Supporting Information

MoS_2 Flash Memory Arrays with Sb Contact for Highly Efficient and Low-Latency Analog In-Memory Searches

Guoyun Gao[1,#], Bo Wen [1,#], Ni Yang[2], Zhiyuan Du[1], Mingrui Jiang[1], Ruibin Mao [1], Yingnan Cao[4], Hongxia Xue[4], Pak San Yip[2], Qihan Liu[5], Yi Wan[7], Dong-Keun Ki[4], Jinyao Tang[3], Paddy K. L. Chan[2], Hao Jiang[5], Han Wang[1,6], Lain-Jong Li[6,7*] and Can Li[1,6,*]

Table of Contents

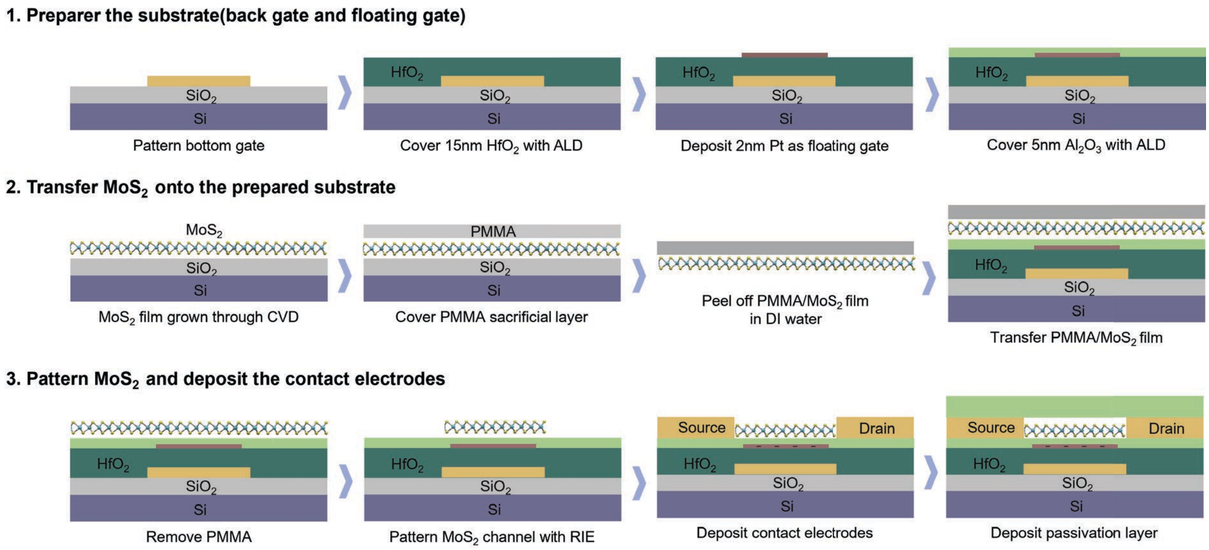
1. Preparer the substrate(back gate and floating gate)
SiO$_2$
Si
Pattern bottom gate
HfO$_2$
SiO$_2$
Si
Cover 15nm HfO$_2$ with ALD
HfO$_2$
SiO$_2$
Si
Deposit 2nm Pt as floating gate
HfO$_2$
SiO$_2$
Si
Cover 5nm Al$_2$O$_3$ with ALD
2. Transfer MoS$_2$ onto the prepared substrate
MoS$_2$
SiO$_2$
Si
MoS$_2$ film grown through CVD
PMMA
SiO$_2$
Si
Cover PMMA sacrificial layer
Peel off PMMA/MoS$_2$ film in DI water
HfO$_2$
SiO$_2$
Si
Transfer PMMA/MoS$_2$ film
3. Pattern MoS$_2$ and deposit the contact electrodes
HfO$_2$
SiO$_2$
Si
Remove PMMA
HfO$_2$
SiO$_2$
Si
Pattern MoS$_2$ channel with RIE
Source
Drain
HfO$_2$
SiO$_2$
Si
Deposit contact electrodes
Source
Drain
HfO$_2$
SiO$_2$
Si
Deposit passivation layer

