nature communications

Article

https://doi.org/10.1038/s41467-025-58005-z

8-bit states in 2D floating-gate memories using gate-injection mode for large-scale convolutional neural networks

Received: 7 July 2024

Accepted: 10 March 2025

Published online: 18 March 2025

Check for updates

Yuchen Cai[1,2], Jia Yang[1,2], Yutang Hou[3], Feng Wang[1,2] ✉, Lei Yin[3], Shuhui Li[1], Yanrong Wang[4], Tao Yan[1], Shan Yan[1], Xueying Zhan[1,2], Jun He[3] & Zhenxing Wang[1,2] ✉

The fast development of artificial intelligence has called for high-efficiency neuromorphic computing hardware. While two-dimensional floating-gate memories show promise, their limited state numbers and stability hinder practical use. Here, we report gate-injection-mode two-dimensional floating-gate memories as a candidate for large-scale neural network accelerators. Through a coplanar device structure design and a bi-pulse state programming strategy, 8-bit states with intervals larger than three times of the standard deviations and stability over 10,000 s are achieved at 3 V. The cycling endurance is over 10^5 and the fabricated 256 devices show a yield of 94.9%. Leveraging this, we carry out experimental image convolutions and 38,592 kernels transplanting on an integrated 9 × 2 array that exhibits results matching well with simulations. We also show that fix-point neural networks with 8-bit precision have inference accuracies approaching the ideal values. Our work validates the potential of gate-injection-mode two-dimensional floating-gate memories for high-efficiency neuromorphic computing hardware.

Machine learning and artificial intelligence based on neural networks (NNs) have shown remarkable capabilities across a wide range of applications, including autonomous driving, weather prediction, speech recognition, and image understanding[1–4]. And it has a substantial demand for accelerators like graphics processing units, which are well-suited for handling large-scale, parallel multiply-and-accumulate operations. However, back-and-forth data movement between the physically separated memory and logic units in the conventional von Neumann architecture and the digital data processing paradigm imposes significant limitations on the system efficiency[5,6]. Consequently, there is a growing interest in high-efficiency neuromorphic computing hardware (NCH), particularly for intelligent edge devices that can process and store data locally and in an analog manner, akin to the human brain[7–9].

At the algorithm level, NNs handle weights with infinite precision, a luxury that NCH cannot afford. To implement NNs at the edge, it is necessary to train and/or infer within device-level nodes that have limited numerical precision. Theoretical simulations have shown that many deep NNs with 8 to 24-bit precision will suffer almost no accuracy degradation compared to a much higher precision, owing to stochastic rounding schemes and the large amounts of parameters they usually contain[10–13]. On the other hand, excessively low precision (such as <8-bit) may lead to performance degradation instead, particularly in small-sized NNs deployed on edge devices that require high energy efficiency, as each parameter has a greater impact on their overall performance[14]. Whether training a fixed-precision NN directly or downloading and quantizing a pre-trained NN to achieve a fixed-precision network, devices

[1]CAS Key Laboratory of Nanosystem and Hierarchical Fabrication, National Center for Nanoscience and Technology, Beijing, P. R. China. [2]Center of Materials Science and Optoelectronics Engineering, University of Chinese Academy of Sciences, Beijing, P. R. China. [3]Key Laboratory of Artificial Micro- and Nanostructures of Ministry of Education, School of Physics and Technology, Wuhan University, Wuhan, P. R. China. [4]Institute of Semiconductors, Henan Academy of Sciences, Zhengzhou, P. R. China. ✉e-mail: wangf@nanoctr.cn; wangzx@nanoctr.cn

capable of supporting many distinguishable conductance levels are crucial.

Non-volatile memories, such as floating-gate memories (FGMs)[15–20], resistive switching memories[8,21–24], phase change memories[25,26], and ferroelectric memories[27–29], have emerged as candidates for NCH. Among them, FGMs are especially promising due to their non-volatile charge-based analog storage mode. When utilized as artificial synapses, FGMs exhibit learning rates that align well with those of visual and auditory signals[15]. Additionally, FGMs offer a large dynamic range and are compatible with standard complementary metal-oxide-semiconductor (CMOS) technology. In addition, the combination of FGMs with emerging two-dimensional (2D) materials to create 2D FGMs holds great promise for highly integrated NCH[30–33]. This is because the atomic thickness of 2D materials offers them exceptional gate control capability and large storage windows, and the van der Waals surface feature facilitates the feasibility of hetero-integration and compatibility with CMOS processes. Nevertheless, the high sensitivity of 2D materials to interfacial states and defect-related instabilities of dielectrics often result in bad long-term stability, poor endurance, and memory states of fewer than one hundred for 2D FGMs[31,34–38]. This poses a significant challenge for NCH based on 2D FGMs.

Here, we report gate-injection-mode (GIM) 2D FGMs with 8-bit states as candidates for large-scale NCH. Through a coplanar device structure design, the control gate (CG), floating gate, and channel are decoupled, and storing charges are programmed and erased from the CG through the shared tunneling layer. By adopting a bi-pulse state programming strategy, highly distinguishable (with intervals larger than three times of the standard deviations) and stable (with retention times longer than 10,000 s) 8-bit conductance states are achieved at 3 V programming voltage. This high state number as well as the small operation voltage is better than other types of nonvolatile memories based on field-effect transistors (FETs), including normal 2D FGMs, Si-Flash cells, and ferroelectric field-effect transistors (FeFETs). The devices also show symmetry state programming tendency and good endurance of over 10^5 cycles. In addition, fabricated 256 devices exhibit a 94.9% yield, good uniformity, and repeatability. Leveraging the above findings, we then carry out experimental image convolutions and project 38,592 convolutional kernel parameters on a 9 × 2 device array with results matching well with that of simulations. Finally, we show that fixed-point NNs with 8-bit precision have inference accuracies approaching the ideal values. Our work demonstrates the potential of GIM 2D FGMs for high-performance neuromorphic computing accelerators.

Results

8-bit-precision programming

GIM 2D FGMs with a device structure shown in Fig. 1a were designed to realize numerous distinguishable conductance levels. Here, monolayer/few-layer MoS_2, 5-nm Pt, and 8 nm Al_2O_3 were used as channel,

