nature sensors

Supplementary information

https://doi.org/10.1038/s44460-026-00122-3

Direct light-to-token conversion with integrated 2D photosensitive memory

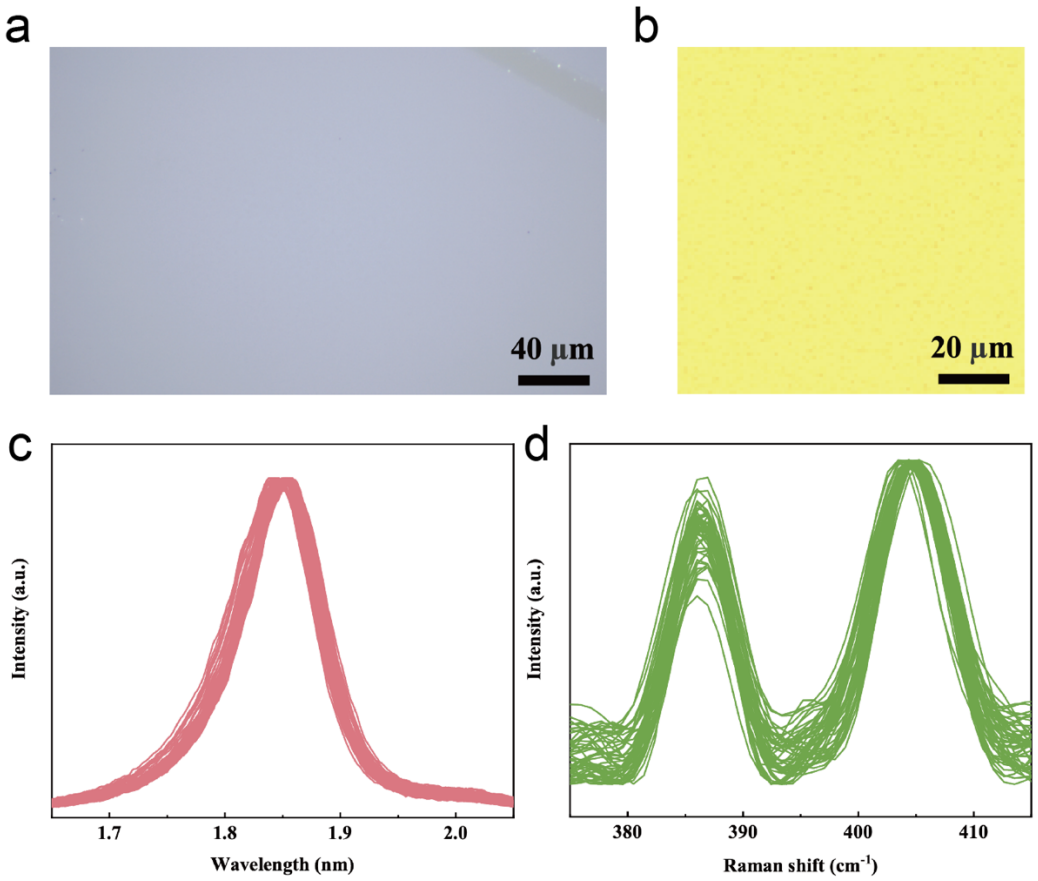
In the format provided by the authors and unedited

Supplementary Information

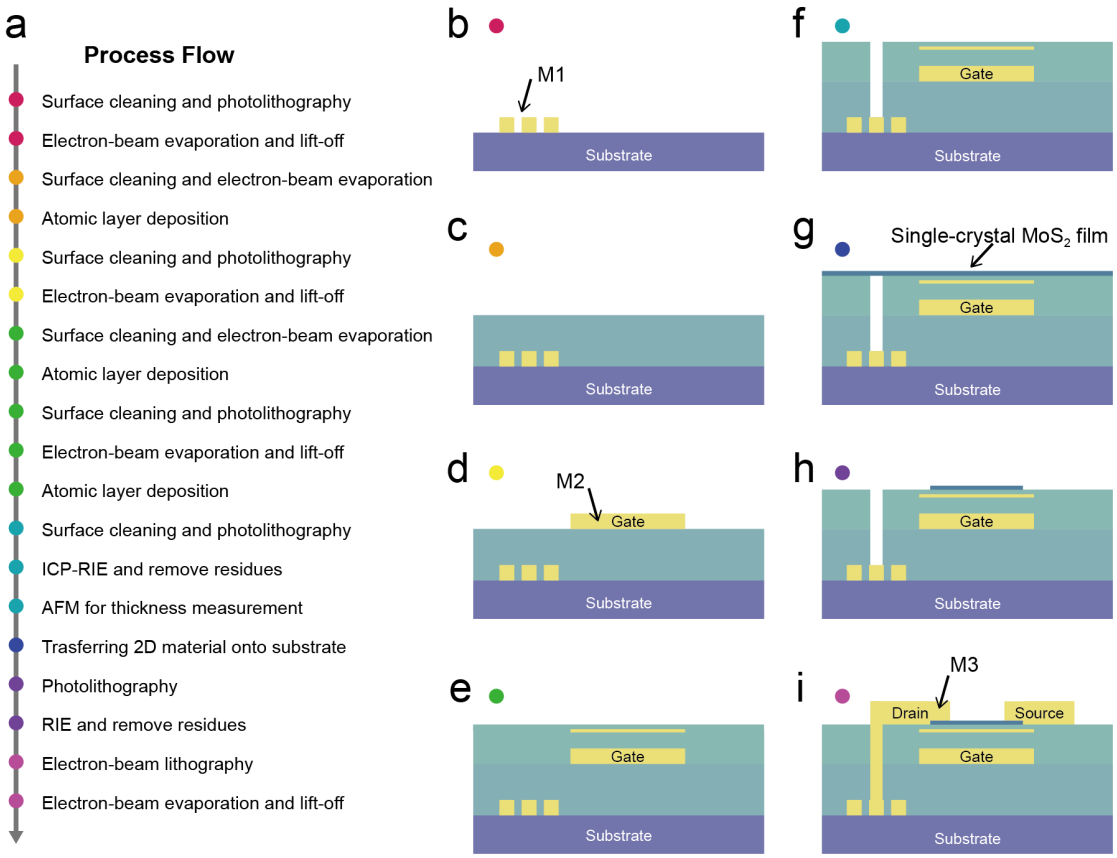
Table of Contents

Section 5: Hardware resource consumption

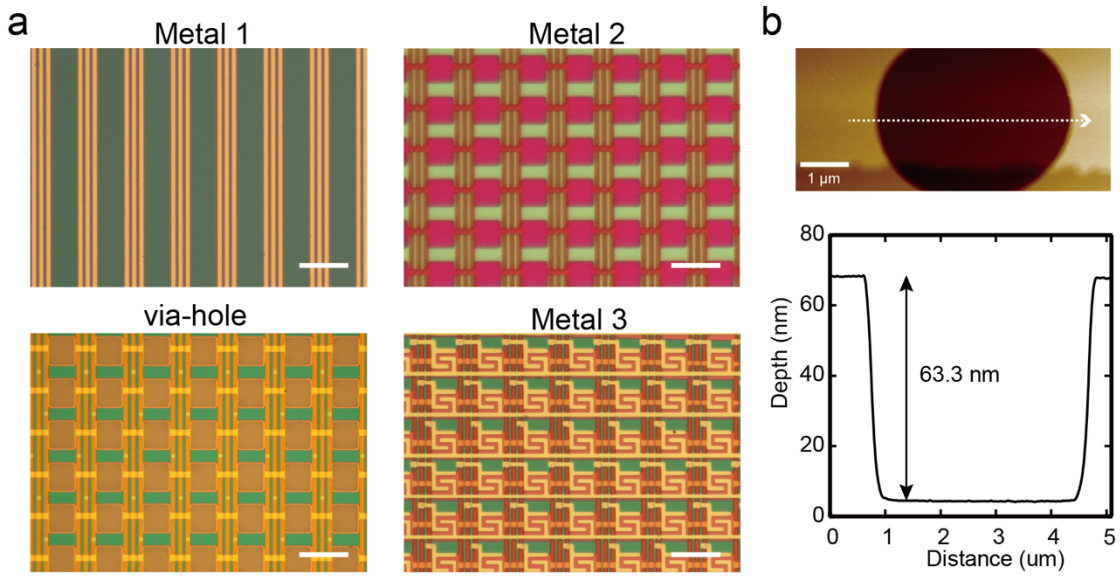
Section 1: Fabrication and Characterization of photosensitive memory array



Supplementary Fig. 1. Single-crystal MoS_2 film characterizations. a, Optical microscope image of single-crystal MoS_2 film. **b,** Polarized second harmonic generation (SHG) mapping for single-crystal MoS_2 film. **c,** Photoluminescence **(**PL) and **d,** Raman spectrum measured and collected over the area of 200×200 μm^2.



Supplementary Fig. 2. Fabrication process flow. a, Schematic of the fabrication process, with 19 steps marked by 8 colored circles. **b-i**, Cross-sectional illustrations corresponding to different fabrication stages, each labeled with a circle matching the respective color in **a**. This process flow is CMOS-compatible and can benefit from ongoing advances in wafer-scale 2D-material growth and wafer-level integration technologies.



Supplementary Fig. 3. Optical and AFM characterization. a, Optical microscope images representing four key stages of the process: Metal 1, Metal 2, via-hole, and Metal 3, with each image featuring a 25 μm scale bar. b. An AFM image acquired after via-hole formation, demonstrating that the vias effectively connect the interlayer between M1 and M3 and exhibit a smooth and clean surface.

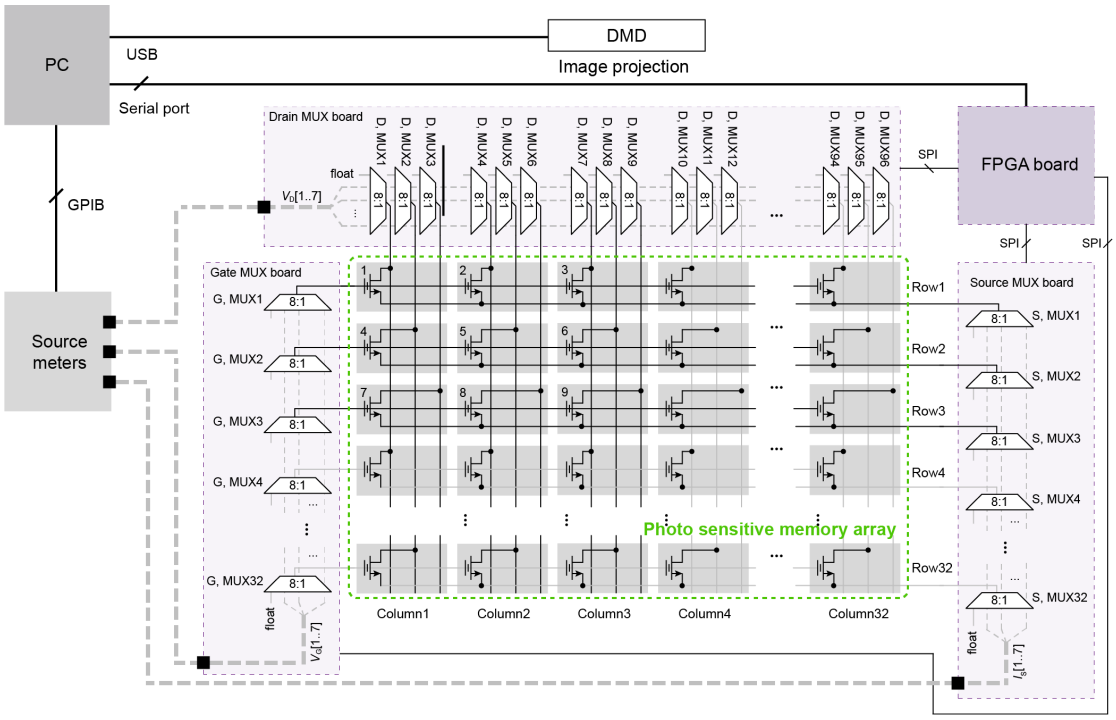
Supplementary Note 1. Array Fabrication. The fabrication process plays a crucial role in determining the yield and performance of large-scale 2D-material array. To achieve scalable fabrication of a 32×32 monolayer MoS_2 array with three-layer metal interconnections, we systematically optimized the process flow in three key aspects:

1. Minimizing Resist Residue. We first introduced an additional surface cleaning procedure prior to each photolithography (PL) and atomic layer deposition (ALD) process using oxygen plasma treatment. These optimizations enhanced substrate cleanliness and flatness, leading to nearly 100% material retention on the substrate after transfer and etching. Furthermore, etching residues were effectively removed through immersion in heated acetone, followed by mild ultrasonication, as detailed in the Methods. Finally, to improve the patterning accuracy and ensure high-quality contact between the 2D material and M3, electron beam lithography (EBL) was used instead of conventional PL.