Fig. S1. The schematic diagram of MoS$_2$ flash memory fabrication procedure. (1) Prepare the substrate(back gate and floating gate): 5/10nm Ti/Au was patterned as the bottom gate by photolithography and e-beam evaporator. 15nm HfO$_2$ was deposited by ALD as a blocking layer. 2nm Al or Pt was patterned as a floating gate by photolithography and e-beam evaporator. 5nm Al$_2$O$_3$ was deposited by ALD as a tunneling layer. For charge-trapping flash memories, 10nm Al$_2$O$_3$ was deposited by ALD as a blocking layer. 4-6nm HfO$_2$ was deposited by ALD as a charge-trapping layer. 5nm Al$_2$O$_3$ was deposited by ALD as a tunneling layer. (2) Transfer MoS2 to the prepared substrate: Monolayer MoS$_2$ continuous film was synthesized by CVD and transferred by wet method. Here PMMA was spin-coated or dropped onto MoS$_2$ film, then made it dry by baking at 60°C for 30min. PMMA/MoS$_2$ film was peeled off from the substrate in DI water slowly and then transferred onto the target substrate. After baking the film at 60°C for 1h or drying it at room temperature overnight, PMMA was removed by emerging the substrate into NMP and Aceton for 1h, respectively. (3) Pattern MoS$_2$ and deposit the contact electrodes and passivation layer. MoS$_2$ continuous film was first patterned by photolithography and RIE. Then Sb/Au contact electrodes and fanout line were patterned by EBL, photolithography, and thermal evaporator, followed by the liftoff process. 40nm Al$_2$O$_3$ was deposited by ALD as a passivation layer. For photolithography, a double-layer photoresist (LOR/AZ5214) was used, which was developed with TMAH. PMMA 950 A4 was used for EBL, with the developer of MIBK: IPA 1:3. NMP was used for all the liftoff processes.

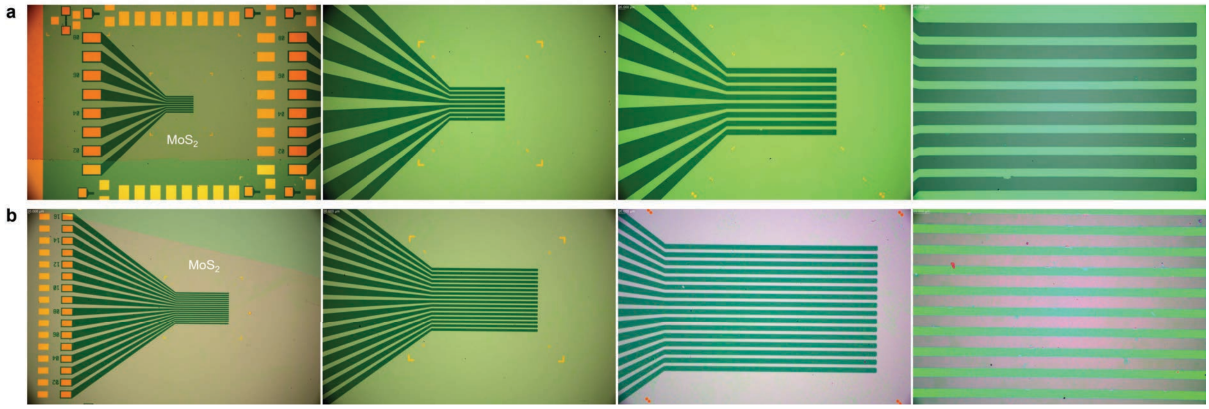

Figure S2. The optical images of MoS$_2$ film after being transferred on the prepared substrate. (a) 8x8 array and (b) 16x16 array.

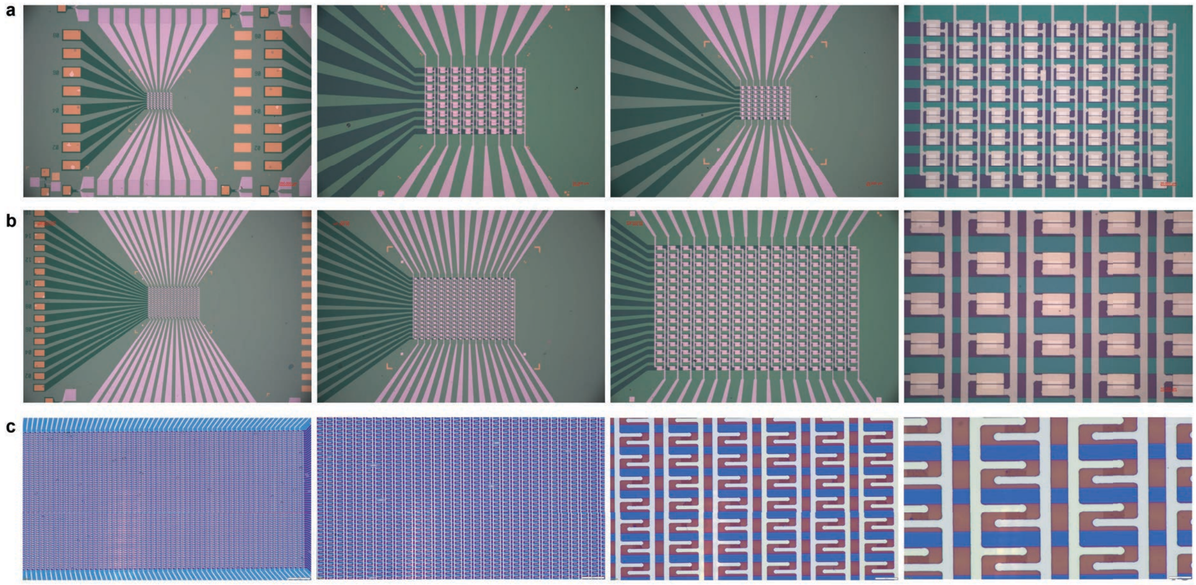

Fig. S3. The optical images of MoS$_2$ flash memory array for analog CAM. (a) 8x8 array(b) 16x16 array, and (c) 64x128 array.