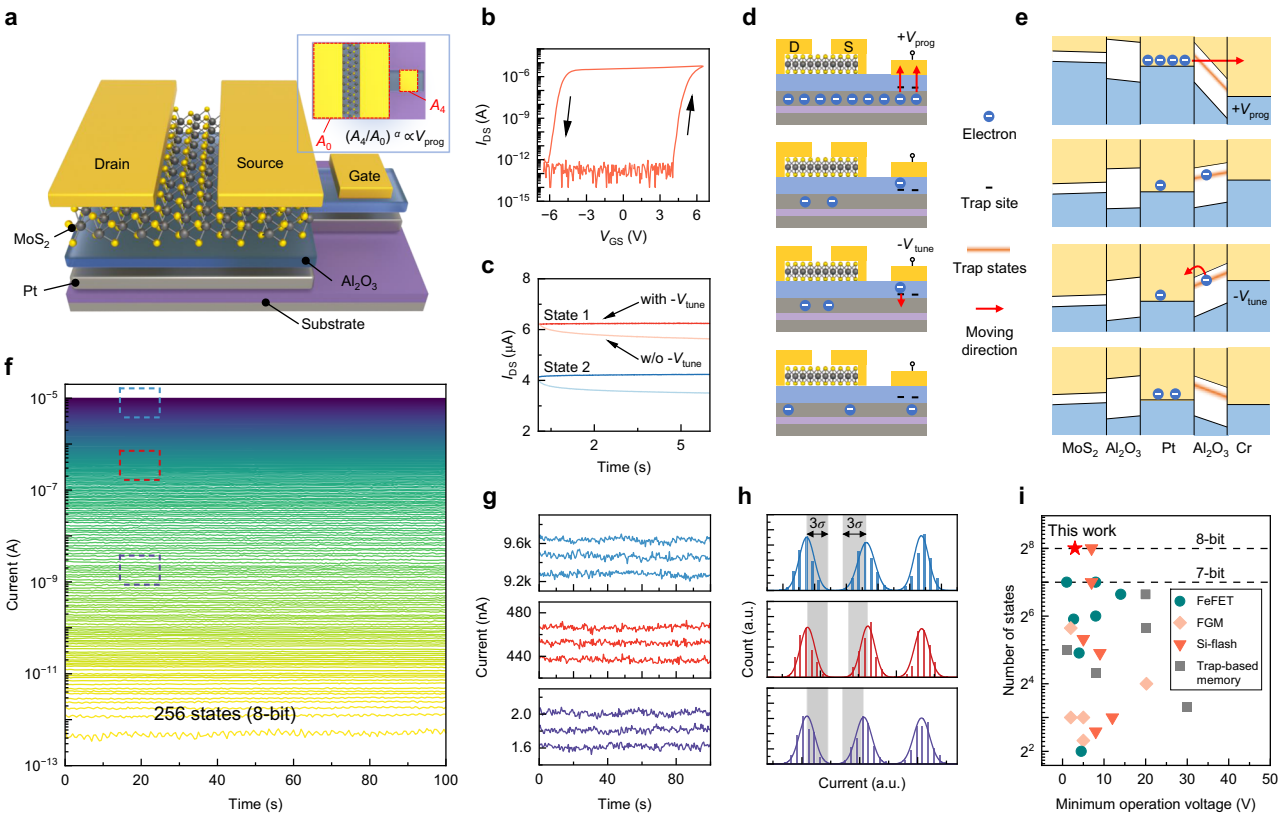

Fig. 1 | Programming of the GIM 2D FGM. a Device structure of the GIM 2D FGM. The inset shows the top view of the structure and the areas of the channel and gate are denoted as A_0 and A_4, respectively. MoS_2, Pt, and Al_2O_3 are used as channel, FG, and tunnelling/blocking layer. **b** Dual-sweep transfer curve that shows a large counterclockwise hysteresis loop. It was tested on the device with gate area $A_4 = 2.31\ \mu m^2$ and channel width/length of 10.37/1.47 μm (as indicated in the OM image of Fig. 2b). **c** Two conductance states after programming with $-V_{tune}$ (deep colors) and without $-V_{tune}$ (shallow colors). **d**, **e** Schematics and band diagrams of the programming and tuning process. Detailed energy values can be seen in the alignment diagram of bands in Supplementary Fig. 2. D drain, S source. **f** 256 states with each sampled for 100 s. The states were programmed using the bi-pulse programming method. **g**, **h** The distinguishable neighboring states at different current levels. **g** Enlarged current-time sampling plots at the corresponding sites denoted in (**f**) (in blue, red and purple, respectively). The corresponding histogram plots of the sampled currents are shown in (**h**). σ is the standard deviation and the fitted curves were attained by fitting with normal distribution function. **i** Benchmarks of the GIM 2D FGM in state number and operation voltage. FeFET ferroelectric field-effect transistor, FGM floating-gate memory. The data are collected from[19,29,30,54–73].

floating gate (FG), and tunnelling/blocking layer. An individual CG coplanar with the source and drain terminals works as both charge programming and erasing electrodes. Although approximately 22% more area may be required compared to a conventional vertical structure, the coplanar design enables the device to support vertical integration with fewer layers of materials (as analyzed in Supplementary Fig. 1). The detailed fabrication processes can be seen in the Methods section. There are several advantages of this design. First, unlike the vertically overlapped structure in a traditional FGM, here channel, FG, and CG are decoupled into two sections: channel-Al_2O_3-FG stack and CG-Al_2O_3-FG stack. Hence, the gate programming voltage can be easily regulated by changing the capacitive coupling ratio that is proportional to the area ratio between the CG and channel (denoted as A_4/A_0 in the inset of Fig. 1a). Second, a state programming strategy combining two sequential gate voltage pulses with opposite signs can be adopted to de-trap the unstable charges captured in the dielectrics, so that highly stable memory states without affecting the channel can be achieved. Third, the state programming is symmetry because of the shared charge tunneling and blocking layer, and the same charge injection and erasing mechanism. These advantages will be discussed in detail in the following sections.

The gate-injection mode is evidenced by a counterclockwise hysteresis in the double-sweep transfer curve as shown in Fig. 1b. Electrons can be injected in or erased from the FG when applying negative or positive voltages with high enough amplitudes on the gate terminal. And the stored charges will non-volatilely change the threshold voltage and the conductance of the MoS_2 FET. The large memory window (about 78% of the sweep range) results from the high-k dielectric layer of Al_2O_3, the ultra-thin channel of MoS_2, and most importantly, the high tunneling efficiency enabled by the optimized gate size, which will be further discussed below. Theoretically, the memory states, i.e., the conductance states of the FGM, should be stable because of the high energy barrier at the Pt/Al_2O_3 interface (~4.7 eV, Supplementary Fig. 2)[39–42]. Nevertheless, the source-drain current (I_{DS}) decreases immediately after voltage programming as seen from the light-colored lines shown in Fig. 1c. This phenomenon is widely observed in 2D FGMs[30,43,44], which is mainly coming from the unstable trapped charges inside the dielectrics during charge injection/erasing process that spontaneously de-trap after programming. And this is the same reason for the well-known bias temperature instability found in many Si-based transistors[45], especially those with high-k dielectrics like Al_2O_3 that have a range of widely distributed trap states near the conduction band[46].

To resolve the above problem, a programming method combining two sequential gate voltage pulses with opposite signs was adopted. Let us use the low-resistance state programming process as an example to illustrate this strategy (Fig. 1c–e). When a positive programming voltage (V_{prog}) is applied, the energy band of the tunneling layer Al_2O_3 can be largely tilted so that a triangle-shaped potential barrier appears (see the first panel of Fig. 1e). Hence, electrons stored in the FG can be erased through Fowler-Nordheim tunneling (FNT, see the first panel of Fig. 1d). The detailed analysis is shown in Supplementary Note 1 and Supplementary Fig. 3. However, some electrons are captured by the trap sites inside the tunneling layer during this process (see the second panels of Fig. 1d, e). After V_{prog} is withdrawn, the trapped electrons will de-trap into the FG by thermal activation in a slow relaxation process, which induces I_{DS} to decrease gradually. Note that, the subthreshold slope (SS) was nearly unchanged during the relaxation process, implying that the trap states were induced within the device fabrication process rather than generated by voltage programming (Supplementary Fig. 4)[38]. By applying a negative tuning pulse ($-V_{tune}$) soon after V_{prog} (see the third panels of Fig. 1d, e), the relaxation process can be largely accelerated through de-trapping the trapped electrons into the FG. Nearly all the trapped electrons can be eliminated after V_{tune} through optimization (see the fourth panels of Fig. 1d, e). As a result, more stable programmed states were attained (stable states in Fig. 1c). The effect of this strategy is obvious and applicable for the whole conductance range as evidenced by the comparison of time-dependent transfer curves between with and without bi-pulse optimizing (Supplementary Fig. 6). Temperature-dependent state retention properties were further studied using the Arrhenius equation (Supplementary Figs. 7–9). The largely decreased stored charge leakage activation energy after applying V_{tune} verifies the detraping effect of the bipolar programming strategy[47,48].

Using the above programming method, the GIM 2D FGM can have up to 256 distinguishable states (Fig. 1f and the output curves are shown in Supplementary Fig. 10, a detailed closed-loop programming method and corresponding parameters see Supplementary Figs. 11 and 12), which is equivalent to an 8-bit precision. The densely distributed states can be recognized from each other with an over-3σ variation (σ, the standard deviation for a state) between neighboring states (Fig. 1g, h). That state number is comparable to the advanced commercial Si-Flash cells and unprecedented among the other previously reported nonvolatile multibit memories based on FETs, including normal 2D FGMs and FeFETs (Fig. 1i and Supplementary Table 1). Note that most of the state numbers from the compared literature come from continuous voltage programming measurements or current-voltage curves rather than current-time curves as used here, which means the state stabilities were actually not well studied. By lowering the state variation to 1σ, even a doubled state number of 512 (9-bit precision) can be achieved (Supplementary Fig. 14). Moreover, the programming voltage can be decreased to a level of 3 V by optimizing the gate size, which is among the lowest ones according to literature (Fig. 1i).