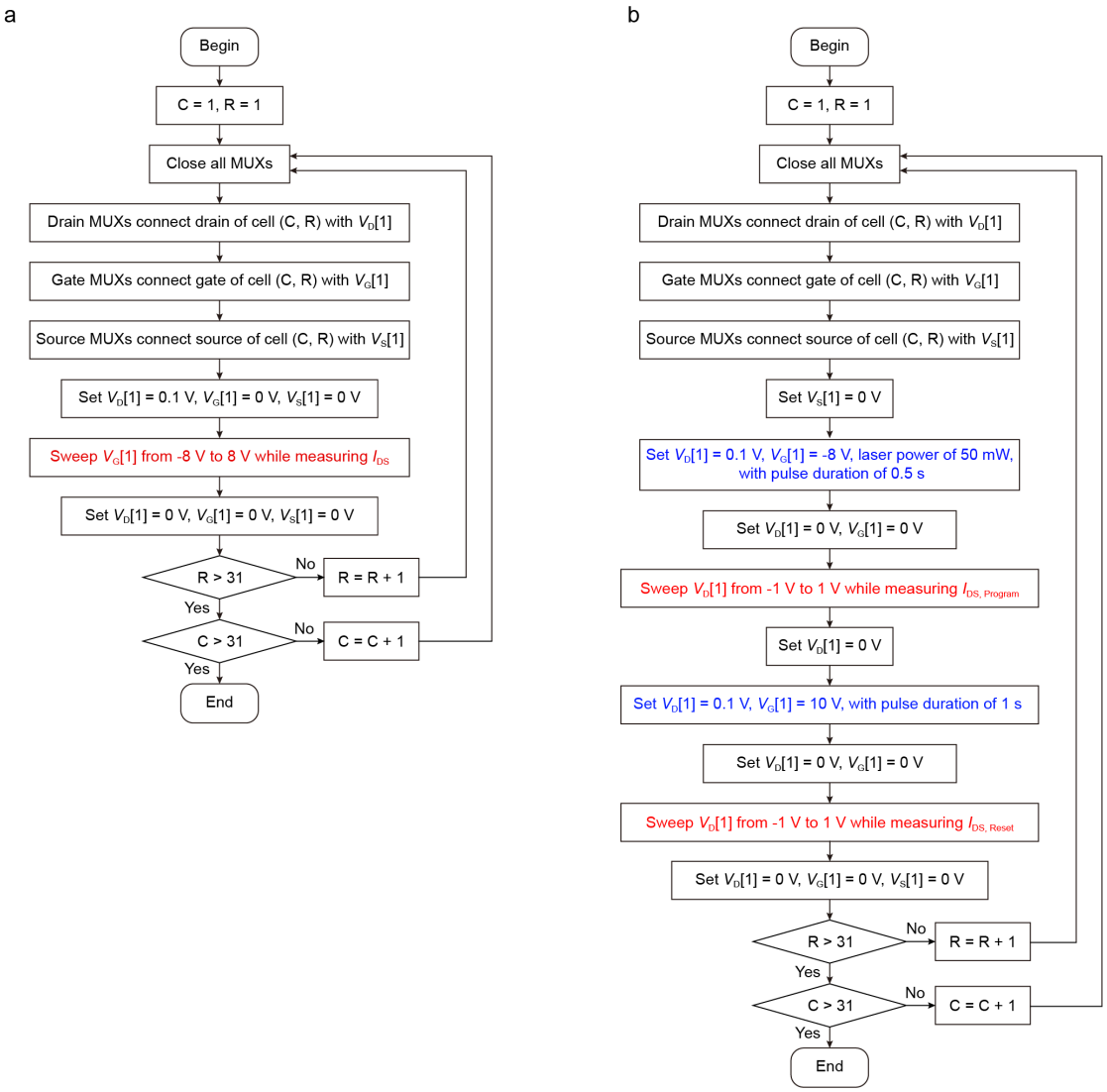
2. Improving the Dielectric Layer Quality. To enhance dielectric layer quality, we employed a global pre-deposition of a 1 nm Al seeding layer using electron-beam evaporation (EBE) to provide uniform nucleation sites for subsequent ALD growth on metal electrodes. Additionally, we optimized the ALD temperature profile by initially depositing an 8 nm oxide layer at 120 ℃ to improve adhesion, followed by a 14 nm high-temperature deposition at 250 ℃ to enhance the interface quality. Experimental results in Supplementary Fig. 10 showed a significant reduction in leakage currents.

3. Dielectric and Electrode Thicknesses. The selection of dielectric and electrode thicknesses is critical for achieving reliable electrical connections and effective gate control. However, these choices are inherently constrained by a trade-off among photoresist thickness, PL resolution, and inductively coupled plasma (ICP) etching rates. The formation of well-defined vias depends on the etching rates of the dielectric layers and the photoresist mask. A thicker photoresist is required for deeper etching, but excessive thickness compromises PL resolution. Conversely, reducing the dielectric thickness helps with vias formation during etching, but may degrade gate control. Balancing these factors, we selected 22 nm for both dielectric layers and 8 nm for the tunneling layer. The metal electrode thicknesses were set to 1 nm Ti/ 25 nm Au for M1, 1 nm Ti/ 30 nm Au for M2, and 30 nm Sb/ 40 nm Au for M3. Achieving robust electrical connection between M1 and M3 required a via depth exceeding 55 nm, along with an M3 thickness greater than the via depth. AFM measurements in Supplementary Fig. 3b confirmed that the via depth reached 63.3 nm, satisfying the interconnect requirement.

Section 2: Experimental setup and control configuration



Supplementary Fig. 4. Experimental hardware setup. Our hardware setup consists of a PC, a DMD, source meters, custom multiplexer (MUX) boards, an FPGA, and a photosensitive memory array. During array measurements, DMD commands are transmitted from the PC via USB and controlled by the vendor-provided software, while FPGA commands are transmitted via a serial protocol and executed by Python scripts to control the drain, gate, and source MUXs through SPI. Measurement-related commands, such as applying voltage or measuring current, are sent to the source meters via the GPIB protocol. Importantly, the DMD and FPGA are controlled by separate programs and do not share a hardware trigger or synchronized clock. Therefore, in the array-level programming experiment, a 0.5 s programming window was used to ensure sufficient temporal overlap among the DMD-projected optical pattern and the electrical programming signal.



Supplementary Fig. 5. Flowcharts for array measurements. a, Flowchart of the transfer characteristics measurement within a 32×32 photosensitive memory array, and **b,** Flowchart of the I-V characteristics measurement after programming and resetting each device in 32×32 array. Here, R and C denote the row and column indices of the array. V_D[1] (V_G[1], V_S[1]) represents that the first voltage channel is selected through the 8:1 Drain (Gate, Source) MUX. Instructions marked in red indicate the steps where current is measured, while instructions marked in blue denote the steps where the device state is modified through program and reset operations.

Section 3: Optoelectronic performance of photosensitive memory array

Supplementary Note 2. Physical Mechanism. The MoS_2 floating-gate phototransistor shows an approximately linear and symmetric dependence of the output current I_{DS} on both the logarithm of the optical power density P_{IN} and the read drain voltage $V_{D,R}$. A band diagram schematic is shown in Supplementary Fig. 6a and 6b.

During the programming stage, the monolayer MoS_2 channel absorbs light (e.g., 660 nm) and generates photogenerated electron-hole pairs. Under a negative programming gate bias V_G together with a drain programming pulse $V_{D,P}$, a strong vertical electric field is established across the Al_2O_3 tunneling dielectric (along with a lateral field along the channel). Driven by this field configuration, the photogenerated holes tunnel through the Al_2O_3 tunneling dielectric into the Au floating gate, where they neutralize part of the stored electrons, thereby modifying the floating-gate charge C_{FG}. After programming, the stored C_{FG} is retained due to dielectric isolation and acts as an additional electrostatic gating effect through capacitive coupling, shifting the effective threshold voltage V_{TH} and modulating the channel carrier concentration/channel conductance (Supplementary Fig. 6b).

The measured I_{DS} shows a sublinear (approximately logarithmic) dependence on P_{IN} (Supplementary Fig. 6c, extracted from Fig. 2e), consistent with a charge-trapping or charge-storage mediated photoresponse where $I_{DS} \propto \log(P_{IN})$[1]. In addition, owing to the good Sb contact[2], the device exhibits near-ohmic and symmetric *I-V* characteristics around zero gate bias (Supplementary Fig. 6d), ensuring an approximately proportional relationship between I_{DS} and $V_{D,R}$. Accordingly, within the defined programming and readout conditions, the device provides proportional I_{DS} response to both $\log(P_{IN})$ and $V_{D,R}$, supporting accurate current summation in the proposed array operation.

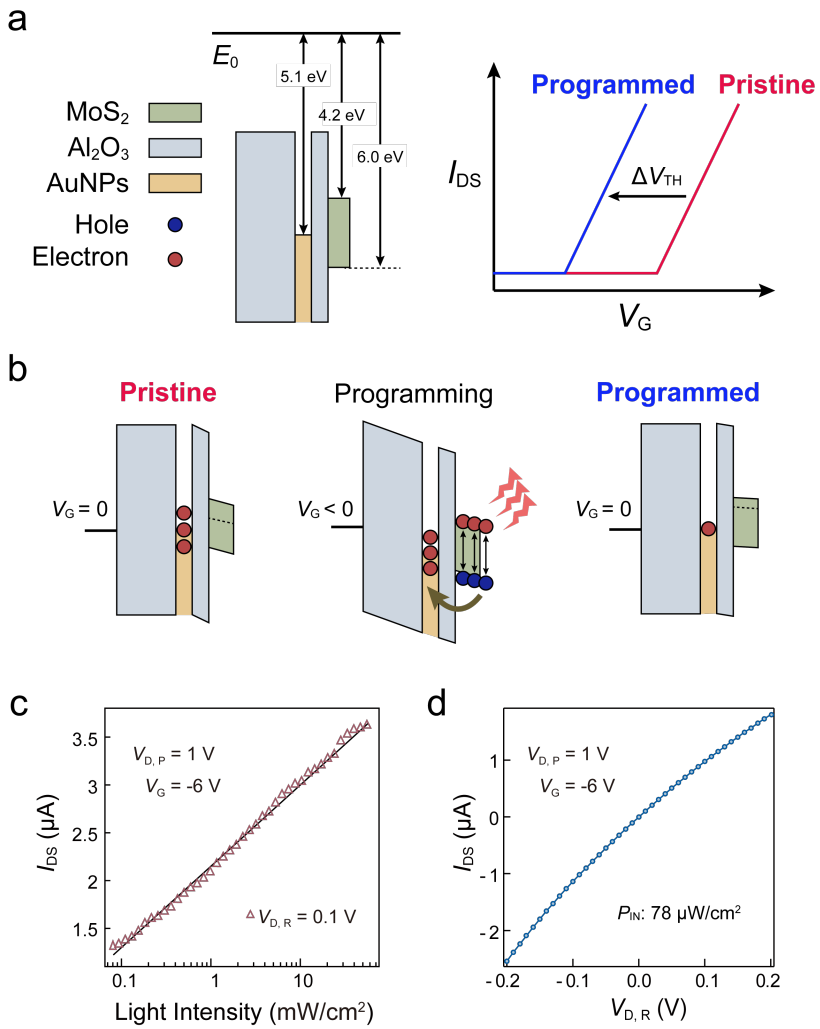
Device-level programming-speed analysis. In the original device-level programming experiment, the selected MoS_2 floating-gate phototransistor was programmed by temporally overlapping an optical pulse with an electrical gate/drain pulse (Fig. 2d and Supplementary Fig. 7). In this setup, the optical pulse was generated by a LabVIEW-controlled Daheng Optics GCI-7103M-B shutter, whereas the electrical pulse was provided by a Keithley 2636 source meter. As a result, the effective programming window was constrained by the temporal overlap between the optical and electrical stimuli.

