nature communications

Article

https://doi.org/10.1038/s41467-024-54218-w

High performance Si-MoS_2 heterogeneous embedded DRAM

Received: 25 March 2024

Accepted: 31 October 2024

Published online: 12 November 2024

Check for updates

Kai Xiao[1,5], Jing Wan[1,5] ✉, Hui Xie[1,5], Yuxuan Zhu[2,5], Tian Tian[1], Wei Zhang[1], Yingxin Chen[1], Jinshu Zhang[2], Lihui Zhou[3], Sheng Dai[3], Zihan Xu[4], Wenzhong Bao[2] ✉ & Peng Zhou[2] ✉

Embedded Dynamic RAM (eDRAM) has become a key solution for large-capacity cache in high-performance processors. A heterogeneous two transistor capacitorless eDRAM (2T-eDRAM) that combines silicon and molybdenum disulfide (MoS_2) is reported to address the short retention issue in conventional gain cell (GC) eDRAMs meanwhile eliminate the pillar capacitor in one transistor and one capacitor (1T1C) eDRAMs. The MoS_2 write transistor with low OFF current (I_{OFF}) enables long data retention, while the Si read transistor offers high drive current and logic compatibility. This combination enhances data retention by 1000 times and sense margin by 100 times respectively compared to full Si and MoS_2 counterparts. A three-dimensional (3D) design stacking MoS_2 on Si is demonstrated with back-end-of-line (BEOL) process to double integration density. With 6000 s data retention, 35 μA/μm sense margin, 5 ns access speeds, 3D integration and CMOS logic compatibility, this Si-MoS_2 eDRAM marks a significant advancement in memory technology.

Modern computer architecture is characterized by a deep memory hierarchy. In order to bridge the gap between high-speed processing units (CPU and GPU) and relatively low-speed main memory (dynamic random access memory, DRAM), cache memory is inserted between the processing unit and DRAM to decrease data access time and reduce latency. The capacity of the cache memory is getting larger and larger in high-performance computer system. Conventional cache using six-transistors static RAM (6T-SRAM) has excellent compatibility with logic process, but suffers from low integration density and high power consumption[1]. To achieve a large capacity cache, novel technologies are being developed by industry and academia. AMD has been developing 3D V-cache technology by stacking SRAM chips to achieve large capacity Level-3 cache[2,3]. Whereas, Intel, TSMC, Samsung, and IBM have been bringing the embedded DRAM (eDRAM) with one-transistor/one-capacitor structure(1T1C-DRAM) in Level-3 and level-4 caches to achieve higher integration density and lower power consumption than SRAM[4–8]. However, due to charge sharing operation and the requirement for additional capacitors, 1T1C eDRAM suffers from limited voltage scaling down, destructive reading, and complex capacitor fabrication[9–12].

In order to achieve high integration density and CMOS logic compatible process, the gain cell eDRAM (GC-eDRAM) has been studied dramatically. GC-eDRAM combines two to four transistors to achieve read, write, and random-access functionalities[13,14]. Its process is fully compatible with the CMOS logic process and exhibits higher integration density than 6T-SRAM. However, the data of GC-eDRAM is stored directly in the gate capacitor of the storage transistor. As the size of the transistor scales down, the storage capacitance is reduced, and the OFF current of the write transistor increases due to short channel effects (SCEs). This significantly degrades the data retention of

[1]School of Information Science and Technology, Fudan University, Shanghai, P. R. China. [2]State Key Laboratory of ASIC and System, Fudan University, Shanghai, P. R. China. [3]Key Laboratory for Advanced Materials and Feringa Nobel Prize Scientist Joint Research Center, Institute of Fine Chemicals, School of Chemistry & Molecular Engineering, East China University of Science and Technology, Shanghai, P. R. China. [4]Shenzhen Sixcarbon Technology, Shenzhen, P. R. China. [5]These authors contributed equally: Kai Xiao, Jing Wan, Hui Xie, Yuxuan Zhu. ✉e-mail: jingwan@fudan.edu.cn; baowz@fudan.edu.cn; pengzhou@fudan.edu.cn

the GC-eDRAM to the level of a microsecond. In the advanced technology node, GC-eDRAM is not applicable due to the short retention time. In order to tackle the challenge of data retention, the incorporation of wide-bandgap semiconductors, such as amorphous oxide semiconductors (AOS), as channel materials have shown promise. The OFF current of the transistor decreases exponentially with increasing bandgap, translating to significantly prolonged data retention times[15–19]. Nonetheless, AOS materials exhibit relatively low carrier mobility, leading to a reduced transistor ON current (I_{ON}) and, consequently, a narrow sense margin. Methods including oxygen plasma annealing, reducing channel width, ultrathin film channels, and dual or triple gate electrostatic control have been proposed to optimize performance, albeit at the cost of additional processing steps and more complex fabrication[17,20–22].

In this work, a heterogenous GC-eDRAM combining silicon (Si) and molybdenum disulfide (MoS_2) is demonstrated to overcome the short data retention and further improve the integration density of GC-eDRAM. As one of the most widely studied two-dimensional (2D) transition metal disulfides (TMDs) materials, the MoS_2 has many advantages, such as atomic layer thickness, material properties close to conventional silicon, low off-state current and relatively mature wafer scale material growth[23–25]. Due to these advantages, MoS_2 has been demonstrated for applications in advanced CMOS, optoelectronics, and artificial intelligence[26–31]. In addition, the combination of low thermal budgets and stackability makes the integration of merging 2D material with mature Si technology a promising avenue for three-dimensional (3D) integration with high density, simple processing, and multifunctionality compared to traditional 3D integration of Si technologies, which suffer from high parasitic capacitance and residual heat/mechanical stress during stacking[32,33]. To date, milestone works have emerged, such as heterogeneous photonic devices[34], heterogeneous CFETs[32,35], and graphene-Si analog-digital circuits[36], exemplifying the potential for applications and industrialization. Our research has been dedicating to integrating MoS_2 with Si chip using a back-end-of-line (BEOL) process and demonstrating attractive applications combining merits from both materials[35,37,38]. In this work, a heterogenous GC-eDRAM with two-transistors (2T-eDRAM) is demonstrated. The MoS_2 is introduced to replace Si and serves as the channel material of the write transistor, Which offers subthreshold slopes that approach the theoretical limit[39–42]. Due to its atomic layer thickness and relatively wide bandgap, the MoS_2 transistor dramatically suppresses SCEs even beyond the 2 nm node, including drain-induced barrier lowering and gate-induced drain leakage caused by band-to-band tunneling effect[43,44]. Thus, the OFF current of the MoS_2 transistor is much lower than Si, which is helpful to prolong the data retention of eDRAMs. Furthermore, the availability of wafer-level transfer techniques for chemical vapor deposition (CVD)-grown MoS_2, with controllable thicknesses, makes TMDs well-suited for large-area integration on Si wafer[45,46]. The data storage and reading are still based on conventional Si transistor, in order to obtain high drive current, resulting in a high sense margin and enable excellent Si CMOS compatibility. Thanks to the merits from both materials, the Si-MoS_2 heterogenous eDRAM shows data retention up to 6000 s and a sense margin of 35 μA/μm, which are improved 1000 times and 100 times compared to the simultaneously fabricated full Si and MoS_2 eDRAMs, respectively.