Programmability and reliability

To investigate the programmability of the GIM 2D FGMs, we adopted a device circuit configuration shown in Fig. 2a. Here, V_{prog} was applied on a selected gate, a small source-drain bias (V_{DS}) of 0.1 V was applied on the drain terminal while the source was kept grounded, and the two equivalent capacitors of channel/FG and FG/gate 1 (C_0 and C_1) were connected in series. In that configuration, the voltage drop on the tunneling layer can be calculated by $V_{tunnel} = V_{prog}C_0/(C_0 + C_1)$. As a result, the programming efficiency during a single programming operation is strongly related to the capacitive coupling ratio between the equivalent lateral configurated capacitors, that is, the ratio C_i/C_0 ($i = 1, 2, 3, \ldots$) in Fig. 2a. To systematically investigate the gate-area-dependent programmability, we fabricated GIM 2D FGMs with varying-area multi gates (Fig. 2b, see Supplementary Fig. 15 for the fabrication process and Supplementary Fig. 16 for detailed geometric parameters).

It's worth noting that because the gates share the same oxide layer and FG, and the capacitor value is calculated by $C = \varepsilon A/(4\pi k d_{ox})$, where ε, A, k, d_{ox} are dielectric constant, effective area, electrostatic force constant and oxide layer thickness, the capacitor ratio C_i/C_0 can be directly calculated by the area ratio A_i/A_0 (the area ratio between gate i and the channel). As demonstrated in Fig. 2c, d, the dual-sweep transfer curves show an obvious area ratio dependency of the memory window, with the largest memory window of 10.3 V and the smallest memory window of 0.46 V. This difference is a direct result of the area-controlled partial voltage on the gate-Al_2O_3-FG stack. Simulated voltage potential distributions given in Supplementary Fig. 17 show similar results, validating the above analysis. The device can behave more like a transistor with a steep switch and a negligible memory window when the area ratio is very large, such as the case with an area ratio of 0.457 in Fig. 2c. That kind of device can be implemented as node selectors or activation function hardware in NNs.

The programming voltage can be decreased while maintaining a large memory window by using a smaller gate area, as shown in Fig. 2e. This dependency is consistent well with the simulation results (Supplementary Note 2 and Supplementary Fig. 18). The programming

https://doi.org/10.1038/s41467-025-58005-z

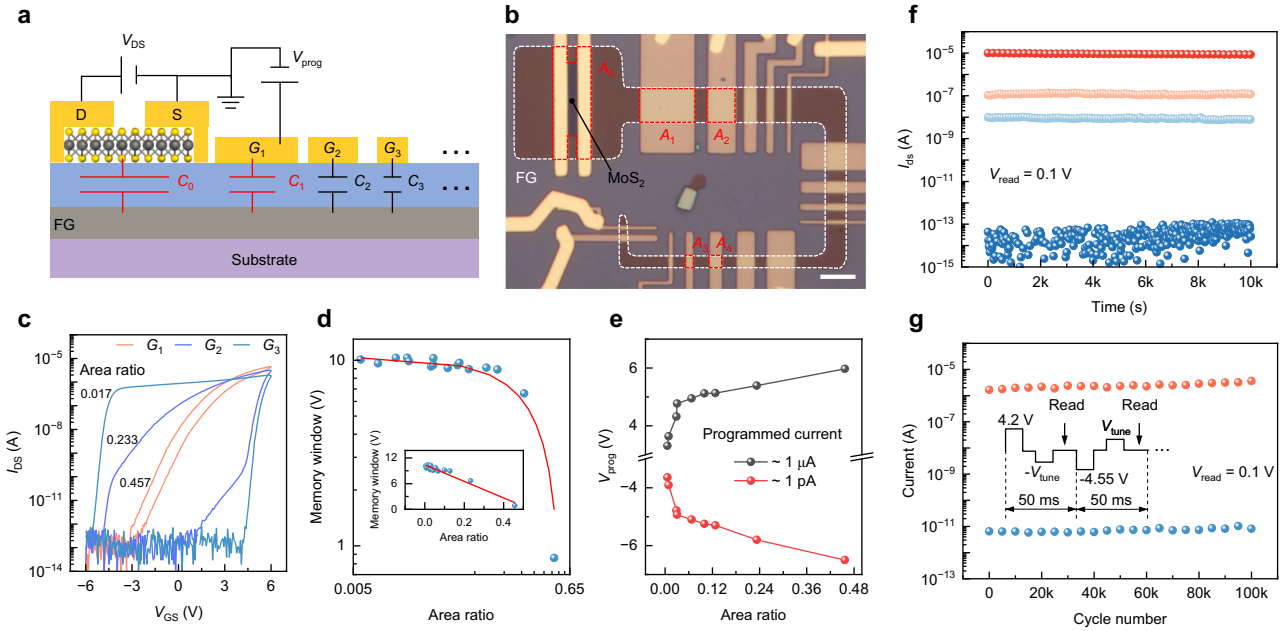

Fig. 2 | Performances of the GIM 2D FGMs. a Schematic diagram of the multi-gate configurated GIM 2D FGM and the equivalent circuit when V_{DS} and V_{prog} are applied to the corresponding terminals. The capacitor value (C_i) between a specific gate and the FG is strongly correlated with the gate area. **b** OM image of the multi-gate device. The channel MoS_2 thickness was determined to be about 2.6 nm by atomic force microscope (see Supplementary Fig. 23), equivalent to four layers of MoS_2[74]. The white-dashed and red-dashed areas denote the floating gate and channel/gate overlapping regions, respectively. The channel's width/length is 10.37/1.47 μm (see Supplementary Fig. 16 for detailed geometric parameters). Scale bar: 5 μm. **c** Dual-sweep transfer curves of selected three gates G_1, G_2, and G_3, which are corresponding to areas A_1, A_2, and A_3 in (**b**). **d** Memory window as a function of the area ratio A_i/A_0 with a linear fitting curve. Inset: the linear scale plot. **e** V_{prog} during carrier erasing/injection operation as a function of area ratio. The erasing and injection operations were conducted to alternatively change I_{DS} between ~1 uA and ~1 pA. **f** Retention of 4 exponentially separated states for 10,000 s. The states were read by a source-drain voltage of V_{read} = 0.1 V. **g** Endurance performance during 10^5 program/erase cycles. 5-us-width programming pulse and tuning pulse were applied with a gap of 200 μs in a single programming/erasing operation. Cycle period: 100 ms. The programming schematic is shown as the inset. V_{tune} = 2 V.

voltage can be as low as 3 V, showing potential in low-power applications. In addition, towards realizing the implementation of this device as the basic unit for NCH, the ability to update the device's weights (conductance states) in a small range under the guidance of a back-propagation algorithm is important for on-chip training processes. That ability was also proved by the quite symmetric state updating in positive and negative directions, which is because of the identical charge injection and erasing mechanism through the coplanar GIM design (Supplementary Fig. 19).

The device also showed stable programmed states for over 10,000 s while maintaining the largest on-off ratio of over 1×10^8 (Fig. 2f). Given the uniform oxide thickness in the channel and gate regions, the device's retention properties exhibit a clear dependence on the overlap areas between the floating gate and the drain, source, and gate electrodes (Supplementary Figs. 20–22). To further enhance the retention property, an additional blocking layer can be introduced below the source and drain regions to suppress this charge leakage pathway. And a good endurance performance of 10^5 cycles was also observed, which shows the reliability of being utilized for high-frequency weight update operations for on-chip training (Fig. 2g).

Repeatability of the 8-bit programming ability

We have fabricated 256 devices using a large-scale MoS_2 film grown by chemical vapor deposition (CVD) to study the repeatability of the 8-bit programming ability (Fig. 3, see Methods for the detailed fabrication process). The optical microscope (OM) image of the devices is shown in Fig. 3a, in which a typical area ratio is calculated to be 0.084 (see Supplementary Fig. 24 for geometric parameters). Of the 256 devices, about 13 devices were broken, which might be due to the discontinuous sites on the large-scale MoS_2 film introduced during the material transfer process, resulting a total yield rate of 94.9% (243 out of 256 devices). Apart from that, large hysteresis windows and the evenly distributed 9 programmed states can be observed from the electrical tests (Fig. 3b, c).