Since the shutter response delay is approximately 80 ms (Supplementary Fig. 7a), a 100 ms programming pulse width was adopted in the original measurements as a conservative protocol to ensure sufficient overlap and reproducible programming. Therefore, the 100 ms pulse width mainly reflects the timing limitation of the laboratory measurement setup, rather than the intrinsic programming-speed limit of the device.

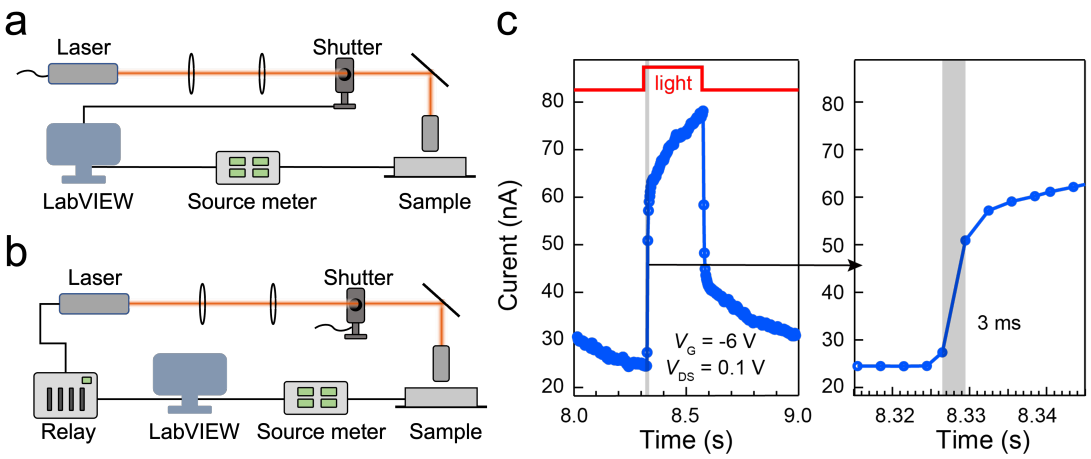
To probe the intrinsic device response more directly, we used an alternative faster optical modulation path, in which a 647 nm laser diode was modulated by a LabVIEW-controlled relay under fixed drain and gate bias (Supplementary Fig. 7b). The measured transient photoinduced current is shown in Supplementary Fig. 7c, with the zoomed-in rising edge indicating a response time of approximately 3 ms. These results confirm that the intrinsic device photoresponse can be significantly faster than the 100 ms programming pulse used in the original setup.

Electrical reset-speed analysis. In the proof-of-concept measurements, relatively long reset pulses were used to ensure complete resetting of the floating-gate state and to maintain consistency with the corresponding programming protocols. These reset durations should therefore be regarded as conservative experimental protocols rather than intrinsic reset-speed limits. Unlike optoelectronic programming, the reset operation is performed purely in the electrical domain and can be much faster.

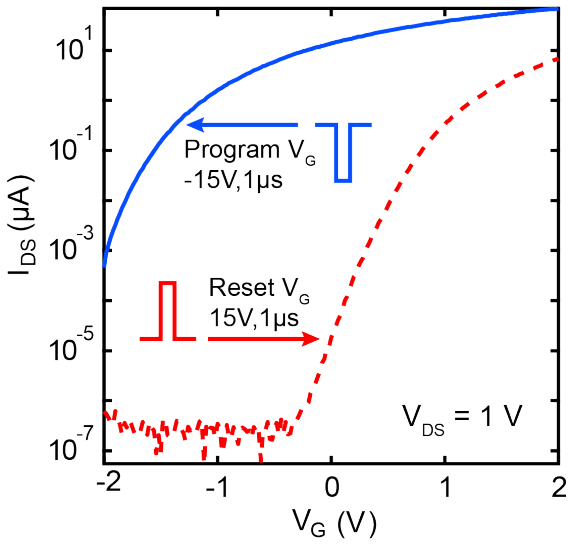
To verify this, we carried out an additional electrical program/reset experiment, in which a 1 μs gate pulse of -15 V was first used to program the device state, followed by a 1 μs gate pulse of +15 V to reset the device. The transfer characteristics measured after each pulse are shown in Supplementary Fig. 8. The clear transfer-curve shift between the programmed and reset states confirms that the floating-gate state can be electrically reset within 1 μs. This value should not be interpreted as the intrinsic speed limit of the reset operation, but rather as an experimentally verified microsecond-scale reset condition in the present setup.



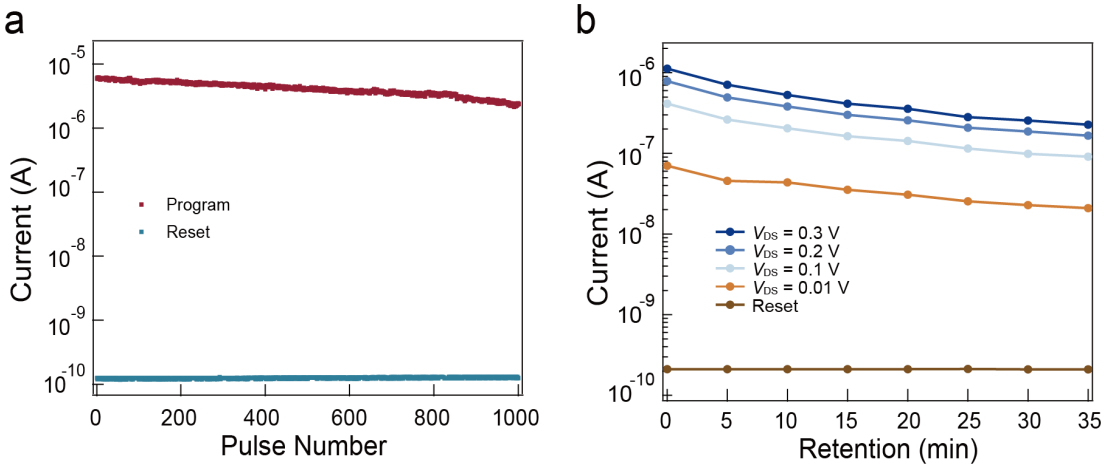
Supplementary Fig. 6. Physical mechanism of MoS_2 floating-gate phototransistor. **a,** Device schematic (left) and illustrative transfer-curve shift before/after programming (right). b, Schematic band diagram of the device in the pristine, programming, and programmed (readout) states. c, Extracted I_{DS} - log(P_{IN}) characteristic from Fig. 2e. **d,** I_{DS}-$V_{D,R}$ curve demonstrating near-ohmic and symmetric behavior.



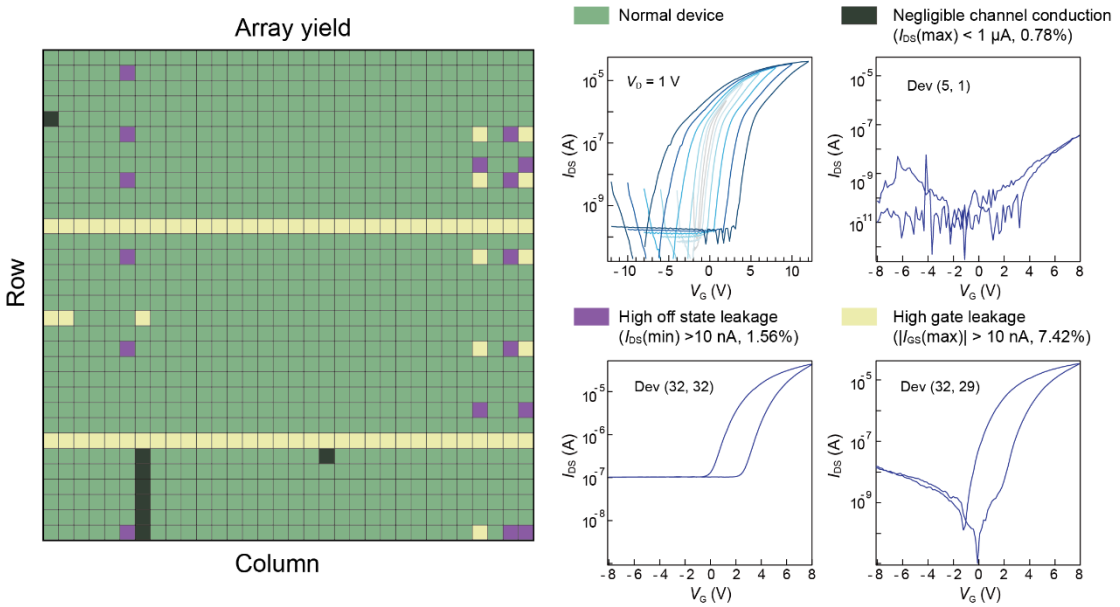
Supplementary Fig. 7. Device-level programming-speed analysis. a, Shutter-based optoelectronic programming setup used for Fig. 2d, where the optical pulse is generated by a LabVIEW-controlled Daheng Optics GCI-7103M-B shutter with an approximately 80 ms response delay. **b,** Alternative faster optical modulation path using a LabVIEW-controlled relay to probe the transient photoresponse of the device under fixed drain and gate bias. **c,** Transient photoinduced current of the MoS_2 floating-gate phototransistor measured under fixed drain and gate bias. The right panel is a zoomed-in view of the rising edge in the left panel, showing a photoinduced current rise time of approximately 3 ms.



Supplementary Fig. 8. Electrical program/reset speed of the MoS_2 floating-gate phototransistor. Transfer characteristics of the device measured after applying a 1 μs electrical programming pulse and a 1 μs electrical reset pulse. The programmed and reset states were generated by gate pulses of -15 V and +15 V, respectively, with V_{DS} fixed at 1 V during readout. The transfer-curve shift confirms that the floating-gate state can be electrically reset on a microsecond timescale.



Supplementary Fig. 9. Optoelectronic programming performance. a, Endurance test over 1000 cycles conducted on a single device under a programming gate pulse of -8 V, drain pulse of 1 V, light intensity of 5.6 mW/cm^2, and duration of 100 ms, followed by a reset gate pulse of 10 V, drain pulse of 1 V, and duration of 100 ms. All currents were measured at V_{DS} = 1 V. **b,** Retention characteristics measured after the endurance test, where the device was programmed and its charge retention was monitored for 2000 s under varying V_{DS} conditions.



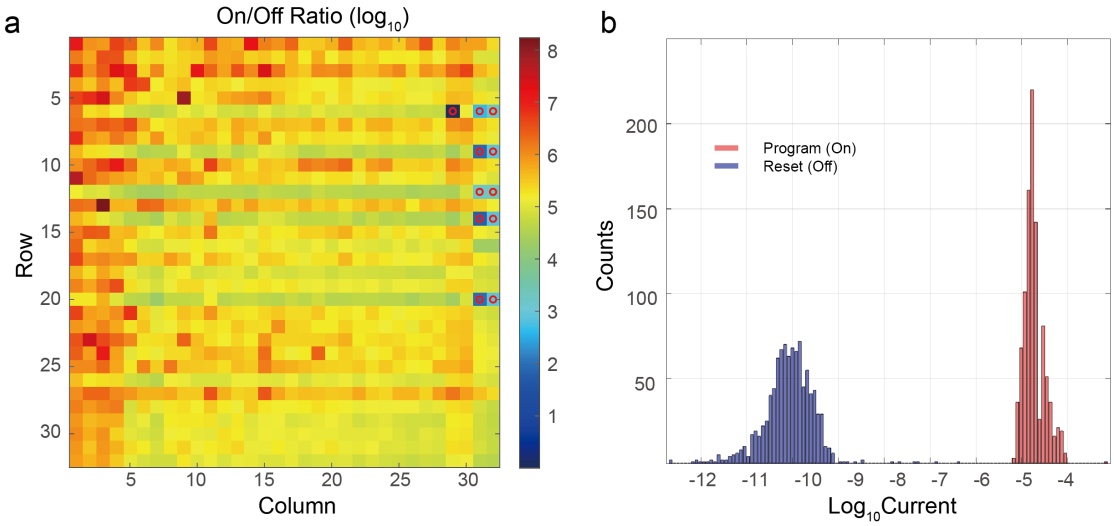
Supplementary Fig. 10. Electrical characterizations of 1024 devices. We measured the transfer characteristics of all 1024 devices in the 32×32 photosensitive memory array. The device yield was determined based on the analysis of transfer curves and categorized into four groups, each represented by a distinct color: normal device (green), negligible channel conduction (black), high off-state leakage (purple), and high gate leakage (yellow). The left panel illustrates the spatial distribution of these categories across the array, while the right panel presents representative transfer curves for each type. A detailed analysis of the yield classification criteria is provided in Supplementary Note 3.