Besides, the access time of the Si-MoS_2 eDRAM is down to 5 ns, which fully satisfy the requirement of a high-level cache application. By stacking the MoS_2 onto Si transistor using a BEOL-compatible process, the 3D Si-MoS_2 eDRAM has also been demonstrated with augment integration density. With long retention, high sense margin, fast access speed, and high integration density, the demonstrated CMOS-compatible Si-MoS_2 eDRAM is expected to revolutionize the large-capacity cache technology and break the bottleneck of the modern high-performance computer systems.

Results

Device structure and fabrication processes

Figure 1a illustrates the three-dimensional structure of the heterogeneous eDRAM, which is fabricated with a silicon-on-insulator (SOI) substrate. Both Si and MoS_2 transistors are n-type and are placed on a buried oxide layer (BOX), they are used to read and write the data respectively. The corresponding equivalent circuit diagram is presented in Fig. 1b. Within this heterogeneous architecture, MoS_2 transistor governs the write operations, while an Si transistor governs the read functionality. The source of the write transistor (MoS_2) connects to the gate of the read transistor (Si), leveraging the latter's gate capacitance for charge storage. The MoS_2 write transistor's low OFF current plays a pivotal role in securing long-term data retention. Its gate serves as the word line for writing (WWL), whereas its drain is designated as the write bit line (WBL) port. Silicon-based read transistors are employed for reading the data stored at the storage node (SN), with their drain and source acting as the read word line (RWL) and read bit line (RBL), respectively. Unlike traditional 1T1C configurations, 2T-eDRAM cells are seamlessly integrated with existing silicon logic processes and have independent read and write functions. Monitoring the read transistor's output enables the discernment of the memory state without perturbing the charge at the SN, i.e., non-destructive read. The data retention is determined by the MoS_2 transistor's OFF current. Whereas, the sense margin of the memory is decided by the on-current of the Si transistor. Thus, combining MoS_2 and Si transistors for write and read, respectively, can achieve 2T-eDRAM with long data retention and high sense margin. The details of the heterogenous fabrication process can refer to Supplementary Fig. 1 and Supplementary Note 1. Notably, this low-temperature wafer-level MoS_2 transfer technology exhibits excellent compatibility with the back-end-of-line (BEOL) process of very large-scale integrated circuits (VLSI). The Raman spectra in Fig. 1c exhibit two characteristic peaks of MoS_2: the in-plane E^1_{2g} vibrational mode at 383.0 cm^{-1}, and the out-of-plane A_{1g} vibrational mode at 407.0 cm^{-1}. The frequency difference between E^1_{2g} and A_{1g}, approximately 24.0 cm^{-1}, corresponds to 2L-MoS_2, in line with the results from scanning transmission electron microscopy (STEM) characterization. Figure 1d depicts the SEM image of the 2T-eDRAM, while STEM cross-sections of Si and MoS_2 regions marked by orange and purple dots, respectively, are presented in Fig. 1e–g. In Fig. 1d, from left to right, they correspond to the Si read transistor, the storage node, and the MoS_2 write transistor. Figure 1e, f provide STEM of the gate metal and source/drain metal regions of the Si transistor and MoS2 transistors, respectively, clearly highlighting the metal contact regions of the read and write transistors and the isolation layer in between. Figure 1g focuses on the blue rectangular area in Fig. 1f, revealing a bilayer (2L) MoS_2 with a unique layered structure. The energy dispersive spectroscopy (EDS) mapping images captured at the interface between the source/drain and metal contact of FETs are shown in Supplementary Fig. 2. Supplementary Fig. 3 provides a performance comparison of FETs using different layers of MoS_2 material as the channel, ranging from monolayer (1L) to four layers (4L). Based on the requirements for long retention time, high write speed, and low static power consumption in DRAM applications, achieving a balance between low OFF current, high ON current, and an appropriate threshold voltage (V_{th}) for the write transistor poses a significant challenge. Based on related researches[39,47,48] and experimental results detailed in Supplementary Fig. 3, bilayer MoS_2 demonstrates a balanced combination of V_{th}, I_{ON}, and I_{OFF} simultaneously, therefore, it is chosen as the channel material for the write transistor.

Si-MoS_2 device and 2T-eDRAM electrical characterization

After device fabrication, comprehensive electrical measurements are conducted under constant temperature of 25 °C in the ambient atmosphere. In this part, the transfer and output characteristics of Si

https://doi.org/10.1038/s41467-024-54218-w

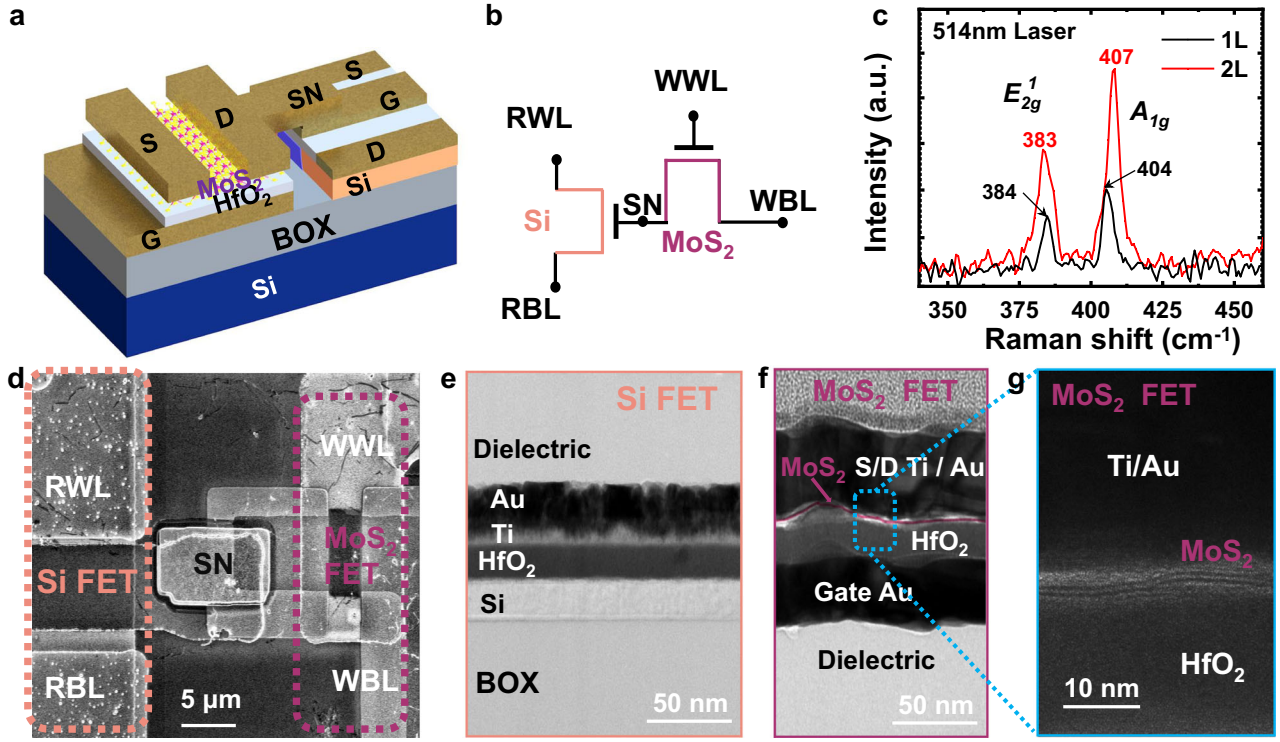

Fig. 1 | Schematic and characterization of the Si-MoS_2 2T-eDRAM. a The schematic illustration of the heterogenous 2T-eDRAM, in which MoS_2 serves as the write transistor, and Si is used as the read transistor. The source of the MoS_2 transistor is connected to the gate of the Si transistor. **b** The corresponding equivalent circuit diagram. **c** The normalized Raman spectra of the monolayer and bilayer MoS_2 films. **d** The SEM image of the 2T-eDRAM. **e** The zoom-in scanning transmission electron microscopy (STEM) image of the Si FET structure in the orange rectangle indicated in Fig. 1d. **f** The zoom-in STEM image of the MoS_2 FET structure in the purple rectangle indicated in Fig. 1d. **g** The zoom-in STEM image in the blue rectangle of Fig. 1f. All STEM images were acquired in High Angle Annular Dark Field (HAADF) mode.

