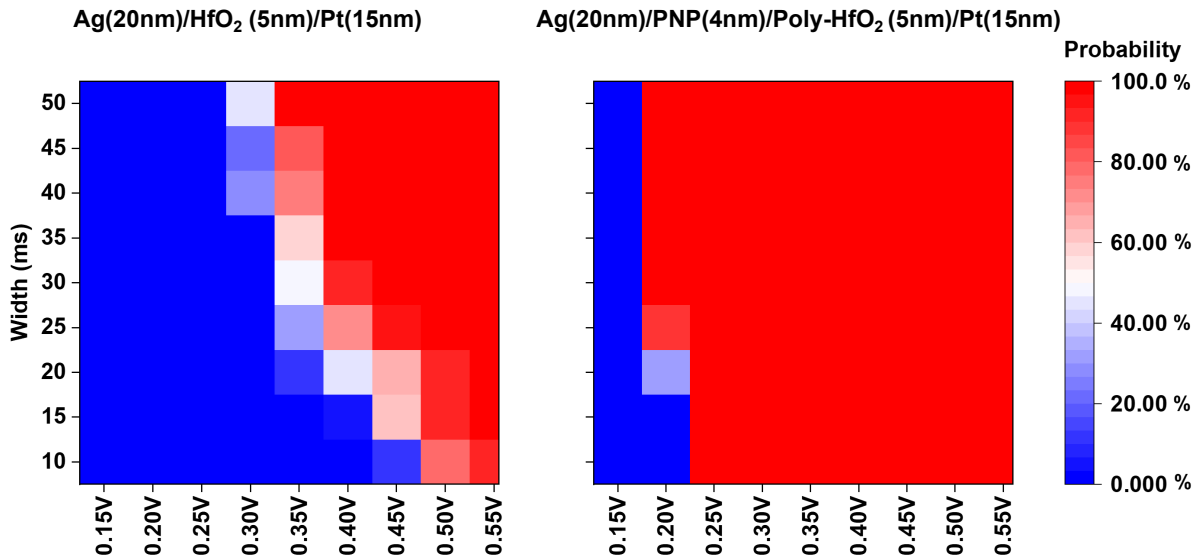
Supplementary Figure 2 | Uniform *I-V* curves of Cu(20 nm)/PNP/SiO_2(7 nm)/Pt(15 nm) memristors with varying PNP layer thickness. Overlaid 100 consecutive switching cycles of the Cu/PNP/SiO_2/Pt memristors with different thickness of the PNP layer, highlighted by the averaged curve.



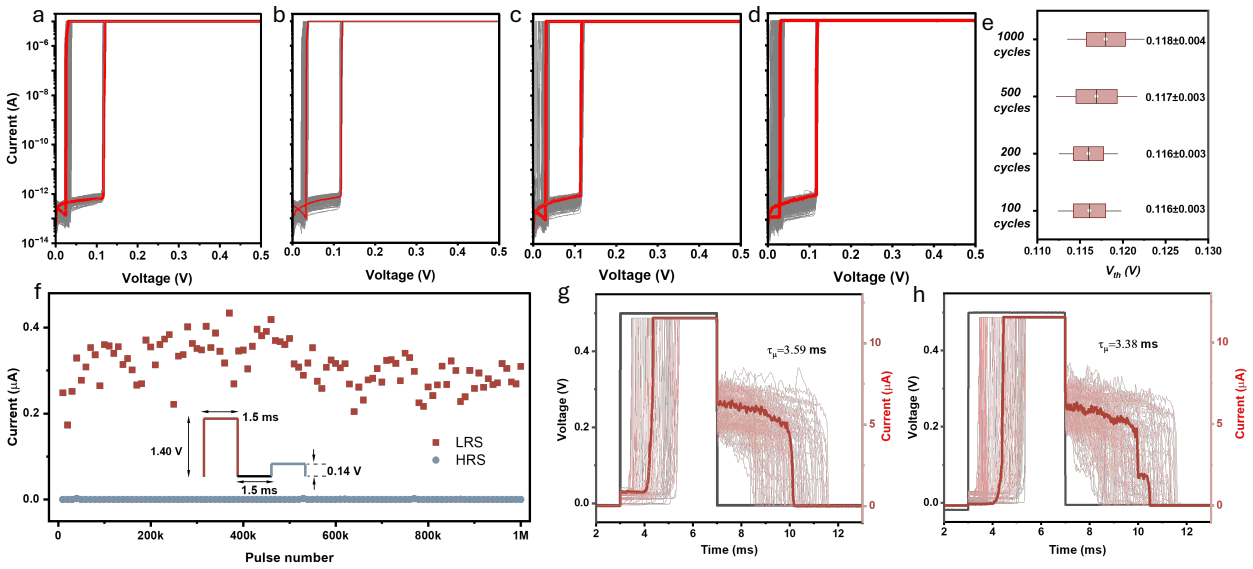
Supplementary Figure 3 | Statistical distribution of V_{th} of Cu(20 nm) /PNP/SiO_2(7 nm)/Pt(15 nm) memristors with varying PNP layer thicknesses. The mean ± SD values decrease gradually from 0.61 ± 0.11 V (no PNP) to 0.54 ± 0.06 V, 0.44 ± 0.05 V, 0.40 ± 0.02 V, and 0.14 ± 0.01 V at 4 nm PNP, indicating both a lower switching voltage and a tighter distribution with increased PNP thickness.



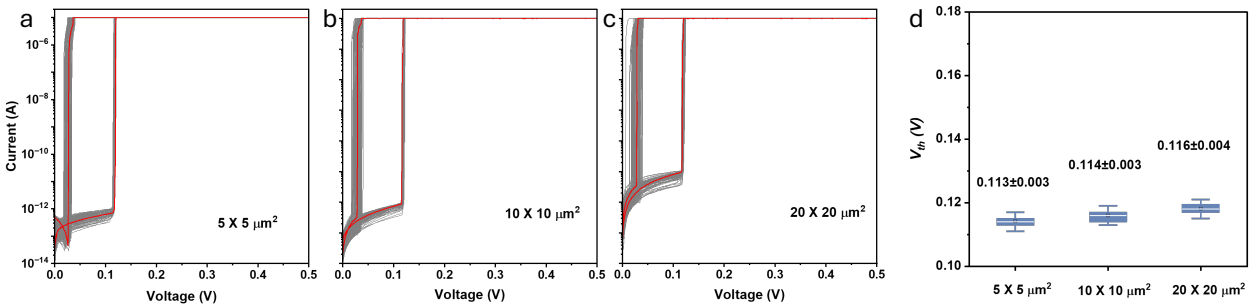
Supplementary Figure 4 | Uniform *I-V* curves of Ag(20 nm)/poly-HfO_2(5 nm)/Pt(15 nm) memristors with varying ALD temperature. Overlaid 100 consecutive switching cycles of the Ag/poly-HfO_2/Pt memristors were highlighted with the averaged curves. The switching slope decreases from 31.36 ± 11.67 mV dec^{-1} for HfO_2 deposited at 200 °C to 1.21 ± 0.98 mV dec^{-1} at 350 °C



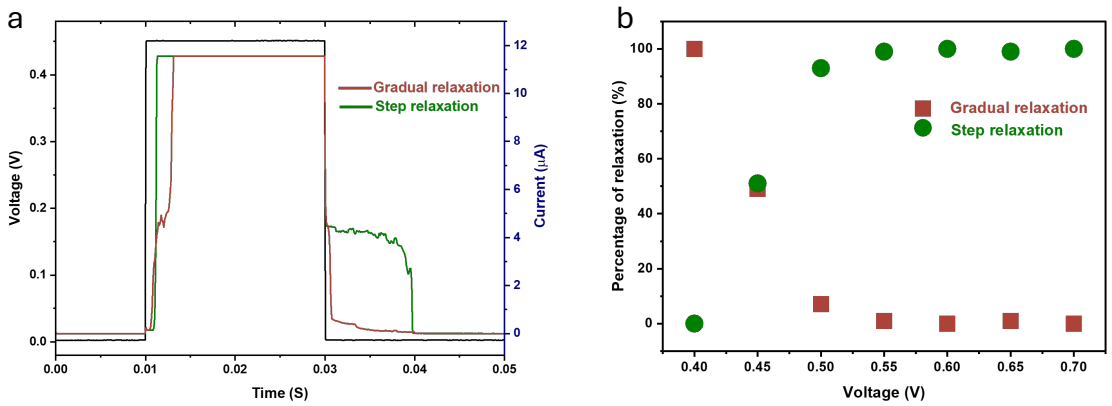
Supplementary Figure 5 | Switching probability enhancement. Switching-probability maps of (left) the control memristor (Ag/α-HfO_2/Pt) and (right) the optimized device (Ag/PNP/poly-HfO_2/Pt) as a function of input-pulse amplitude and pulse width. The control device exhibits a broad, probabilistic switching zone with weak dependence on programming conditions. In contrast, the optimized memristor shows an abrupt and deterministic transition boundary, indicating strongly confined filament nucleation enabled by the PNP layer and grain-boundary-guided ion transport.



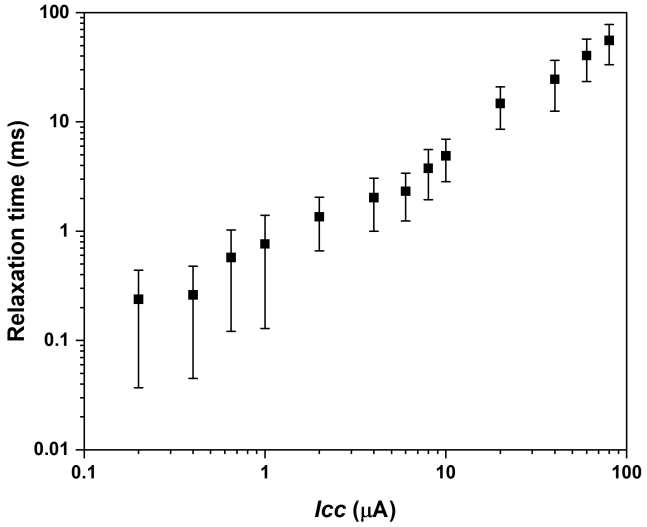
Supplementary Figure 6 | Endurance and relaxation stability of the optimized diffusive memristor. a–d, Overlaid I–V characteristics of the Ag/PNP/poly-HfO_2/Pt device measured over consecutive voltage sweeps, showing stable threshold switching from 100(a) to 200(b), 500(c) and 1000(d) cycles. **e,** Statistical distributions of threshold voltage V_{th} extracted from the I-V sweeps in a-d, indicating no obvious drift and narrow variation during extended cycling. **f,** Pulse endurance measurement of the device connected in series with a 2 MΩ resistor to limit the switching current to approximately 300 nA. The insert shows the pulse waveform used for endurance testing. **g and h,** One hundred relaxation curves measured after 500K switching cycles (g) and 1M switching cycles (h), confirming that the volatile relaxation behaviour is preserved after long-term operation.