Supplementary Note 3. Yield Analysis. Since the performance of our photosensitive memory array depends on both the optical response and the electrical charge storage capability of MoS_2 floating-gate phototransistors, we evaluate device yield using two complementary analyses. First, we assess yield based on the transfer characteristics of each device according to the following four criteria:

1. Maximum channel ON current ($I_{DS(max)}$) > 1 μA
2. Minimum channel OFF current ($I_{DS(min)}$) < 10 nA
3. Maximum gate leakage current ($|I_{GS(max)}|$) < 10 nA
4. Minimum memory window > 2 V

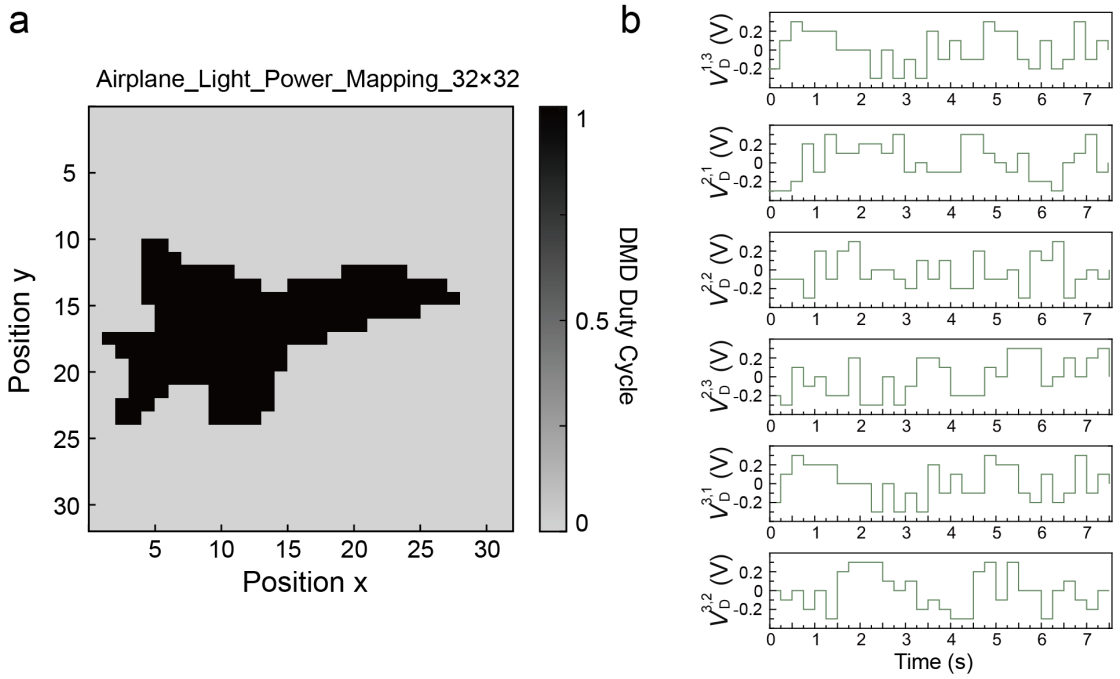
Based on these criteria, we assessed the device yield and its spatial distribution, as shown in Supplementary Fig. 10. Among the 1024 devices, 8 (0.78%) exhibited negligible channel conduction, with dev (5,1) serving as a representative example, failing to meet criterion 1. This issue is likely caused by poor metal contact. A total of 16 devices (1.56%) exhibited excessively high OFF-state leakage, with dev (32,32) serving as a representative example, failing to meet criterion 2. This issue is likely attributable to interface roughness, which weakens gate control, or poor material adhesion. Additionally, 76 devices (7.42%) exhibited high gate leakage, as illustrated by dev (32,29), failing to meet criterion 3. This high leakage, the dominant factor affecting yield, is likely attributable to compromised dielectric quality. In particular, the "rabbit ear" effect-resulting in irregular electrode edges that can promote tip discharge and hinder uniform dielectric deposition-as well as residual photoresist, likely contribute to the degraded dielectric performance. Furthermore, among the normal devices (90.23%) that did not fall into the aforementioned failure categories, as illustrated in Fig. 2f, the memory windows of the MoS_2 phototransistors remain within 2-4 V and symmetrically centered around 0 V, demonstrating the array's scalability and reliability. It is attributed to the coordinated optimization of multiple fabrication processes, ensuring a high on-off ratio for photosensitive memory and minimizing undesired sneak path.

Then, we evaluate each device's optical response and charge storage performance under light stimulation, following the measurement protocol detailed in Supplementary Fig. 5b, with results presented in Fig. 2g and Supplementary Fig. 11. It is worth noting that even in devices failing to meet the standard yield criteria 1-4, the photogenerated current in these devices dominates over the dark current and leakage, ensuring a clear photocurrent signal. Such discrepancy arises because the channel material almost covers the full active region, and most of them maintain leakage currents below 10 nA at zero gate bias. This indicates that when the yield analysis criteria are based solely on the photo-charge storage performance, achieving an ON/OFF ratio exceeding 10^4, the array yield can reach approximately 98.92%.

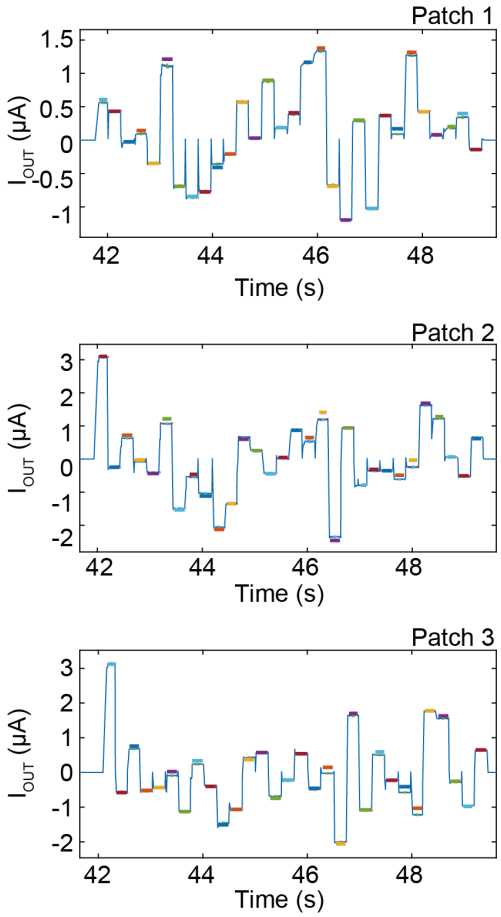


Supplementary Fig. 11. Optoelectronic characterizations of 1024 devices. We measured the I_{DS}-V_{DS} characteristics of all 1024 devices before and after optoelectronic programming (as shown in Fig. 2d) and extracted the readout current at V_D = 1 V and V_G = 0 V. **a,** The spatial distribution of the switching ratio within the array, where red circles indicate devices with an On/Off ratio lower than 10^4. **b,** Distribution of readout currents in the Programming state and Reset states across the array.

Section 4: Patch embedding



Supplementary Fig. 12. Patch embedding experiment. a, The original airplane image from CIFAR-10 is binarized for the experiment. **b,** This panel shows remaining six 1×30 vectors of a 3×3×30 embedding matrix (represented as nine 1×30 vectors), while three vectors ($V^{1,1}$, $V^{1,2}$, $V^{3,3}$) are shown in Fig. 3i.



Supplementary Fig. 13. Patch embedding results analysis. We analyze the discrepancy between experimental results (Fig. 3j) and theoretical values to demonstrate the array's capability in patch embedding. After programming all devices in the 3×3 patch, the theoretical values were calculated by measuring I_{DS} as a function of V_{DS} for each device and performing the same patch embedding computations on the PC. Our experimental data follow the same trend as the theoretical prediction, with a slightly lower magnitude that can be attributed to device variations and minor retention during sequential readout. The sharp, abrupt current variations in Fig. 3j originate from transient switching of the multiplexers during patch addressing and do not affect the embedding values, which are extracted from the steady-state current after settling.

Supplementary Note 4. Effect of input-state precision on the token fidelity

The multi-level conductance states of the MoS_2 floating-gate phototransistors are not only a device-level characteristic, but also directly determine the input-state precision of the physical tokenizer. In the physical tokenizer, the incident optical image is stored in the photosensitive memory array as nonvolatile conductance states, and each selected image patch is multiplied with the embedding-matrix voltage pattern to generate patch embeddings. Therefore, the number of distinguishable optical memory states affects how much grayscale information is retained before patch embedding and, consequently, the fidelity of the generated token maps.

To evaluate this effect, we used a representative 224×224 image derived from the CIFAR-10 dataset and quantized it into three input precision levels: 2-level, 41-level, and 256-level inputs (Supplementary Fig. 14). The 2-level input represents a strongly quantized binary case, the 41-level input corresponds to the experimentally demonstrated optical input-state precision in this work, and the 256-level input is used as a high-precision reference. The same trained embedding-matrix slices shown in Fig. 4c were then applied to these three input images. Specifically, each 224×224 image was divided into 196 non-overlapping patches with a patch size of 16×16. For each selected embedding-matrix slice, every 16×16 patch was multiplied with the corresponding 16×16 weight pattern and summed to generate one token value, producing a 14×14 token map.

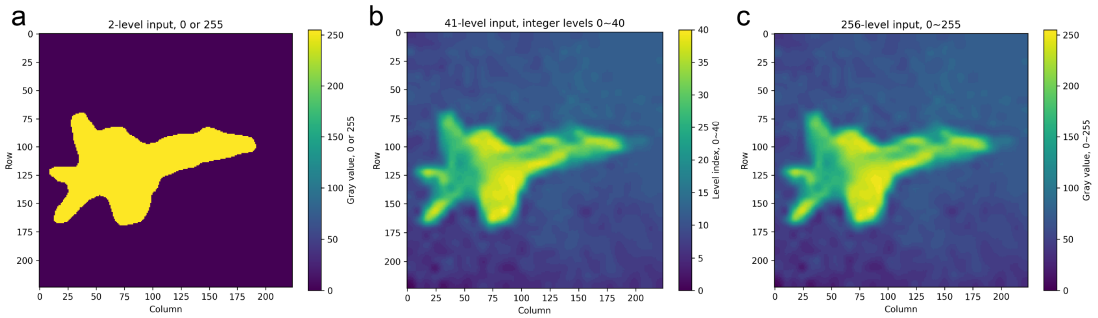
The generated token maps are shown in Supplementary Fig. 15. The 2-level input produces visibly distorted token maps compared with the 256-level reference, because the binary representation loses most of the grayscale information in the original image. In contrast, the token maps generated from the 41-level input preserve the major spatial structures and show much closer visual agreement with the 256-level reference. This comparison indicates that the experimentally demonstrated 41-level optical memory states provide sufficient input precision to retain meaningful image information during physical patch embedding.

