nature communications

Article

https://doi.org/10.1038/s41467-025-59815-x

# Two-dimensional materials based two-transistor-two-resistor synaptic kernel for efficient neuromorphic computing

Received: 6 November 2024

Accepted: 6 May 2025

Published online: 09 May 2025

Check for updates

Qian He[1,5], Hailiang Wang[1,5], Yishu Zhang[1,2], Anzhe Chen[1], Yu Fu[3], Guodong Xue[4], Kaihui Liu[4], Shiman Huang[1], Yang Xu[1] & Bin Yu[1]

Neuromorphic computing based on two-dimensional materials represents a promising hardware approach for data-intensive applications. Central to this new paradigm are memristive devices, which serve as the essential components in synaptic kernels. However, large-scale implementation of synaptic matrix using two-dimensional materials is hindered by challenges related to random component variation and array-level integration. Here, we develop a 16 × 16 computing kernel based on two-transistor-two-resistor unit with three-dimensional heterogeneous integration compatibility to boost energy efficiency and computing performance. We demonstrate the 4-bit weight characteristics of artificial synapses with low stochasticity. The synaptic array demonstration validates the practicality of utilizing emerging two-dimensional materials for monolithic three-dimensional heterogeneous integration. Additionally, we introduce the Gaussian noise quantization weight-training scheme alongside the ConvMixer convolution architecture to achieve image dataset identification with high accuracy. Our findings indicate that the synaptic kernel can significantly improve detection accuracy and inference performance on the CIFAR-10 dataset.

Computing for data-intensive artificial intelligence spans multiple domains, such as image processing, natural language processing, smart transportation, and medical diagnosis[1–5]. The increasing complexity and sheer scale of the application scenarios have come with a voracious appetite for computing power[6,7]. Conventional hardware is reaching its limits in the planar scale, exposing inherent bottlenecks that hinder performance[8,9]. In this context, neuromorphic computing, which mimics the operation of biological neural networks, presents a promising solution to deal with the exponential growth in information data[10,11]. The established artificial neural network architectures, such as feedforward neural networks (FNNs)[12], convolutional neural networks (CNNs)[13], recurrent neural networks (RNNs)[14], and Spiking Neural Networks (SNNs)[15], are inherently reliant on the programmable weight matrix computations executed by synaptic computing kernels (SCKs). Synaptic arrays, which emulate biological synapses, are functional units responsible for carrying out matrix-vector multiplications and weight-update operations critical to neural network algorithms. The dynamical reconfiguration of synaptic weights is pivotal for enabling learning, inferencing, and decision-making capabilities with the bio-inspired computing architectures[16,17]. Thus, developing high-performance, energy-efficient, and scalable synaptic arrays is of paramount importance to the practical implementation and widespread adoption of artificial neural network models. Moreover, advancing device technologies and circuit architectures that faithfully reproduce the programmable synaptic functions remain a central focus in neuromorphic computing.

[1] College of Integrated Circuits, Zhejiang University, Hangzhou, Zhejiang, China. [2] ZJU-Hangzhou Global Scientific and Technological Innovation Center, Hangzhou, Zhejiang, China. [3] Department of Physics, Key Laboratory of Quantum State Construction and Manipulation (Ministry of Education), Renmin University of China, Beijing, China. [4] State Key Laboratory for Mesoscopic Physics, Frontiers Science Center for Nano-optoelectronics, School of Physics, Peking University, Beijing, China. [5] These authors contributed equally: Qian He, Hailiang Wang. e-mail: zhangyishu@zju.edu.cn; yu-bin@zju.edu.cn

In recent years, memristors have emerged as a compelling choice for SCK hardware architectures, attributed to their suitability for matrix multiplication calculations. However, traditional memristive crossbar arrays have exposed defects in the neuromorphic computation task of dense SCKs, such as leakage currents[18,19] and coupling between reading and writing operations[20]. To address these challenges, novel synaptic matrix architectures like one-transistor-one-resistor (1T1R)[21–23], two-transistor-one-capacitance (2T1C)[24], one-transistor-four-resistor (1T4R)[25], and two-transistor-two-resistor (2T2R)[26–28] have been proposed. The 1T1R configuration offers independent control of the access transistor and the non-volatile synapse, enabling selective addressing and updating of individual synaptic weights while not disturbing the state of neighboring synapses[21–23]. This capability enhances the flexibility and practicability of neural computation. Further, the monolithic integration of two-dimensional (2D) materials with memristors in the 1T1R configuration can significantly enhance performance without increasing overhead, which is critical to minimizing spatial and energy budgets in neuromorphic systems. The 2T2R architecture, which supports differential weight representation, expands the design space for synaptic computing by allowing additional degrees of freedom in weight programming and modulation. This configuration potentially increases dynamic range and precision, making it well-suited for artificial neural network models with enhanced computational capabilities[26–30]. However, its inherent complexity of functionality needs more stringent performances, posing major integration challenges, particularly in three-dimensional (3D) heterogeneously stacked architectures.

With the growing demand for compact and stackable 2T2R SCKs, 2D materials have attracted attention due to their atomic thinness and potential in low-stress integration[25,31–34]. Amidst this context, the maturity of large-scale fabrication for $MoS_2$ has positioned the 2D material as an enabler for heterogeneous integration and the realization of 3D computing architectures[9,35–37]. $MoS_2$ exhibits good carrier transport properties, high drive current, and continuously improving wafer-scale uniformity via chemical vapor deposition (CVD)[38], facilitating the integration of 2D materials with functional components to implement advanced multifunctional devices and systems. Lu et al.[36] reported ten $MoS_2$ logic circuit tiers for monolithic 3D integration system by repeating the van der Walls (vdW) lamination process in the vertical direction. Xie and co-workers[25] demonstrated scale-level integration of 1T4R unit, although it remains a separate configuration that does not form a complete system. Kang et al.[34] mimicked the vertical heterogeneous integration with a $MoS_2$ (1T)−$WSe_2$/h-BN (1R) structure, validating the feasibility of the SCK monolithic 3D heterogeneous integration process. However, there is a notable lack of research on reliability and yield of these integrated systems. The endurance and variability of neuromorphic devices based on heterogeneous integration of traditional materials and emergent 2D materials face significant scaling challenges. These challenges stem from the need for high-quality large-area synthesis, effective material transfer, and improved fabrication techniques[39], which have hindered the progression of neuromorphic applications. Therefore, achieving high-reliability 2D heterogeneous stacked 2T2R SCK is of paramount significance.

In this work, we introduce a heterogeneously integrated SCK based on 2D materials for multi-bit storage and image detection. Our approach leverages two $MoS_2$ field-effect transistors (FETs) as selectors and two $Al_2O_3$ analog memristors to implement a 2T2R structure comprising two differential 1T1R units sharing the common source electrode. The 1.8 eV bandgap of $MoS_2$ facilitates low-leakage switching, making it a good selector for memory devices. Additionally, the compatibility of 3D oxide-based memristors with semiconductor fabrication processes enables the vertical stacking of multiple layers of memory cell layers. We use the 2T2R units to experimentally demonstrate 4-bit signed weighting characteristics of artificial synapses and low device variations. Furthermore, we successfully realize a 16 × 16 array incorporating the 2T2R structure, achieving a yield of 91.2% with 15 discrete conductance states. Yield is defined that two sets of 1T1R in each 2T2R can be written and erased normally, and then 15 signed weights of 2T2R units are achieved through differential implementation. This configuration employs pulse width, number, and amplitude to effectively represent analog input signals, resulting in a functional analog-conductance matrix that supports a non-von-Neumann architecture for the storage and computation layer. The testing of the 16 × 16 2T2R array validates the system's reliability. With high device uniformity, we construct a QuantConvMixer neural network (QCMNN) for the CIFAR-10 recognition task, achieving an accuracy of over 85%, comparable to the ideal accuracy of 89.3%. This work demonstrates the feasibility of monolithic 3D heterogeneous integration of 2D materials with 3D oxides at low temperatures (<200 °C) and provides insights for developing multifunctional customized mega data computing hardware.