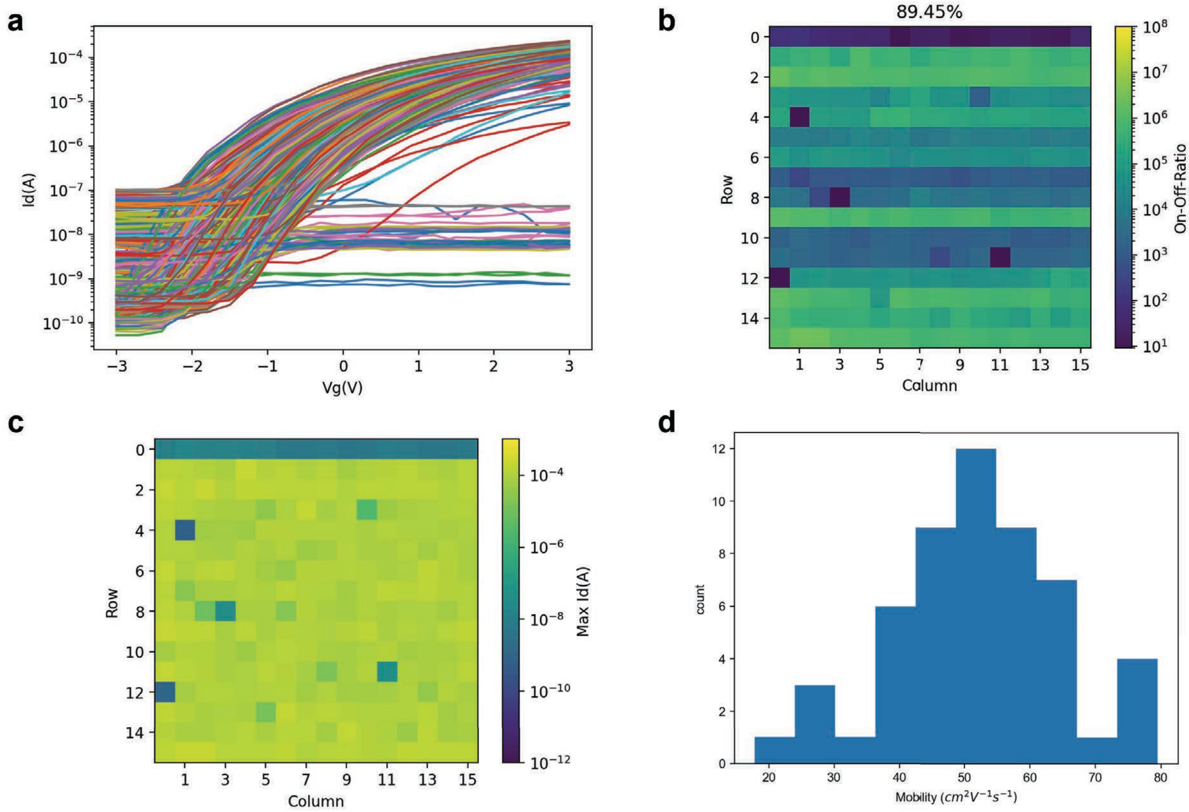

Fig. S4. The electrical performance of MoS_2 flash memory array measured with a probe card. (a) I_D-V_D curves of 16×16 array with a total of 256 MoS_2 flash memory devices with V_D=1V (L_{CH}= 500nm, W_{CH}=10um). Even though the contact issue of one pin of the probe leading to no gate controllability for one column devices, most of device still work normally with a large enough ON-OFF for analog CAM inference application. (b-c) Statistics of on-off ratio and of readout current for the 256 devices with a yield (on-off-Ratio>10^3) up to 89.45%. If not counting those devices that caused by probe card pin issue, our yield would be much higher than this value. For most devices, readout current can reach over 100uA. (d) Statistics of extracted filed-effect mobility for the 50 devices at V_D=1V and V_G=2V.