We further quantified the token-map fidelity by calculating the pixel-wise absolute error using the 256-level token maps as the high-precision reference:

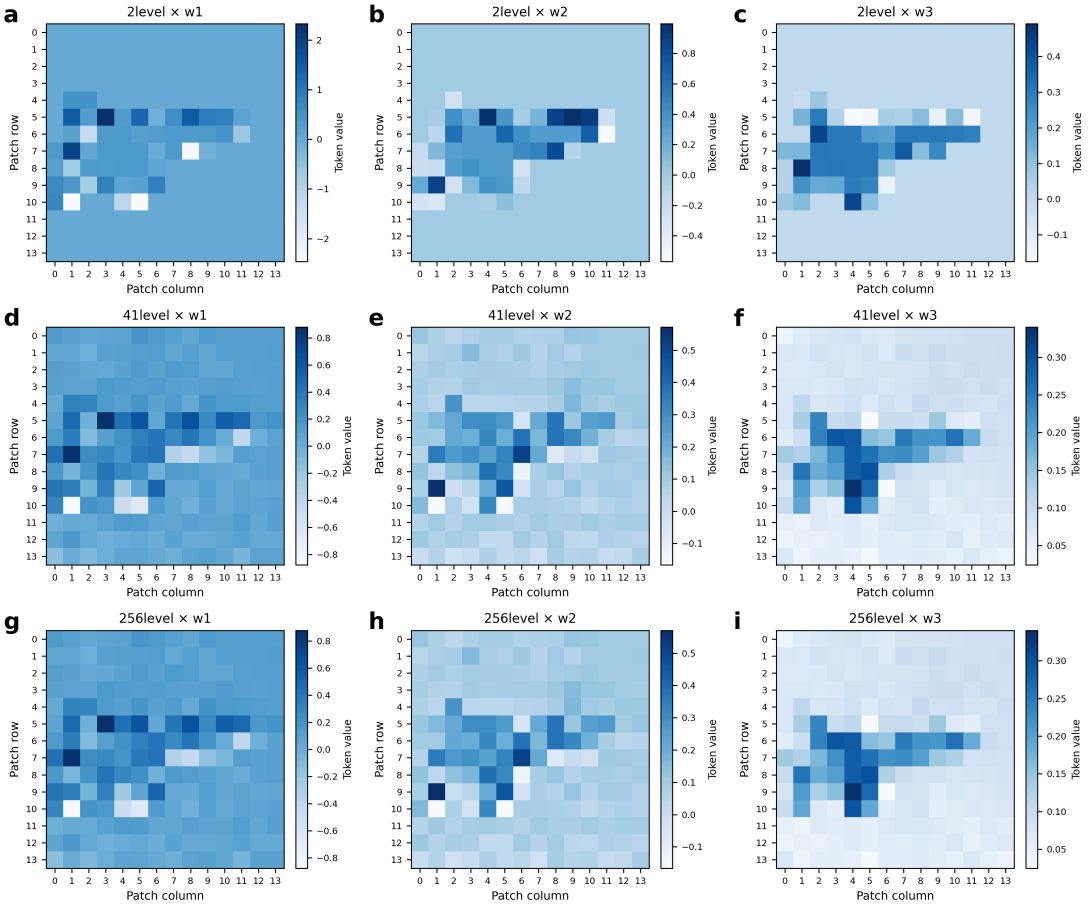
$$\mathrm{E} = |Token_{N-level} - Token_{256-level}|$$

where N is 2 or 41. As shown in Supplementary Fig. 16, the absolute errors of the 41-level token maps are substantially smaller than those of the 2-level token maps for all representative embedding-matrix slices. These results demonstrate that the multi-level optical memory states directly improve the fidelity of the generated tokens by preserving more grayscale information before the analog patch-embedding operation. Therefore, the multi-state characteristics of the photosensitive memory are functionally exploited in the physical tokenizer, rather than remaining only a device-level characterization. This improved token fidelity is also consistent with the software-

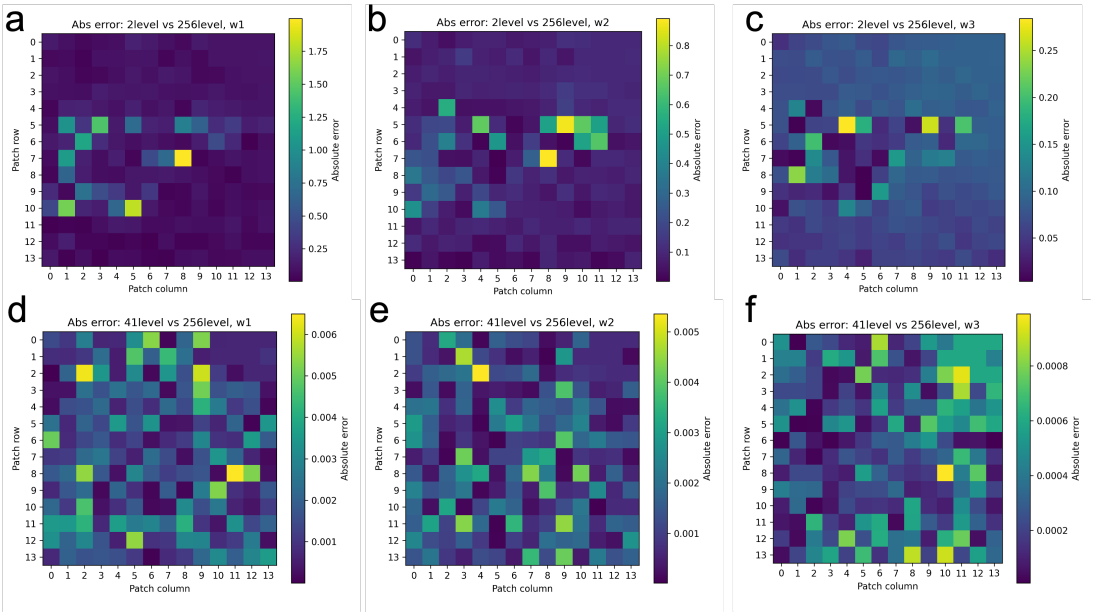
comparable classification accuracy obtained using the experimentally measured device response in the physical tokenizer evaluation.



Supplementary Fig. 14. Representative input images with different input-state precision. A representative image derived from the CIFAR-10 dataset was quantized into three input precision levels: a, 2-level input; b, 41-level input; and c, 256-level input. The 41-level input corresponds to the experimentally demonstrated optical input-state precision of the MoS_2 floating gate phototransistor, while the 256-level input is used as a high-precision reference.

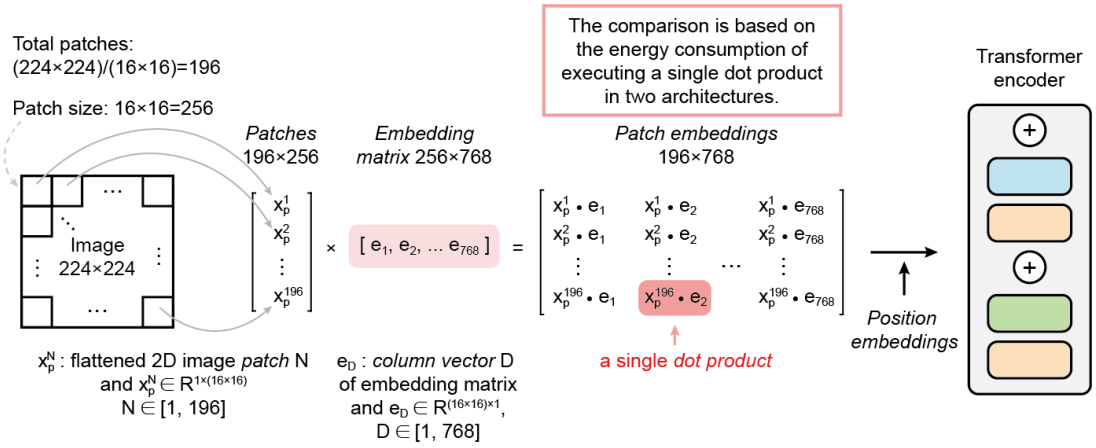


Supplementary Fig. 15. Token maps generated under different input-state precisions. The 2-level, 41-level, and 256-level input images were divided into 16×16 patches and multiplied with representative 16×16 embedding-matrix slices extracted from the trained embedding matrix shown in Fig. 4c. Each patch-wise multiplication and summation produces one token value, resulting in a 14×14 token map. Compared with the 2-level input (a-c), the 41-level input (d-f) produces token maps that more closely match those generated from the 256-level reference (g-i), indicating improved token fidelity enabled by multi-level optical memory states.



Supplementary Fig. 16. Absolute error analysis of token maps. Absolute error maps were calculated using the 256-level token maps as the high-precision reference. a-c, Absolute errors between the 2-level and 256-level token maps. d-f, Absolute errors between the 41-level and 256-level token maps. The error was calculated pixel-wise as$|Token_{N\text{-}level}-Token_{256\text{-}level}|$, where N is 2 or 41. The smaller errors in d-f show that the experimentally demonstrated 41-level input states better preserve grayscale information during physical patch embedding, thereby improving the fidelity of the generated tokens.

Section 5: Hardware resource consumption



Supplementary Fig. 17. Schematic of a single dot product operation in ViT. The comparison focuses on the energy consumption of executing a single dot product in two architectures, which are detailed in Supplementary Note 5.

Supplementary Note 5. Energy Consumption Comparison.

As shown in Supplementary Fig. 17, the input image has a typical size of 224×224, which is divided into 14×14 patches, each of size 16×16. If the Transformer's input dimension is 768, an image requires (14×14) ×768 dot products, each involving a 256-length vector. The comparison focuses on the energy required to execute a single dot product in each architecture.

Supplementary Fig. 18a illustrates the schematic of the energy comparison boundary between the proposed physical tokenizer and the conventional digital architecture. The energy comparison in this work focuses on the tokenization stage highlighted on the left of the red dashed line, where physical tokenizer integrates sensing and processing in the photosensitive memory array, whereas the conventional approach separates image sensing and a digital dot-product engine with intermediate data movement. The detailed energy breakdowns for the two paths are provided in Supplementary Fig. 18b and Supplementary Fig. 18c. The blocks on the right of the red dashed line are shared by the two architectures and use the same Transformer encoder, so their energy contributions are identical and are not counted repeatedly in our comparison.

Physical tokenizer. The proposed physical tokenizer is shown in Supplementary Fig. 18b, including (1) DACs, (2) drain buffers, (3) drain-side and source-side MUXs, (4) the photosensitive memory array, (5) a current bias, (6) a current-to-voltage converter (I-to-V), and (7) an ADC. The time required to execute a single dot product in the physical tokenizer is determined by the settling time of the output current after drain MUXs switching, which is experimentally measured to be 4 μs in the 3×3 patch-embedding prototype, as shown in Supplementary Fig. 19c.

Although the energy analysis is performed for a 16×16 patch configuration, this 4 μs value is adopted as a practical timing target because the measured delay in the present prototype is mainly limited by board-level RC settling. Tighter integrated implementations with shorter interconnects and reduced parasitics are expected to support this operating window at the 16×16 scale. Thus, the speed requirements for peripheral circuits are set to 4 μs, and their energy consumption is calculated based on their respective operation speeds.

(1) **DACs**. The time required to execute one dot product corresponding to one patch-embedding element is 4 μs. For a fixed embedding-vector component, the same DAC outputs can be reused while scanning all 14 × 14 patches. Therefore, completing this operation for all patches requires 14 × 14 × 4 μs = 784 μs, and the DACs only need to be updated at a frequency of 1/784 μs, i.e., approximately 1.28 kHz.