## Results

### Structure and mechanism of synaptic array

We implement the SCK by exploiting 2T2R units with two FETs and two memristors using bilayer $MoS_2$ as the channel material. This design leverages the high electrostatic sensitivity of ultra-thin 2D semiconductors combined with the stable analog state storage capabilities of oxide-based memristors. To enable compact array configurations, we use pairs of 1T1R configurations with shared source electrodes, allowing for signed kernel weight representation through precise row and column selection in the memory units. The architecture features 16 × 16 2T2R basic elements organized with 32 word lines (WLs), 32 bit lines (BLs), and 16 source lines (SLs), as depicted in Fig. 1a and Supplementary Fig. 1. To form the transistor channels, we sequentially transfer and etch two layers of wafer-scale CVD-grown monolayer $MoS_2$. A 40 nm gold (Au) is patterned and evaporated on top to form the drain-source contacts for the FETs, which are controlled by Au gates fabricated using a top-gate approach with $Al_2O_3$ serving as the dielectric. A 1 nm aluminum (Al) seed layer deposited via electron beam evaporation improves the dielectric layer growth through atomic layer deposition, facilitating exposure and packaging of the 2D channels, thus improving yield. Vias in the $Al_2O_3$ layer allow electrical connections between the bottom drain and top memristor electrodes, with the memristors comprising a 3 nm $Al_2O_3$ layer sandwiched between Au and 3 nm/40 nm Ti/Au configuration. Further fabrication details are outlined in the Supplementary Fig. 2.

To address the challenge of representing negative weights in neural networks, some studies employ the conventional strategy: mapping weight values to differential pairs of 1T1R cells connected to different SLs, with subtraction performed digitally in peripheral circuits. Here, we utilize the 2T2R configuration to achieve analog-domain subtraction. Both 1T1R cells are connected to the same SL, enabling direct processing of differences in the analog domain. In this array, control voltages ($V_{GP}$ and $V_{GN}$) govern the switching for two columns of 1T1Rs separately. During programming, the transistors are selectively enabled and disabled to ensure precise memristor programming, while remaining fully conductive during computation. The signed kernel weight ($G$) is represented as a differential conductance between the positive and negative memristors, measured at contrasting voltages. The output signal is derived from the current of a source line summing the contributions from a row of 2T2R currents, as illustrated in Fig. 1b. Figure 1c exemplifies how a typical element contributes to the output current in response to oppositely signed voltages applied to the drain electrodes. Figure 1d schematically illustrates the proposed QCMNN architecture and its implementation based on the array configuration. A close-up optical microscopy image of the synapse array is shown in Fig. 1e, while Supplementary Fig. 3 provides a scanning electron microscopy image of the unit plane structure. The cross-sectional high-resolution transmission electron microscopy (HR-TEM) and corresponding energy dispersive spectrometry (EDS) of the memristor cell and FET gate are depicted in Fig. 1f. To assess the reliability of the array, the variability measurements across all units are presented in Fig. 1g. The switching of all $MoS_2$ transistors can be controlled normally by the gate. On this basis, we investigate the erasing and writing of the 1T1R (transistor drain in series with memristor). Both ON and OFF currents of 1T1Rs were recorded at $V_{DS} = 0.5$ V (the $V_{DS}$ of positive 1T1R is $V_{DP}$-$V_S$; the $V_{DS}$ of negative 1T1R is $V_S$-$V_{DN}$), revealing a favorable distribution across the array, confirming that $MoS_2$ FETs function effectively as good selectors for memristors and demonstrating an impressive yield of 91.2% in a 2D material array. We further validate the functionality of the array by constructing a modified QCMNN for image detection. The robust structure of the units enables the realization of both positive and negative weights, which aligns well with the requirements of the convolutional network. This architecture processes Gaussian noise in the convolutional and fully connected layers, utilizing weights derived from the discrete conductive states of 2T2R. Detailed implementation specifics will be provided in the subsequent sections.

### Two-dimensional $MoS_2$-channel transistor

Figure 2a illustrates the conceptual framework and operational principles of the $MoS_2$-based 2T2R synapse unit. In this configuration, input voltage signals ($V_{DN}$ and $V_{DP}$) are applied to two neighboring columns with the same amplitude but opposite polarities. This setup utilizes two uniformly fabricated 1T1R cells to create a 2T2R unit that enables symmetry in positive and negative weight representation through subtraction. Specifically, the $MoS_2$ FETs function as selectors, while the analog memristors act as artificial synapses in this synergistic setup. To enable effective gating operations for the memristors, reliable $MoS_2$ FETs were initially developed. From an industrial mass production perspective, achieving uniform 2D FETs is paramount; however, conventional post-growth transfer methods often lead to issues such as film cracking, surface wrinkling, and polymer residue