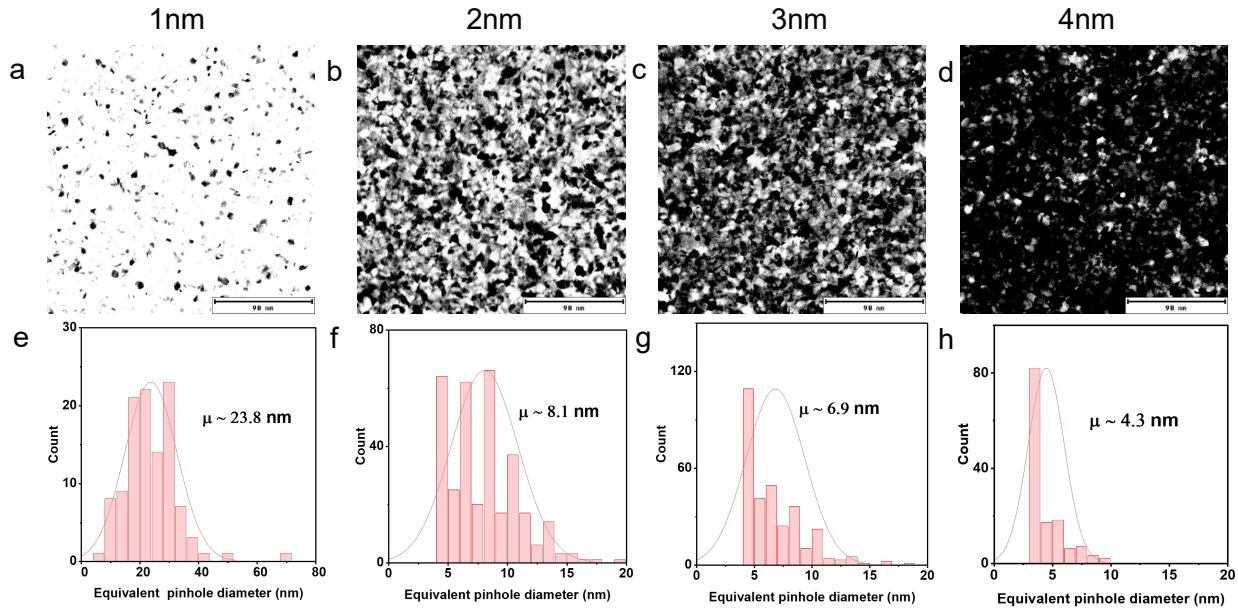
Supplementary Fig. 7 | Device-area dependence of threshold switching in Ag/PNP/poly-HfO_2/Pt diffusive memristors. a-c, Overlaid I-V characteristics of devices with junction areas of 5 × 5, 10 × 10 and 20 × 20 μm^2, respectively. **d,** Statistical distributions of threshold voltage for the three device areas, showing negligible area dependence of V_{th}.



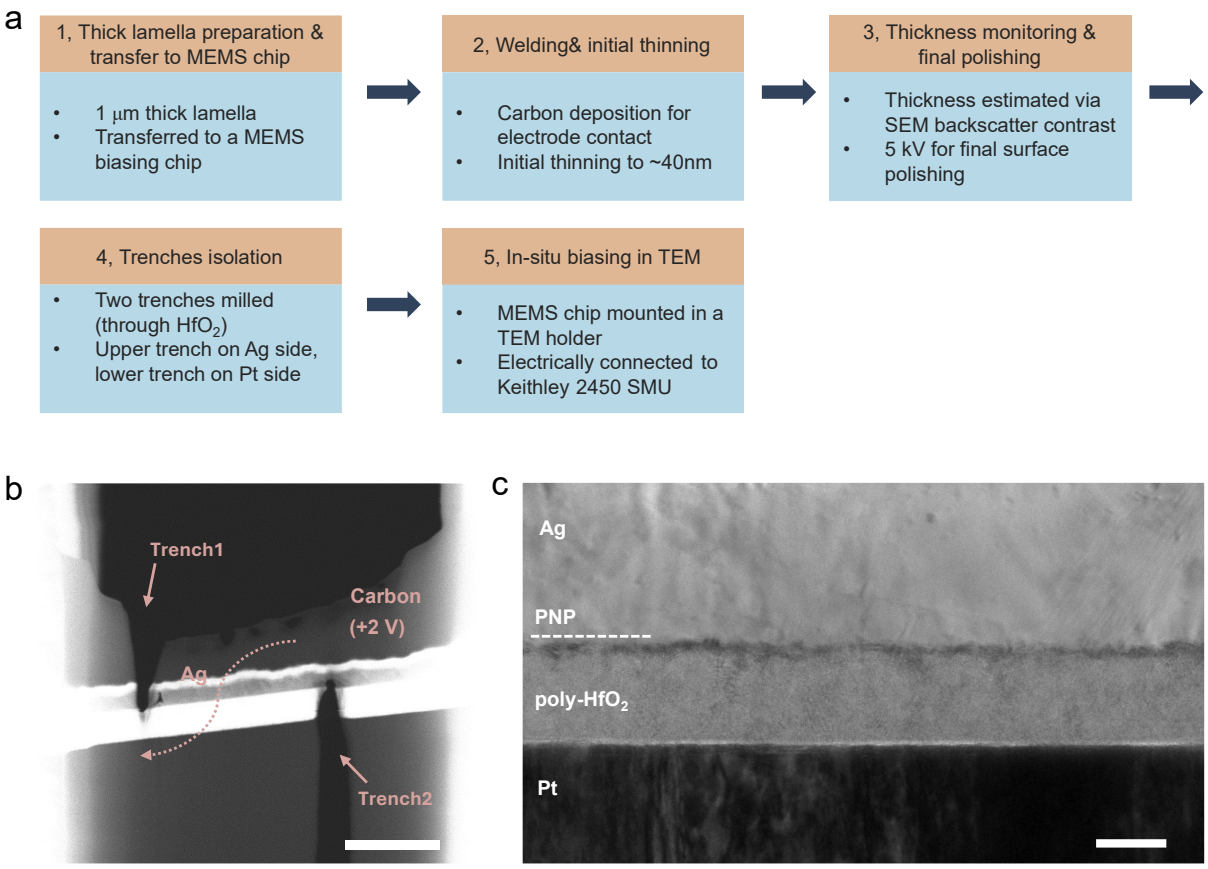
Supplementary Figure 8 | Analysis of two types of relaxation behaviour. a, The two representative current responses of the step relaxation and gradual relaxation with I_{CC} = 11.5 μA (voltage pulses 0.45 V for 20 ms, black curve). **b**, the transition between the step relaxation and gradual relaxation with increasing of pulse voltage. The absolute step relaxation can be achieved by modulating input pulse voltage. Each condition is extracted from 100 repetitive voltage pulses.



Supplementary Figure 9 | Compliance-current-dependent tuning of relaxation time. Relaxation time of the Ag/PNP/poly-HfO_2/Pt diffusive memristor as a function of effective compliance current I_{cc}. Current-limited conditions were obtained using calibrated series resistors, and the effective I_{cc} was extracted from the measured switching current. The relaxation time increases systematically with I_{cc}, spanning from sub-millisecond to tens of milliseconds.

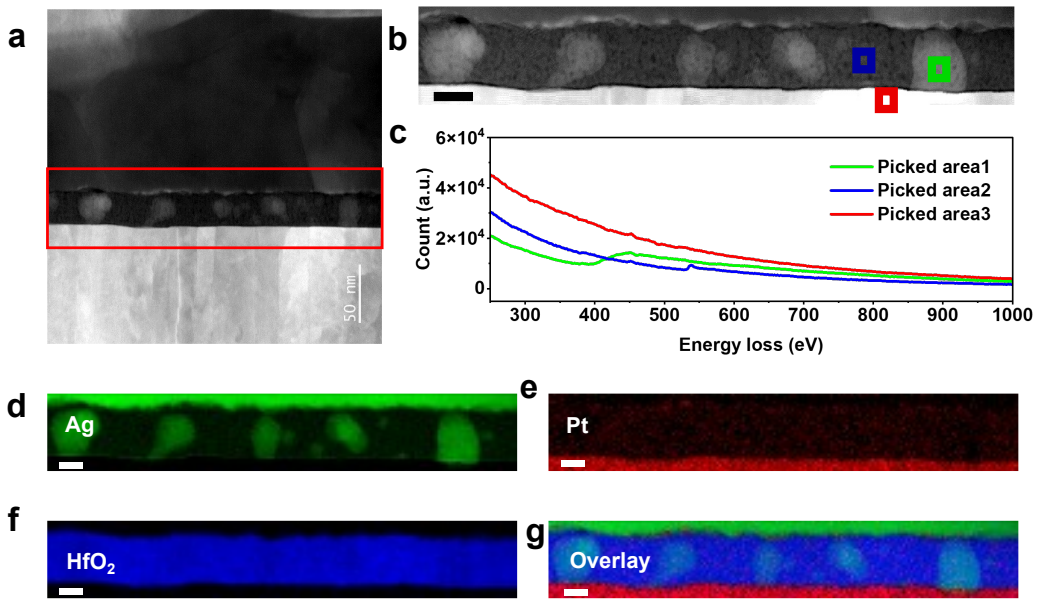


Supplementary Figure 10 | Quantitative TEM analysis of PNP morphology with varying Pt thickness. a-d, TEM images of PNP layers with increasing Pt thickness deposited on HfO_2, showing the evolution from isolated Pt islands to a percolating network and finally to a nearly continuous Pt film. Scale bars, 90 nm. **e-h,** Corresponding distributions of pinhole equivalent diameter extracted from TEM images in a-d.