Moreover, we programmed 120 out of 137 devices with a yield of 87.6% into 256 (8-bit) distinct states, ranging from a current level of 1 pA to 100 nA (the original data are shown in Supplementary Figs. 27–31). The statistics of state current as a function of device number and state number are presented in Fig. 3d. These 120 devices exhibit an overall low device-to-device variation of below 4% for the programmed states (Fig. 3e, f), which can be largely attributed to the accurate programming method employed and the wide memory windows of the devices.

The 8-bit states, low programming voltage, good stability and endurance, good repeatability and scalability shown above demonstrate the potential of GIM 2D FGMs for NCH.

Hardware convolutions based on device arrays

Vector-matrix multiplications are the most important operation in NNs, like the representation transformation processes between neighboring layers and kernel filtering processes in convolution layers for feature extraction. In this section, we fabricated a 9 × 2 array, which is comparable to other configured arrays for analog computing (see Supplementary Table 3)[29–31,49–53], and carried out hardware convolutions to demonstrate the potential of GIM 2D FGMs for NCH. The optical image of an array bonded on a chip carrier is shown in Fig. 4a (see Methods and Supplementary Fig. 32 for the array fabrication process, see Supplementary Fig. 33 for geometric parameters). A homemade test system was used to experimentally run the convolution process as shown in Supplementary Fig. 34. The gate lines were

https://doi.org/10.1038/s41467-025-58005-z

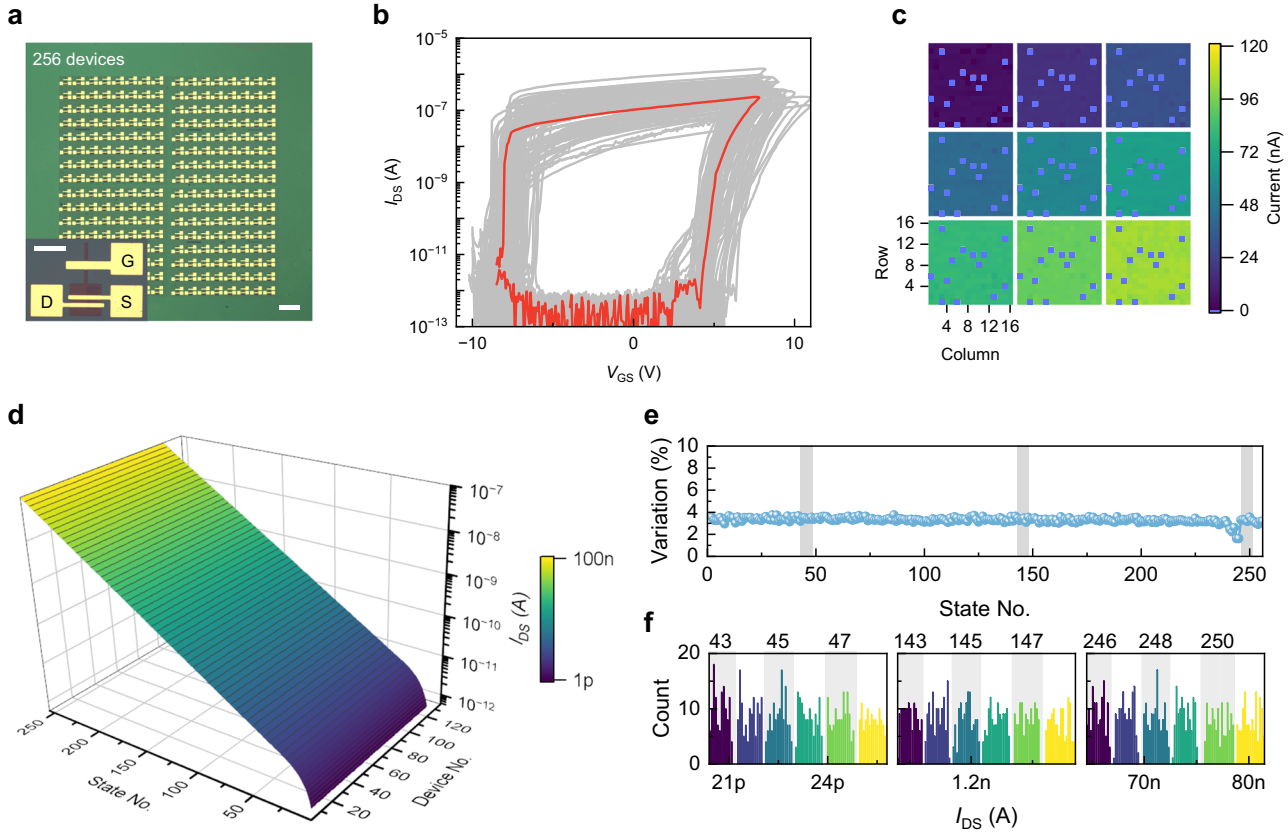

Fig. 3 | Uniformity and repeatability of 8-bit programming. a OM image of the fabricated 256 devices. Scale bar: 0.2 mm. Inset: OM image of a typical device among the 256 devices (scale bar: 30 µm). The channel's width/length is 11.22/3.11 µm (see Supplementary Fig. 24 for detailed geometric parameters). **b** Dual-sweep transfer curves of the devices. About 92.9% of devices have an on-state current exceeding 100 nA (Supplementary Fig. 25). The uniformity of transfer curves is comparable with previous works (Supplementary Fig. 26)[31,75]. **c** Current maps of the programmed 9 separate states. Sites in blue denote the 13 broken devices. **d** 256 programmed states for 120 devices. **e** device-to-device variation as a function of state number. **f** current distributions of selected adjacent states extracted from (**d**). Read voltage: 1 V. The programming method and parameters are the same as those used in Supplementary Figs. 11 and 12.

wired out for the programming operation, while the rows and columns were wired out and connected to every device's drain and source terminals respectively. As a 3 × 3 convolution kernel, the first column stores positive kernel weights and the second column stores negative kernel weights. That kernel configuration can eliminate possible parasitic currents (as analyzed in Supplementary Fig. 35) And the device structure also shows small parasitic capacitances and device-to-device interferences as thoroughly analyzed in Supplementary Note 3 and Supplementary Figs. 36–39. We adopted a parallel programming method for weight (conductance states) updating (Fig. 4b), i.e., devices in a selected row were programmed simultaneously by gate voltages with the common drain terminals grounded. And a row-by-row validation scheme was used to validate the kernel programmed (Supplementary Fig. 40). Additional discussions on the limitations when operating the device array can be found in Supplementary Note 4.

Figure 4c uses the convolution operation of image '0' in the MNIST dataset as an example to illustrate the inference process. The image pixels were converted into voltages based on greyscale and grouped into 3 × 3 patches. Then the pixels in each patch were imported as drain inputs to the array and the output currents on the source terminals were collected as the convolution results. With different kinds of kernels that were separately programmed onto the device array (Fig. 4d and Supplementary Fig. 41), the output images after convolutions show different features (Fig. 4e). The convolution results of another image from the Fashion MNIST dataset and the convolution results with large current outputs are also shown in Supplementary Figs. 42 and 43. The experimental output images show almost the same distributions with that of software-based convolutions (Fig. 4f), demonstrating the array works well as physical kernels for feature extraction.

Considering the 8-bit states realized on GIM 2D FGFESTs, more complex kernels can be mapped onto the 9 × 2 array for high-level feature extractions. Take the convolutional base of the large-scale convolutional neural network (CNN) VGG16 as an example. It contains a 5-block convolutional base, with each block containing several convolution layers and a pooling layer (Fig. 4g). All the 38,592 kernel parameters in the first block were mapped onto the 9 × 2 array kernel-by-kernel, as shown in Fig. 4h. We see the hardware-based kernels' weights show almost the same landscapes as the software-based values (Fig. 4i). A more direct comparison can be seen from the distributions of conductance and weight values (Fig. 4j, k). The above result implies the hardware integration capability for vector-matrix multiplication, and brings us the concept of incorporating GIM 2D FGMs in the whole body of large-scale NNs to validate the potential of constructing advanced NCH.