and MoS_2 FETs under the same channel size (width 5 μm, length 3 μm) are compared in Fig. 2a, b. The transfer characteristics (I_D-V_G) of both transistors show excellent ON/OFF ratio. The Si nFET transistor has low OFF current under low $V_D = 1$ V. As V_D increases, the gate-induced-drain-leakage (GIDL) effect in Si nFET becomes apparent as shown in Fig. 2a. The MoS_2 FET, however, has low OFF current below the lower limit of our measurement equipment even under high V_D bias, as shown in Fig. 2b. The output characteristics show the drive current of these two FETs. The ON current of MoS_2 nFET reaches 6.31 μA/μm whereas the Si nFET reaches 165.97 μA/μm at the same bias conditions. Figure 2c compares our heterogeneous eDRAM and traditional AOS DRAM in terms of I_{ON_R} of read transistor and I_{ON_R}/I_{OFF_W} ratio between read and write transistors[49,50]. The x-axis indicates the ON current of the read FET and the y-axis shows the ratio between the ON current of the read FET and the OFF current of the write FET. Low OFF current of write FET and high ON current of read FET are critical to obtain long data retention and large sense margin in 2T-eDRAM. By using Si-MoS_2 heterostructure, we achieved an I_{ON_R} of 280 μA (channel length-width normalized to 1 μm) and an I_{ON_R}/I_{OFF_W} ratio up to 10^{14}. This outperforms AOS and pure MoS_2 FETs significantly, as shown in Fig. 2c. The ON current of the MOSFET is mainly determined by its size, carrier mobility, and parasitic resistance. A simplified model[51] of the ON current is expressed as:

$$I_D = \frac{u_0 \cdot C_{OX} \cdot W}{2L} \cdot (V_G - V_{th} - I_D \cdot R)^2 \quad (1)$$

where parameters W, L, C_{ox}, and R are channel width, gate length, the capacitance density, and parasitic resistance, respectively. V_G and V_{th} are gate voltage and threshold voltage. For technical accuracy and relevance, the data are normalized for devices with a channel length of 1 μm and width of 1 μm. The I_{ON_R}/I_{OFF_W} ratio, critical for device performance, is deduced from the transfer characteristics of each device, specifically the maximum ON current and the minimum OFF current from the transfer characteristics. It's important to note that this comparison effectively demonstrates the competitive performance and potential of heterogeneous DRAM read and write transistors utilizing varied channel materials. Notably, the ON current of the read transistor and I_{ON_R}/I_{OFF_W} ratio in Si-MoS_2 heterogeneous 2T-eDRAM transistors are two to four orders and one to six orders of magnitude higher than other devices.

We subsequently performed write and read tests on the 2T-eDRAM. Figure 2d illustrates timing diagrams detailing the write and read operations for the 2T-eDRAM bit cell. During these tests, the RBL voltage was held at 1 V, while the RWL voltage was set to 0 V in order to continuously monitor the stored charge in the SN node. In real eDRAM operation, the read signals are only applied during the read phase to reduce static power consumption. During the write "1" operation, the peak voltages for the WBL and the WWL were pulsed up to 4 V and 5 V, respectively. Monitoring the output current (I_{RBL}) of the Si read FET allows us to retrieve data stored at SN during write and read operations, as shown in Fig. 2e. The observed current difference between the "0" and "1" states represents the memory sense margin. After the write "1" operation, the charge is stored in the SN and increases the voltage at the storage node (V_{SN}). The read FET turns on, and the current increases significantly to the ON state. The write FET during the hold operation is turned off to keep the charge stored in the SN node. However, the storage node slowly discharges due to the OFF current of the write FET, which causes the slow drop of the I_{RBL}. During the write