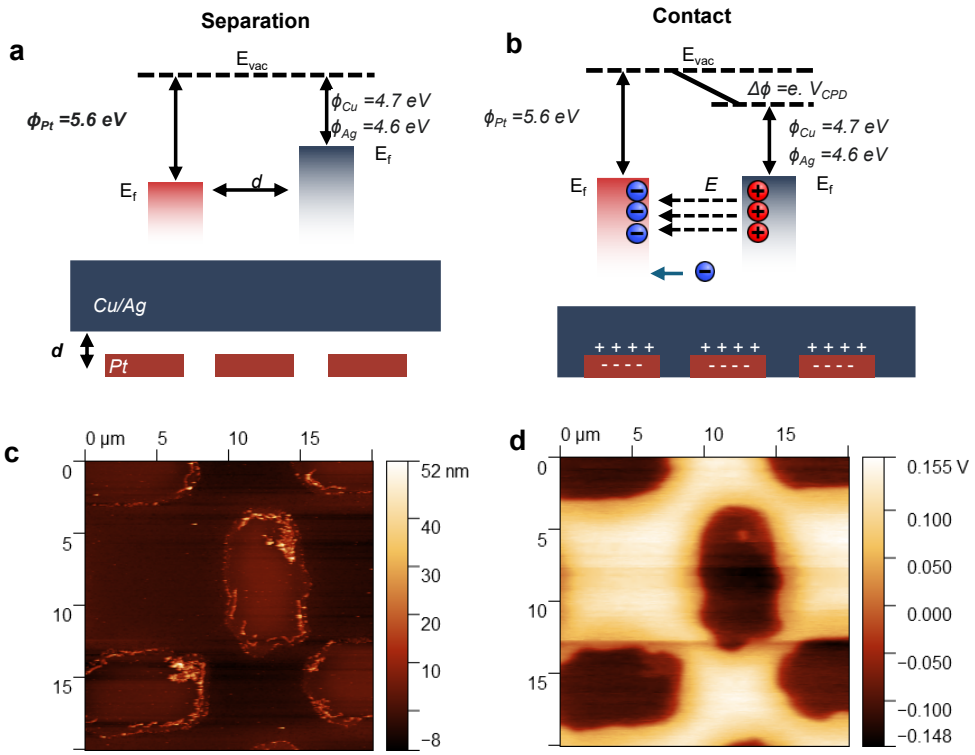


Supplementary Figure 11 | Lamella preparation protocol for *in-situ* TEM biasing.

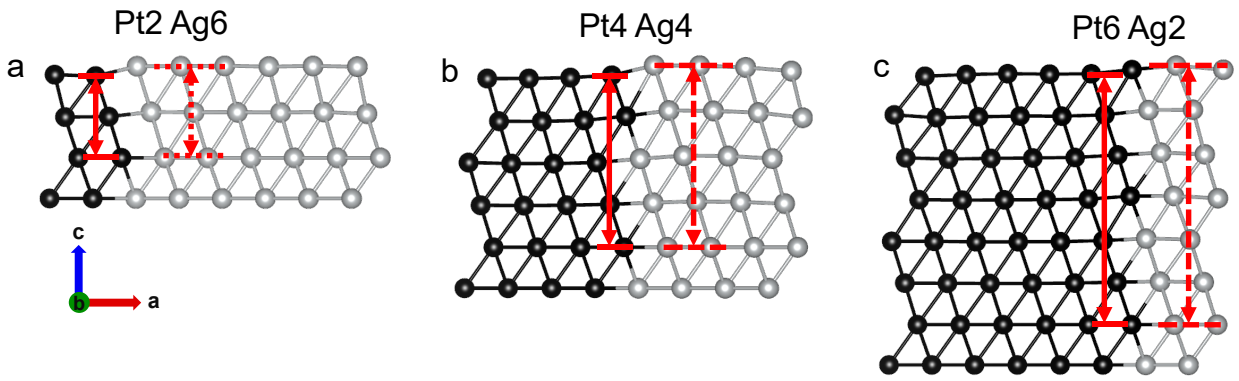
a, Schematic workflow illustrating the FIB lift-out and transfer process onto a MEMS biasing chip. The procedure involves transferring a ~1 μm thick lamella, welding and progressive thinning to electron transparency (~40 nm), followed by a trench isolation step to electrically separate the top (Ag) and bottom (Pt) electrodes. **b,** SEM image of the fabricated lamella ready for *in situ* testing. The pink dashed arrow indicates the direction of the electric field and Ag cation migration path under a positive bias (+2 V). The device is mechanically supported and electrically contacted via deposited carbon. Scale bar, 500 nm. **c,** Cross-sectional TEM image for Ag/PNP/poly-HfO_2/Pt stack, the line marks the PNP at the Ag/HfO_2 interface. Scale bar: 20 nm.



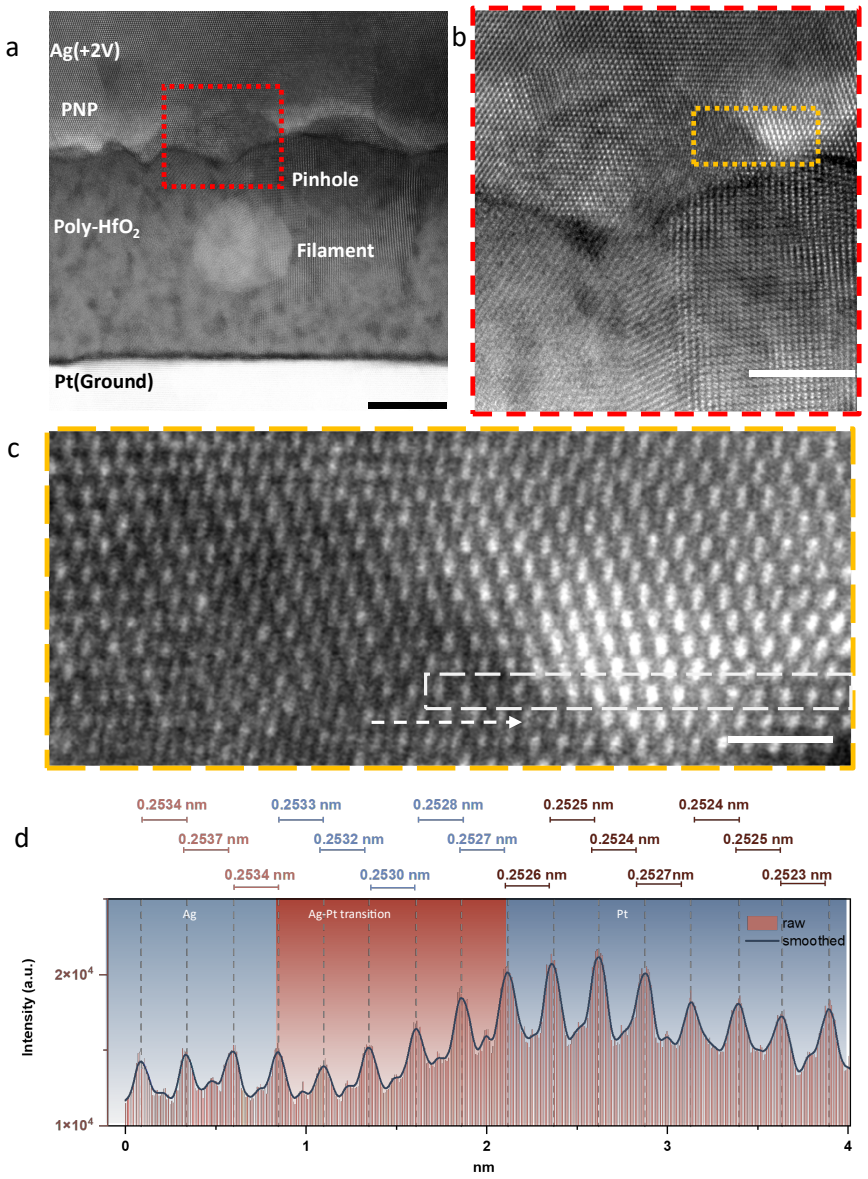
Supplementary Figure 12 | Electron energy loss spectroscopy mapping of the Ag/PNP/poly-HfO_2/Pt region. **a,** HAADF-STEM overview of the cross-section. The red box marks the interface used for EELS analysis. Scale bar, 50 nm. **b,** Magnified view of the boxed region, colored squares mark three sampling positions, scale bar: 10nm. **c,** EELS spectra from the picked areas. Intensity differences across the Ag-M, Pt-N, and O-K/Hf-M edges distinguish Ag inclusions from the oxide matrix. **d–f,** Elemental maps extracted from the EELS data. **d,** Ag distribution, green; **e,** Pt distribution, red; **f,** HfO_2 distribution, blue; **g,** RGB overlay confirms Ag filaments are formed inside HfO_2 matrix. The scale bars for d-g are 10 nm.



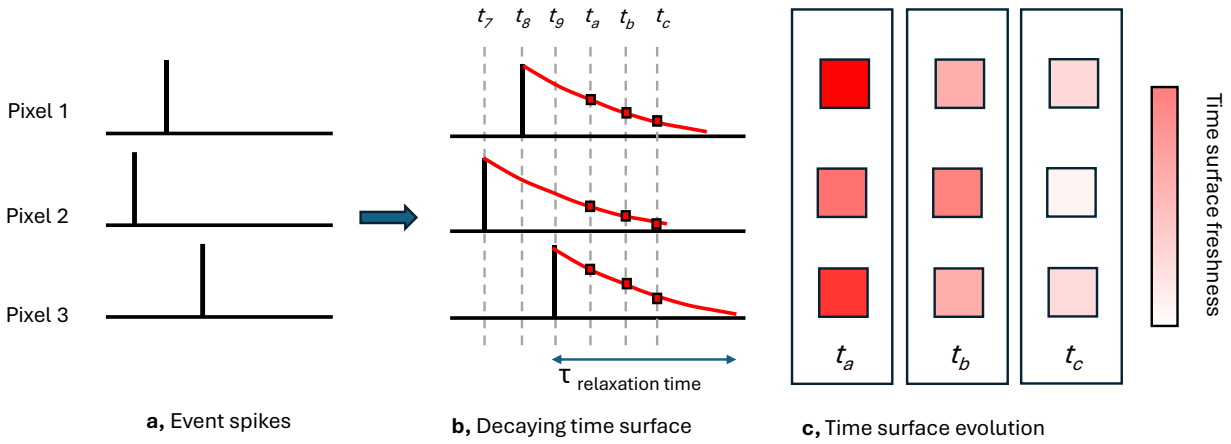
Supplementary Figure 13 | Charge transfer and build-in electric field generation at interfacial layer. a and b, Schematic energy band diagrams of the Cu/Ag electrode and Pt interfacial layer in the separated state (a) and in electrical contact (b). The higher work function of Pt (ϕPt = 5.6 eV) compared to Cu (ϕCu = 4.7 eV) and Ag (ϕAg = 4.6 eV) induces electron transfer upon contact. This equilibration creates a dipole layer and a built-in electric field, resulting in theoretical contact potential differences (CPD) of ~0.9 V for Pt/Cu and ~1.0 V for Pt/Ag. **c,** AFM topography of Pt sputtered on an ultra-smooth Cu substrate, showing PNP features. **d,** KPFM map of the same area reveals CPD contrast between PNP and contacted Cu regions.