Convolutional neural networks with 8-bit precision

The accuracy of NNs with limited numerical states (fixed-point NNs) is an important issue for the practical application of NCH. We note that downloading a pre-trained NN to a local NCH and quantizing the weights with limited numerical states (quantization after training) is generally a more energy-efficient approach. Therefore, to demonstrate the potential of GIM 2D FGMs array for NCH (Fig. 5a), pre-trained large-scale convolutional neural networks (CNNs) were used for ImageNet

https://doi.org/10.1038/s41467-025-58005-z

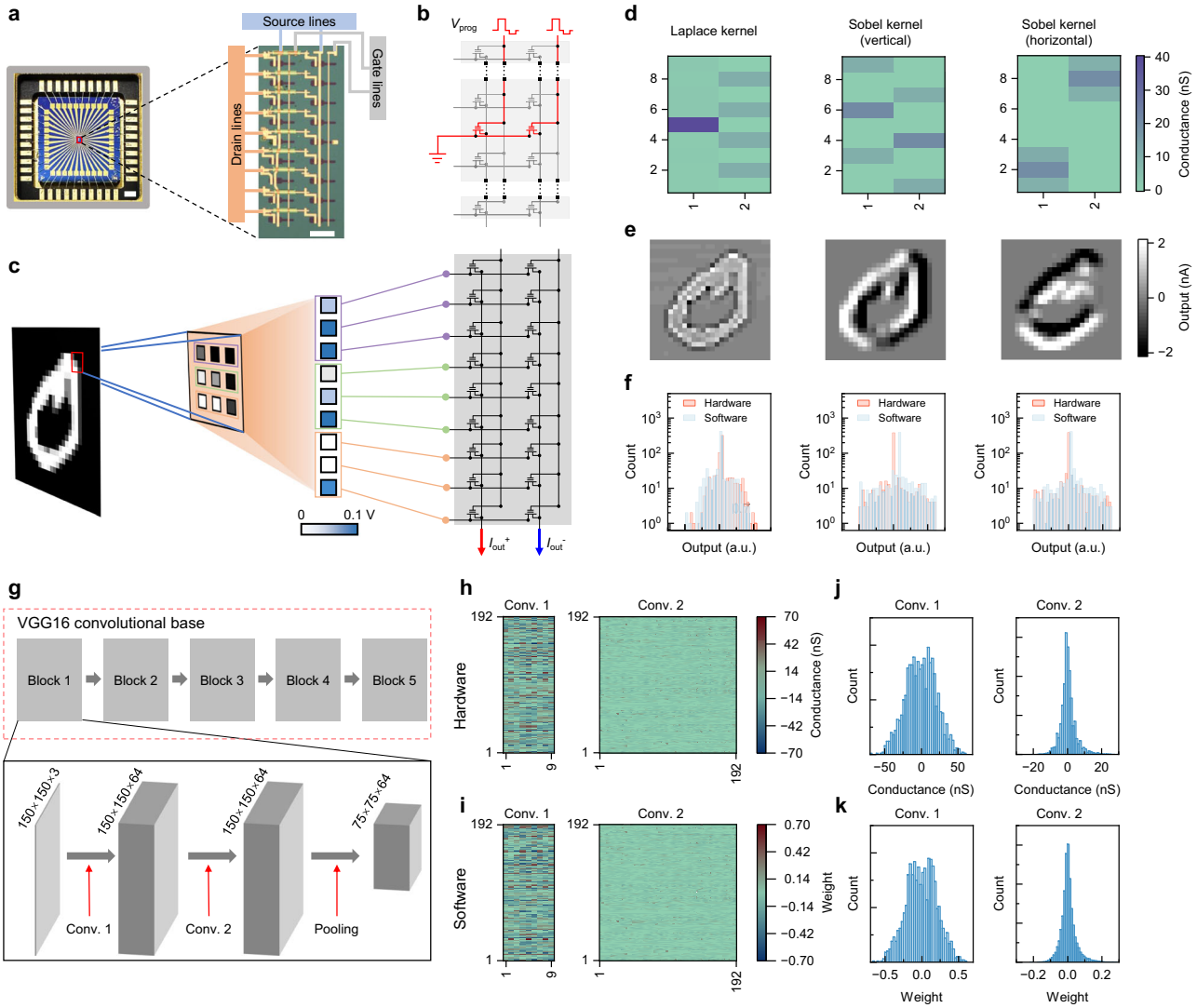

Fig. 4 | Hardware convolutions using GIM 2D FGMs array. a Photo (left, scale bar: 1 mm) and OM image (right, scale bar: 40 μm) of the wired 9 × 2 array. The electrode lines are also shown. The channel's width/length is 3.95/2.47 μm (see Supplementary Fig. 33 for detailed geometric parameters). **b** Parallel programming method for a selected row of the array. **c** Illustration of the vector-matrix multiplication operation for image convolution. The kernel weights were mapped as conductance states of the device array before each convolution process. During the convolution process, 3 × 3 patches of pixels were converted to drain voltage inputs patch-by-patch, with the patches sliding through the whole image row-by-row. **d** Conductance maps of three kinds of kernels that were mapped to the array. **e** The corresponding convolution results mapped into the source-drain current. **f** Comparisons between output current distributions in (**e**) (hardware) and software-based convolution results. The results have been normalized. **g** Illustration of the VGG16 convolutional base structure. There are 5 convolution blocks with each containing several convolution layers and a pooling layer. **h**, **i** Hardware-based conductance maps of the two convolution layers in block 1 (**h**) and the corresponding software-based weight maps (**i**). **j**, **k** The corresponding histogram plots of hardware-based conductances (**j**) and software-based weights (**k**).

dataset recognition (Fig. 5b). Here, the large number of parameters in these CNNs were quantized to the 8-bit states of the GIM 2D FGM using a nearest-rounding method. According to the simulation results with different bit precisions (the 4-bit, 5-bit, 6-bit, and 7-bit states adopted are shown in Supplementary Fig. 44), a 8-bit precision is sufficient for CNNs to achieve high recognition accuracy compared to their unlimited-precision version (Fig. 5c and Supplementary Fig. 45). It's important to note that while 8-bit precision achieves a higher recognition accuracy (89.43%) compared to lower precisions (such as 88.96% for 7-bit) for the smallest MobileNet model, 7-bit precision is sufficient for the larger Xception model. This suggests that larger CNNs can operate effectively with lower bit precision. However, from a practical perspective, deploying small-sized NNs on edge devices is typically more energy-efficient. Therefore, the higher 8-bit precision storage for these small-sized CNNs is crucial for enhancing their performance.

An alternative approach involves directly training fixed-point NNs on the NCH with limited states (quantization during training). Even though this approach consumes much more energy and time compared to quantization after training, which is mainly due to the large-scale weight updating, it offers greater flexibility by adapting to specific tasks through weight fine-tuning. Through simulation of quantization during training (Fig. 5d), we observed that the advantage of 8-bit precision over lower precisions is still very obvious for both MobileNet and Xception models. This is further supported by results from a simpler model for MNIST recognition (Supplementary Fig. 46). However, an overall accuracy decrease is observed across all fixed precisions compared to quantization after training, likely due to the