https://doi.org/10.1038/s41467-024-54218-w

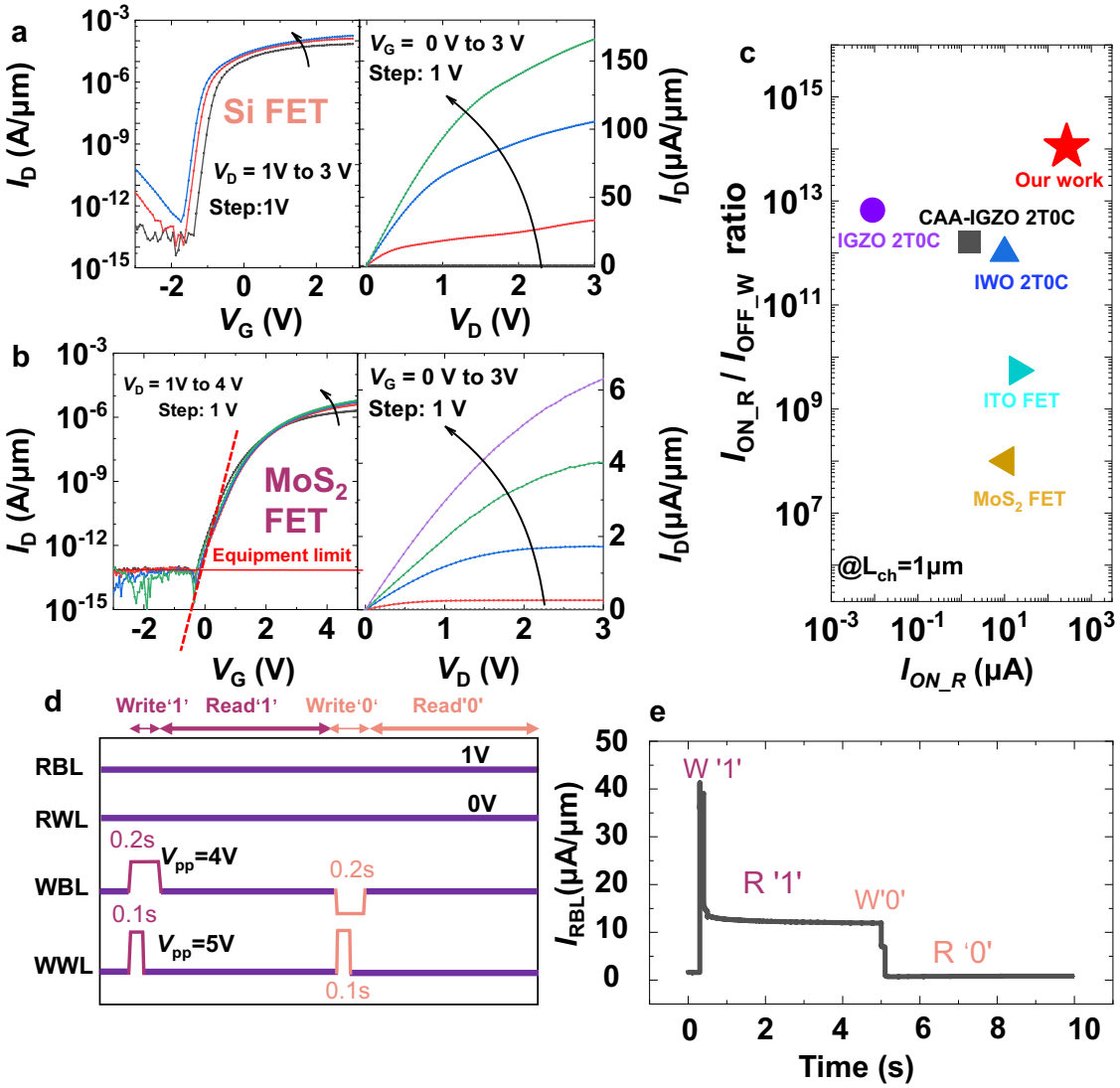

Fig. 2 | Electrical performances of the Si read transistor, the MoS_2 write transistor and the heterogenous 2T-eDRAM. a The I_D-V_G transfer characteristics and the I_D-V_D output characteristics of Si nFET. **b** The I_D-V_G transfer characteristics and the I_D-V_D output characteristics of bilayer MoS_2 nFET. **c** Benchmark of ON current of read transistor (I_{ON_R}) and ratio of ON current of read transistor and OFF current of write transistor (I_{ON_R}/I_{OFF_W}) of 2T-eDRAM and conventional AOS DRAM normalized at $L = 1$ μm, $W = 1$ μm and $V_D = 1$ V[17,19,49,50,56]. **d** Timing diagram of applied biases during write and read operations for 2T-eDRAM. **e** The I_{RBL}-Time characteristic of the fabricated 2T-eDRAM.

"0" operation, the peak voltages for the WBL and the WWL were pulsed up to −4 V and 5 V, respectively. This turns on the MoS_2 write FET and discharges the storage node and turns off the read FET. After the WWL drops back to −1 V during the hold phase, the write FET is turned off, and the I_{RBL} remains low, as shown in Fig. 2d, e. Thus, the eDRAM is fully functional with write and read operations.

High performances of Si-MoS_2 2T-eDRAM

To evaluate the memory performance of Si- MoS_2 heterogeneous 2T-eDRAM, data retention time, sense margin, and access speed are characterized and compared. The variation of I_{RBL} with time is recorded after the write "1" and write "0" operations, as shown in Fig. 3a. After the write "1" operation, the charges are stored at the SN and drive the V_{SN} to a higher level. The sense margin between logic "1" and "0" reaches 35μA/μm after the write operation, thanks to the high ON current of the Si read FET. High sense margin is helpful for the amplifier to distinguish different logic states. The high current with logic "1" stays stable until 1000 s, after which it drops sharply. Conventional AOS-DRAMs have low mobility, resulting in a much smaller sense margin between the "0" and "1" states than silicon. A 90% drop in maximum state current or a 0.1 V drop in node voltage is commonly used to define the retention of the memory cell at state "1"[17–19,49]. In order to compare the retention times with those of AOS-DRAM under different definitions, both I_{RBL} and V_{SN} are considered in this paper. As shown in Fig. 3b, the I_{RBL} difference between logic "1" and "0" stays above 10% of the maximum current difference for more than 3400 s. We then derive V_{SN} - I_{RBL} curves from the read FET prior to DRAM testing, and by employing a polynomial fit to the curves, the retention time can be extracted from the drop in node voltage from the change in I_{RBL} as shown in Fig. 3c. Figure 3d shows the V_{SN} discharge for our Si-MoS_2 heterogeneous 2T-eDRAM devices. Considering a voltage-drop $\Delta V_{SN} = 0.1$ V as the retention failure criterion, the retention of our device reaches over 1000 s, and the OFF current is roughly 3.2×10^{-18} A/μm, which is calculated from the retention characteristic, as the formula shown in Figure 3d[17]. The detailed deviation and the measured C_{SN}-V_{SN} curve can be found in Supplemental Note 2 and Supplemental Fig. 4. Concurrently, we performed a write-speed test for the 2T-eDRAM. In an ideal scenario, the amount of charge stored in the SN