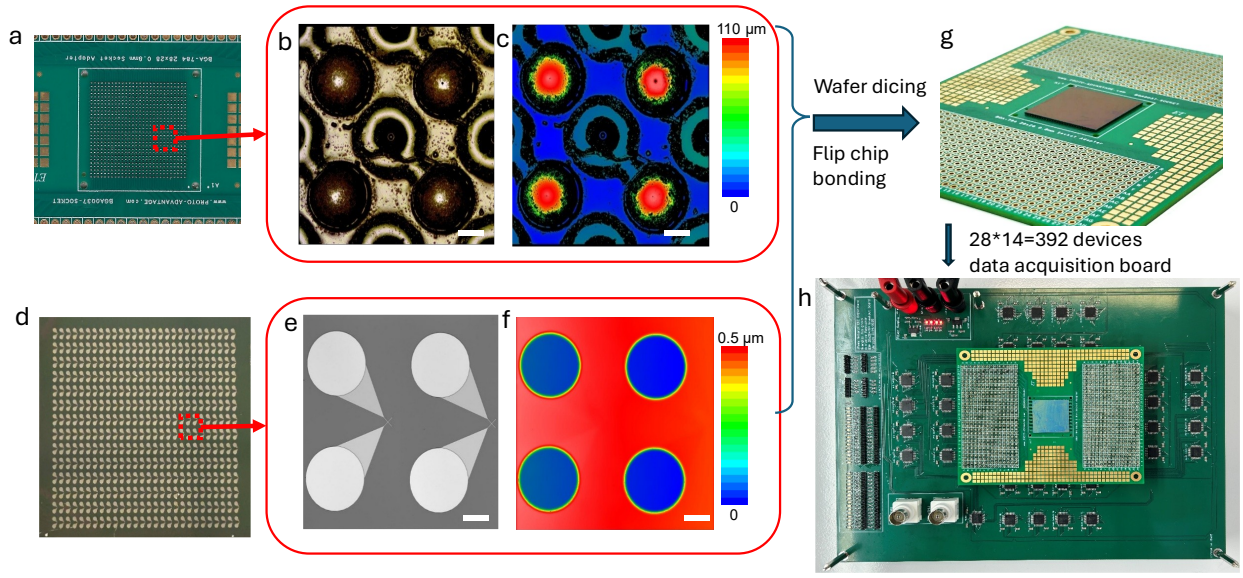
Supplementary Figure 14 | Calculation of lattice mismatch at Ag(111)/Pt(111) interfaces with increasing Pt coverages. Schematics of the computational models: **a,** Pt2Ag6; **b,** Pt4Ag4 and **c,** Pt6Ag2, showing the c interlayer distance. The interfacial volumetric lattice mismatch was calculated to be -18.2%, -21.1%, and -23.6% for Pt2Ag6, Pt4Ag4, and Pt6Ag2, respectively.



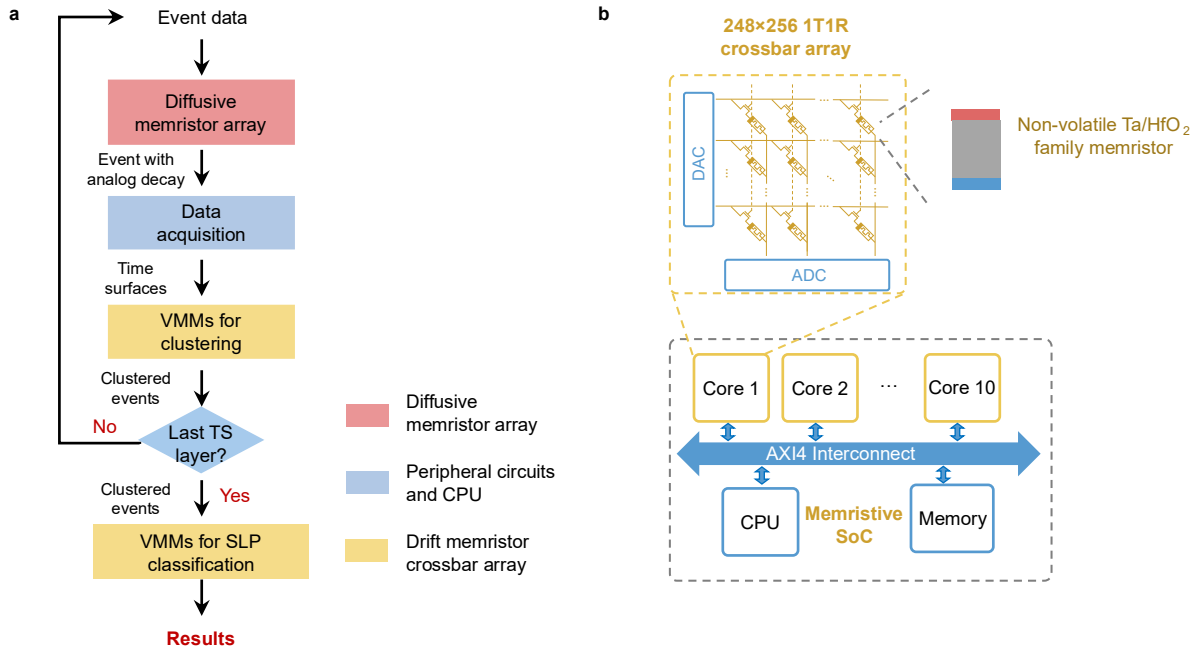
Supplementary Figure 15 | *In situ* STEM observation of Ag filament formation and local lattice structure near a PNP opening. a, STEM images of the pinhole area after applying +2 V to the Ag electrode, with the bottom Pt electrode grounded. An Ag filament forms beneath an opening in the PNP. **b,** Magnified view of the red-boxed region in **a**. **c,** Atomic-resolution image of the yellow-boxed Ag/Pt interfacial region in **b**. The white rectangle and arrow indicate the sampling region and direction for the intensity profile. **d,** Corresponding raw and smoothed intensity profiles across the Ag-rich, Ag/Pt interfacial, and Pt regions. Annotated values indicate the projected spacings between adjacent intensity peaks. Scale bars, 10 nm (**a**), 5 nm (**b**), and 1 nm (**c**).



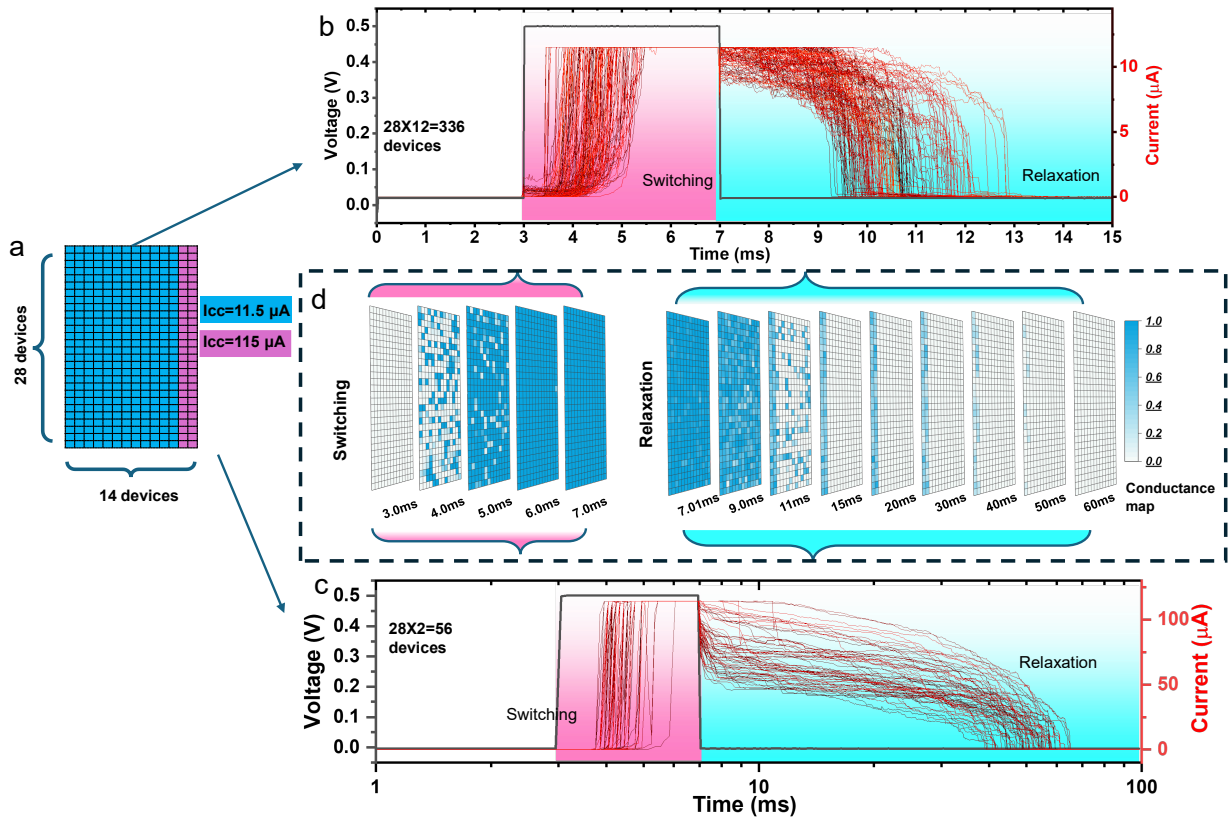
Supplementary Figure 16| Schematic of time surface generation. **a,** asynchronous event spikes. **b**, transformation into decaying time surfaces with relaxation time τ. **c,** resulting time surface at each sampling time, where color intensity represents the freshness of the event.