To estimate the DAC energy conservatively, we refer to a reported 10-bit, 40 kHz sample rate SAR ADC with a total power consumption of 72 nW, in which the 10-bit

DAC accounts for approximately 25% of the total ADC power [IEEE Journal of Solid-State Circuits, 2013, 48(12), 3011-3018]. Under this assumption, the energy consumed by one DAC conversion is estimated as

$$72\ \text{nW} / 40\ \text{kHz} * 0.25 = 0.45\ \text{pJ}$$

Since a 16 × 16 patch requires 256 DAC channels, the total DAC energy for one dot-product operation is:

$$16 \times 16 \times 0.45 = 115.2\ \text{pJ}$$

(2) **Drain buffers**. During the voltage-settling phase of the drain buffers, the other circuit blocks can be temporarily disabled to reduce unnecessary power consumption. Therefore, the drain buffers do not need to operate at a high speed. Here, we estimate the buffer energy using the specifications of the TI LPV802 low-power operational amplifier. Under a 1.8 V supply, the LPV802 consumes a quiescent current of 320 nA and provides a gain-bandwidth product (GBW) of 8 kHz [quiescent current data and GBW data are adopted from https://www.ti.com.cn/product/LPV802], allowing the output voltage to settle within approximately 500 μs. The energy consumed during this voltage-settling phase is therefore:

$$1.8\ \text{V} \times 320\ \text{nA} \times 500\ \mu\text{s} = 288\ \text{pJ}$$

Once the buffer voltage is established, it can be reused for the subsequent 14 × 14 patch-embedding operations. Therefore, the voltage-settling energy amortized to each patch is:

$$288 / (14 \times 14) = 1.4694\ \text{pJ}$$

During the 4 μs dot-product operation, the energy consumed by each buffer is:

$$320\ \text{nA} \times 1.8\ \text{V} \times 4\ \mu\text{s} = 2.304\ \text{pJ}$$

Therefore, the total energy consumed by one buffer to generate one patch-embedding element is:

$$1.4694 + 2.304 = 3.7734\ \text{pJ}$$

Since a 16 × 16 patch requires 256 drain-buffer channels, the total buffer energy before load-capacitance normalization is:

$$(1.4694 + 2.304) \times 16 \times 16 = 965.99\ \text{pJ}$$

It should be noted that the LPV802 is specified for driving a load capacitance of 20 pF. In a future chip-level implementation of the physical tokenizer, the drain buffer is not expected to drive such a large load. In our architecture, the total capacitive load driven by each buffer, including 14 transmission gates and 16 photosensitive memory devices,

is estimated to be 7.35 pF (evaluated using the Cadence Virtuoso design environment with a foundry-provided CMOS process design kit). For an operational amplifier with the same topology, phase margin, GBW, and transistor overdrive condition, the required bias current generally increases approximately linearly with the load capacitance. Therefore, as a first-order conservative estimate, we normalize the buffer energy by the load-capacitance ratio of 7.35 pF / 20 pF. The resulting total energy consumption of the 16 × 16 drain buffers is:

$$965.99 \times 7.35 / 20 = 355.0 \text{ pJ}$$

(3) **Drain-side and source-side MUXs**. The drain-side and source-side MUXs require custom design. Their energy consumption is evaluated using the Cadence Virtuoso design environment with a foundry-provided CMOS process design kit (PDK). The simulation schematic is shown in Supplementary Fig. 20.

For a 16 × 16 patch, a 256 × 1 transmission-gate array is required to route the corresponding drain-line signals. The energy consumption of the transmission-gate array mainly originates from the dynamic current generated when the control registers drive the capacitive loads of the transmission gates during switching. The energy (E_{drMUX}) is calculated by integrating the supply current I_{VDD} over the 4 μs operation window and multiplying it by the supply voltage:

$$\mathrm{E_{drMUX}} = VDD \int_{t_0}^{t_0+4\mu s} I_{VDD}(t)dt$$

As shown in Supplementary Fig. 21, the simulated energy consumption of the 256 × 1 drain-side transmission-gate array is:

$$E_{drMUX} = 380.9 \text{ pJ}.$$

In the adopted operation sequence, the source-side MUXs are switched at one-fourteenth of the drain-side MUX switching rate. Therefore, the source-side MUX energy is estimated as one fourteenth of the drain-side MUX energy:

$$380.9 / 14 = 27.2 \text{ pJ}.$$

(4) **Photosensitive memory array**. The energy consumption of the photosensitive memory array is determined from experimental data, as shown in Supplementary Fig. 19b. In the 3 × 3 patch-embedding prototype, the total current of nine photosensitive memory devices is approximately 1 μA. Scaling this value to a 16 × 16 patch gives an estimated array current of (16×16) / 9 × 1 μA. Since the device current is supplied by the drain buffers, the supply voltage is taken as 1.8 V. Therefore, the energy consumed by the 16 × 16 photosensitive memory array during the 4 μs dot-product operation is:

$$(16\times16) / 9 \times 1\ \mu A \times 1.8\ V \times 4\ \mu s = 204.8\ pJ$$

(5) **Current bias**. The current bias I_{bias} is used to ensure that the patch-embedding current after current-to-voltage conversion remains within a positive voltage range suitable for ADC sampling. Therefore, I_{bias} is set to exceed the maximum expected output current of the photosensitive memory array.

In the measurement condition shown in Supplementary Fig. 19a, the device bias is switched from half of the maximum positive bias to half of the maximum negative bias. Accordingly, the maximum total output current of the 3 × 3 patch can be conservatively estimated as approximately 2×1 μA. Scaling this value to a 16 × 16 patch gives the maximum expected output current $2 \times (16\times16) / 9 \times 1\ \mu A$. Therefore, under a 1.8 V supply and a 4 μs operation window, the energy consumed by the current bias is estimated as:

$$2 \times (16\times16) / 9 \times 1\ \mu A \times 1.8\ V \times 4\ \mu s = 409.6\ pJ$$

(6) **Current-to-voltage converter**. The current-to-voltage conversion must be completed within the 4 μs operation window. To estimate the energy required for current-to-voltage conversion, we implement the current-to-voltage converter using a transimpedance-amplifier structure. We use the specifications of the TI OPA338 operational amplifier as a conservative reference. Its gain-bandwidth product of 62.5 MHz (data from https://www.ti.com/product/OPA338) is sufficiently higher than the required operating speed, ensuring that the amplifier bandwidth is not the limiting factor in the current-to-voltage conversion.

The OPA338 has a quiescent current of 0.525 mA. Under a 3.6 V supply, the energy consumed during a 4 μs operation window is:

$$0.525\ mA\times3.6\ V\times4\ \mu s=7.56\ nJ$$

It should be noted that the OPA338 is designed to drive a load capacitance larger than 100 pF. In the proposed physical tokenizer, however, the load of the current-to-voltage converter is mainly the input capacitance of the following ADC, which is typically on the order of 2 pF (adopted from ADI AD9057, https://www.analog.com/en/products/ad9057.html). For an operational amplifier with the same topology, phase margin, GBW, and transistor overdrive condition, the required bias current generally scales approximately with the load capacitance. Therefore, as a first-order load-capacitance-normalized estimate, the energy of the current-to-voltage converter is scaled by the ratio of 2 pF / 100 pF:

$$7.56\ nJ \times 2\ pF\ /100\ pF= 151.2\ pJ$$

(7) **ADC**. In the proposed physical tokenizer, the ADC only digitizes the current-to-

voltage converted output corresponding to one dot-product result, i.e., one element of the patch embedding. Therefore, only one ADC conversion is required for each dot-product operation. To preserve the analog information contained in the patch-embedding current as much as possible, we adopt a 12-bit ADC for digitizing the current-to-voltage converted output.

The energy required for a single ADC conversion is estimated from the reported ADC energy-per-conversion data summarized in Supplementary Fig. 22. Based on these representative ADC results, we use 100 pJ as a typical ADC energy per conversion. Therefore, the ADC energy contribution for one dot-product operation in the physical tokenizer is 100 pJ.

Therefore, the total energy consumption of the physical tokenizer is:

$$115.2 + 355.0 + 380.9 + 27.2 + 204.8 + 409.6 + 151.2 + 100 = 1743.9 \text{ pJ}$$

Conventional architecture with 41-level CIS readout precision. For comparison, the typical conventional architecture is shown in Supplementary Fig. 18c. It consists of (1) vertical and horizontal scanners, (2) a CMOS image sensor array and column current sources, (3) amplifiers, (4) ADCs, (5) I/O buffer for data transfer, and (6) a digital dot product chip.

To make the comparison conservative and favorable to the conventional baseline, the data-movement energy is simplified as the energy required for the I/O buffer to drive the PAD capacitance and board-level interconnect capacitance. This simplified treatment does not include additional overheads from practical I/O design or data formatting. Therefore, the resulting comparison should be regarded as a conservative estimate of the energy advantage of the proposed physical-tokenizer architecture.

To maintain the same computational throughput as the physical tokenizer, each 16×16 patch in the conventional architecture must be sampled and processed within the same 4 μs operation window. This timing requirement is used to determine the energy consumption of each circuit block in the conventional baseline.

(1) **Vertical and horizontal scanners**. The energy consumption of the vertical and horizontal scanners is evaluated using the Cadence Virtuoso design environment with a foundry-provided CMOS PDK. Simulation results reveal that their energy consumption is 1.670 pJ. Since this contribution is relatively small compared with the other circuit blocks, we neglect it to make the comparison more favorable to the conventional baseline.

(2) **CMOS image sensor array and column current sources**. The column current source consumes a constant current over 4 μs. The current-source parameter is

estimated based on a reported CMOS image sensor design [IEEE Journal of Solid-State Circuits, 2005, 40(5): 1147-1156], in which the transconductance of the current-source transistor is 144 μS. Since the energy consumption is determined by the bias current rather than by the transconductance itself, we convert the reported gm value into the corresponding drain current using the transconductance efficiency, defined as gm/Id. Here, gm is the small-signal transconductance and Id is the drain bias current. Physically, gm/Id is primarily determined by the transistor inversion level, and it is commonly used in analog circuit design. For analog transistors biased in a current-efficient amplification regime, typically around moderate inversion, gm/Id is relatively insensitive to the specific CMOS technology node and commonly falls around 20 V^{-1}. Therefore, assuming gm/Id = 20 V^{-1}, the bias current of each column current source is:

$$144\ \mu S / 20\ V^{-1} = 7.2\ \mu A$$

Using the same 1.8 V supply voltage as in the physical tokenizer, the energy consumed by the 16 column current sources during the 4 μs operation window is:

$$7.2\ \mu A \times 1.8\ V \times 4\ \mu s \times 16 = 829.4\ pJ$$

Since the energy consumption of the CMOS image sensor array itself is negligible, we set this contribution to zero to make the comparison more favorable to the conventional baseline. The total energy of the CMOS image sensor array and column current sources is 829.4 pJ.

(3) **Amplifiers**. In the conventional architecture, the pixel values within a 16 × 16 patch must be read out before the digital dot-product operation can be performed. Therefore, to achieve the same throughput as the proposed physical tokenizer, the amplifier in the conventional readout path must process 16×16 pixel signals within the same 4 μs operation window. Accordingly, the available settling time for each pixel signal is:

$$4\ \mu s / (16 \times 16) = 15.625\ ns$$

To make the estimate favorable to the conventional baseline, we assume the amplifier operates as a unity-gain buffer. This is a highly conservative lower-bound assumption, because in a practical CMOS image-sensor readout path, the amplifier usually provides a voltage gain larger than one to condition the pixel signal and match the ADC input range. For a commonly used voltage-feedback amplifier, the closed-loop bandwidth is approximately given by GBW divided by the closed-loop gain. Therefore, assuming unity gain gives the maximum possible closed-loop bandwidth for a given GBW and thus minimizes the required amplifier GBW, bias current, and energy consumption. Even under this favorable unity-gain assumption, the amplifier must still provide sufficient settling speed within the 15.625 ns time window. Using a conservative five-

time-constant settling criterion, the required characteristic settling rate is estimated as:

$$5 / 15.625 \text{ ns} = 320 \text{ MHz}$$

To make the estimate favorable to the conventional baseline, we choose the TI OPA838 as a low-power reference amplifier. Its GBW of 300 MHz is close to, but slightly below, the estimated requirement of 320 MHz; therefore, using its quiescent current provides a lower-bound estimate of the amplifier energy in the conventional readout path. A strictly higher-GBW amplifier that fully satisfies this settling requirement would consume even more energy. When the supply voltage is set to 3.6 V, consistent with that used in the physical-tokenizer estimate, the OPA838 consumes a quiescent current of 0.96 mA. [quiescent current data and GBW data are adopted from https://www.ti.com.cn/product/OPA838]. Therefore, the energy consumed within the 4 μs operation window is:

$$4 \text{ μs} \times 0.96 \text{ mA} \times 3.6 \text{ V} = 13824.0 \text{ pJ}$$

It should be noted that the OPA838 is specified with a load capacitance of 6 pF. The amplifier only needs to drive the input capacitance of the following ADC, which is assumed to be 2 pF, consistent with the physical-tokenizer estimate. Under the same amplifier topology, phase margin, GBW, and transistor overdrive conditions, the required bias current can be approximately scaled with the load capacitance as a first-order estimate. Therefore, the amplifier energy is normalized by the load-capacitance ratio:

$$13824.0 \text{ pJ} \times 2 \text{ pF} / 6 \text{ pF} = 4608 \text{ pJ}$$

Thus, the amplifier energy consumption in the conventional readout path is estimated to be 4608 pJ.