https://doi.org/10.1038/s41467-024-54218-w

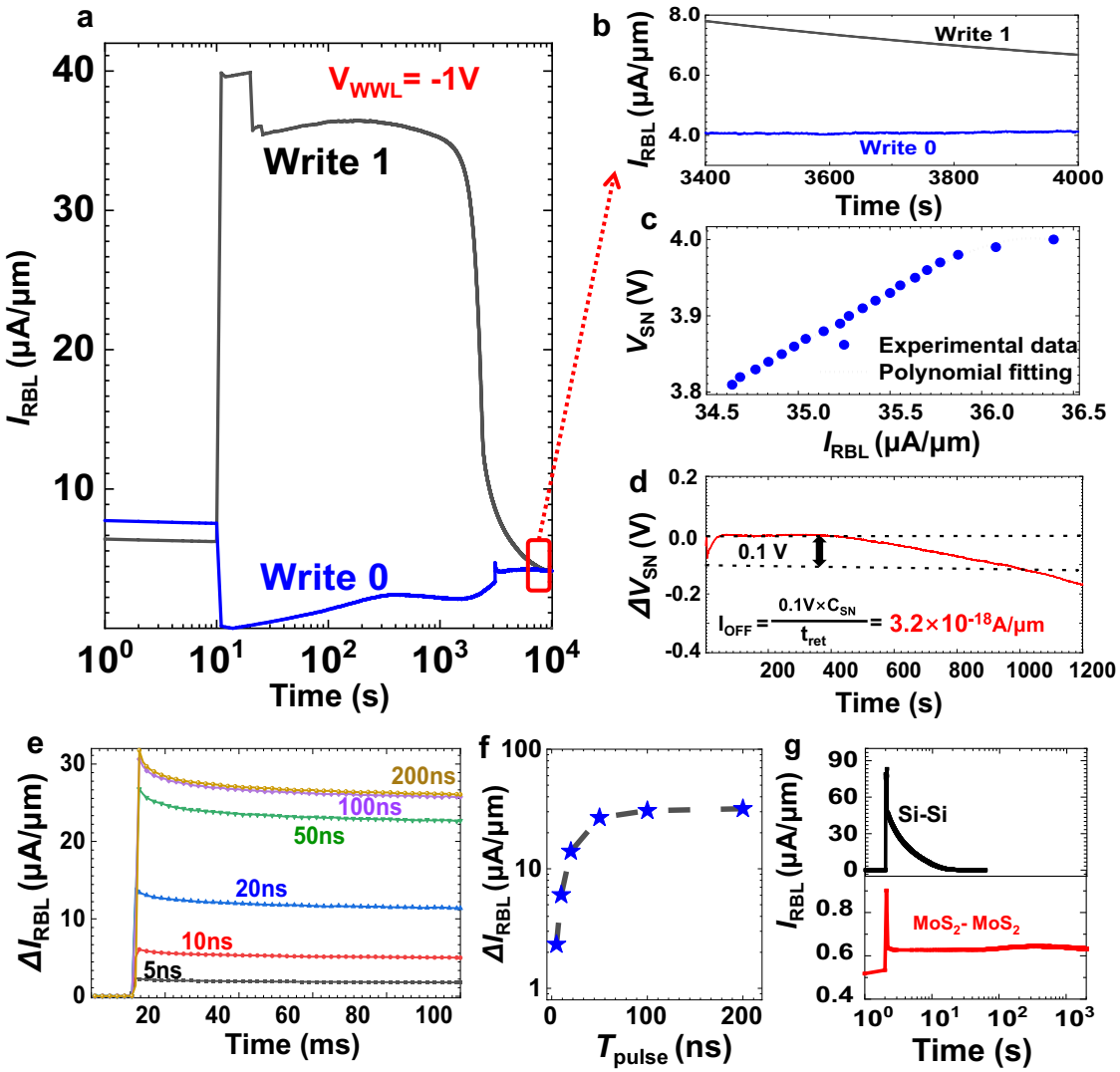

Fig. 3 | Demonstration of advantages of Si-MoS_2 heterogenous 2T-eDRAM with long data retention and high speed access. a Evolution of I_{RBL} showing data retention characteristics after writing "1" and writing "0". **b** The zoom-in I_{RBL}-Time characteristics in the red rectangle of Fig. 3a showing that logic "1" and "0" are still distinguishable even after 3000 s. **c** Polynomial fitting of V_{SN} - I_{RBL} relation from experimental data, which helps to convert the measured I_{RBL} to the voltage at the storage node (V_{SN}). **d** V_{SN}-Time characteristic of the fabricated 2T-eDRAM. **e** ΔI_{RBL} -Time characteristics after writing "1" with pulse width ranging from 5 ns to 200 ns. **f** Relation between ΔI_{RBL} and pulse width. **g** Output current I_{RBL}-Time characterization of the fabricated full-Si and full-MoS_2 eDRAMs after writing "1", which have poor data retention and sense margin, respectively.

node is determined by the ON current of the write FET and the write pulse time. Following a 5 ns write pulse (the limit of our testing apparatus), the I_{RBL} response for data "0" and data "1" exhibit a current difference exceeding 2 μA/μm. The high sense margin of Si FETs also results in distinct current levels of the read FET at different write times, ranging from 5 ns to 200 ns, and these levels remain stable throughout the 100 ms measurement period. This significant state difference enables the potential for multi-level retention capability within the 2T-eDRAM, as illustrated in Fig. 3e. The variation of I_{RBL} saturates with write pulse longer than 100 ns, due to the full charge of the storage node, as shown in Fig. 3f. Simultaneously, we conducted tests for 2T-eDRAMs composed of pure Si FETs and pure MoS_2 FETs for comparison, as illustrated in Fig. 3g. The fabrications of these two eDRAMs are illustrated in Supplementary Figs. 5, 6. Notably, the pure Si DRAM exhibited substantial OFF current, resulting in a rapid decline of I_{RBL} to the 0 state within a few seconds. In contrast, the MoS_2 DRAM exhibited enhanced stability, maintaining a consistent current over an extended period of 1000 s. However, it is critical to note that despite its improved data retention, MoS_2 DRAM is characterized by a notably poor sense margin due to its low drive current.

Longer data retention with 1L-MoS_2 and higher density with 3D integration

1L-MoS_2 has a wider bandgap compared to multilayer MoS_2, which is helpful to further reduce the subthreshold and tunneling leakage currents, as illustrated in the band-diagram in Fig. 4a. The transfer characteristics of MoS_2 FETs with different number of layers for the same channel length of 3 μm are compared in the Supplementary Fig. 3. Using 1L-MoS_2 as the write FET channel material, its V_{th} is shifted to the right enable the heterogeneous 2T-eDRAM to operate even with V_{WWL} held at 0 V, which means exponentially reduced power consumption. And its OFF current is reduced compared to multi-layer MoS_2, which can make the retention time of eDRAM longer, as shown in Fig. 4b. Supplementary Fig. 7 illustrates the variation of I_{RBL} over an extended duration following a write "1" operation. The I_{RBL} difference between logic "1" and "0" stays above 10% of the maximum current

https://doi.org/10.1038/s41467-024-54218-w

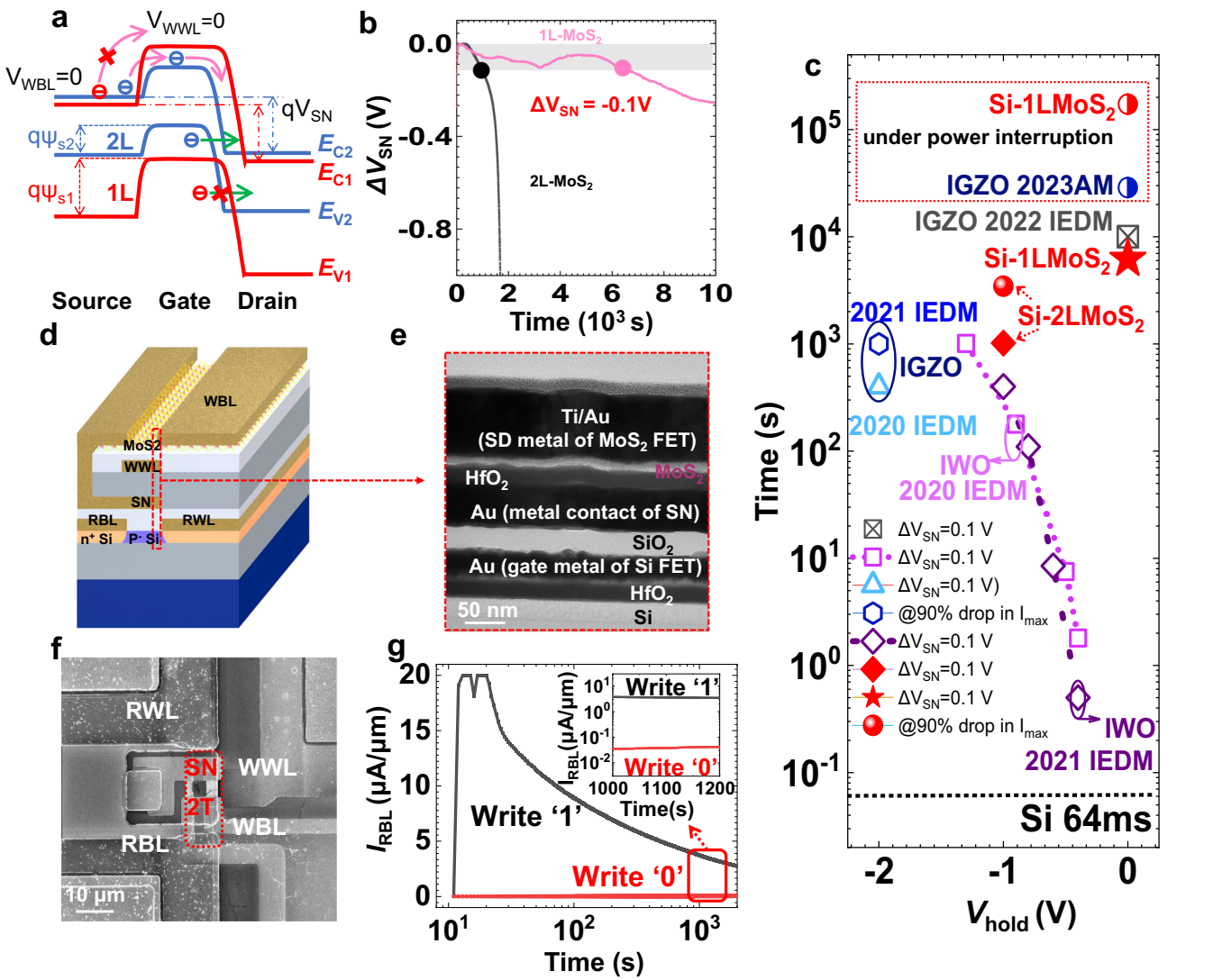

Fig. 4 | Benchmark and application potential of our 2T-eDRAM. a Energy band diagrams obtained in the channel area of 1L-MoS_2 and 2L-MoS_2 FETs. **b** Comparison of polynomial fit and merger of single-layer and bilayer MoS_2 to the V_{SN} -time based on experimental data. **c** Benchmark of retention time versus hold voltage for 2T-eDRAM. All retention data are obtained using the 0.1 V drop in V_{SN} standard [17,18,49,53] except that from ref. 57, which uses a 90% drop in maximum current (I_{max}). **d** The schematic illustration of the Si- MoS_2 heterogenous vertically 2T-eDRAM **e** The zoom-in STEM image in the red rectangle in Fig. 4d. **f** The SEM image of the vertically stacking heterogenous 2T-eDRAM. **g** Vertical stacking structure heterogenous 2T-eDRAM I_{RBL} -Time characteristics after writing "1" and writing "0".