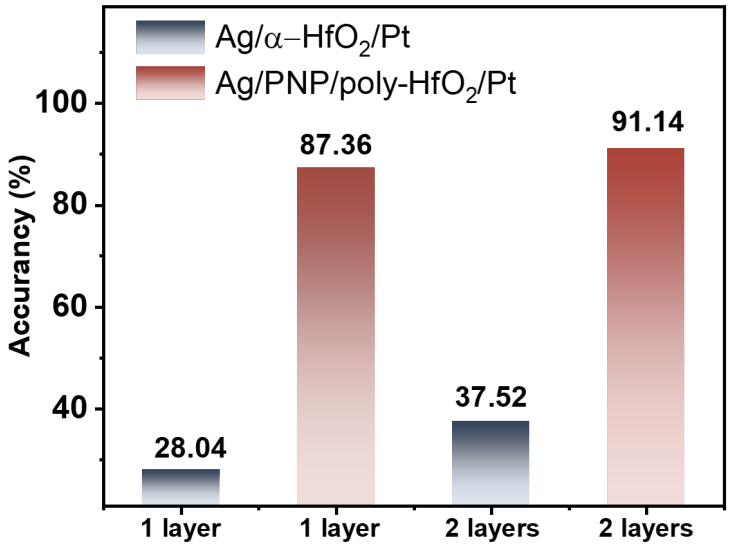
Supplementary Figure 17 | Packaging and assembling diffusive memristor array with the customized acquisition board. a, BGA packaged breakout board. **b-c**, Optical and corresponding profiler images of BGA solder-ball array for electrical connection. scale bar, 50 μm. **d**, Optical image of a diced diffusive memristor array. **e**, SEM image of individual Ag (20 nm)/PNP (4 nm)/poly-HfO_2 (5 nm)/Pt (15 nm) devices. Scale bar, 20 μm. **f**, Laser profiler contour map of the pad area. Scale bar, 20 μm. **g**, Packaged die mounted on the BGA breakout board after wafer dicing and flip-chip bonding. **h**, Assembled acquisition circuit integrating with the packaged die and peripheral electronics for readout of 392 diffusive memristors.



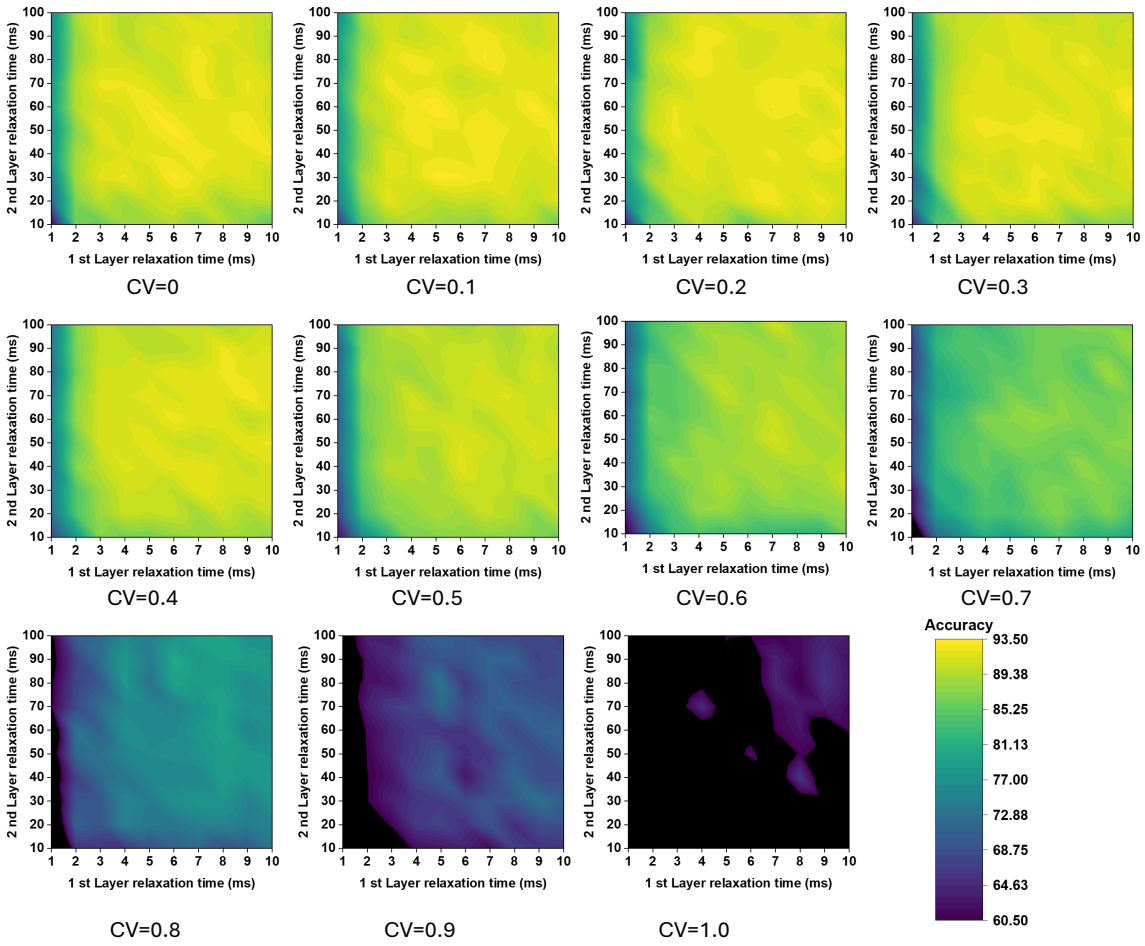
Supplementary Figure 18 | Hardware workflow of the HOTS implementation. a, Hardware implementation flow of the HOTS algorithm using diffusive memristor arrays and drift memristor crossbar arrays. The binary event data are fed into the diffusive memristor array to generate the time surfaces. The generated time surfaces are collected by the peripheral circuits and then transmitted to a memristive SoC for clustering, in which the core operations are vector-matrix multiplications (VMMs). The same process is repeated for each layer, while the clustered time surfaces are ultimately classified by a single-layer perceptron (SLPs), whose primary operations are also VMMs. **b,** A memristive system-on-a-chip (SoC) integrating drift memristor crossbar arrays, a CPU, and peripheral circuits. The SoC includes ten computing cores for analog VMMs (8 bit precision), one RISC-V CPU for general-purpose functions, and a data bus for on-chip data transmission. Each computing core includes a 248 × 256 one-transistor–one-memristor (1T1R) crossbar array and essential peripheral circuits, such as DACs and ADCs. The memristor at each crosspoint of the array is a nonvolatile Ta/HfO_2 family drift memristor.



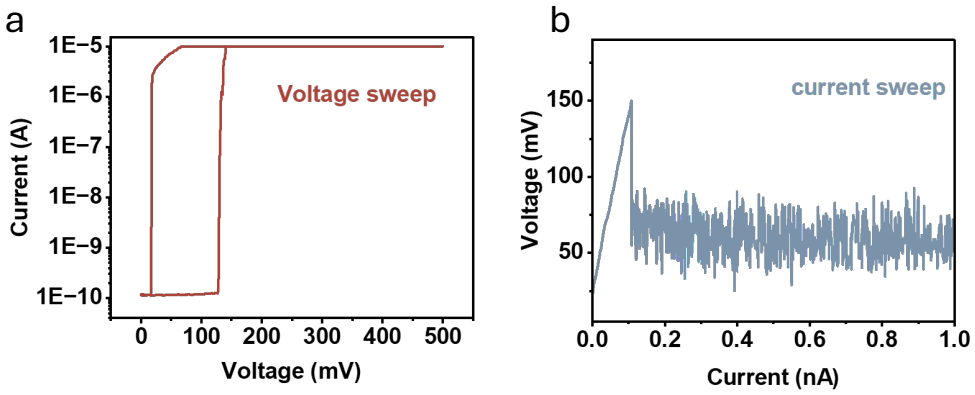
Supplementary Figure 19| Programming the array into two relaxation windows for two-layer time-surface generation. a, the 28×14 array (392 devices) is partitioned into two groups by compliance current. **b,** 28×12 = 336 devices programmed at I_{cc}= 11.5 μA (0.5 V, 4 ms) to realize the short-decay (1st layer) window. **c,** 28×2 = 56 devices programmed at I_{cc} = 115 μA (0.5 V, 4 ms) to realize the long-decay (2nd layer) window. **d,** the time-lapse conductance maps confirm synchronous switching and indicate two well-separated relaxation bands (~3–6 ms and ~35–50 ms).



Supplementary Figure 20| Comparison of HOTS classification using the optimized Ag/PNP/poly-HfO_2/Pt device and the Ag/α-HfO_2/Pt control device. The same HOTS pipeline was used for both devices. The optimized device achieves 91.14% classification accuracy, compared with 37.52% for the control device, demonstrating the importance of uniform threshold switching and relaxation dynamics for reliable temporal processing.



Supplementary Figure 21| Robustness of HOTS classification accuracy against relaxation time variability. The evolution of classification accuracy heat maps as a function of the first (τ1) and second (τ2) layer relaxation times under increasing levels of stochasticity. The variability is quantified by the coefficient of variation (CV, defined as σ/μ) of the relaxation time, ranging from the ideal case (CV=0) to extreme variability (CV=1.0). The high-accuracy region (yellow, >90%) remains extensive and largely preserved up to CV ≈ 0.6. This demonstrates that the HOTS algorithm possesses inherent tolerance to device-to-device and cycle-to-cycle fluctuations. Significant performance degradation is only observed when the variability reaches unrealistic levels (CV > 0.7), which far exceeds the intrinsic variability of our optimized Ag/PNP/poly-HfO_2 devices (CV=0.253).



Supplementary Figure 22 | Comparison of voltage- and current-controlled sweeps. a, Representative *I-V* curve via voltage-controlled sweep on an optimized memristor (Ag/PNP/poly-HfO_2/Pt). **b**, Representative *V-I* curve via current-controlled sweep on the same device. Comparison between voltage- (Fig. S22a) and current-controlled (Fig. S22b) sweeps highlights the controllability of current regulation. Voltage biasing results in abrupt switching driven by positive thermal feedback, making intermediate filament states unstable. In contrast, the current sweep reveals negative differential resistance (NDR) with continuous evolution.

Supplementary Note 1 | Theoretical modeling of relaxation kinetics and compliance current regulations.

This section provides a comprehensive physical and mathematical model to elucidate the mechanism by which compliance current (I_{cc}) governs the relaxation time (τ_{relax}) in our Ag/PNP/poly-HfO_2/Pt diffusive memristors. While previous studies have separately discussed filament growth[1] and instability[2], here we integrate these frameworks to derive a unified model that quantitatively explains the precise tunability and enhanced uniformity observed in our experiments.