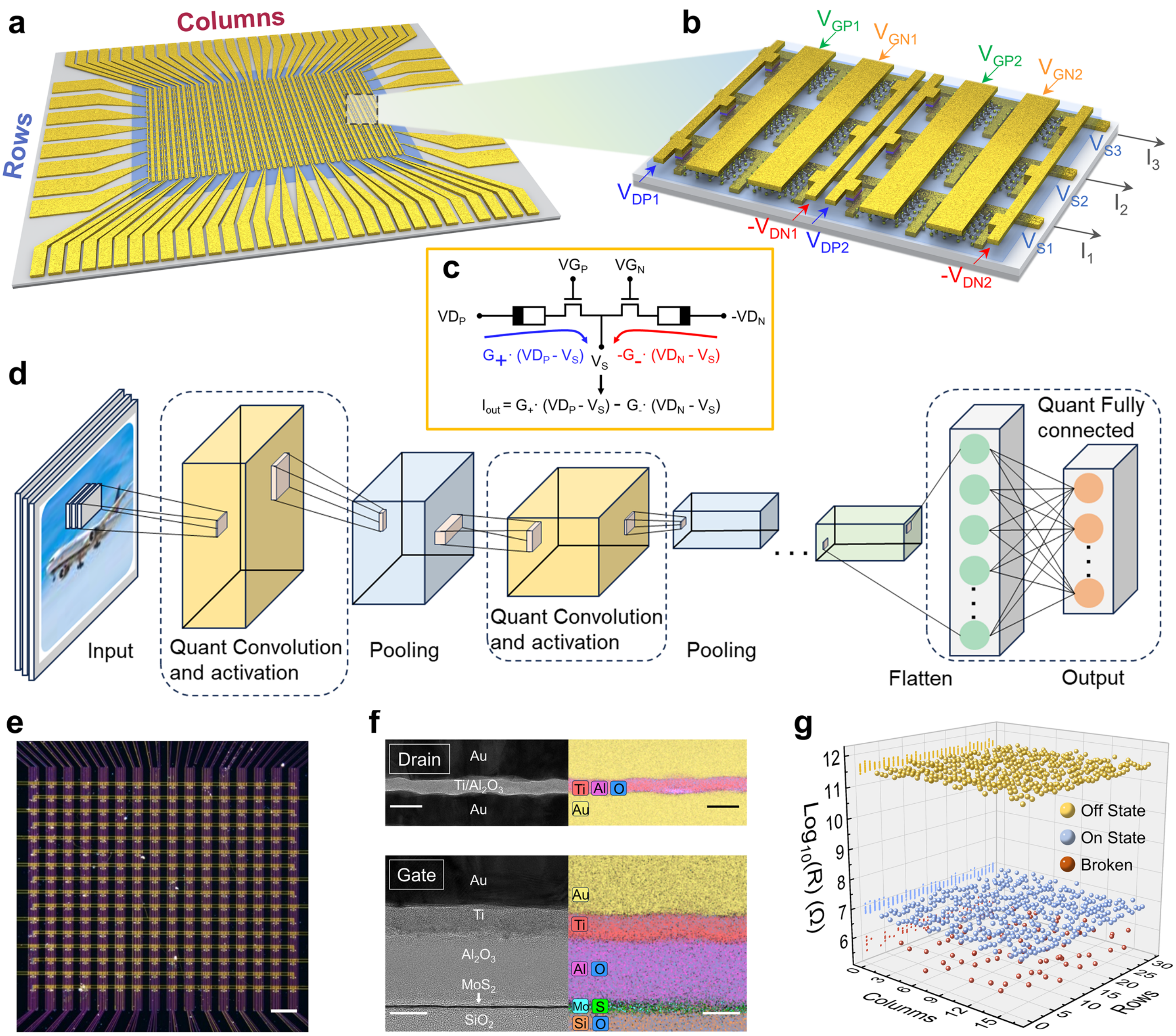

**Fig. 1 | Structure and characterization of 16 × 16 2T2R array. a** Three-dimensional rendering of the array. **b** Zoomed-in three-dimensional rendering on the array. Every two FETs share the same source and connect in rows for source row selection. The gate and drain are connected in columns, respectively. The drain-source current output from the right side. **c** This illustration describes the current of the 2T2R unit, which is equal to the differential current flowing through the positive and negative units. The output current value reflects the conductivity value of the 2T2R unit. **d** The schematic diagram of the QCMNN network for the 2T2R array. **e** Optical image of the 2T2R array. Scale bar, 400 µm. **f** Transmission electron microscope and corresponding energy dispersive spectrometer image of the 2T2R unit, including $MoS_2$ FET gate and Au/Ti/$TiO_x$/$Al_2O_3$/Au memristor with drain connected in series. Scale bar, 20 nm (up), 10 nm (down). **g** The 3D mapping plot displays 16 × 16 2T2R units turn on resistance and turn off resistance mappings, with the red dots indicating the broken units which stuck in a high conductivity state.

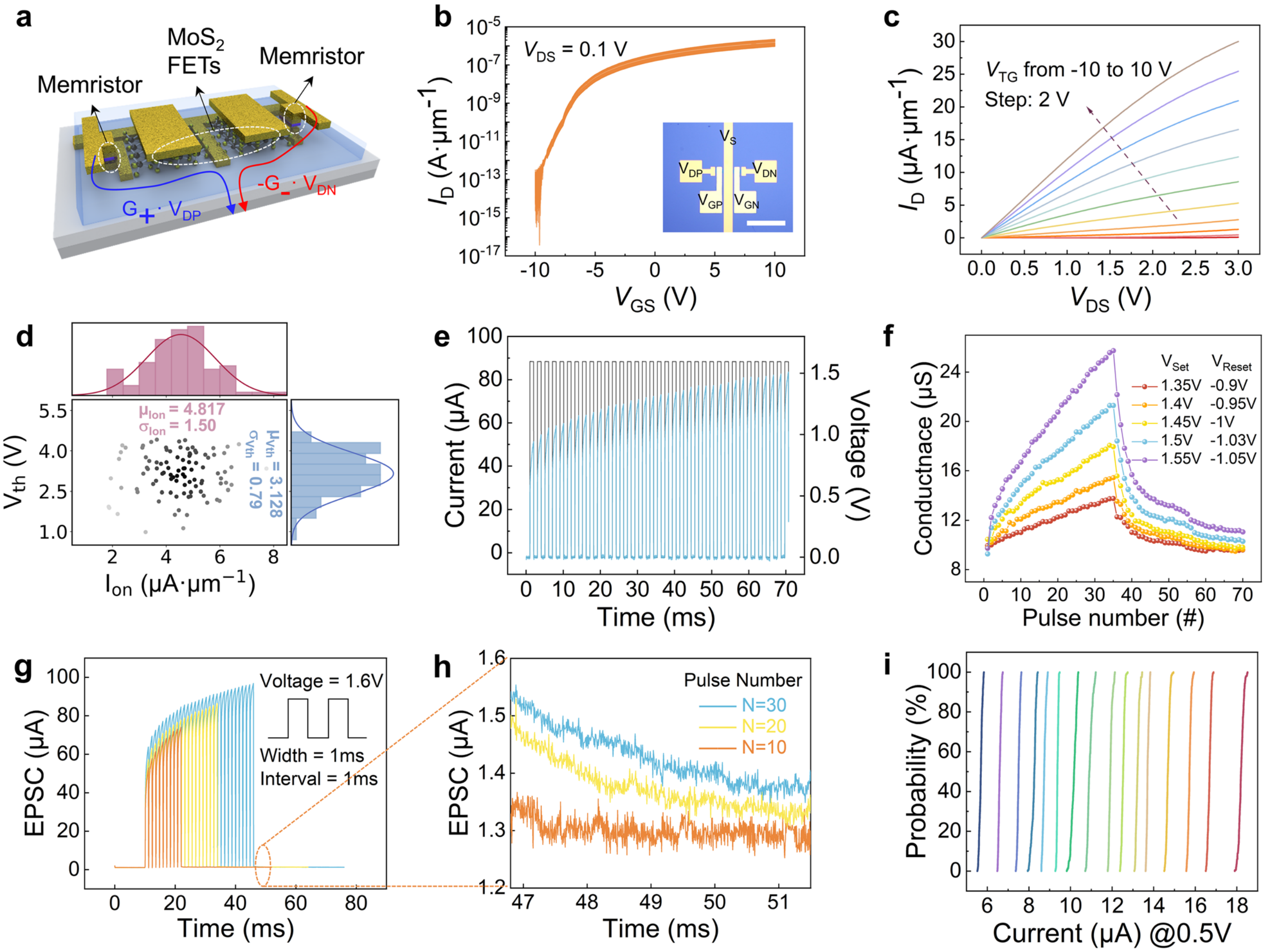

**Fig. 2 | $MoS_2$ FET and analog memristor properties.** **a** 3D schematic diagram of a single 2T2R unit. **b** 200 cycles of $I_D$-$V_{GS}$ transfer curve of a typical $MoS_2$ FET, which is under $V_{DS}$ = 0.1 V, exhibiting a large on/off ratio and good uniformity. The inset shows the optical micrograph of the 2T2R unit. Scale bar, 100 µm. **c** The output characteristic curve of $MoS_2$ FET shows high current density, with $V_{TG}$ ranging from −10 V to 10 V and a step of 2 V. **d** On-current and threshold voltage statistics from 50 $MoS_2$ FETs, showing good device-to-device variation. **e** Continuous conductivity changes of the memristor under 35 identical pulse trains, with pulse amplitude of 1.6 V, pulse width of 1 ms, and duty cycle of 50%. **f** Long-term potentiation and long-term depression of the memristor under identical pulse training. The conductivity value of each point is read by applying a 1 ms 0.5 V small amplitude pulse after each pulse. By sequentially increasing the amplitude of the pulses (from red to purple), the conductivity training speed can be changed. **g** Excitatory postsynaptic current of the memristor increases with the number of pulses. **h** Local magnification of (**g**), shows that excitatory postsynaptic current has storage property after delayed leakage. **i** The memristor can be programmed to 16 conductivity states, showing retention of more than 1000 s, reflecting the non-volatile property of the device.