difference for more than 10000 s. Notably, the V_{SN} decreases by only 0.1 V over an extraordinary duration of 6000 s with V_{WWL} held at 0 V, demonstrating exceptional data retention with 1L-MoS_2 in the MoS_2-Si eDRAM in Fig. 4b. Furthermore, we also investigated the data retention of 2T-eDRAM under power interruption conditions[52,53]. The measurement schematic is detailed in the timing diagram, as depicted in Supplementary Fig. 8. Under power interruption mode, we observed an unprecedented long data retention time of 48 h, longer than that reported in AOS eDRAM[52,53]. Data retention performance is further analyzed in Fig. 4c, which presents the retention time versus hold voltage (V_{hold}) benchmark. Our research employing 1L-MoS_2 as the channel material in the write FET. This design effectively shifts the threshold voltage to the right while drastically reducing OFF current, thereby substantially enhancing data retention. Impressively, our DRAM cells, with V_{WWL} held at 0 V, demonstrate an improvement in data retention by nearly five orders of magnitude compared to existing models. This advancement represents a significant breakthrough in eDRAM technology, addressing key limitations and setting a higher benchmark for performance.

Further, we introduced a vertically stacked 3D configuration. In this structure, after the fabrication of Si read FET, MoS_2 write FET is fabricated on top of the first layer. The fabrication process of this device is presented in the Supplementary Fig. 9, and it is fully compatible with the BEOL process in the conventional CMOS process. The 3D MoS_2-Si 2T-eDRAM bit cell largely reduces the cell area and further augments the integration density. The schematic illustration of this architecture can be seen in Fig. 4d, while Fig. 4e provides a corresponding STEM image of the 3D stacked structure. Figure 4f further illustrates the 3D stacked eDRAM with an SEM image. It can be observed that the area occupied by the cell is drastically reduced compared to the parallel heterogeneous structure. The energy dispersive spectroscopy (EDS) mapping images captured at the interface between the source/drain and metal contact of FETs are shown in Supplementary Fig. 10. Our experiments validate the efficacy of this design. In the vertically stacked structure, the sense margin between state "0" and state "1" is still high. The I_{RBL} difference between logic "1" and "0" stays above 10% of the maximum current difference for more than 1000 s. This difference spans two orders of magnitude, as shown

in Fig. 4g. The capacitance of SN and parasitic of the stacked 2T-eDRAM is also measured, with the results shown in Supplementary Fig. 11. The SN capacitance is found to be 0.28 pF, while the parasitic capacitance is found to be 0.033 pF. This innovative stacking approach makes our device a strong contender in the pursuit of ultra-high density and high performances eDRAM. To further verify the scaling capability and retention performance of our 3D heterogeneous 2T-eDRAM, we performed computer-aided design simulations based on the definition of 1 nm technology node with a gate length of 10 nm, and the simulated structure and results are shown in the Supplementary Figs. 12–14. A comprehensive comparison between our work, including planar and stacked 2T-eDRAMs, and other related works can be found in Supplementary Table 1. Our work distinguishes itself for its inspiring performance.

Discussion

In summary, we report a groundbreaking heterogeneous 2T-eDRAM comprising Si-based read FET and MoS_2-based write FET. Our heterogeneous Si-MoS_2 2T-eDRAM demonstrates quasi-nonvolatile data retention, with a node voltage drop of only 0.1 V sustained for over 6,000 s at $V_{WWL} = 0$ V. This remarkable achievement surpasses the performance of previously reported eDRAMs based on full-Si CMOS, full-2D materials. Moreover, our 2T-eDRAM exhibits write times as short as 5 ns, which leads to high-speed operation. A 3D stacking architecture is further demonstrated for high integration density with CMOS BEOL-compatible process. With long retention, high speed, large sense margin, and high integration density, the Si-MoS_2 2T-eDRAM technology presenting a viable alternative to SRAM in cache applications, significantly reducing the footprint of storage units. Notably, its revolutionary 6000 s data retention is more than five orders of magnitude higher than traditional silicon-based DRAM. This enhancement represents a significant leap forward, contributing to a substantial reduction in power consumption. Meanwhile, the core idea of combining the merits of two different material systems, utilizing innovative materials with low OFF current as write transistors and mature silicon with high ON current as read transistors, marks a significant milestone for memory technology, offering insights for more heterogeneous DRAM candidates, such as Si-2D and Si-AOS hybrid DRAMs.

Methods

Fabrication of Si-based FET

We utilized Silicon-On-Insulator (SOI) substrates featuring a top silicon layer with a thickness of 28 nm and a Buried Oxide (BOX) layer consisting of SiO_2 with a thickness of 145 nm. The fabrication process involved a sequence of key steps. Initially, we employed photolithography in conjunction with wet etching, utilizing a TMAH (Tetramethylammonium Hydroxide) to H_2O ratio of 2:5 by volume, to delineate the active regions and channels on the Si substrate. Subsequently, we introduced n^+ doping into the active regions through phosphorous-ion implantation. To activate this doping, rapid thermal annealing was conducted at 950 °C. For the establishment of source/drain contacts, we deposited a metal stack of Titanium (Ti) and Gold (Au) layers, with thicknesses of 5 nm and 45 nm, respectively, utilizing an electron beam evaporator (EBE). To enhance the contact properties, we carried out rapid thermal annealing at 350 °C for 5 min. To complete the device structure, we deposited 18 nm of Hafnium Oxide (HfO_2) using Atomic Layer Deposition (ALD) to serve as the gate dielectric, followed by the deposition of 50 nm of Gold (Au) by EBE to form the gate electrode.

Synthesis of MoS_2

The growth of monolayer MoS_2 was carried out in the two-temperature-zone CVD system[54]. The precursors used in zone 1 and zone 2 were sulfur (Alfa Aesar 99.95%) and MoO_3 (Alfa Aesar 99.95%) powders, respectively. The distance between the two zones was 30 cm and the synthesis temperature for zone 1 and zone 2 was 180 °C and 650 °C. The sapphire substrate was placed downstream and face-down above the MoO_3 in zone 2. Using argon as the carrier gas, continuous monolayer MoS_2 films were formed on sapphire substrates at atmospheric pressure with a sulfuration time of 10 min.