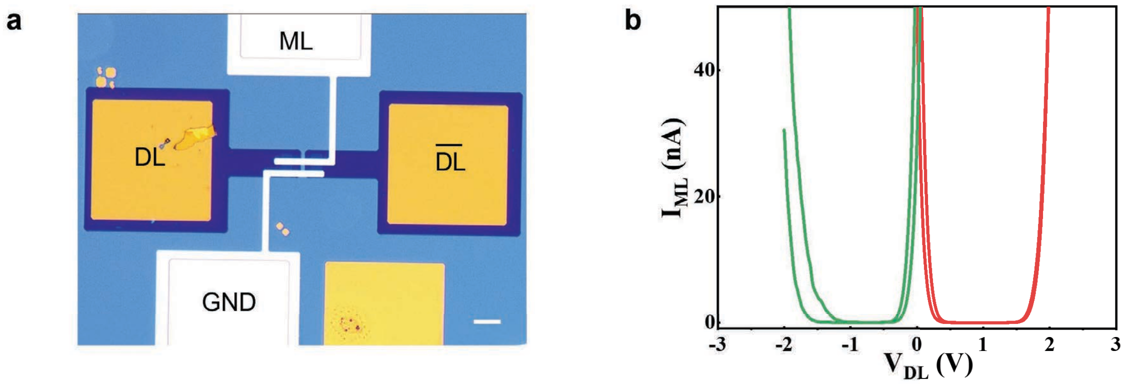

Fig. S5. A MoS_2 analog CAM single cell for range search. (a) The optical image of one MoS_2 analog CAM single cell. (b) The two programmed match range for range search operation.

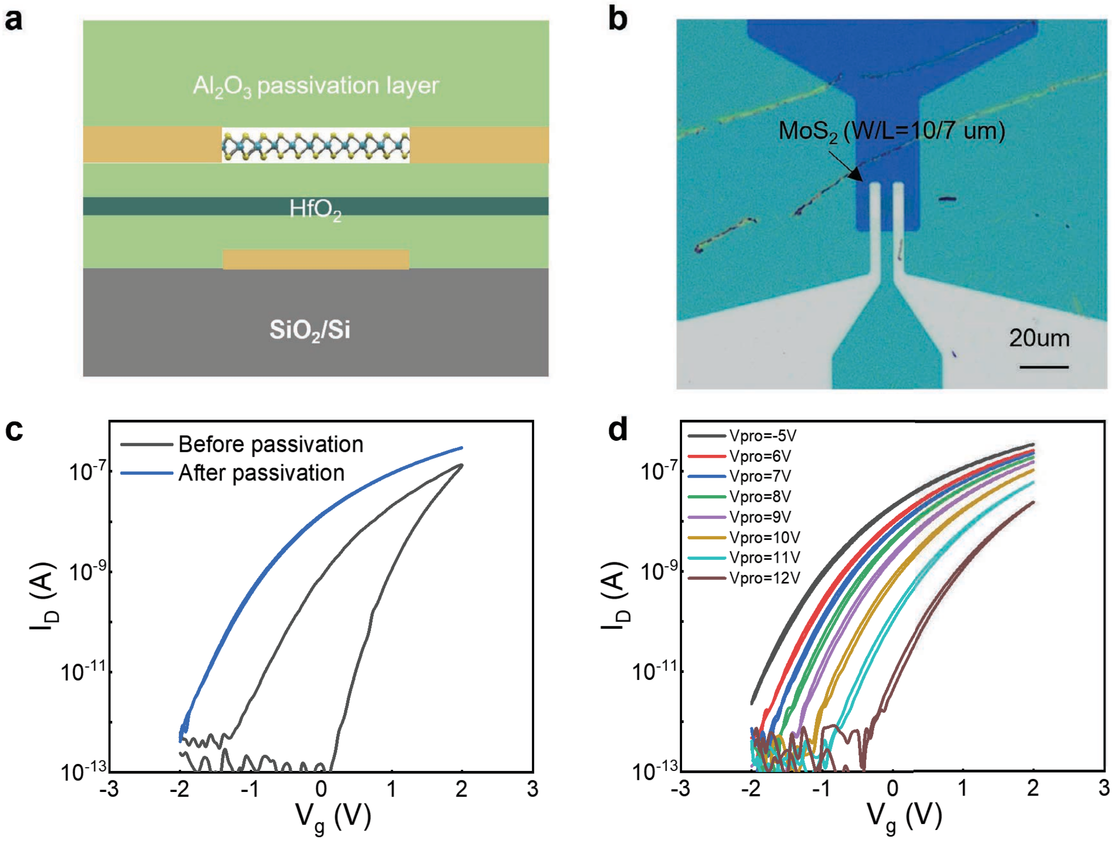

Fig. S6. Device passivation. (a) The schematic diagram of MoS_2 flash memory with 40nm Al_2O_3 passivation layer, deposited by ALD at low temperature (138°C) using O_3 and TMA as the precursors. (b) The optical image of an individual device. (c) I_D-V_D curves of the device with and without a passivation layer were measured in the ambient environment. The passivation layer can depress the hysteresis window obviously, making V_{th} more stable and Improving 2D device electrical stability in the air. (d) I_D-V_G curves of eight programmed states with a program voltage ranging from 6V to 12V.