Based on the experimental data, τ_{relax} exhibits a non-linear increase with I_{cc} but is less sensitive to pulse width (t_{pulse}). This identifies I_{cc} as the primary extrinsic control parameter regulating the filament's initial geometry. Fundamentally, the dissolution of the Ag filament is governed by surface-diffusion-driven Rayleigh-Plateau instability. A higher I_{cc} induces a larger and more stable filament, which consequently takes longer to undergo spontaneous breakdown. The dynamics are determined by the interplay of three key processes:

1, Filament nucleation and growth (thermally activated ion hopping).

2, Self-limiting growth via negative feedback (controlled by I_{cc}).

3, Filament dissolution (Relaxation physics).

I. **Filament nucleation and growth model.** The rate of change of the filament diameter is described as a thermally activated ion hopping process. The probability of an ion overcoming the energy barrier to migrate follows an Arrhenius relationship, modulated by the electric field:

$$\frac{d\phi}{dt} = A \cdot \exp\left(-\frac{E_A}{k_B T}\right) (1)$$

Where A is the pre-exponential factor. k_B is the Boltzmann constant, and T is the local temperature of the filament; E_A is lowered by the electric field (Poole-Frenkel-like effect):

$$E_A = E_{A0} - \alpha q V (2)$$

Substituting (2) into (1) yields the growth rate equation:

$$\frac{d\phi}{dt} = A \cdot \exp\left(-\frac{E_{A0} - \alpha q V}{k_B T}\right) (3)$$

The local temperature of the filament can be expressed as:

$$T = T_0 + \frac{V^2}{8\rho k_{th}} (4)$$

Where k_{th} is the thermal conductivity and ρ is the conductivity. The local temperature of the filament depends only on the applied voltage V across the filament, independent of the filament's diameter.

II. Negative feedback and self-limiting growth controlled by *Icc*. The self-limiting growth fundamentally arises from a negative feedback loop, governed by *Icc*. Initially, the high voltage across the high-resistance device drives rapid nucleation. However, as the filament connects and resistance drops, the series resistance from the measurement circuit (or transistor in an array) absorbs the voltage drop. This limits the effective voltage across the memristor once the current hits the *Icc* limit. Since the growth rate (Eq. 3) is exponentially dependent on voltage (V) and temperature ($T \propto V^2$), this voltage reduction instantly suppresses further radial expansion, locking the filament radius at an equilibrium value primarily set by *Icc*.

III. Relaxation Physics via Rayleigh-Plateau Instability. Once the external voltage is removed, this filament is thermodynamically unstable because of the Gibbs free energy minimization via surface diffusion, leading to a rupture of a conductive filament.
The breakup kinetics were rigorously solved by Nichols and Mullins in 1965[2].

$$\tau \approx \frac{k_B T \cdot r^4}{B \cdot D_s \cdot \gamma \cdot \Omega^{4/3}} (5)$$

Where r is the filament radius; B is a dimensionless geometric factor that depends on the wavelength of the disturbance. D_s is the surface diffusion coefficient of Ag atoms. γ is the surface energy of the Ag/HfO_2 interface; Ω is the atomic volume of Ag and k_BT is the thermal energy. Ignoring the influence of the constant term, we can express relaxation time as:

$$\tau \propto r^4 (6)$$

Hence, it is the result of the interaction between the circuit's voltage division feedback and Arrhenius exponential dynamics, the feedback loop automatically leads to a self-limiting growth defined by *Icc*. Correspondingly, *Icc* serves as a robust control method to regulate the relaxation time τ.

Supplementary Note 2 | Lattice mismatch calculation at Ag(111)/Pt(111) interfaces in three models

To quantify the compressive strain of Ag deposited close to Pt, we first find the DFT-optimized unit-cell volume of pure Ag to be $V_0 = (2.892\ \text{Å})^3$. For each model, the effective Ag lattice parameter along the c-direction is obtained by averaging the two Ag columns adjacent to Pt at the two Ag/Pt interfaces (arising from periodic boundary conditions). This average is taken from the third atomic layer below the surface to the surface region, while the top two frozen layers are excluded from the analysis. The averaged Ag c-lattice parameters are 2.580, 2.486, and 2.408 Å for Pt2Ag6, Pt4Ag4, and Pt6Ag2, respectively. Because Ag is constrained to the Pt lattice in the in-plane directions, $a = b = 2.77$ Å, the Ag unit-cell volume in each model is $V = c \times 2.77 \times 2.77\ \text{Å}^3$, and the interfacial volume distortion can be calculated as

$$\varepsilon = \frac{V - V_0}{V_0}$$

Using this definition, the Ag layer exhibits progressively increasing compressive strain as the Pt fraction increases and the Ag layer thickness decreases, with $\varepsilon = -18.2\%$, -21.1%, and -23.6% for Pt2Ag6, Pt4Ag4, and Pt6Ag2, respectively, indicating that Pt2Ag6 is the least strained and Pt6Ag2 the most strained configuration.

Supplementary Note 3 | Energy comparison of hardware implementation for the inference of HOTS algorithm

To demonstrate the energy efficiency of the time surface generation to conventional digital platforms, we estimate the energy consumption of the diffusive memristor array, and compare the digital decay generation on Intel Hala Point, a state-of-the-art neuromorphic hardware platform built on the 4 nm CMOS node.

$$\frac{E_{\text{decay_total}}}{E_{\text{TS_diffusive}}} = 11.65\,\mu\text{J} / 533.1\,n\text{J} \approx 21.9\times$$

where the $E_{\text{decay-total}}$ denotes the estimated energy consumption of the decay generation implemented using Intel Hala Point, whereas $E_{\text{TS-diffusive}}$ denotes the estimated energy of time surface generation using diffusive memristor array.

To demonstrate the energy efficiency of the proposed system relative to conventional digital platforms, we estimate the energy consumption of the two-layer HOTS algorithm implemented on our analogue hardware platform comprising a diffusive memristor array and a memristive SoC fabricated with the 65 nm CMOS node integrating drift memristor crossbar arrays, and compare it with that of Intel Hala Point, The system-level energy reduction enabled by the proposed hardware is

$$\frac{E_{\text{total_digital}}}{E_{\text{total_analog}}} = 14.606\,\mu\text{J} / 4.040\,\mu\text{J} \approx 3.62\times$$

where the $E_{\text{total_digital}}$ denotes the estimated energy consumption of the two-layer HOTS algorithm implemented using Intel Hala Point, whereas $E_{\text{total_analog}}$ denotes the estimated energy consumption of the two-layer HOTS algorithm implemented using the proposed analog hardware platform comprising the diffusive memristor array, the memristive SoC, and essential peripheral circuits. The system-level energy consumption is expected to improve further as we continue our ongoing work to implement the system by integrating memristors with advanced nodes such as 22 nm and 12 nm CMOS, and beyond.

The details of the energy consumption estimation are provided as follows.

1. **Energy estimation of the two-layer HOTS algorithm implemented on our analogue hardware platform**

As shown in panel (a) of Supplementary Figure 18, within each layer of the HOTS algorithm, the diffusive memristor array is used to generate the time surfaces. The generated time surfaces are

collected by the peripheral circuits and then transmitted to a memristive SoC for clustering, in which the core operations are VMMs. The same process is repeated for each layer, while the clustered time surfaces are ultimately classified by a single-layer perceptron (SLP), whose primary operations are also VMMs. To estimate the system-level energy consumption of the two-layer HOTS implementation using a diffusive memristor array and drift memristor crossbar arrays, we calculate the contributions from the following three major hardware blocks:

1. The diffusive memristor array that generates time surfaces through intrinsic volatile conductance relaxation.
2. A memristive SoC integrating ten computing cores with drift memristor crossbar arrays for analog VMMs.
3. Peripheral circuits for analog data acquisition and control logic.

The total energy consumption for HOTS inference can therefore be estimated as the sum of the energy consumed by these three components:

$$E_{\text{total_analog}} = E_{\text{TS_diffusive}} + E_{\text{VMM_SoC}} + E_{\text{total_periphery}}$$

- **Energy estimation of time surface generation with diffusive memristor array**

Assume that N_1 and N_2 denote the numbers of generated time surfaces in the first and second layers, respectively, for one input pattern. The total energy consumption for the time surface generation during inference of the two-layer HOTS algorithm is given by:

$$E_{\text{TS_diffusive}} = N_1 E_{\text{diff},1} + N_2 E_{\text{diff},2}$$

where $E_{\text{diff},1}$ and $E_{\text{diff},2}$, represent the energy consumption per time surface generation in the first and second layers, respectively.

The energy consumed by a diffusive memristor during the time surface generation is calculated from the measured voltage-current integral:

$$E_{\text{diff}} = \int_0^T V(t)I(t)dt$$

For the short-timescale diffusive memristors used for the time surface generation in the first layer, and programmed with $I_{\text{cc}} = 11.5\ \mu\text{A}$ under a 0.5 V, 4 ms pulse condition, the measured energy consumption per time surface generation is

$$E_{\text{diff},1} \approx 0.636\ \text{nJ}$$

For the longer-timescale diffusive memristors used for the time surface generation in the second layer and programmed with a higher compliance current, $I_{cc} = 115\ \mu A$, to achieve a longer relaxation window, the measured energy consumption per time surface generation is

$$E_{diff,2} \approx\ 4.695\ \text{nJ}$$

If we assume that each input pattern contains 100 time surfaces per layer, corresponding to $N_1 = N_2 = 100$, the estimated energy consumption of the diffusive memristors across both layers is

$$E_{TS_diffusive} = 100 \times 0.636\ \text{nJ} + 100 \times\ 4.695\text{nJ} = 533.1\ \text{nJ}$$

- **Energy estimation of clustering and classification with the memristive SoC**

As shown in panel (b) of Supplementary Figure 18, a memristive SoC integrating ten computing cores, a CPU, and essential peripheral circuits is used for the VMM operations in clustering and classification. The measured average power consumption of the memristive SoC was 11.922 mW as an integrated system. Therefore, we estimated the energy consumption in the two-layer HOTS algorithm as the product of the measured power consumption, $P_{VMM_SoC} = 11.922\ \text{mW}$, and the estimated inference time, $t_{inference}$, on the SoC.

$$E_{VMM_SoC} = P_{VMM_SoC} \times t_{inference}$$

The total inference time, $t_{inference}$, is calculated by considering both the VMM execution time and the data transfer time associated with the VMM operations required for clustering in both layers and for classification using the SLP:

$$t_{inference} = N_{VMM} \times (t_{VMM} + t_{transfer})$$

where N_{VMM} is the total number of VMM operations for the two-layer HOTS algorithm.