Transfer of MoS_2

After the synthesis of MoS_2 and the fabrication of the SOI FET, the transfer process for precise layers of MoS_2 was carried out[35,55]. First, a double layer Methyl methacrylate/polymethyl methacrylate (MMA/PMMA) (Aladdin, 99%) was separately spin-coated on monolayer MoS_2 and baked 180 °C for 3 min. Then, the MMA/PMMA/ MoS_2 films were then firmly attached with thermal release tape (TRT) in order to remove easily from the substrate with deionized water. The whole peeled MMA/PMMA/ MoS_2 films were utilized to adhere and exfoliate another monolayer MoS_2 repeatedly until the desired number of layers was obtained. By means of the above, we prepared the TRT/photoresists/MoS_2 and TRT/photoresists/bilayer MoS_2 samples. The two samples were then transferred to the target SOI substrates in a customized vacuum transfer setup by heating at 130 °C for 30 min to release TRT, respectively. Next, the MMA/PMMA was washed out in acetone and isopropanol successively. Finally, the monolayer MoS_2/SOI and the bilayer MoS_2/SOI were obtained after annealing at 200 °C for 2 h under vacuum (~10^{-3} Pa), which improves the adhesion between transferred MoS_2 and SOI substrates.

Fabrication process of the 2T-eDRAM

The detailed fabrication process scheme is given in Supplementary Fig. 1 and Supplementary Note 1 in the Supplementary Information.

Material characterization and electrical measurement

Atomic force microscope (AFM, Dimension Edge, Bruker, USA) at tapping mode was employed to characterize the thickness of the as-prepared few-layer MoS_2. Raman spectra (Horiba, Japan) were measured under a 532 nm laser as an excitation source at room temperature. The cross-sectional high-resolution transmission electron microscopy (HRTEM, Titan 80-300, Thermo-Fisher, USA) with an EDS analyzer was carried out to confirm the 3D structure of the device. The electrical performances of the devices were measured by a semiconductor analyzer (FS-Pro 380, PRIMARUS, China) at room temperature.

Data availability

All data needed to evaluate the findings of this study are available within the Article. Source Data file has been deposited in Zenodo under accession code https://doi.org/10.5281/zenodo.13894399.

Code availability

The codes used for simulation and data plotting are available from the corresponding authors under request.

References

1. Mittal, S. A survey of cache bypassing techniques. *J. Low Power Electron. Appl.* **6**, 5 (2016).
2. Wuu, J. et al. 3D V-Cache: the implementation of a hybridbonded 64MB stacked cache for a 7nm x86-64 CPU. In *2022 IEEE International Solid-State Circuits Conference* (ISSCC) (2022).
3. Munger, B. et al. "Zen 4": The AMD 5nm 5.7GHz x86-64 Microprocessor core. In *2023 IEEE International Solid-State Circuits Conference* (ISSCC) (2023).
4. Xie, S. et al. 16.2 eDRAM-CIM: Compute-in-memory design with reconfigurable embedded-dynamic-memory array realizing adaptive data converters and charge-domain computing. In *2021 IEEE International Solid-State Circuits Conference* (ISSCC) (2021).

5. Huang, K. C. et al. A high-performance, high-density 28nm eDRAM technology with high-K/metal-gate. In *2011 IEEE International Electron Devices Meeting* (IEDM) (2011).
6. Rosenblatt, S. et al. Field tolerant dynamic intrinsic chip ID using 32 nm high-K/metal gate SOI embedded DRAM. *IEEE J.* **48**, 940–947 (2013).
7. Pilo, H. & Zhang, K. Session 19 overview: High-performance embedded memory. In *2010 IEEE International Solid-State Circuits Conference* (ISSCC) (2010).
8. Hamzaoglu, F. et al. A 1 Gb 2 GHz 128 GB/s Bandwidth embedded DRAM in 22 nm Tri-gate CMOS technology. *IEEE J. Solid State Circuits* **50**, 150–157 (2015).
9. Barth, J. et al. A 500 MHz Random cycle, 1.5 ns latency, SOI embedded DRAM macro featuring a three-transistor micro sense amplifier. *IEEE J.* **43**, 86–95 (2008).
10. Chun, K. C., Jain, P., Kim, T. H. & Kim, C. H. A 667 MHz Logic-compatible embedded DRAM featuring an asymmetric 2T gain cell for high speed on-die caches. *IEEE J.* **47**, 547–559 (2012).
11. Yoo, T., Kim, H., Chen, Q., Kim, T. T. H. & Kim B. A logic-compatible 4T dual embedded DRAM array for in-memory computation of deep neural networks. In 2019 IEEE/ACM Int. Symp. Low Power Electron. Des. (ISLPED) (2019).
12. Ramaswamy, N. et al. NVDRAM: A 32Gb Dual Layer 3D Stacked Non-volatile Ferroelectric Memory with Near-DRAM Performance for Demanding AI Workloads. In 2023 IEEE International Electron Devices Meeting (IEDM) (2023)
13. Chun, K. C., Jain, P., Lee, J. H. & Kim, C. H. A 3T Gain cell embedded DRAM utilizing preferential boosting for high density and low power On-Die Caches. *IEEE J.* **46**, 1495–1505 (2011).
14. Chun, K. C., Zhang, W., Jain, P. & Kim, C. H. A 2T1C Embedded DRAM macro with no boosted supplies featuring a 7T SRAM based repair and a cell storage monitor. *IEEE J.* **47**, 2517–2526 (2012).
15. Samanta, S. et al. Amorphous IGZO TFTs featuring extremely-scaled channel thickness and 38 nm channel length: Achieving record high Gm,max of 125 μS/μm at VDS of 1 V and ION of 350 μA/μm. In *2020 IEEE Symposium VLSI Technology* (2020).
16. Han, K. et al. First Demonstration of oxide semiconductor nanowire transistors: a Novel digital etch technique, IGZO channel, nanowire width Down to ~ 20 nm, and Ion Exceeding 1300 μA/μm. In *2021 IEEE symposium VLSI technology*. (2021).
17. Belmonte, A. et al. Capacitor-less, long-retention (>400s) DRAM cell paving the way towards low-power and high-density monolithic 3D DRAM. In *2020 IEEE International Electron Devices Meeting* (IEDM) (2020).
18. Belmonte, A. et al. Tailoring IGZO-TFT architecture for capacitorless DRAM, demonstrating > 103s retention, >1011 cycles endurance and Lg scalability down to 14nm. In *2021 IEEE International Electron Devices Meeting (IEDM)* (2021).
19. Duan, X. et al. Novel vertical channel-all-around (CAA) In-Ga-Zn-O FET for 2T0C-DRAM with high density beyond 4F2 by monolithic stacking. *IEEE Trans. Electron Devices* **69**, 2196–2202 (2022).
20. Si, M. et al. Scaled indium oxide transistors fabricated using atomic layer deposition. *Nat. Electron.* **5**, 164–170 (2022).
21. Matsubayashi, D. et al. 20-nm-Node trench-gate-self-aligned crystalline In-Ga-Zn-Oxide FET with high frequency and low off-state current. In *2015 IEEE International Electron Devices Meeting* (IEDM) (2015).
22. Si, M. et al. Why in2O3 can make 0.7 nm atomic layer thin transistors. *Nano Lett.* **21**, 500–506 (2021).
23. Novoselov, K. S. et al. Electric field effect in atomically thin carbon films. *Science* **306**, 666–669 (2004).
24. Chhowalla, M., Jena, D. & Zhang, H. Two-dimensional semiconductors for transistors. *Nat. Rev. Mater.* **1**, 16052 (2016).
25. Franklin, A. D. Nanomaterials in transistors: From high-performance to thin-film applications. *Science* **349**, aab2750 (2015).
26. Qiu, C. et al. Scaling carbon nanotube complementary transistors to 5-nm gate lengths. *Science* **355**, 271–276 (2017).
27. Qiu, C. et al. Dirac-source field-effect transistors as energy-efficient, high-performance electronic switches. *Science* **361**, 387–392 (2018).
28. Akinwande, D. et al. Graphene and two-dimensional materials for silicon technology. *Nature* **573**, 507–518 (2019).
29. Peng, L.-M., Zhang, Z. & Qiu, C. Carbon nanotube digital electronics. *Nat. Electron.* **2**, 499–505 (2019).
30. Liu, Y. et al. Promises and prospects of two-dimensional transistors. *Nature* **591**, 43–53 (2021).
31. Das, S. et al. Transistors based on two-dimensional materials for future integrated circuits. *Nat. Electron.* **4**, 786–799 (2021).
32. Guan, S.-X. et al. Monolithic 3D integration of back-end compatible 2D material FET on Si FinFET. *Npj 2D Mater. Appl.* **7**, 9 (2023).
33. Kang, J.-H. et al. Monolithic 3D integration of 2D materials-based electronics towards ultimate edge computing solutions. *Nat. Mater.* **22**, 1470–1477 (2023).
34. Deng, J. et al. MoS 2/silicon-on-insulator heterojunction field-effect-transistor for high-performance photodetection. *IEEE Electron Device Lett.* **40**, 423–426 (2019).
35. Tong, L. et al. Heterogeneous complementary field-effect transistors based on silicon and molybdenum disulfide. *Nat. Electron.* **6**, 37–44 (2023).
36. Hong, S. K., Kim, C. S., Hwang, W. S. & Cho, B. J. Hybrid integration of graphene analog and silicon complementary metal–oxide–semiconductor digital circuits. *ACS Nano* **10**, 7142–7146 (2016).
37. Deng, J. et al. MoS2/Silicon-on-insulator heterojunction field-effect-transistor for high-performance Photodetection. *IEEE Electron Device Lett.* **40**, 423–426 (2019).
38. Deng, J. et al. MoS2/HfO2/Silicon-on-insulator dual-photogating transistor with ambipolar photoresponsivity for high-resolution light wavelength detection. *Adv. Funct. Mater.* **29**, 1906242 (2019).
39. Radisavljevic, B., Radenovic, A., Brivio, J., Giacometti, V. & Kis, A. Single-layer MoS2 transistors. *Nat. Nanotechnol.* **6**, 147–150 (2011).
40. Sarkar, D. et al. A subthermionic tunnel field-effect transistor with an atomically thin channel. *Nature* **526**, 91–95 (2015).
41. Shen, P.-C. et al. Ultralow contact resistance between semimetal and monolayer semiconductors. *Nature* **593**, 211–217 (2021).
42. Wu, F. et al. Vertical MoS2 transistors with sub-1-nm gate lengths. *Nature* **603**, 259–264 (2022).
43. Desai, S. B. et al. MoS2 transistors with 1-nanometer gate lengths. *Science* **354**, 99–102 (2016).
44. Liu, L. et al. Uniform nucleation and epitaxy of bilayer molybdenum disulfide on sapphire. *Nature* **605**, 69–75 (2022).
45. Li, N. et al. Large-scale flexible and transparent electronics based on monolayer molybdenum disulfide field-effect transistors. *Nat. Electron.* **3**, 711–717 (2020).
46. Zhu, J. et al. Low-thermal-budget synthesis of monolayer molybdenum disulfide for silicon back-end-of-line integration on a 200 mm platform. *Nat. Nanotechnol.* **18**, 456–463 (2023).
47. Kayyalha, M., Maassen, J., Lundstrom, M., Shi, L. & Chen, Y. P. Gate-tunable and thickness-dependent electronic and thermoelectric transport in few-layer MoS2. *J. Appl. Phys.* **120**, 134305 (2016).
48. Song, B. et al. Layer-dependent dielectric function of wafer-scale 2D MoS2. *Adv. Opt. Mater.* **7**, 1801250 (2019).
49. Ye, H. et al. Double-gate W-doped amorphous indium oxide transistors for monolithic 3D capacitorless gain cell eDRAM. In *2020 IEEE International Electron Devices Meeting* (IEDM) (2020).
50. Li, S. et al. Nanometre-thin indium tin oxide for advanced high-performance electronics. *Nat. Mater.* **18**, 1091–1097 (2019).
51. Sze, S. M., Li, Y. & Ng, K. K. *Physics of semiconductor devices*. John Wiley & Sons (2021).