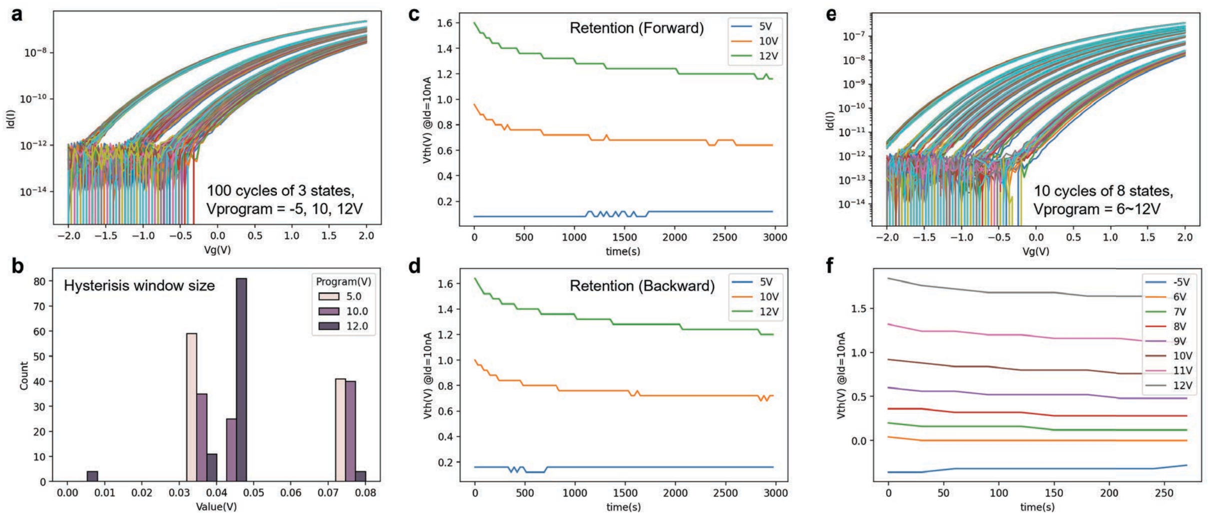

Fig. S7. The program performance of MoS_2 flash memories after passivation. (a) I_D-V_G curves of three stored states programmed by -5V, 10V, and 12V, with 100 times cycle-to-cycle test for each state measured for nearly 3000s in the ambient environment. (b) Statistics of hysteresis window size for each cycle test. All the programed states show a negligible hysteresis window, indicating a good electrical stability in the air due to the encapsulation by the passivation layer. (c-d) The three extracted V_{th} keep well distinguishable after 3000s measurement for both forward and backward sweep. (**e**) 10 times cycle-to-cycle test for eight stored states programmed by 6-12V measured for over 250s in the ambient environment. (f) The eight extracted V_{th} also maintain distinct after 250s measurement.

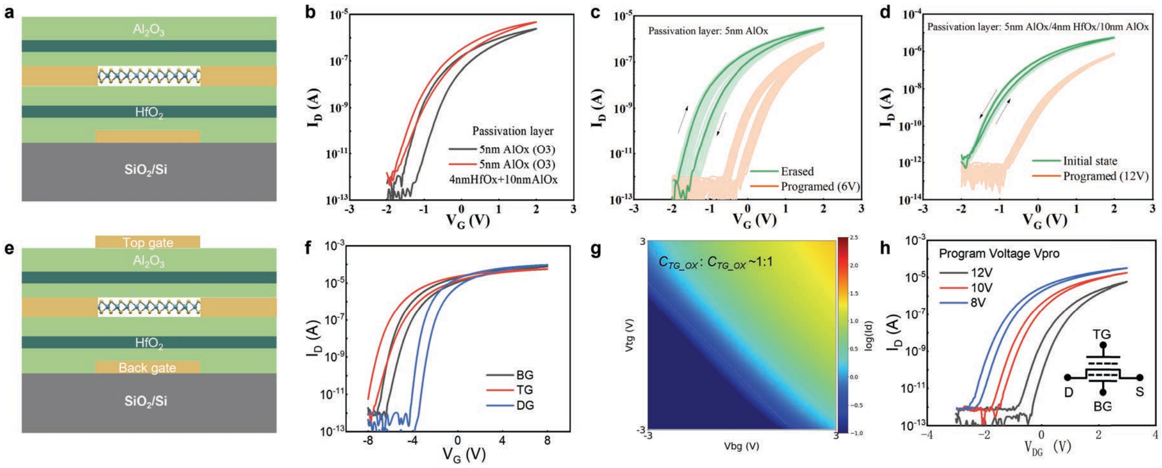

Fig. S8. The electrical performance of MoS_2 dual-gate flash memories. (a) The schematic diagram of MoS_2 flash memory with 5nm Al_2O_3 /4nm HfO_2 /10nm Al_2O_3 top gate dielectric stack as a passivation layer. (b) I_D-V_G curves of the device after encapsulation with 5nm Al_2O_3 and 5nm Al_2O_3 /4nm HfO_2 /10nm Al_2O_3, respectively. (c-d) The corresponding program and erase measurements for 10 times of the two devices in the ambient environment. (e) The schematic diagram of MoS_2 dual-gate flash memory. (f) I_D-V_G curves of the dual-gate device with gate voltage applied on bake gate (BG), top gate (TG), and dual gate (DG), respectively. (g) 2D current mapping *vs.* back gate and top gate voltage. The dash line shows the V_{th} variation *vs.* back gate and top gate voltage, with a slope of -1 indicating a similar capacitance of back gate and top gate. (h) I_D-V_G curves of three stored states with programmed and read by dual gate.