(4) **ADC**. In the conventional architecture, the pixel values within a 16 × 16 patch must be digitized before the digital dot-product operation. Therefore, to achieve the same throughput as the proposed physical tokenizer, the ADC in the conventional readout path must complete 16 × 16 conversions within the same 4 μs operation window. In contrast, the proposed physical tokenizer only requires one ADC conversion after the in-sensor dot-product operation has already been completed.

To make the comparison consistent with the experimentally demonstrated 41 optical states in our physical tokenizer, we set the ADC precision of the conventional CMOS image sensor to the same effective input precision. Since $2^{5.36}=41$, the 41-level optical input corresponds to an effective ADC resolution of 5.36 bits. Here, the fractional bit number is used as an information-equivalent resolution for energy estimation, which further favors the conventional baseline compared with rounding up to a practical 6-bit

ADC.

According to Supplementary Fig. 22, the energy required for one 5.36-bit ADC conversion is estimated to be 1 pJ. Therefore, the ADC energy consumption in the conventional architecture is:

$$16 \times 16 \times 1\ \text{pJ} = 256\ \text{pJ}.$$

(5) **Data movement**. The energy consumption of data movement is evaluated using custom-designed circuits in the Cadence Virtuoso design environment with the foundry-provided CMOS PDK. In practical chip-level systems, signal transfer between the chip and external circuits is typically driven by thick-oxide I/O devices. To make the comparison conservative and favorable to the conventional baseline, we simplify the I/O interface as an inverter-based output driver, as shown in Supplementary Fig. 23.

In this simplified model, one PAD introduces a capacitance of approximately 2 pF. Since data transfer occurs between the CMOS image sensor chip and the digital dot-product chip, the total load capacitance is estimated to be 4 pF. For a 5.36-bit image-sensor readout, transferring the raw pixel values of a 16 × 16 patch requires an equivalent of 16×16×5.36 bit-level transfers. The data-movement energy (E_{mv}) is obtained by integrating the supply current of the inverter driver over the 4 μs operation window and multiplying it by the supply voltage:

$$\mathrm{E_{mv}} = VDD \int_{t_0}^{t_0+4\mu s} I_{VDD}(t)dt$$

As shown in Supplementary Fig. 23, the simulated energy consumption for transferring the 5.36-bit raw-pixel data of a 16 × 16 patch is 38450 pJ.

In the above simulation, the inverter driver is toggled for the equivalent 16×16×5.36 bit-level transfers, corresponding to the worst-case full-switching condition. For practical digital data transfer, however, not every transmitted bit changes its value between consecutive transfers. Assuming random transferred bits, the average switching activity is approximately 50%. Therefore, to avoid overestimating the data-movement energy of the conventional architecture, we further apply a switching-activity factor of 0.5 to the simulated full-switching energy. The effective data-movement energy is thus estimated as:

$$38450\ \text{pJ} \times 0.5 = 19225\ \text{pJ}$$

(6) **Digital dot-product chip**. The energy consumption of the digital dot-product engine is estimated based on the NVIDIA H100 GPU, which is implemented using the TSMC 4N process. According to the reported specifications, the H100 provides an

INT8 Tensor Core throughput of 3341 TOPS under a power consumption of 350 W. Therefore, the energy consumption of one multiply-accumulate operation is estimated as:

$$350\ \text{W} / 3341\ \text{TOPS} = 0.105\ \text{pJ}$$

To generate one element of the patch embedding for a 16 × 16 patch, the digital dot-product engine needs to perform 16 × 16 multiply-accumulate operations. Therefore, its energy consumption is:

$$16 \times 16 \times 350 / 3341 = 26.82\ \text{pJ}.$$

Since this contribution is relatively small compared with the other circuit blocks, we neglect it to make the comparison more favorable to the conventional baseline. Therefore, the total energy consumption for the conventional architecture is:

$$0 + 829.4 + 4608 + 256 + 19225 + 0 = 24918.4\ \text{pJ}$$

The conventional architecture consumes:

$$24918.4\ \text{pJ} / 1743.9\ \text{pJ} = 14.3$$

times the energy of the proposed physical tokenizer.

Conventional architecture with 256-level CIS readout precision. We have experimentally expanded the input precision of the physical tokenizer to an 8-bit optical-input condition, corresponding to 256 optically distinguishable states, as shown in Supplementary Fig. 24. This input precision is comparable to the commonly used 8-bit, 256-level CIS readout baseline. Therefore, we further performed an energy analysis for the conventional architecture under the same 256-level input-precision condition.

Under this condition, the ADC energy and data-movement energy of the conventional architecture increase due to the higher readout precision, while the energy consumption of the other circuit blocks remains unchanged.

ADC. Under the 256-level CIS readout condition, the ADC precision in the conventional architecture needs to be increased to 8 bits. According to Supplementary Fig. 22, the energy required for one 8-bit ADC conversion is 6.24 pJ. Since the conventional architecture must digitize all 16×16 pixel values within one patch, the total ADC energy is:

$$16 \times 16 \times 6.24\ \text{pJ} = 1597.44\ \text{pJ}$$

Data movement. The data-movement energy is evaluated using the same inverter-driver-based I/O simulation setup as that used for the 5.36-bit CIS readout case. The difference is that, under the 8-bit image-sensor readout condition, transferring the raw

pixel values of a 16 × 16 patch requires 16×16×8 bit-level transfers. As shown in Supplementary Fig. 25, the simulated I/O data-movement energy is 57440 pJ under the full-switching condition. To make the estimate more favorable to the conventional baseline, we further assume a 50% bit-level switching activity, meaning that only half of the transferred bits contribute to dynamic I/O switching energy on average. Therefore, the effective data-movement energy is:

$$57440\ \text{pJ} \times 0.5 = 28720\ \text{pJ}$$

Thus, the total energy consumption of the conventional architecture with 256-level CIS readout precision becomes:

$$0 + 829.4 + 4608 + 1597.44 + 28720 + 0 = 35754.84\ \text{pJ}$$

Therefore, under the 256-level input-precision condition, the conventional architecture consumes:

$$35754.84\ \text{pJ} / 1743.9\ \text{pJ} = 20.5$$

times the energy of the proposed physical-tokenizer architecture.

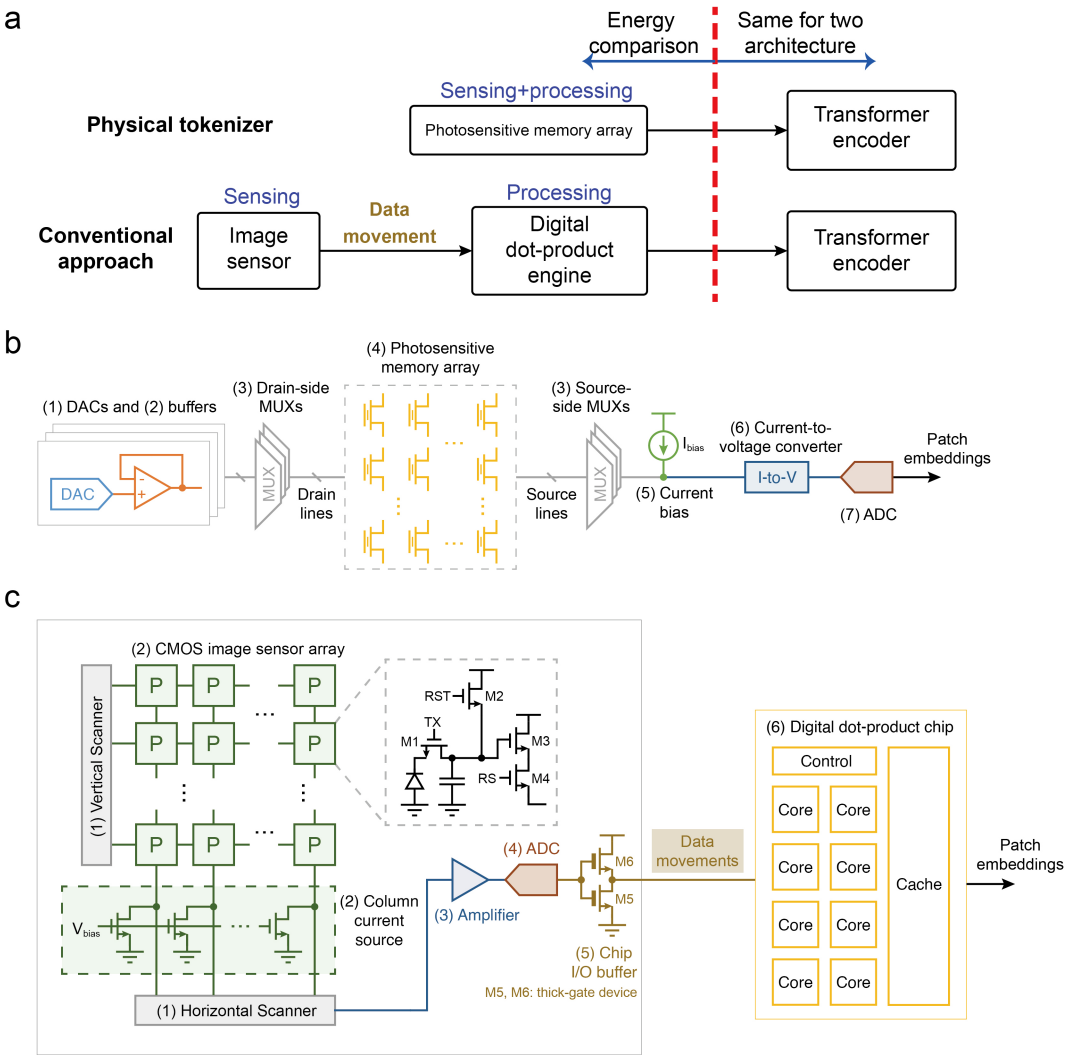
Three key factors contribute to the lower energy consumption of the physical tokenizer:

1) Zero data movement. In the photosensitive memory array, computation is performed at the sensor level, eliminating the need for data movement. In contrast, in the conventional architecture, data movement between the sensor and processing units consumes a substantial amount of energy.