contamination[39], resulting in significant performance variations and potential device failures. To address these issues, we successfully prepared a uniform bilayer $MoS_2$ thin film on a standard Si/$SO_2$ wafer using a harmless transfer method, yielding a smooth and continuous surface with minimal roughness. This was achieved through a two-layer sequential transfer of monolayer $MoS_2$ via thermal release tape (TRT) assisted transfer (see Methods), ensuring residue-free removal of impurities. Atomic force microscope (AFM) measurements indicate an average thickness of ~1.78 nm for the two $MoS_2$ monolayers, consistent with the theoretical expectations. The preserved flat surface confirms the harmless transfer process of $MoS_2$ (Supplementary Fig. 4). Additionally, Raman spectroscopy confirms the presence of two layers of monolayer $MoS_2$ (Supplementary Fig. 5). The resulting smooth surface significantly enhances device yield and minimizes inhomogeneity, as the large-area, plane-to-plane van der Waals contact between the $MoS_2$ stacking layers facilitates efficient charge transport. The utilization of stacked $MoS_2$ layers mitigates damage associated with the individual transfer of monolayers. Furthermore, the etching processes and atomic layer deposition of the $Al_2O_3$ gate dielectric layer inflict minimal damage on the $MoS_2$ film. As depicted in Fig. 2b, the transfer characteristics of the $MoS_2$ FETs exhibit a current density of ~5 µA·µm$^{-1}$ and an on/off ratio exceeding $10^8$ at a drain voltage of 0.1 V. The output characteristic curve, illustrated in Fig. 2c, demonstrates ohmic behavior under different gate voltages, indicating robust switching characteristics, effective gate control, and optimal contact quality. Statistical analyses of 50 devices, depicted in Fig. 2d, reveal low variability in the switching characteristics of the $MoS_2$ FETs. The $MoS_2$ FETs exhibit very low leakage current (<0.1 pA·µm$^{-1}$) in the OFF state and a sufficiently high drive current (average at 4.817 µA·µm$^{-1}$) in the ON state, enabling efficient writing and reading operations for the memristors without substantial voltage drops across the FETs. Our transistor exhibits great performance in CVD $MoS_2$ transistors (Supplementary Table 1). Furthermore, the high uniformity of the $MoS_2$ FETs significantly enhances yield in the synaptic array, demonstrating their potential for applications in neuromorphic computing systems.

## Memristive device

The challenges associated with heterogeneous integration of 2D synapse kernels extend beyond achieving large-scale uniformity in materials but also involve ensuring the reliable programming of memristor devices to stable analog values in conjunction with 2D material FETs[16]. Herein, we demonstrate high-precision analog tuning and control of the memristor cells integrated on the drain of $MoS_2$ FETs. As detailed in the Methods section, the prepared memristor dimensions are 6 × 6 μm. The memristors exhibit analog resistive switching behavior in both the DC current-voltage (*I-V*) sweeping (Supplementary Fig. 6) and identical pulse mode measurements (Fig. 2e, and Supplementary Fig. 7). The gradual set and reset processes indicate that the memory effect in these memristors is primarily governed by the migration of interface oxygen vacancies as shown in Supplementary Fig. 8[40–42]. This non-filament mechanism supports the design of memristors that are well-aligned with 2D FETs through size adjustments. Figure 2f illustrates that the writing and erasing processes can be precisely controlled by varying the voltages in set and reset operations. By applying 35 set pulses, the device can alter its conductance from 10 μS to 30 μS. Besides, the dynamic responses of the memristors have been studied under pulsed electric stimuli with varying amplitudes, durations, and intervals (Supplementary Fig. 9), supporting diverse weight-update operations and dynamical reconfiguration. These rich programming methods provide flexible options for peripheral circuit design. For controlled variable programming, we uniformly program the memristor by regulating the number of write/erase pulses while maintaining a consistent pulse amplitude, width, and duty cycle. An identical pulse programming scheme, specifically the increased pulse number scheme, is employed to achieve the desired target current as illustrated in Fig. 2g. To further demonstrate the nonvolatile properties of the memristors, we applied a longer read pulse (0.05 V, 30 ms) following a serial of write pulses. This reading voltage is verified not to disturb the device's state shown in Supplementary Fig. 10 and Supplementary Fig. 11. Although the current slightly deteriorates over time, it ultimately retains a certain degree of memorization, indicating nonvolatile characteristics after writing pulses (depicted in Fig. 2h and Supplementary Fig. 12). Consequently, we validated the feasibility of memristor with the conductance values pretrained to represent 4-bit multilevel states. As shown in Fig. 2i, resistance exhibits minimal variation across 16 states over 1000 seconds, signifying decent multilevel retention capability and suitability for use as an artificial synapse. High temperature retention and repeated writing cycles are presented in Supplementary Fig. 13 and Supplementary Fig. 14. Additionally, spike-timing-dependent plasticity, an essential unsupervised learning rule in spiking neural networks, is embodied in the memristor (Supplementary Fig. 15). Collectively, these findings highlight the broad conductance tunability and diverse bio-neurological characteristics of memristors, laying a robust foundation for subsequent array-level demonstrations of the synapse kernel.

### Heterogeneously integrated synaptic array

Based on the fundamental components discussed, we have developed a fully integrated and functional reprogrammable SCK based on the 2T2R structure, alongside a customized array interface chip carrier for characterizing this SCK. The 2T2R structure provides several advantages: it minimizes the shared current flowing through the source line due to the differential operation between positive and negative cells, thereby reducing the IR drop[26,28]. This configuration also supports the implementation of both positive and negative weights, significantly increasing the number of available stored states[26]. Additionally, the 2T2R structure has demonstrated higher reliability compared to the conventional configuration under similar device variabilities[27]. For example, it can mitigate programming issues by allowing the conductivity of another 1T1R to compensate when one 1T1R is stuck or non-responsive. This architecture also reduces the integration area while enabling negative conductivity (Supplementary Fig. 16).

The physical layout of the chip is shown in Fig. 3a, where programming and computing are performed in conjunction with external test instruments, with all signals generated off-chip (see "Methods"). Hardware processor architecture is shown in Supplementary Fig. 17. Initially, we validated the performance of two co-sources 1T1R devices within the 2T2R structure, confirming a good match between the $MoS_2$ FETs and $Al_2O_3$ memristors. The desired DC sweep characteristics indicate that the memristor can gradually program nonvolatile conductance states (Fig. 3b and Supplementary Fig. 18). We find that the $MoS_2$ FET can effectively control the multi-stage conductance tuning of the memristor with a high degree of accuracy. When activated, the $MoS_2$ FET allows the connected memristor to be programmed into 8 distinct states by changing the number of applied pulses, as shown in Fig. 3c. The reduced state count per 1T1R unit (from 16 to 8) ensures compatibility with the FET's gate-voltage tuning limits while maintaining sufficient computational granularity for neuromorphic tasks. Given a target conductance and tolerance range, the programming feedback loop adjusts the number of the applied voltage accordingly. The programmed conductivity states are non-overlapping, and each state exhibits a small standard deviation (<0.004) across repeated programming of cycles (Supplementary Fig. 19). To validate inter-unit uniformity, we conducted a repeatability involving 1000 cycles of writing and erasing for each conductance state, resulting in a reasonable spread of individual conductance levels and high cycling uniformity across the units. Besides, the retention of 8 states demonstrates its good non-volatile characteristics (Supplementary Fig. 20). Figure 3d displays the initial conductance of all the 16 × 16 × 2 1T1R units rendered in a heatmap. While some units remain stuck at high conductance (synapse broken), others fall within a desirable conductivity range, facilitating further state regulation. The measurement used a read voltage of 0.5 V applied to the selected 1T1R unit, with unselected WL and BL floating, highlighting the process-induced device-to-device variation. This approach also paves the way for future implementation of array-level conductance programming.