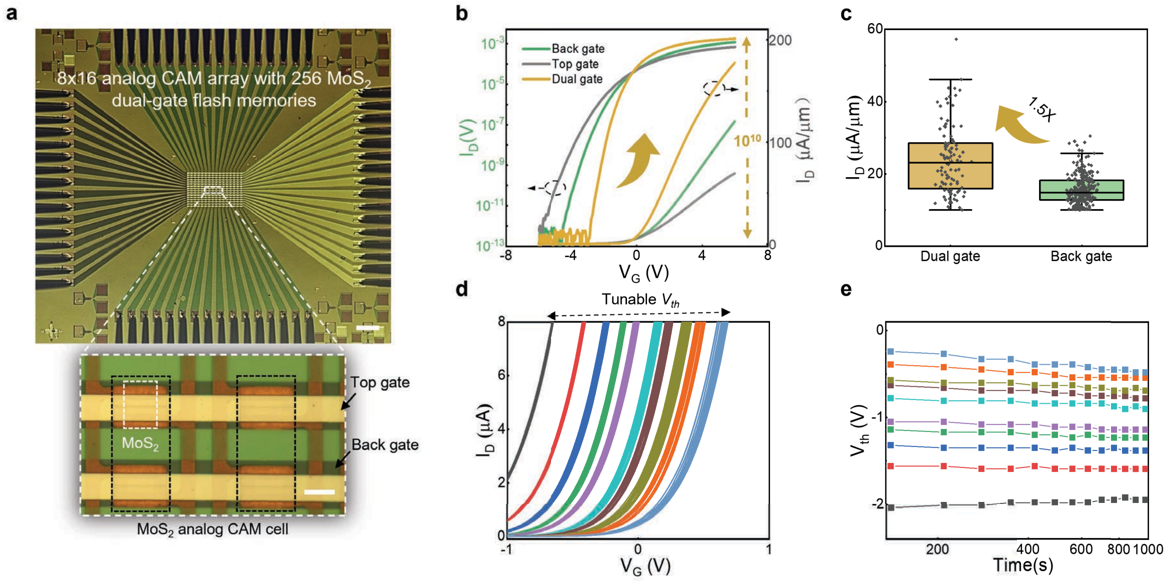

Fig. S9. The electrical performance of MoS$_2$ flash memory array. (a) The optical image 16x16 MoS$_2$ dual-gate flash memory array, which can be used as an 8x16 analog CAM array with 256 MoS$_2$ dual-gate flash memories (L_{CH}/W_{CH}=0.5/10μm). The scale bar is 200 um. Inset is a zoomed-in image, showing two analog CAM cells with four MoS$_2$ dual-gate flash memories. The scale bar is 10 um. (b) I_D-V_G curves of the dual gate device, showing increased current ON/OFF ratio (~10^{10}), $I_{ON,}$ and steeper subthreshold slop (SS), indicating that the dual gate configuration can enhance electrostatic control, facilitate additional carrier accumulation and improve the carrier transfer efficiency. (c) Statistics of readout current at V_G=3V and V_D=1V for 104 dual gate and 586 back gate MoS$_2$ flash memories with L_{CH} of 500nm, showing a 1.5-time improvement of average readout current by dual gate configuration. (d) 10-time cycles-to-cycle test for ten programmed states with a programming voltage of 7~12V, showing a cycle-to-cycle uniformity. (e) The ten extracted V_{th} maintain distinct after 1000s cycles-to-cycle measurement.