52. Hu, Q. et al. True nonvolatile high-speed DRAM cells using tailored Ultrathin IGZO. *Adv. Mater.* **35**, 2210554 (2023).
53. Hu, Q. et al. Optimized IGZO FETs for capacitorless DRAM with retention of 10 ks at RT and 7 ks at 85 °C at zero Vhold with Sub-10 ns speed and 3-bit operation. In *2022 IEEE International Electron Devices Meeting* (IEDM) (2022).
54. Xu, H. et al. High-performance wafer-scale MoS2 transistors toward practical application. *Small* **14**, 1803465 (2018).
55. Guo, X. et al. Stacking monolayers at will: A scalable device optimization strategy for two-dimensional semiconductors. *Nano Res* **15**, 6620–6627 (2022).
56. Smets, Q. et al. Ultra-scaled MOCVD MoS2 MOSFETs with 42nm contact pitch and 250μA/μm drain current. In *2019 IEEE International Electron Devices Meeting* (IEDM) (2019).
57. Zhao, Z. et al. Computational associative memory based on monolithically integrated metal-oxide thin film transistors for update-frequent search applications. In *2021 IEEE International Electron Devices Meeting (IEDM)* (2021).

Acknowledgements

This work was supported by the National Key R&D Program of China under Grant 2021YFA120050; the National Natural Science Foundation of China under Grant 62474052; Shanghai Science and Technology Commission "explorer project" under Grant 21TS1401300; the Pioneering Project of Academy for Engineering and Technology, Fudan University under Grant gyy2021-001; Natural Science Foundation of Shanghai under Grant 23ZR1405900; Guangdong Province Research and Development in Key Fields from Guangdong Greater Bay Area Institute of Integrated Circuit and System under Grant No. 2021B0101280002; and Guangzhou City Research and Development Program in Key Field under Grant No. 20210302001.

Author contributions

J.W., W.B., and P.Z. conceived and supervised the project. K.X. wrote the initial manuscript, H.X completed the final manuscript, and J.W., Y.Z., W.B., and P.Z. revised the manuscript. J.W. designed the device, H.X. and Y.Z. fabricated the device. K.X. performed the electrical measurements and data processing with assistance from J.W. and H.X. J.W., and H.X. contributed to the technology computer-aided design simulation. T.T., W.Z., Y.C., and J.Z. contributed to the sample fabrication. Y.Z. performed the Raman spectra and photoluminescence spectra characterizations. H.X., Y.Z., L.Z., and S.D. performed the scanning electron microscope and transmission electron microscopy characterization and analysis. Z.X. performed CVD growth and a part of transfer work of MoS_2. All authors discussed the results and commented on the manuscript.

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/s41467-024-54218-w.

Correspondence and requests for materials should be addressed to Jing Wan, Wenzhong Bao or Peng Zhou.

Peer review information *Nature Communications* thanks Yann-Wen Lan, Jawar Singh, Suraj Cheema, Peng Wu, and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.