As each time surface is clustered using one VMM operation and the histogram-counting features after all clustering operations in both layers are classified by a SLP, which also requires one VMM operation, the total number of VMM operations can be estimated under the previous assumption that N_1 and N_2 denote the numbers of generated time surfaces in the first and second layers, respectively, for one input pattern.

$$N_{VMM} = N_1 + N_2 + 1$$

where the additional 1 corresponds to the VMM operation required for the SLP. With the same assumption that each input pattern contains 100 time-surfaces per layer, corresponding to $N_1 = N_2 = 100$, the total number of VMM operations is

$$N_{VMM} = 100 + 100 + 1 = 201$$

For the memristive SoC used in this work, one VMM operation requires 384 clock cycles when the SoC operates at 300 MHz. Thus, the VMM time is

$$t_{\mathrm{VMM}} = 384 \times \frac{1}{300\ \mathrm{MHz}} = 1.28\ \mu s$$

This duration includes the analog VMM operation and the analog-to-digital conversions associated with the on-chip peripheral circuits, including transimpedance amplifiers, ADCs, DACs, and digital function blocks.

For the transfer overhead between VMM operations, assuming that all outputs from 256 columns with 8-bit precision are transferred core-to-core through the on-chip 64-bit AXI4 data bus, the transfer time is

$$t_{\mathrm{transfer}} = \frac{256 \times 8\ \mathrm{bits}}{64\ \mathrm{bits}} \times \frac{1}{300\ \mathrm{MHz}} = 0.11\ \mu s$$

The total inference time for the VMM operations in the two-layer HOTS inference is therefore

$$t_{\mathrm{inference}} = 201 \times (1.28 + 0.11)\ \mu s = 279.39\ \mu s$$

Accordingly, under the assumption of 100 time surfaces per layer, the total energy consumption of the two-layer HOTS algorithm is

$$E_{\mathrm{VMM_SoC}} = 11.922\ \mathrm{mW} \times 279.39\ \mu s = 3.331\ \mu \mathrm{J}$$

- **Energy estimation of peripheral circuits for data acquisition**

To capture the time surfaces generated by the diffusive memristor array, peripheral circuits, including ADC, analog switches, and digital control circuits are required. For each input event, one event pulse is applied to the corresponding diffusive memristor. Then the analog switches connect the diffusive memristors within the local time surface window for ADC digitization, after which digitalized data are collected and transmitted to the memristive SoC for clustering.

The two-layer HOTS algorithm is configured with 7 × 7 and 3 × 3 time-surface windows for the first and second layers, respectively. Therefore, 49 ADC conversions are required for each event in the first layer, whereas 9 ADC conversions are required for each event in the second layer. Under the assumption that $N_1 = N_2 = 100$, the total number of ADC conversions is

$$N_{ADC} = 100 \times 49 + 100 \times 9 = 5800$$

For the estimation of ADC energy consumption, we use the same ADC architecture adopted in the memristive SoC, where the measured ADC power consumption is 3.45 mW for 256 ADCs within a single computing core. Since ADC conversion accounts for most of the execution time in one VMM operation, we further assume that the ADC conversion time is approximately equal to t_{VMM}. For the estimation of energy consumption associated with the analog switches and digital control

circuits, we use the measured power consumption of the analog routing circuit and digital control circuit in the memristive SoC to approximate the power consumption of the analog switches and digital control circuits in the proposed system. The measured power consumption of the 256-column analog routing circuit and digital control circuit is 0.65 mW and 1.98 mW, respectively. Since the analog switches operate simultaneously with ADC readout, their operating duration is assumed to be equal to the ADC conversion time. Therefore, the energy consumed per event pulse, including the analog switch, ADC conversion, and digital control circuit operations, is estimated as

$$E_{\text{periphery}} = P_{\text{periphery}} \times t_{\text{VMM}} = \frac{3.45\ \text{mW} + 0.65\ \text{mW} + 1.98\ \text{mW}}{256} \times 1.28\ \mu\text{s} = 30.4\ \text{pJ}$$

Accordingly, the total energy consumption of the peripheral circuits for data acquisition is

$$E_{\text{periphery_total}} = E_{\text{periphery}} \times N_{ADC} = 30.4\ pJ \times 5800 = 176.32\ \text{nJ}$$

Considering the energy consumption of all three components involved in the hardware implementation of the two-layer HOTS algorithm, the total estimated energy consumption is

$$E_{\text{total_analog}} = E_{\text{TS_diffusive}} + E_{\text{VMM_SoC}} + E_{\text{periphery_total}}$$

$$= 533.1\ \text{nJ} + 3.331\ \mu\text{J} + 176.32\ n\text{J} = 4.040\ \mu\text{J}$$

2. Energy estimation of hardware implementation in Intel Hala Point

Since all computations are implemented digitally in Intel Hala Point, the total energy consumption of the digital implementation is

$$E_{\text{total_digital}} = E_{\text{decay_total}} + E_{\text{VMM_total}}$$

- **Energy estimation of digital decay generation on Intel Hala Point**

In digital computing systems, generating a time surface decay kernel $K(t) = a \cdot e^{-t/\tau} + b$ typically requires approximating the exponential function using a Taylor series expansion:

$$e^{-x} \cong \sum_{i=0}^{n} \frac{(-x)^i}{i!}$$

The suggested pseudocode for estimating e^{-x} using the Taylor series is shown as follows:

```
function taylor_series(x, n):
    result = 1.0
    last term = 1.0
    for i from 1 to n-1:
        last term = last term * x
        last term = last term / i
        result = result + last term
    return result
```

As derived from the pseudocode, the inner for-loop performs three floating-point operations (FLOPs) per iteration. Therefore, estimating e^{x} with n Taylor terms requires $3n - 3$ FLOPs. The Intel Hala Point neuromorphic system, which incorporates Loihi 2 neuromorphic processors, sustains approximately 20,000 teraFLOPs at 2600 W. Accordingly, the theoretical time required to compute a single data point is

$$t = \frac{297}{20 \times 10^{15}} \cong 0.01485 \text{ ps}$$

Assuming that $n = 100$, corresponding to 297 FLOPs per data point, and a curve-rendering density of 100 data points/ms, the energy consumed to generate one decay corresponding to the diffusive memristors used in the first layer (decay time ≈ 3.61 ms) and second layers (decay time≈ 26.60 ms) is estimated as

$$E_1 \approx 2600\ W\ \times\ 0.01485 \text{ ps} \times 3.61 \times 100 \approx 13.94\ n\text{J}$$

$$E_2 \approx 2600W \times 0.01485 \text{ ps} \times 26.60 \times 100 \approx 102.71\ n\text{J}$$

Under the same assumption used in the analog implementation, namely $N_1 = N_2 = 100$, corresponding to 100 time surfaces per layer, the total energy consumption associated with digital decay generation is

$$E_{\text{decay_total}} = 100 \times E_1 + 100 \times E_2 \approx 11.65\ \mu\text{J}$$

- **Energy estimation of clustering and SLP on Intel Hala Point**

In the digital implementation, clustering in the two-layer HOTS algorithm corresponds to a VMM operation between a flattened time-surface vector and the cluster-weight matrix. The two-layer

HOTS algorithm is configured with 7×7 and 3×3 time-surface windows for the first and second layers, respectively. Accordingly, the input vector dimensions are 7×7=49 and 3×3=9, respectively. In addition, because there are 64 and 192 centroids in HOTS layers 1 and 2, respectively, the clustering weight matrices have dimensions of 49×64 and 9×32×192, respectively. Since there are 10 output classes, the weight matrix of the SLP has dimensions of 96 × 10.

By counting each VMM operation as 2 × (input × output) FLOPs, the total number of FLOPs required for the two-layer HOTS algorithm under the assumption of 100 time surfaces per layer is

$$FLOPs_{\text{total}} = 100 \times 2 \times 49 \times 64 \; + \; 100 \times 2 \times 9 \times 64 \times 192 + 2 \times 192 \times 10 \approx 22.749 \; \times \; 10^{6}$$

Treating the digital VMM operations as serial floating-point workloads at the same operating point of 20 PFLOPS at 2600 W (i.e., 7.69 TFLOPS/W), the total energy consumption under the assumption of 100 time-surfaces per layer is

$$E_{\text{VMM_total}} = 2600W \times (22.749 \times 10^{6}/20 \times 10^{15}) \approx 2956.2 \text{ nJ}$$

By combining the energy contributions from decay generation and VMM operations, the total HOTS inference energy per pattern on Intel Hala Point is

$$E_{\text{digital_total}} \; = \; E_{\text{decay_total}} \; + \; E_{\text{VMM_total}} \approx 11.65 \; \mu\text{J} + 2.956 \; \mu\text{J} \; \approx \; 14.606 \; \mu\text{J}$$

Supplementary Table 1 | Relaxation time benchmark of 26 representative volatile and threshold-switching devices and the present work. The devices are grouped by their reported recovery or relaxation window, and the table identifies whether a dedicated external timing capacitor or resistor is used.