With the compatibility of $MoS_2$ FETs and $Al_2O_3$ memristors, we have successfully realized the core characteristics of the 2T2R structure: storing the subtraction weights of two 1T1Rs. We configured the 2T2R to achieve a zero-conductance state, exemplifying its differential conductance function (Fig. 3e). The high uniformity of the devices, enabled two 1T1Rs to be programmed to similar conductivity values, resulting in a 2T2R conductivity value of zero obtained after the differential operation. Additionally, we accessed and precisely adjusted the conductivity of the memristor units through a pulse sequence. With 8 distinct conductivity states available for each 1T1R, we successfully obtained 15 differential conductivity states for 2T2R. The experimental measurements of these 15 conductances were statistically analyzed and plotted (Fig. 3f), with box plots illustrating 50 programming iterations for each resistive state.

To evaluate device-to-device reproducibility, we configured a 16 × 16 array with the previously mentioned 15 states, converting conductances into image pixel intensities. The distribution of states over the entire array is depicted in Fig. 3g, which includes readout conductance matrices after programming the array, showcasing the logo of Zhejiang University. Spatial uniformity was maintained across all 256 units of the array (Supplementary Fig. 21). A yield rate of 91.2% demonstrates high array reliability. In the context of hardware implementation of the 2D synapse kernel, previous studies have demonstrated multi-bit memristor arrays to perform the computation of the convolutional network. The high yield and device variation in this work do not have an adverse effect on the accuracy of algorithm applications (Supplementary Tables 5 and 6). Our work represents the successful demonstration of an integrated signed-weighted synapse kernel based on 2D materials (Supplementary Table 2). The advantages of the 2T2R structure are summarized in Supplementary Table 3. Supplementary Table 4 benchmarks the overall performance metrics of various 2D materials-based heterogeneous integrated arrays, with our $MoS_2$ array exhibiting good performance in terms of integration

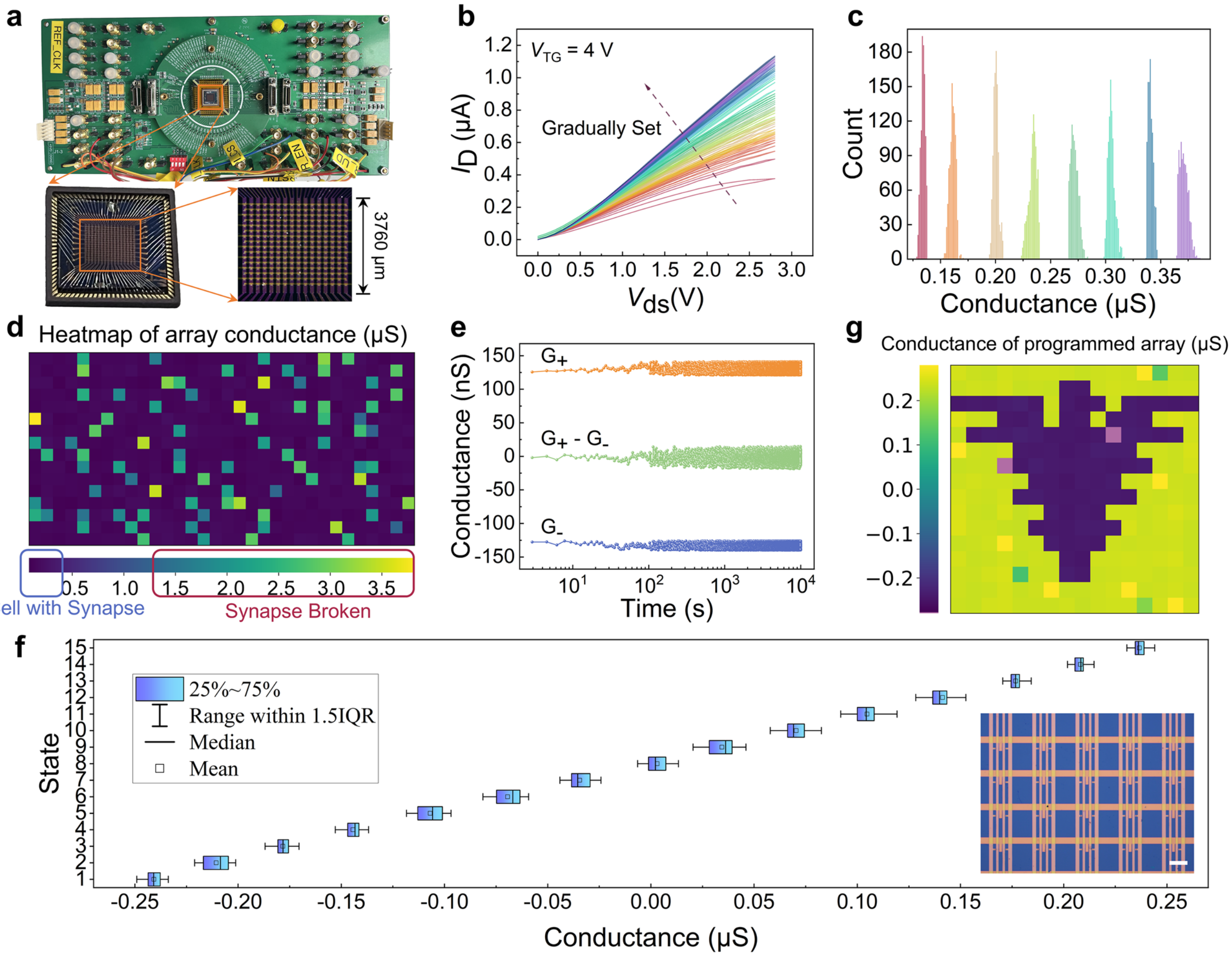

**Fig. 3 | 16 × 16 2T2R array characteristics analysis. a** Diagram of the array being bound to the chip carrier. **b** Typical DC testing of 1T1R in the 2T2R unit. The sweeping voltage of $V_{DS}$ is scanned to 2.8 V several times under $V_{TG}$ = 4 V, the 1T1R $I_D$ is gradually increased. **c** 1T1R unit was programmed to 8 conductivity states 1000 cycles. **b**, **c** embody the good matching between $MoS_2$ FET and memristor, realizing the writing and erasing of memristor controlled by $MoS_2$ FET. **d** Thermal mapping of the initial conductivity state of each 1T1R unit in the array, reflecting good yield and uniformity of the array. **e** A typical example is that the conductance of the 2T2R unit is equal to the difference between two 1T1R conductances. The conductivity maintains good stability over time. **f** Statistics on the programmed 15 conductance values of 50 2T2Rs. The inset is partial physical image of 2T2Rs. Scale bar in illustration, 100 μm. **g** Overall programming of 16 × 16 array, presenting the logo of Zhejiang University.

size, signed weights, and multiple conductance states compared to other works. Notably, our proposed fabrication processes emphasize transfer-clean and low-temperature techniques, providing good compatibility with thermally constrained 3D monolithic integration and flexible electronics. We also demonstrated a larger-scale array to prove the scalability of the array preparation process (Supplementary Fig. 22). The quantized kernel weights are accurately transferred to the corresponding memristor cells with minimal discrepancies between the actual and target conductance values. Due to the retention properties of the conductance states, the values exhibit minimal fluctuation post-mapping. We anticipate that the system energy efficiency could be further enhanced by employing synapses operating in a high-resistance range, characterized by a linear *I-V* relationship and stable multilevel states alongside the use of reduced voltage inputs and/or shorter pulses. Though a direct comparison is not asserted, our system demonstrates higher resistance in 2D multi-state synaptic kernels, presenting a clear advantage in minimizing power consumption.