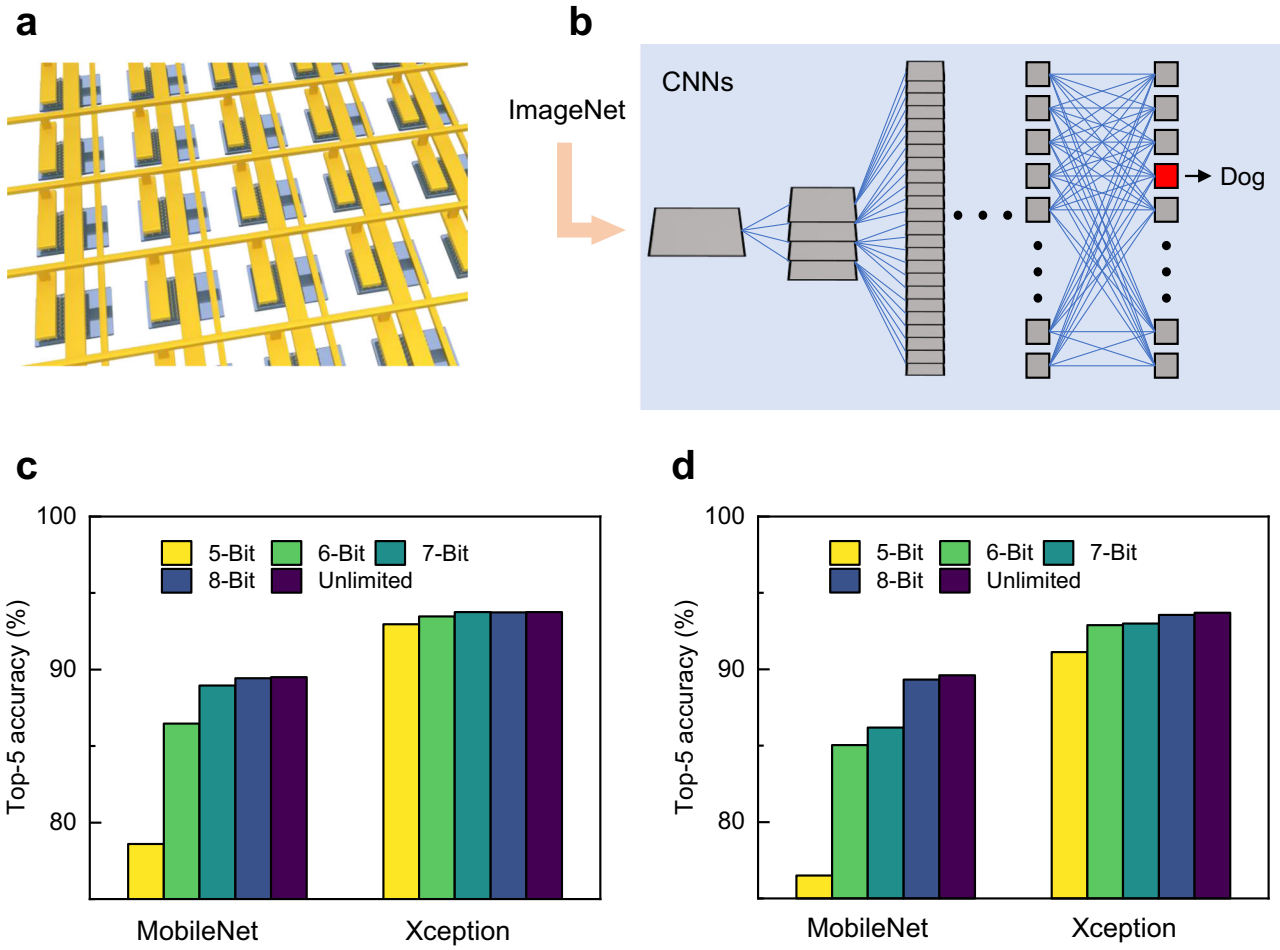

Fig. 5 | Image recognition using CNNs with different precisions. a Schematic of a large-scale GIM 2D FGM array for vector-matrix multiplication in neural networks. **b** Schematic of CNNs for ImageNet image recognition. **c**, **d** Comparison of top-5 accuracy for CNNs quantized with different precisions after (**c**) and during (**d**) training. The quantization process used the nearest rounding scheme. The numbers of parameters are denoted in the brackets following the models' names (4.3 million of MobileNet, 22.9 million of Xception).

reduced efficiency of the training process caused by inaccurate weight updates at lower precisions.

Another important point is the choice of rounding scheme. In the above simulations, a nearest-rounding scheme was adopted. However, according to previous reports[10–13], a stochastic rounding scheme can enhance NNs performance. To validate this, we have reconducted the simulations using a stochastic rounding scheme (Supplementary Fig. 47), and the results showed an obvious accuracy increase for all the fixed precisions, especially for lower precisions such as 5-bit and 6-bit, confirming the benefits of stochastic rounding. Combined with the demonstrated capabilities of vector-matrix multiplication and high repeatability of 8-bit programming, GIM 2D FGMs show great promise for system-level-integrated vector-matrix multiplication arrays in NN accelerators.

Discussion

To sum up, we have designed 2D floating-gate memories working in a gate-injection mode as potential device units for large-scale NCH. The CG, floating gate, and channel are decoupled through this design, so that a bi-pulse state programming strategy could be adopted to realize 8-bit conductance states. This is because the subsequent tuning voltage can promote the de-trapping process of unstable charges captured by the dielectric defects that have a lower potential barrier. The states are highly distinguishable with intervals larger than three times the standard deviations and very stable with retention times longer than 10,000 s. The devices also show good endurance of over 10^5 cycles. In addition, because charges are injected and erased from the CG through the shared Al_2O_3 layer via FNT, the state programming is almost symmetry. And through changing the capacitance ratio by varying the aera of the CG, a 3 V programming voltage can be achieved. Moreover, the fabricated 256 devices exhibit a 94.9% yield, good uniformity and repeatability. Then, a 9 × 2 device array was fabricated and experimental image convolutions were carried out with results matching well with that of software simulations. Leveraging the device's multi-state programming capability, we successfully transferred 38,592 convolutional kernel parameters from a pre-trained VGG16 network to the array. Finally, we studied the image recognition accuracies of fixed-point NNs containing different levels of precisions. Notably, no matter whether NNs designed by downloading pre-trained networks or directly training networks locally, the inference accuracies at 8-bit precision could approach the ideal values. Our work validates the potential of GIM 2D FGMs for high-performance neuromorphic computing accelerators.

Methods

Device fabrication

A p-doped silicon substrate with 300 nm thermal-oxidized SiO_2 was firstly coated with poly(methyl methacrylate) (PMMA) and baked for 2 min at 150 °C. After that, the floating gate Pt was patterned and deposited by electron beam lithography (EBL) and electron beam

evaporation, respectively. After a standard lift-off process, a layer of Al_2O_3 with 8-nm thickness was deposited on the floating gate by atomic layer deposition (ALD). The ALD was processed at 150 °C, using water and trimethylaluminum as precursors. Then, mechanically exfoliated MoS_2 (purchased from Shanghai Onway Technology Co., Ltd.) with Scotch tape was transferred onto the top surface of the Al_2O_3/Pt stack by a standard wet-transfer method, using polypropylene carbonate and polydimethylsiloxane as holders. At last, source, drain and gate electrodes of Cr/Au (8/80 nm) were patterned and deposited using EBL and thermal evaporation. To fabricate the 256 devices, a large-scale few-layer MoS_2 was grown by CVD on a 1 × 0.5-cm-sized sapphire substrate. The CVD-grown material was transferred with PMMA, patterned through EBL and etched with CF_4 and O_2 through reactive ion etching.

Array fabrication

Before the fabrication of the wired 9 × 2 array, the CVD-grown MoS_2 was transferred onto a substrate on which the wiring metal patterns were pre-deposited, following the transfer process illustrated in Supplementary Fig. 48. During the fabrication process, a 25-nm-thick layer of ALD-deposited HfO_2 was used as the insulating layer for the isolation of the overlapped drain and gate lines in the array. The other array fabrication steps were the same as the abovementioned device fabrication process.

Electronic measurements

Except for the 9 × 2 array, the electronic performance of the as-fabricated devices was tested on a probe station (Lakeshore, TTP4) under a high vacuum condition ($<10^{-6}$ Torr), which is equipped with Keysight B1500A semiconductor analyzer system. All tests on the 9 × 2 array were conducted on a homemade probe station equipped with an electrical testing system (National Instruments, cDAQ-9189) under atmosphere conditions.