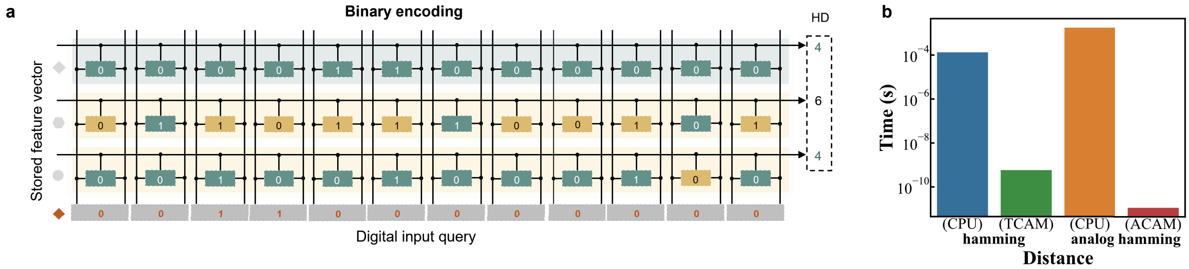

Fig. S10. (a) TCAM used classification applications using k-nearest neighbor (KNN, k=3) search in analog CAM. The embedded digital data after the binarization encoding, and distance computing results for a given digital input query. The hamming distance which can be computed by TCAM is used after the binarization of data, while with limited accuracy. (b) KNN inference latency for each sample with Hamming distance or analog Hamming distance on CPU or CAM. The latency is averaged over 10 times on the 4 datasets. TCAM is a traditional 45 nm node 16T CMOS. Compared with hamming distance, the analog hamming distance costs more time in CPU but can efficiently be accelerated by ACAM, about 10^8 faster.

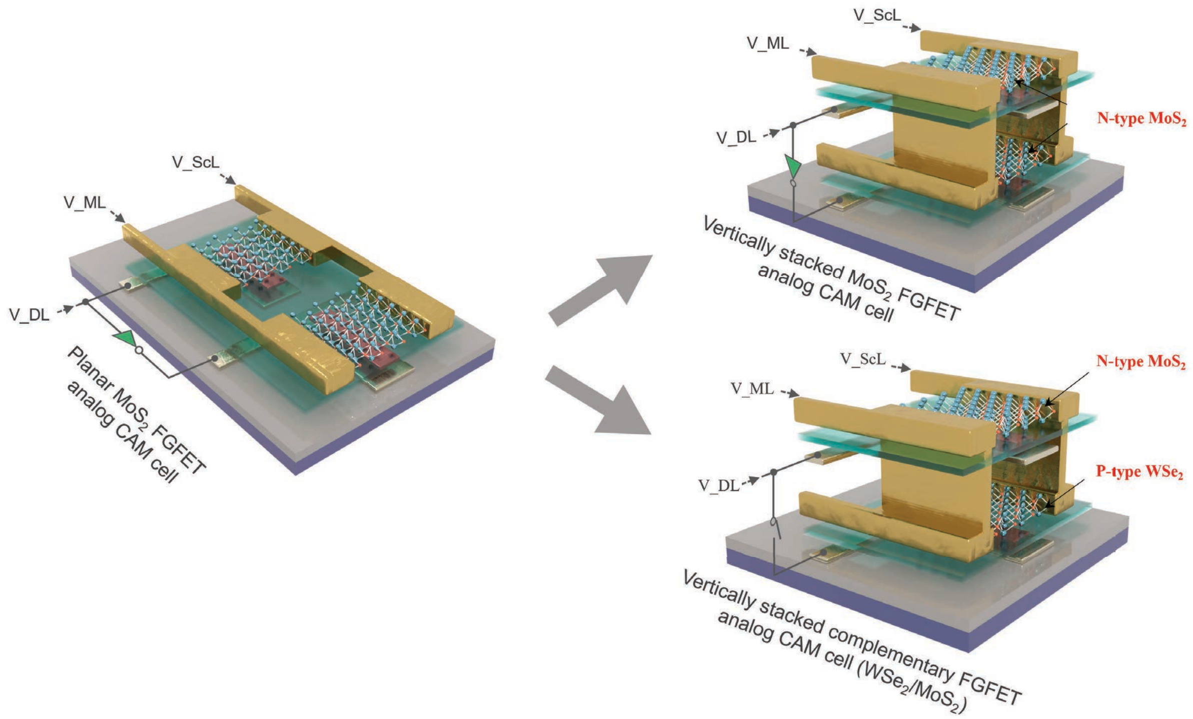
V_ScL
V_ML
V_DL
Planar MoS_2 FGFET analog CAM cell
V_ScL
V_ML
V_DL
N-type MoS_2
Vertically stacked MoS_2 FGFET analog CAM cell
V_ScL
V_ML
V_DL
N-type MoS_2
P-type WSe_2
Vertically stacked complementary FGFET analog CAM cell (WSe_2/MoS_2)

Fig. S11. The schematic diagram for different structures of one MoS_2 analog CAM cell. Compared with the planar one, vertical structure shows better area efficiency, with improved integration density and shorter interconnection that further reduce the latency. The schematic diagram of monolithic integration of complementary (N- type MoS_2 and P-type WSe_2) flash memories for one analog CAM cell.