## Synaptic kernel for quantized ConvMixer simulation

To evaluate the feasibility and effectiveness of the array operation and weight mapping schemes for the synapse kernel, we conducted a modified convolution experiment utilizing the 2T2R array for image detection. The ConvMixer architecture presents a straightforward yet powerful CNN that has demonstrated superior performance compared to models such as Vision Transformers (ViT), MLP-Mixer variants, and traditional vision models like ResNet[43], while maintaining similar parameter counts and dataset size. In hardware implementations of convolutional models, low-memory networks such as quantized neural networks (QNNs) are advantageous, as they minimize the burden of device state control through weight quantization. We adapted the official implementation of ConvMixer[43] to develop a quantized version termed Quantized ConvMixer convolution neural network. Figure 4a displays the overall structure of the QCMNN, featuring the 2T2R array. Here, “Conv” and “FC” refer to the convolutional and fully connected layers, while “qConv” and “qFC” denote our quantized counterparts. We quantized the weight of both convolutional and fully connected layers into the conductance levels of 1T1R (eight levels) or 2T2R (fifteen levels) frameworks, incorporating runtime conductance deviations based on experimental data to better simulate the quantized and stochastic behavior of the devices. The distributions of initial and trained weights are illustrated in Fig. 4b, c, respectively. All quantized models were trained from quantization-aware training (QAT), as detailed in the Methods section. Upon completing the training, the weights were quantized and mapped to the synaptic computing

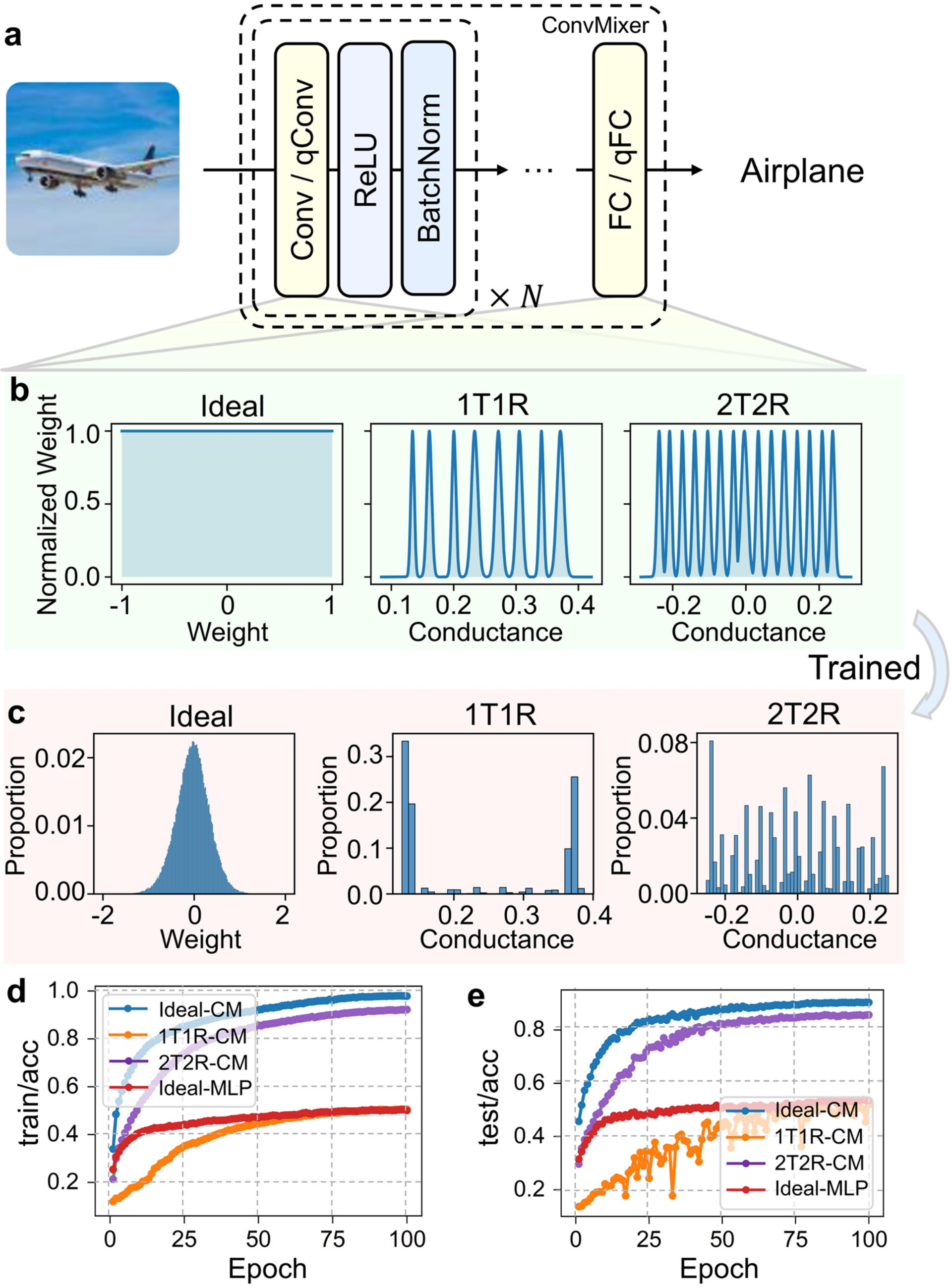

**Fig. 4 | 2T2R SCK for QCMNN. a** Network structure. We use ConvMixer (a CNN-based neural network) as the baseline and quantize the weight and bias in convolutional (Conv) and fully connected (FC) layers. The airplane photo is sourced from the CIFAR-10 dataset. **b** Initial weight affinity of ideal and quantized linear layers. Ideal weight has no constraints, "1T1R" has 8 distinct conductance levels, "2T2R" has 15 distinct conductance levels. **c** Weight distributions of ideal and quantized ConvMixer models after training. **d** Training and **e** validation accuracy of software-based ideal ConvMixer, quantized ConvMixer with our device (both 1T1R and 2T2R structures), and software-based ideal multiple-layer-perceptron (MLP). All the models have comparable size of trainable parameters.

kernel. As shown in Fig. 4d, e, the 1T1R implementation supports only positive weights under same hardware condition, resulting in a significant reduction in convergence speed and classification performance on the CIFAR−10 dataset. In contrast, the 2T2R implementation accommodates signed weights, achieving classification performance comparable to that of the ideal software-based ConvMixer. In addition, the simulation also shows that 2T2R has better robustness, reflected in its stronger tolerance to read interference and retention shrinkage (Supplementary Tables 7 and 8). Consequently, the integration of 2D heterogeneous stacked 2T2R SCK facilitates an increased number of weight states while enhancing the classification accuracy of neural networks (Supplementary Table 3).