Simulation of large-scale CNNs

The adopted large-scale CNNs are pre-trained models loaded from the Keras platform and they were coded with Python scripts for convenient handling of the internal weights. The ImageNet samples incorporated here for evaluation were all collected from the ILSVRC2012 validation data set, which contains 50,000 images with each labelled with its class. Before evaluation, all the pre-trained weights of CNNs were replaced by the normalized conductance states with the corresponding bit precisions. During the evaluation process for each kind of CNN, the 50,000 images were clipped to a certain size of 224 × 224 and taken as the inputs of the model one by one. The output scores were translated to the recognized class for every image and all the correctly recognized images were summed for the calculation of the evaluated final recognition accuracy on this data set. The three-layer FCNN was also constructed on the Keras platform layer-by-layer. The relu function was used as the activation function, the cross-entropy method was used to calculate the loss function, and a learning rate of 10^{-3} was adopted for model training.

Data availability

The data that support the findings of this study are available within the paper and supplementary information. Source data are provided with this paper.

Code availability

The code of neural network training and evaluation used in this study is available at https://doi.org/10.7910/DVN/BDVBVC.

References

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
2. Mozaffari, S., Al-Jarrah, O. Y., Dianati, M., Jennings, P. & Mouzakitis, A. Deep learning-based vehicle behavior prediction for autonomous driving applications: a review. *IEEE Trans. Intell. Transp. Syst.* **23**, 33–47 (2022).
3. Bi, K. et al. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **619**, 533–538 (2023).
4. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).
5. Ambrogio, S. et al. Equivalent-accuracy accelerated neural-network training using analogue memory. *Nature* **558**, 60–67 (2018).
6. Zidan, M. A., Strachan, J. P. & Lu, W. D. The future of electronics based on memristive systems. *Nat. Electron.* **1**, 22–29 (2018).
7. Mead, C. Neuromorphic electronic systems. *Proc. IEEE* **78**, 1629–1636 (1990).
8. Xia, Q. & Yang, J. J. Memristive crossbar arrays for brain-inspired computing. *Nat. Mater.* **18**, 309–323 (2019).
9. Burr, G. W. et al. Neuromorphic computing using non-volatile memory. *Adv. Phys. -X* **2**, 89–124 (2017).
10. Gupta, S., Agrawal, A., Gopalakrishnan, K. & Narayanan, P. Deep learning with limited numerical precision. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)* (eds Bach, F. & Blei, D.) 1737–1746 (JMLR, 2015).
11. Merolla, P., Appuswamy, R., Arthur, J., Esser, S. K. & Modha, D. Deep neural networks are robust to weight binarization and other non-linear distortions. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1606.01981 (2016).
12. Na, T., Ko, J. H., Kung, J. & Mukhopadhyay, S. On-chip training of recurrent neural networks with limited numerical precision. In *International Joint Conference on Neural Networks (IJCNN)* (eds Choe, Y., Jayne, C., Hammer, B., & King, I.) 3716–3723 (IEEE, 2017).
13. Sakr, C., Kim, Y. & Shanbhag, N. Analytical guarantees on numerical precision of deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)* (eds Precup, D. & The, Y. W.) 3007–3016 (JMLR, 2017).
14. Bianco, S., Cadene, R., Celona, L. & Napoletano, P. Benchmark analysis of representative deep neural network architectures. *IEEE Access* **6**, 64270–64277 (2018).
15. Hasler, P., Diorio, C., Minch, B. A. & Mead, C. Single transistor learning synapses. In *Proceedings of the 7th International Conference on Neural Information Processing Systems (NeurIPS)* (eds Tesauro, G., Touretzky, D. S., & Leen, T. K.) 817–824 (MIT Press, 1994).
16. Diorio, C., Hasler, P., Minch, A. & Mead, C. A. A single-transistor silicon synapse. *IEEE Trans. Electron Devices* **43**, 1972–1980 (1996).
17. Diorio, C., Hasler, P., Minch, B. A. & Mead, C. A. A floating-gate MOS learning array with locally computed weight updates. *IEEE Trans. Electron Devices* **44**, 2281–2289 (1997).
18. Bayat, F. M. et al. Redesigning commercial floating-gate memory for analog computing applications. In *2015 IEEE International Symposium on Circuits and Systems (ISCAS)* (eds Fernandes, J. & Serdijn, W.) 1921–1924 (IEEE, 2015).
19. Gu, X., Wan, Z. & Iyer, S. S. Charge-trap transistors for CMOS-only analog memory. *IEEE Trans. Electron Devices* **66**, 4183–4187 (2019).
20. Danial, L. et al. Two-terminal floating-gate transistors with a low-power memristive operation mode for analogue neuromorphic computing. *Nat. Electron.* **2**, 596–605 (2019).
21. Chua, L. Memristor-The missing circuit element. *IEEE Trans. Circuit Theory* **18**, 507–519 (1971).
22. Ielmini, D. & Wong, H. S. P. In-memory computing with resistive switching devices. *Nat. Electron.* **1**, 333–343 (2018).
23. Jo, S. H. et al. Nanoscale memristor device as synapse in neuromorphic systems. *Nano Lett.* **10**, 1297–1301 (2010).