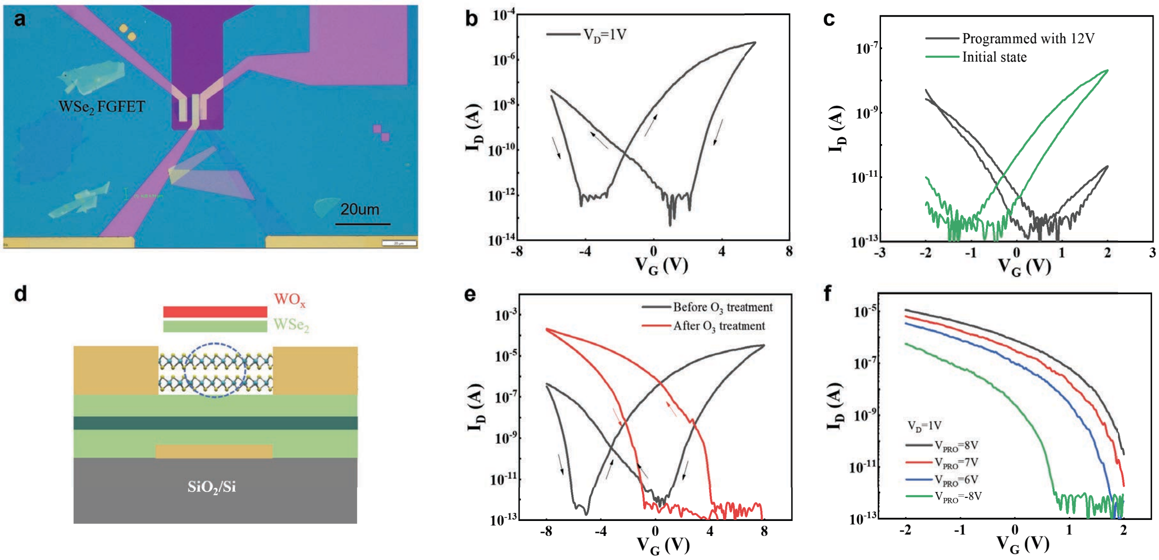

Fig. S12. The program performance of p-type WSe_2 flash memories. (a) The optical image of the fabricated back gate WSe_2 flash memory with contact metal Sb/Au. (b) I_D-V_G curves of WSe_2 flash memory. The as-fabricated device shows an ambipolar behavior with a large memory window for both n and p branches. (c) Two programmed states of WSe_2 flash memory. (d) The schematic diagram of one WSe_2 flash memory with WOx p-doping effect through UV-O3 treatment. (e) P branch can be enhanced by O_3 treatment due to the p-doping effect of WO_x, by exposing the WSe_2 channel in O_3 environment for 15min. (d) Four programmed states of WSe_2 flash memory with program voltage of -8, 6, 7, and 8V.

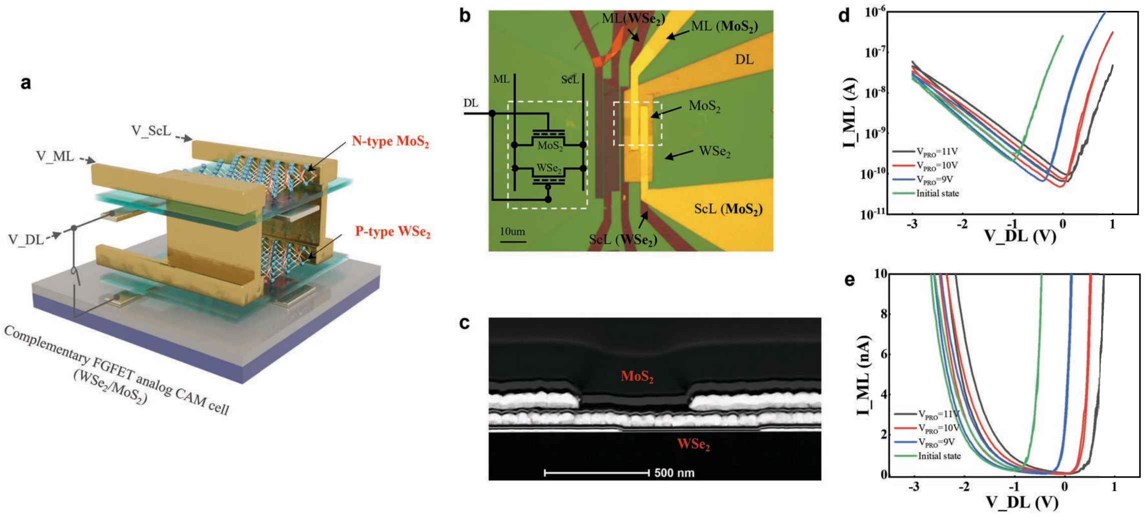

Fig. S13. Experimentally demonstrate the search operations performance of one analog CAM cell with 3D-stacked complementary 2D flash memory devices. (a)The schematic diagram of one analog CAM cell with monolithic integration of complementary flash memories (N-type MoS2 and P-type WSe2). (b) The optical image of the monolithic integration of complementary flash memories. The inset shows the circuit diagram. (c) Cross-sectional HAADF-STEM image of the fabricated device. (d-e) The I_ML-V_DL (I_D-V_G) curves in log and liner scale, showing a tunable match range.