## Discussion

In summary, we experimentally demonstrated a heterogeneously integrated array using 2D materials and oxide-based memristors as the synaptic kernel. The defect-controlled material transfer and low-temperature fabrication result in high yield (>91.2%). The low-temperature heterogeneous integration approach represents a viable system-level solution for monolithic 3D heterogeneous integration, addressing the computational challenges posed by the explosive growth of data. The core 2T2R unit in the array, mapping weight to the conductance difference between two memristors, facilitates the implementation of a signed weight SCK that is well-suited for CNNs. We constructed a QCMNN for the CIFAR−10 recognition task, achieving an accuracy exceeding 85%. Benchmarking results indicate that the 2T2R-based synaptic kernel outperforms the conventional structures in terms of accuracy and training speed. Moreover, this synaptic kernel has the potential to be extended to other matrix multiplication-based neuromorphic hardware, providing an efficient route to enhance the overall system performance.

## Methods

### $MoS_2$ chemical vapor deposition synthesis

A silicon chip was prepared by immersing it into an aqueous solution of $Na_2MoO_4$ at a concentration of 12 mg·mL$^{-1}$, followed by air drying. Subsequently, this chip and a ZnS crystal plate were sequentially positioned above a sapphire substrate, using mica spacers for each layer, and then collectively inserted into a CVD furnace. The furnace chamber was purged with Ar (300 sccm) and heated to the designated growth temperature of ~800 °C. Throughout the growth period, the pressure within the system was maintained at ~120 Pa, and the growth duration was controlled between 30 and 60 min. After the completion of the growth process, the system was allowed to cool down naturally to room temperature.

### $MoS_2$ film transfer

Initially, a PMMA layer was spin-coated onto $MoS_2$/sapphire at ~1500 rpm for 1 min and subsequently baked in air at 120 °C for 5 min. Following this, a TRT piece, pre-punched with a hole, was attached to the PMMA/$MoS_2$/sapphire assembly. Then the TRT/PMMA/$MoS_2$ composite was detached from the sapphire in water and allowed to dry in air before being laminated onto the $SiO_2$/Si substrate. The TRT was removed by heating to the designated release temperature and then peeled off. To further improve the interaction between $MoS_2$ and the $SiO_2$/Si, the PMMA/$MoS_2$/$SiO_2$/Si assembly was baked at 180 °C for 10 min. It was then washed with acetone and annealed in Ar (300 sccm) at 400 °C for 4 h to remove the PMMA, leaving $MoS_2$ on $SiO_2$/Si.

### Heterogeneous integrated array fabrication

The fabrication process originates with a p-doped silicon wafer featuring a 285 nm $SiO_2$. After transfer, bilayer $MoS_2$ with wafer annealed at 400 °C for 8 h, under $H_2$/Ar (20/150 sccm), ~1 Torr. Patterning channels with Direct Write Lithography (DWL) and the exposed region etched by Inductively Coupled Plasma (ICP) with $O_2$/Ar. SLs and bottom electrodes patterned by DWL, a 40 nm Au deposited by Electron-Beam Evaporation (EBE). Post the lift-off process, a 1 nm Al coated by EBE as a seed layer and annealed at 95 °C for 5 min in ambient air. Plasma and ozone pretreatment increase the surface hydroxyl concentration of the seed layer. Twenty nanometer $Al_2O_3$ blocking layer grown by ALD at 200 °C. Annealing at 350 °C for 30 min activates chemical passivation (filling dangling bonds) and field passivation (fixing negative charges) of $Al_2O_3$. Contact vias etched through the $Al_2O_3$ layer at the drain region, preparatory to depositing 3 nm $Al_2O_3$ as the memristor functional layer. Finally, 3 nm/40 nm Ti/Au stack patterned and evaporated on top, forming the BLs and WLs.

### Electrical characterization

For the implementation of SCK hardware system, the 2T2R array was connected to the package shell with Al wires through wire bonding, on the PCB-FPGA control system. The electrical measurements of the 2T2R array are performed using a Keithley 4200A-SCS Semiconductor analyzer in atmospheric pressure at room temperature. Source measurement units are used for DC measurements, and pulse measurement units are used for pulse measurements.

### Hyperparameters of ConvMixer

We use depth = 6 (6 consecutive depth-wise convolutional blocks), dimension = 128, kernel size = 5, patch size = 2. This yields around 0.12 M trainable parameters, which is less than a typical two-layer MLP ($784 \times 128 + 128 \times 10 \approx 0.227$ M) for MNIST classification. For detailed explanation of these parameters, please refer to the original paper of ConvMixer[43].

### Quantization aware training

Given $N$ available conductance levels of device $\mathbf{c} = [c_1, c_2, \ldots, c_N]$ and the deviation at each conductance level $\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \ldots, \sigma_N]$, in the network forward pass, we first quantized the unconstraint weights $w_i \in \mathbf{W}$ into the nearest conductance level, we then add gaussian noise to mimic the reading noise of noise:

$$w_i^q = c_j + \mathcal{N}(0, \sigma_i) \tag{1}$$

where $j = \mathrm{argmin}_{1 \le k \le N}(w_i - c_k)$

Then the output of convolutional or fully connected layers is computed from the quantized weights. Take the fully connected layers as example, we have:

$$\mathbf{y} = \mathbf{W}^q \cdot \mathbf{x} + \mathbf{b} \tag{2}$$

In the backward pass, the gradient descents are conducted on the original unconstraint weight, with gradient estimated via straight through estimator (STE)[44], i.e., $\partial\mathcal{L}/\partial\mathbf{W} = \partial\mathcal{L}/\partial\mathbf{W}^q$:

$$\mathbf{W} = \mathbf{W} - \lambda \frac{\partial\mathcal{L}}{\partial\mathbf{W}^q} = \mathbf{W} - \lambda \frac{\partial\mathcal{L}}{\partial\mathbf{y}} \cdot \mathbf{x}^{\mathrm{T}} \tag{3}$$

where $\lambda$ is the learning rate, $\mathcal{L}$ is the final loss.

After the training is finished, we export the quantized weights $\mathbf{W}^q$ for inference.

## Data availability

All of the data generated in this study can be found in the main text and in the Supplementary Information. Any other data that support the findings of this study are available from the corresponding authors upon request. The CIFAR–10 dataset used in this study can be accessed at https://www.cs.toronto.edu/~kriz/cifar.html[45].

## Code availability

The algorithms and codes supporting the findings of this study are available at https://doi.org/10.5281/zenodo.10775253.

## References

1. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
2. Maloca, P. M. et al. Unraveling the deep learning gearbox in optical coherence tomography image segmentation towards explainable artificial intelligence. *Commun. Biol.* **4**, 170 (2021).
3. Alqahtani, T. et al. The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. *Res. Soc. Adm. Pharm.* **19**, 1236–1242 (2023).
4. Xu, Y. et al. Artificial intelligence: a powerful paradigm for scientific research. *Innovation* **2**, 100179 (2021).
5. Zhang, C. & Lu, Y. Study on artificial intelligence: the state of the art and future prospects. *J. Ind. Inf. Integr.* **23**, 100224 (2021).
6. Thompson N., Greenewald K., Lee K., Manso G. F. The Computational Limits of Deep Learning. In: *Ninth Computing within Limits* 2023 (2023).
7. Conklin, A. A. & Kumar, S. Solving the big computing problems in the twenty-first century. *Nat. Electron.* **6**, 464–466 (2023).
8. Williams, R. S. What's next? [The end of Moore's law]. *Comput. Sci. Eng.* **19**, 7–13 (2017).
9. Bishop, M. D., Wong, H. S. P., Mitra, S. & Shulaker, M. M. Monolithic 3-D integration. *IEEE Micro* **39**, 16–27 (2019).
10. Barabasi, A. L. Scale-free networks: a decade and beyond. *Science* **325**, 412–413 (2009).