24. Song, W. et al. Programming memristor arrays with arbitrarily high precision for analog computing. *Science* **383**, 903–910 (2024).
25. Suri, M. et al. Phase change memory as synapse for ultra-dense neuromorphic systems: application to complex visual pattern extraction. In *2011 International Electron Devices Meeting (IEDM)* (eds Ishimaru, K., Misra, V. & Ghani, T.) 4.4.1–4.4.4 (IEEE, 2011).
26. Burr, G. W. et al. Experimental demonstration and tolerancing of a large-scale neural network (165 000 synapses) using phase-change memory as the synaptic weight element. *IEEE Trans. Electron Devices* **62**, 3498–3507 (2015).
27. Aziz, A. et al. Computing with ferroelectric FETs: devices, models, systems, and applications. In *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)* (eds Madsen, J. & Coskun, A. K.) 1289–1298 (IEEE, 2018).
28. Khan, A. I., Keshavarzi, A. & Datta, S. The future of ferroelectric field-effect transistor technology. *Nat. Electron.* **3**, 588–597 (2020).
29. Ning, H. et al. An in-memory computing architecture based on a duplex two-dimensional material structure for in situ machine learning. *Nat. Nanotechnol.* **18**, 493–500 (2023).
30. Migliato Marega, G. et al. Logic-in-memory based on an atomically thin semiconductor. *Nature* **587**, 72–77 (2020).
31. Migliato Marega, G. et al. A large-scale integrated vector-matrix multiplication processor based on monolayer molybdenum disulfide memories. *Nat. Electron.* **6**, 991–998 (2023).
32. Liu, L. et al. Ultrafast non-volatile flash memory based on van der Waals heterostructures. *Nat. Nanotechnol.* **16**, 874–881 (2021).
33. Cao, W., Kang, J., Bertolazzi, S., Kis, A. & Banerjee, K. Can 2D-nanocrystals extend the lifetime of floating-gate transistor based nonvolatile memory? *IEEE Trans. Electron Devices* **61**, 3456–3464 (2014).
34. Giusi, G., Marega, G. M., Kis, A. & Iannaccone, G. Impact of interface traps in floating-gate memory based on monolayer MoS_2. *IEEE Trans. Electron Devices* **69**, 6121–6126 (2022).
35. Sup Choi, M. et al. Controlled charge trapping by molybdenum disulphide and graphene in ultrathin heterostructured memory devices. *Nat. Commun.* **4**, 1624 (2013).
36. Bertolazzi, S., Krasnozhon, D. & Kis, A. Nonvolatile memory cells based on MoS_2/graphene heterostructures. *ACS Nano* **7**, 3246–3252 (2013).
37. Kaczer, B. et al. A brief overview of gate oxide defect properties and their relation to MOSFET instabilities and device and circuit time-dependent variability. *Microelectron. Reliab.* **81**, 186–194 (2018).
38. Zafar, S., Kumar, A., Gusev, E. & Cartier, E. Threshold voltage instabilities in high-κ gate dielectric stacks. *IEEE Trans. Device Mater. Reliab.* **5**, 45–64 (2005).
39. Wu, H. et al. Multifunctional half-floating-gate field-effect transistor based on MoS_2-BN-graphene van der Waals heterostructures. *Nano Lett.* **22**, 2328–2333 (2022).
40. Lee, S., Peng, R., Wu, C. & Li, M. Programmable black phosphorus image sensor for broadband optoelectronic edge computing. *Nat. Commun.* **13**, 1485 (2022).
41. Hu, W. et al. Ambipolar 2D semiconductors and emerging device applications. *Small Methods* **5**, e2000837 (2021).
42. Liu, X., Choi, M. S., Hwang, E., Yoo, W. J. & Sun, J. Fermi level pinning dependent 2D semiconductor devices: challenges and prospects. *Adv. Mater.* **34**, e2108425 (2022).
43. Wu, E. et al. Tunable and nonvolatile multibit data storage memory based on $MoTe_2$/boron nitride/graphene heterostructures through contact engineering. *Nanotechnology* **31**, 485205 (2020).
44. Wang, Y. et al. Band-tailored van der Waals heterostructure for multilevel memory and artificial synapse. *InfoMat* **3**, 917–928 (2021).
45. Cartier, E., Linder, B. P., Narayanan, V. & Paruchuri, V. K. Fundamental understanding and optimization of PBTI in nFETs with SiO_2/HfO_2 gate stack. In *2006 International Electron Devices Meeting (IEDM)* (ed. Mistry, K.) 1–4 (IEEE, 2006).
46. Vais, A. et al. On the distribution of oxide defect levels in Al_2O_3 and HfO_2 high-k dielectrics deposited on InGaAs metal-oxide-semiconductor devices studied by capacitance-voltage hysteresis. *J. Appl. Phys.* **121**, 144504 (2017).
47. Molas, G. et al. Reliability of charge trapping memories with high-k control dielectrics (Invited Paper). *Microelectron. Eng.* **86**, 1796–1803 (2009).
48. Bersuker, G. et al. Mechanism of electron trapping and characteristics of traps in HfO_2 gate stacks. *IEEE Trans. Device Mater. Reliab.* **7**, 138–145 (2007).
49. Mennel, L. et al. Ultrafast machine vision with 2D material neural network image sensors. *Nature* **579**, 62–66 (2020).
50. Tong, L. et al. 2D materials–based homogeneous transistor-memory architecture for neuromorphic hardware. *Science* **373**, 1353–1358 (2021).
51. Pan, C. et al. Reconfigurable logic and neuromorphic circuits based on electrically tunable two-dimensional homojunctions. *Nat. Electron.* **3**, 383–390 (2020).
52. Lee, M.-P. et al. Silicon–van der Waals heterointegration for CMOS-compatible logic-in-memory design. *Sci. Adv.* **9**, eadk1597 (2023).
53. Huang, X. et al. An ultrafast bipolar flash memory for self-activated in-memory computing. *Nat. Nanotechnol.* **18**, 486–492 (2023).
54. Luo, Z. D. et al. Dual-ferroelectric-coupling-engineered two-dimensional transistors for multifunctional in-memory computing. *ACS Nano* **16**, 3362–3372 (2022).
55. Zhou, Y. et al. A reconfigurable two-WSe_2-transistor synaptic cell for reinforcement learning. *Adv. Mater.* **34**, 2107754 (2022).
56. Wang, S. et al. Two-dimensional ferroelectric channel transistors integrating ultra-fast memory and neural computing. *Nat. Commun.* **12**, 53 (2021).
57. Kim, M.-K., Kim, I.-J. & Lee, J.-S. CMOS-compatible compute-in-memory accelerators based on integrated ferroelectric synaptic arrays for convolution neural networks. *Sci. Adv.* **8**, eabm8537 (2022).
58. Kim, K. H. et al. Scalable CMOS back-end-of-line-compatible AlScN/two-dimensional channel ferroelectric field-effect transistors. *Nat. Nanotechnol.* **18**, 1044–1050 (2023).
59. Soliman, T. et al. First demonstration of in-memory computing crossbar using multi-level Cell FeFET. *Nat. Commun.* **14**, 6348 (2023).
60. Li, Y. et al. Low-voltage ultrafast nonvolatile memory via direct charge injection through a threshold resistive-switching layer. *Nat. Commun.* **13**, 4591 (2022).
61. Liu, C. et al. A semi-floating gate memory based on van der Waals heterostructures for quasi-non-volatile applications. *Nat. Nanotechnol.* **13**, 404–410 (2018).
62. Wu, L. et al. Atomically sharp interface enabled ultrahigh-speed non-volatile memory devices. *Nat. Nanotechnol.* **16**, 882–887 (2021).
63. Won, U. Y. et al. Multi-neuron connection using multi-terminal floating-gate memristor for unsupervised learning. *Nat. Commun.* **14**, 3070 (2023).
64. Wang, W. et al. A memristive deep belief neural network based on silicon synapses. *Nat. Electron.* **5**, 870–880 (2022).
65. Lee, S. T., Kwon, D., Kim, H., Yoo, H. & Lee, J. H. NAND flash based novel synaptic architecture for highly robust and high-density quantized neural networks with binary neuron activation of (1, 0). *IEEE Access* **8**, 114330–114339 (2020).
66. Lee, S. T. et al. Operation scheme of multi-layer neural networks using NAND flash memory as high-density synaptic devices. *IEEE J. Electron Devices Soc.* **7**, 1085–1093 (2019).
67. Choi, H. S. et al. 3-D floating-gate synapse array with spike-time-dependent plasticity. *IEEE Trans. Electron Devices* **65**, 101–107 (2018).

68. Gao, S. et al. Programmable linear RAM: a new flash memory-based memristor for artificial synapses and its application to speech recognition system. In *2019 IEEE International Electron Devices Meeting (IEDM)* (eds Takayanagi, M., Datta, S. & Grasser, T.) 14.11.11–14.11.14 (IEEE, 2019).
69. Yang, F.-S. et al. Oxidation-boosted charge trapping in ultra-sensitive van der Waals materials for artificial synaptic features. *Nat. Commun.* **11**, 2972 (2020).
70. Xiang, D., Liu, T., Zhang, X., Zhou, P. & Chen, W. Dielectric engineered two-dimensional neuromorphic transistors. *Nano Lett.* **21**, 3557–3565 (2021).
71. Sangwan, V. K. et al. Multi-terminal memtransistors from polycrystalline monolayer molybdenum disulfide. *Nature* **554**, 500–504 (2018).
72. Nam, J. H. et al. Low power MoS_2/Nb_2O_5 memtransistor device with highly reliable heterosynaptic plasticity. *Adv. Funct. Mater.* **31**, 2104174 (2021).
73. Judy, M. et al. A digitally interfaced analog correlation filter system for object tracking applications. *IEEE Trans. Circuits Syst. I-Regul. Pap.* **65**, 2764–2773 (2018).
74. Radisavljevic, B., Radenovic, A., Brivio, J., Giacometti, V. & Kis, A. Single-layer MoS_2 transistors. *Nat. Nanotechnol.* **6**, 147–150 (2011).
75. Jiang, Y. et al. A scalable integration process for ultrafast two-dimensional flash memory. *Nat. Electron.* **7**, 868–875 (2024).

Acknowledgements

This work was supported by National Key R&D Program of China (2021YFA1201500), National Natural Science Foundation of China (Nos. U24A2055, U24A20302, 62374048, 62274046, 12204122, 62204217). The authors also gratefully acknowledge the support of Youth Innovation Promotion Association CAS and CAS Key Laboratory of Nanosystem and Hierarchical Fabrication.

Author contributions

Y.C., F.W., and Z.W. conceived the idea and designed experimental schemes. Y.H. and L.Y. provided the CVD-grown materials. Y.C. and J.Y. fabricated the devices and carried out electrical measurements. S.L., Y.W., T.Y., and S.Y. assisted with the array fabrication. Y.C. and F.W. analyzed the data and co-wrote the manuscript in consultation with all the other authors. F.W., S.L., Z.W., X.Z., and J.H. provided funding support. Z.W. supervised the project. All authors participated in manuscript revision and refinement.

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/s41467-025-58005-z.

Correspondence and requests for materials should be addressed to Feng Wang or Zhenxing Wang.

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