Ref.	Device stack	Switching/relaxation mechanism	Reported recovery or relaxation window	Temporal-computing application	Dedicated external R/C
I. Picosecond-to-nanosecond Ultrafast threshold / Mott devices					
4	Au/VO2/Au	Mott transition	Hundreds of ps; sub-ns recovery	Ultrafast device benchmark	N/A
5	Pt/Ti/NbOx/Pt/ Ti	Mott transition	On-to-off <25 ns; off-to-on <50 ns	Spiking neuron and integrated neural network	Yes; load resistor and membrane capacitor
6	Pt/Ti/NbOx/Pt/ Ti	Mott transition	Return to off state <50 ns	Rate and time-to-first-spike visual coding	N/A
7	Au/VO2/Au	Mott transition	Approximately 30 ns turn-on and 50 ns turn-off	Crossmodal sensory neuron	Yes; 10 nF membrane capacitor
8	Pt/CoO/ITO	Trap-filled space-charge-limited conduction	Approximately 40 ns	Selector	N/A
9	Ag nanodots/HfO_2/ Pt	Confined Ag filament	Approximately 240 ns	X-point selector	N/A
10	Ag/GeTe8/W	Ag conductive filaments in $GeTe_8$	Approximately 240 ns	High-current selector	N/A
11	Ag/WSe2/Ag	Se-vacancy-related filamentary conduction	Approximately 300 ns	Monolithic 3D 1T1R memory selector	N/A
II. Microsecond-to-subsecond temporal devices					
12	Ag/SiOx/hBN/ Au/Pt	Volatile Ag filament	Approximately 1.2 microseconds	Stochastic edge detection	N/A
13	Au/Ti/h-BN/Au	Generation of boron vacancies	Approximately 144 microseconds	Short-term synaptic dynamics	N/A
14	Pt/Ag/SiO_2:Ag/ Pt	Ag conductive filaments	Approximately 100 microseconds	True random-number generation	N/A
15	Pt/HfO_2/TiN	Electron trapping produces threshold switching	Approximately 800 microseconds	True random-number generation	N/A
16	Ag/IGZO/ITO	Defect-mediated variable-range hopping	0.75 ms after one pulse; up to 11 ms after repeated pulses	Short-term synaptic plasticity	N/A
17	Cu/SiO_2/Pt	Cu conductive filaments	Approximately 1 ms	Volatile/nonvolatile device benchmark	N/A
18	Pt/Ag:HfO_2/Pt	Ag conductive filaments	1.13 microseconds-1.25 ms using external R or C	N-MNIST	Yes; external R/C used to extend range
19	Pd/Ag/TiOx/Ti	Interface/dual-ion dynamics	6-9 ms; two pulse-width-dependent time constants	Hardware HOTS; Poker-DVS	N/A

20	$MoS_2/HfO_2/Pt/HZO/Pt/Cr$; three-terminal phototransistor	Ferroelectric-modulated photocarrier trapping/detrapping	Approximately 7 ms–6.88 s	Multi-speed motion perception	No
21	$Al_2O_3/MoS_2/Al_2O_3$; local-bottom-gate phototransistor	Photocarrier generation and charge trapping/detrapping	Approximately 10^1–10^6 ms; gate-voltage-dependent	Motion perception and action recognition	No
22	Pt/TiOx/Ti	Oxygen-anion redistribution	Approximately 40 ms intrinsic relaxation	Artificial neuron and neuromorphic computing	No
23	Pd/WOx/W;	Oxygen-vacancy redistribution	Approximately 50 ms	Reservoir computing; temporal prediction and MNIST	No
III. Second-to-minute-scale ionic, interfacial and fluidic relaxation					
24	$Au/Cs_3Sb_2Br_9/Au$	Br-vacancy migration	Approximately 0.6 s	Reservoir computing; Temporal information processing	N/A
25	Pt/ZnO:Ag/Ti/PET	Ag conductive filaments	Approximately 1 s	Flexible threshold switch / synaptic behavior	N/A
26	$Pt/SrTiO_3/Nb:SrTiO_3$	Oxygen vacancy redistribution	Fast ms component plus slow seconds-scale relaxation	Neuromorphic volatile-state analysis	N/A
27	Three terminal P3HT/ethylviologen redox memtransistor	Reversible redox gating changes the channel polaron density	Approximately 0.2 s to at least 100 s with rehearsal	Short- and long-term synaptic plasticity	N/A
28	Two-dimensional nanofluidic memristor	Applied current converts confined Bjerrum pairs into a conductive polyelectrolyte	Demonstrated intervals of 0.25-1,200 s	Programmable short-term depression	N/A
29	$Cr/Pt-SrTiO_3-Ti$	Interfacial oxygen-vacancy redistribution	Approximately hundreds of seconds	Multiple synaptic mechanisms and neural networks	N/A
Present work					
This work	Ag/PNP/poly-HfO_2/Pt; confined diffusive Ag filament; 28 x 14 array	Ag conductive filaments	0.47-118 ms device range; approximately 0.5-50 ms application window	Two-layer hardware HOTS; N-MNIST	No

Supplementary Table 2| Simulated recognition accuracy for N-MNIST with varying time-surface size, cluster number and classifier (single layer perceptron (SLP) and multiple layer perceptron (MLP)). The results show increased accuracy in all conditions from Layer1 to Layer2, and higher cluster number, nonlinear classifier will increase the accuracy as well.

TS Size for Layer1	TS Size for Layer2	Cluster Number for Layer1	Cluster Number for Layer2	Classifier	Layer1 Accuracy	Layer2 Accuracy
7x7	3x3	64	192	SLP	84.19%	87.69%
7x7	3x3	64	192	2-Layer MLP	91.16%	93.12%
5x5	3x3	64	192	SLP	82.41%	84.49%
5x5	3x3	64	192	2-Layer MLP	91.92%	92.75%
7x7	3x3	32	96	SLP	77.38%	83.08%
7x7	3x3	32	96	2-Layer MLP	90.52%	92.50%
5x5	3x3	32	96	SLP	75.43%	83.27%
5x5	3x3	32	96	2-Layer MLP	88.19%	90.63%

To assess the effect of spatial downsampling, we compared cropped 28 × 28 N-MNIST inputs with 14 × 14 inputs obtained by integer division of the event coordinates by two. Both software evaluations used two HOTS layers with 64/192 centroids, 7 × 7 / 3 × 3 time-surface windows, and a two-layer MLP classifier. The recognition accuracy was 93.40% for the 28 × 28 inputs and 93.12% for the downsampled 14 × 14 inputs. Downsampling therefore reduced the pixel count fourfold with an accuracy decrease of 0.28 percentage points, supporting the use of 14 × 14 inputs to accommodate the available diffusive memristor array under the tested configuration. The hardware recognition accuracy of 91.14% was obtained using an SLP classifier on the downsampled 14 × 14 N-MNIST inputs.

Supplementary References:

1 Ielmini, D. Modeling the Universal Set/Reset Characteristics of Bipolar RRAM by Field- and Temperature-Driven Filament Growth. *IEEE Trans. Electron Devices* **58**, 4309–4317 (2011).

2 Nichols, F. A. & Mullins, W. W. Morphological Changes of a Surface of Revolution due to Capillarity-Induced Surface Diffusion. *J. Appl. Phys.* **36**, 1826–1835 (1965).

3 Intel. Taking neuromorphic computing to the next level with Loihi2. Intel Technology Brief. https://download.intel.com/newsroom/2021/new-technologies/neuromorphic-computing-loihi-2-brief.pdf. (2021).

4 Schmid, S. W. et al. Picosecond Femtojoule Resistive Switching in Nanoscale VO2 Memristors. Acs Nano 18, 21966–21974 (2024).

5 Duan, Q. X. et al. Spiking neurons with spatiotemporal dynamics and gain modulation for monolithically integrated memristive neural networks. Nat. Commun. 11 (2020).

6 Li, F. F. et al. An artificial visual neuron with multiplexed rate and time-to-first-spike coding. Nat. Commun. 15 (2024).

7 Li, Z. Y. et al. Crossmodal sensory neurons based on high-performance flexible memristors for human-machine in-sensor computing system. Nat. Commun. 15 (2024).

8 Saitoh, S. & Kinoshita, K. Oxide-based selector with trap-filling-controlled threshold switching. Appl. Phys. Lett. 116 (2020).

9 Hua, Q. et al. A Threshold Switching Selector Based on Highly Ordered Ag Nanodots for X-Point Memory Applications. Adv Sci 6, 1900024 (2019).

10 Zhang, S. et al. Breakthrough in high ON-state current based on Ag–GeTe8 selectors. J. Mater. Chem. C 8, 2517–2524 (2020).

11 Sivan, M. et al. All WSe2 1T1R resistive RAM cell for future monolithic 3D embedded memory integration. Nat. Commun. 10, 5201 (2019).

12 Song, L. et al. Lightweight error-tolerant edge detection using memristor-enabled stochastic computing. Nat. Commun. 16, 4550 (2025).

13 Shi, Y. et al. Electronic synapses made of layered two-dimensional materials. Nat. Electron. 1, 458–465 (2018).

14 Jiang, H. et al. A novel true random number generator based on a stochastic diffusive memristor. Nat. Commun. 8, 882 (2017).

15 Woo, K. S. et al. A True Random Number Generator Using Threshold-Switching-Based Memristors in an Efficient Circuit Design. Adv. Electron. Mater. 5, 1800543 (2019).

16 Franco, M. et al. Inkjet printed IGZO memristors with volatile and non-volatile switching. Sci. Rep. 14, 7469 (2024).

17 Chen, W., Barnaby, H. J. & Kozicki, M. N. Volatile and Non-Volatile Switching in Cu-SiO2 Programmable Metallization Cells. IEEE Electron Device Lett. 37, 580–583 (2016).

18 Ye, F., Kiani, F., Huang, Y. & Xia, Q. Diffusive Memristors with Uniform and Tunable Relaxation Time for Spike Generation in Event-Based Pattern Recognition. Adv. Mater. 35, 2204778 (2022).

19 Cho, Y. et al. Tunable Dynamics via Dual-Ion Modulation for Event-based Data Processing Using a Highly Uniform and Self-Rectifying Memristor Array. Adv Sci, e75594 (2026).

20 Wang, Y. et al. Temporal programmable in-sensor system for multi-speed motion perception. Nat. Sens. 1, 740–748 (2026).

21 Chen, J. et al. Optoelectronic graded neurons for bioinspired in-sensor motion perception. Nat. Nanotechnol. 18, 882–888 (2023)

22 Park, S.-O., Jeong, H., Park, J., Bae, J. & Choi, S. Experimental demonstration of highly reliable dynamic memristor for artificial neuron and neuromorphic computing. Nat. Commun. 13, 2888 (2022).

23 Du, C. et al. Reservoir computing using dynamic memristors for temporal information processing. Nat. Commun. 8, 2204 (2017).

24 Mao, J.-Y. et al. Lead-free monocrystalline perovskite resistive switching device for temporal information processing. Nano Energy 71, 104616 (2020).

25 Park, Y., Han, U.-B., Kim, M.-K. & Lee, J.-S. Solution-Processed Flexible Threshold Switch Devices. Adv. Electron. Mater. 4, 1700521 (2018).

26 Hellwig, J. et al. Resolving the Relaxation of Volatile Valence Change Memory. Adv. Electron. Mater. 10, 2400062 (2024).

27 Sagar, S., Udaya Mohanan, K., Cho, S., Majewski, L. A. & Das, B. C. Emulation of synaptic functions with low voltage organic memtransistor for hardware oriented neuromorphic computing. Sci. Rep. 12, 3808 (2022).

28 Ismail, A. et al. Programmable memristors with two-dimensional nanofluidic channels. Nat. Commun. 16, 7008 (2025).

29 Weilenmann, C. et al. Single neuromorphic memristor closely emulates multiple synaptic mechanisms for energy efficient neural networks. Nat. Commun. 15, 6898 (2024).