11. Marković, D., Mizrahi, A., Querlioz, D. & Grollier, J. Physics for neuromorphic computing. *Nat. Rev. Phys.* **2**, 499–510 (2020).
12. Baldi, P. & Vershynin, R. The capacity of feedforward neural networks. *Neural Netw.* **116**, 288–311 (2019).
13. Khan, A., Sohail, A., Zahoora, U. & Qureshi, A. S. A survey of the recent architectures of deep convolutional neural networks. *Artif. Intell. Rev.* **53**, 5455–5516 (2020).
14. Schmidt R. M. Recurrent neural networks (RNNs): a gentle introduction and overview. *arXiv*, https://doi.org/10.48550/arXiv.41912.05911 (2019).
15. Roy, K., Jaiswal, A. & Panda, P. Towards spike-based machine intelligence with neuromorphic computing. *Nature* **575**, 607–617 (2019).
16. Zhang, W. et al. Neuro-inspired computing chips. *Nat. Electron.* **3**, 371–382 (2020).
17. Joshi, V. et al. Accurate deep neural network inference using computational phase-change memory. *Nat. Commu.* **11**, 2473 (2020).
18. Jeon, K. et al. Self-rectifying resistive memory in passive crossbar arrays. *Nat. Commun.* **12**, 2968 (2021).
19. Li, J. et al. Sparse matrix multiplication in a record-low power self-rectifying memristor array for scientific computing. *Sci. Adv.* **9**, eadf7474 (2023).
20. Li Y., Ang K.-W. Hardware implementation of neuromorphic computing using large-scale memristor crossbar arrays. *Adv. Intell. Syst.* **3**, 2000137 (2020).
21. Li, C. et al. Analogue signal and image processing with large memristor crossbars. *Nat. Electron.* **1**, 52–59 (2017).
22. Yao, P. et al. Fully hardware-implemented memristor convolutional neural network. *Nature* **577**, 641–646 (2020).
23. Feng, Y. et al. Memristor-based storage system with convolutional autoencoder-based image compression network. *Nat. Commun.* **15**, 1132 (2024).
24. Wang, Y. et al. An in-memory computing architecture based on two-dimensional semiconductors for multiply-accumulate operations. *Nat. Commun.* **12**, 3347 (2021).
25. Xie, M. et al. Monolithic 3D integration of 2D transistors and vertical RRAMs in 1T-4R structure for high-density memory. *Nat. Commun.* **14**, 5952 (2023).
26. Liu Q. et al. 33.2 a fully integrated analog ReRAM based 78.4TOPS/W compute-in-memory chip with fully parallel MAC computing. In *Proc. 2020 IEEE International Solid-State Circuits Conference (ISSCC)* (IEEE, 2020).
27. Zhou Z. et al. A new hardware implementation approach of BNNs based on nonlinear 2T2R synaptic cell. In *Proc. 2018 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2018).
28. Zhang, W. et al. Edge learning using a fully integrated neuro-inspired memristor chip. *Science* **381**, 1205–1211 (2023).
29. Yang, R. et al. Ternary content-addressable memory with $MoS_2$ transistors for massively parallel data search. *Nat. Electron.* **2**, 108–114 (2019).
30. Gao, B. et al. Memristor-based analogue computing for brain-inspired sound localization with in situ training. *Nat. Commun.* **13**, 2026 (2022).
31. Wang C.-H. et al. 3D Monolithic stacked 1T1R cells using monolayer $MoS_2$ FET and hBN RRAM fabricated at low (150 °C) temperature. In Proc. *2018 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2018).
32. Sivan, M. et al. All $WSe_2$ 1T1R resistive RAM cell for future monolithic 3D embedded memory integration. *Nat. Commun.* **10**, 5201 (2019).
33. Su, C. J. et al. 3D Integration of vertical-stacking of $MoS_2$ and Si CMOS featuring embedded 2T1R configuration demonstrated on full wafers. In Proc. 2020 IEEE International Electron Devices Meeting (IEDM) (IEEE, 2020).
34. Kang, J. H. et al. Monolithic 3D integration of 2D materials-based electronics towards ultimate edge computing solutions. *Nat. Mater.* **22**, 1470–1477 (2023).
35. Wang, S. et al. Two-dimensional devices and integration towards the silicon lines. *Nat. Mater.* **21**, 1225–1239 (2022).
36. Lu, D. et al. Monolithic three-dimensional tier-by-tier integration via van der Waals lamination. *Nature* **630**, 340–345 (2024).
37. Guo, Y. et al. Van der Waals polarity-engineered 3D integration of 2D complementary logic. *Nature* **630**, 346–352 (2024).
38. Akinwande, D. et al. Graphene and two-dimensional materials for silicon technology. *Nature* **573**, 507–518 (2019).
39. Choi, S. H. et al. Large-scale synthesis of graphene and other 2D materials towards industrialization. *Nat. Commun.* **13**, 1484 (2022).
40. Yang, J. J. et al. Memristive switching mechanism for metal/oxide/metal nanodevices. *Nat. Nanotechnol.* **3**, 429–433 (2008).
41. Sun, W. et al. Understanding memristive switching via in situ characterization and device modeling. *Nat. Commun.* **10**, 3453 (2019).
42. Andreeva, N. V., Ryndin, E. A., Mazing, D. S., Vilkov, O. Y. & Luchinin, V. V. Organismic memristive structures with variable functionality for neuroelectronics. *Front. Neurosci.* **16**, 913618 (2022).
43. Trockman A., Kolter J. Z. Patches are all you need? *Transactions on Machine Learning Research* (2023).
44. Courbariaux M., Bengio Y., David J.-P. BinaryConnect: training deep neural networks with binary weights during propagations. In *Proc. 29th International Conference on Neural Information Processing Systems* Vol. 228, 3123–3131 (ACM, 2015).
45. Krizhevsky, A. Learning multiple layers of features from tiny images. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf (2009).

## Acknowledgements

This research was supported by the NSFC of Zhejiang Province (Grant Nos. DT23F0401, DT23F04008, and DT23F04009), the NSFC (Grant No. 92264106) and the open research fund of Suzhou Laboratory (Grants No. SZLAB-1208-2024-TS012). We appreciate the support of the ZJU Micro-Nano Fabrication Center. The authors also acknowledge Dr. Jiabao Sun of ZJU Micro-Nano Fabrication Center for his professional and enthusiastic assistance. Thank Prof. Kechao Tang from Peking University for the fruitful discussion.

## Author contributions

Q.H. and H.W. contributed equally to this paper. B.Y., Y.Z., and Y.X. guided the research. Q.H. and H.W. designed the experiment. Q.H. and H.W. fabricated and tested the devices. A.C. and S.H. helped in the neural network simulation. Y.F., G.X., and K.L. helped in large-scale fabrication. All the authors revised the paper.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-59815-x.

**Correspondence** and requests for materials should be addressed to Yishu Zhang or Bin Yu.

**Peer review information** *Nature Communications* thanks Lei Ye and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at http://www.nature.com/reprints