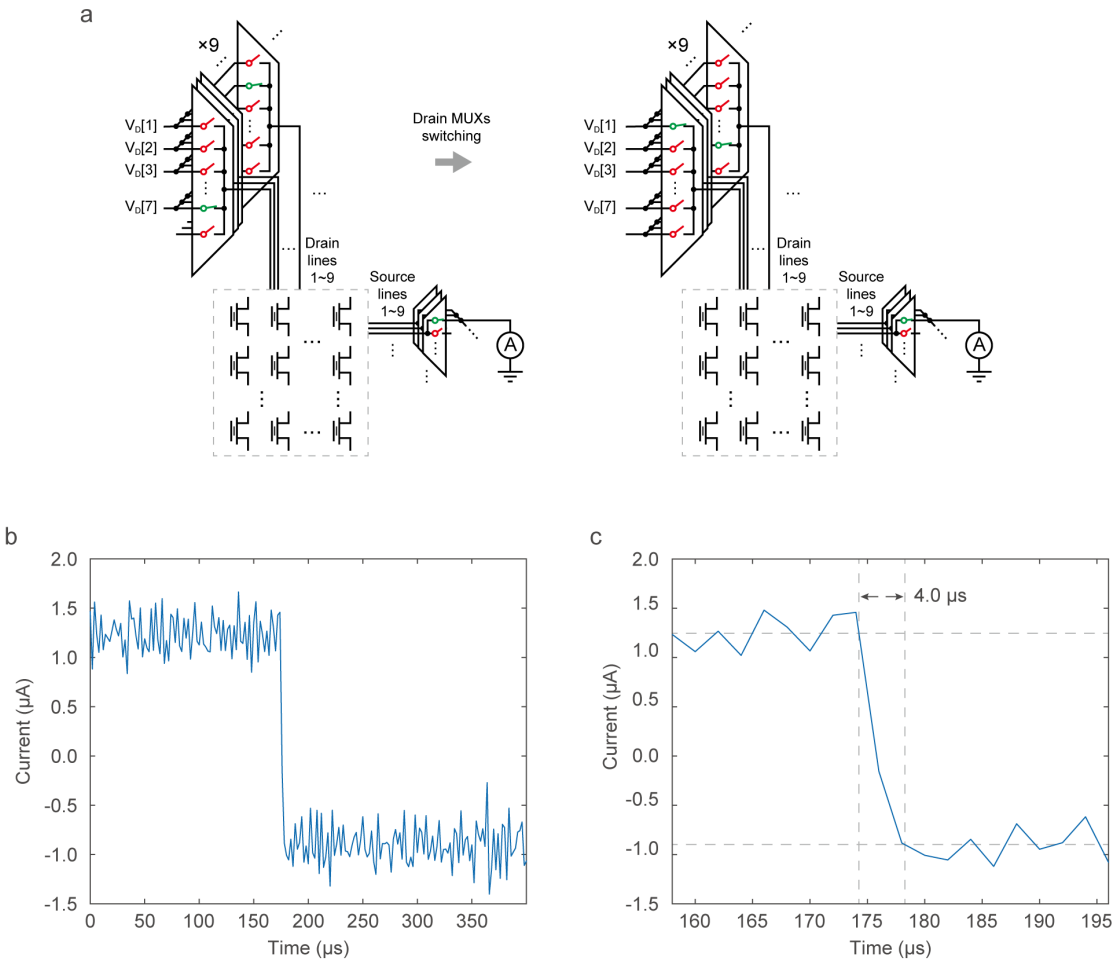
2) Parallel in-sensor dot product reduces the speed requirement for the amplifier and ADC. In the physical tokenizer, the dot product is completed inside the photosensitive memory array. Therefore, the I-to-V converter and ADC only need to process one final dot-product result for each patch-embedding element, and the number of high-speed analog readout operations scales as $O(1)$ with respect to the patch size. In contrast, the conventional architecture must sequentially read out and digitize all pixels in an N×N patch before performing the digital dot product. Therefore, the number of high-speed amplifier and ADC operations scales as $O(N^2)$. For a 16 × 16 patch, this corresponds to a 256-fold higher speed requirement for the amplifier and ADC in the conventional architecture, leading to substantially higher energy consumption.

3) Operation sequence optimization minimizes the operating speed of high-power analog circuits. Analog components that are used in large quantities, such as DACs and drain buffers, consume significant energy when operating at a rate of 1/(4 μs). By optimizing the output order of patch embeddings, we lower the operation speed of these analog components and thereby reduce their energy consumption. The frequent switching is instead shifted to the low-power digital circuits such as drain MUXs, where

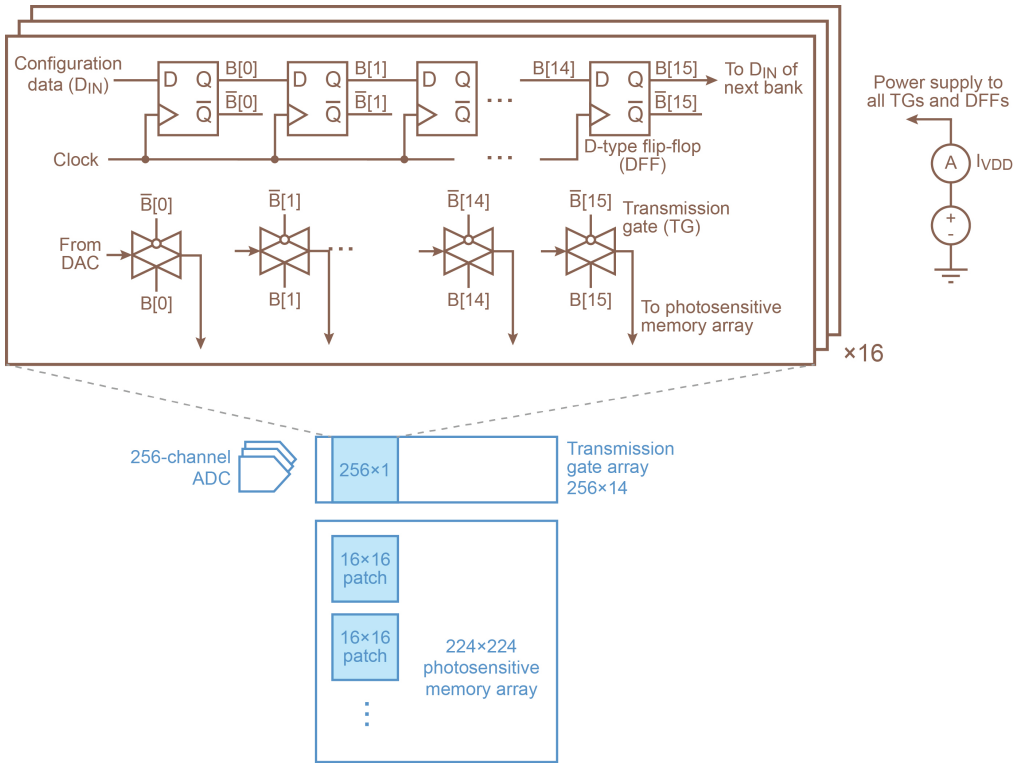
the additional power overhead is minimal. As a result, the overall energy consumption of the system remains low.



Supplementary Fig. 18. Energy consumption comparison in two architectures. a, Schematic of the energy comparison boundary (red dashed line) between the physical tokenizer and conventional approaches. **b,** Schematic of physical tokenizer. **c,** Schematic of conventional CMOS image sensor and Von Neumann-based architecture.



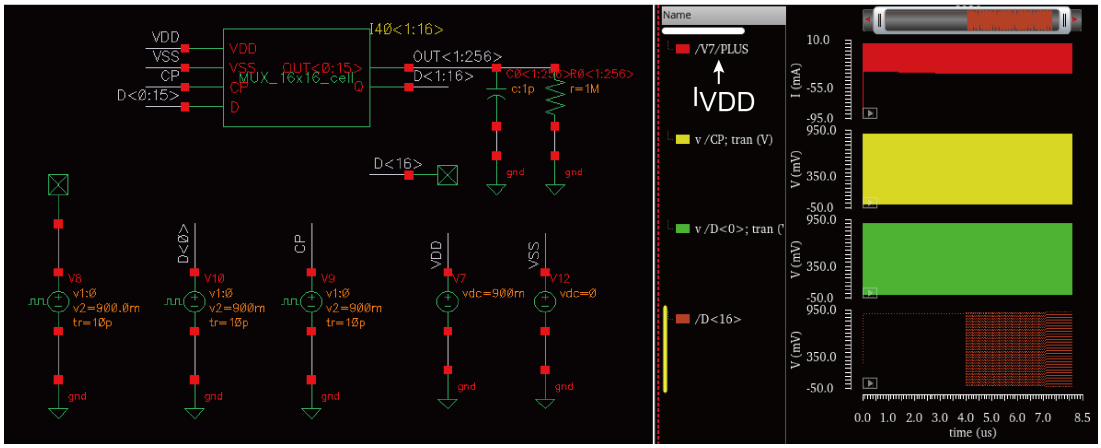
Supplementary Fig. 19. Measured settling time after drain MUX switching. a. Schematic of drain MUX switching in physical tokenizer. **b.** Measured output current over time after drain MUX is switched. **c.** A zoomed-in view of the transition region, indicating the setup time of 4 μs.



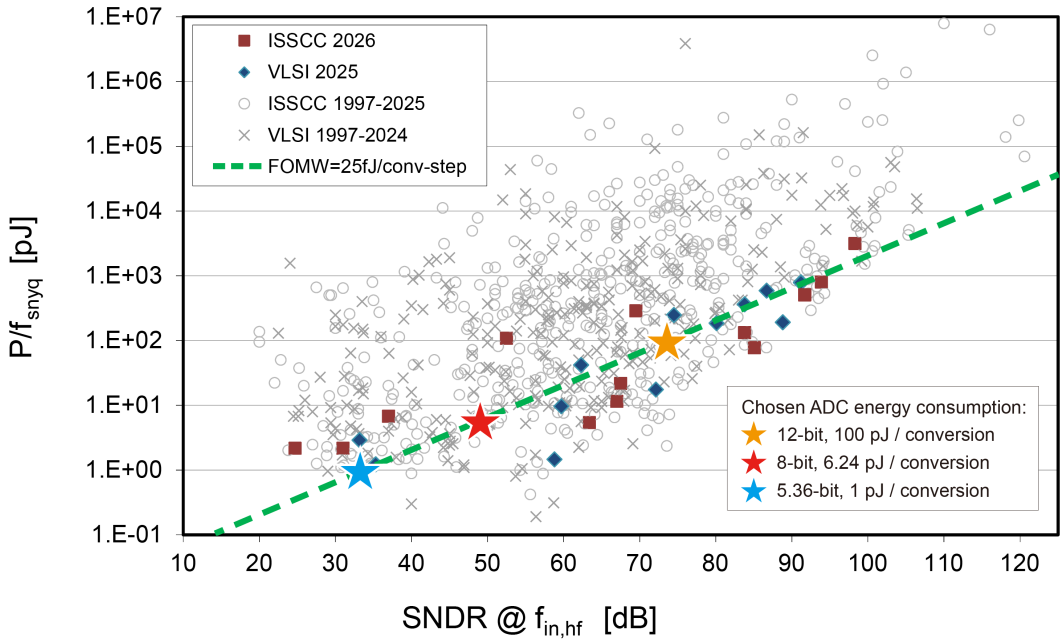
Supplementary Fig. 20. Schematic for the energy-consumption simulation of the drain MUXs.

$$E_{drMUX} = VDD \int_{t_0}^{t_0+4\mu s} I_{VDD}(t)dt$$

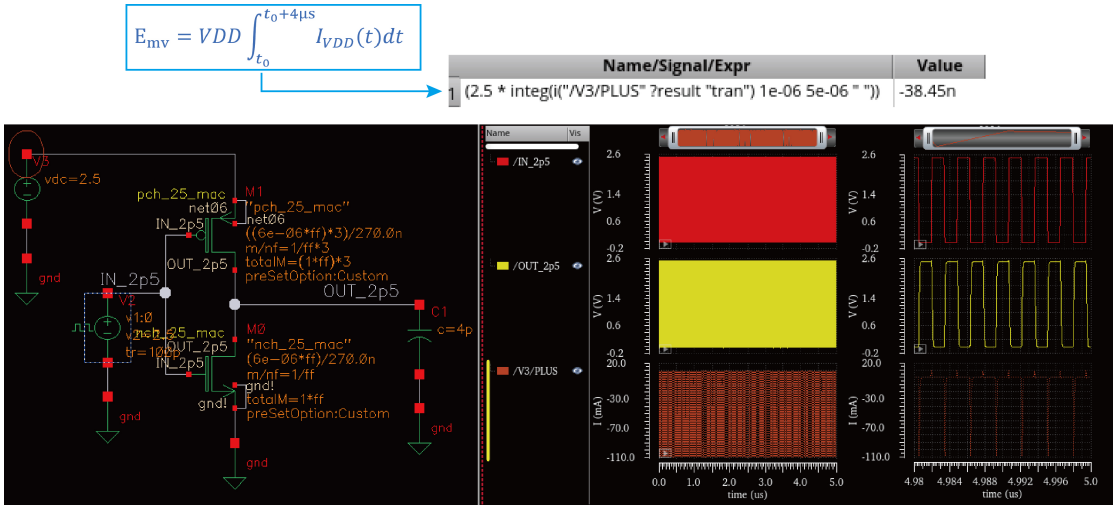
	Name/Signal/Expr	Value
1	/V7/PLUS	wave
2	(integ(i("/V7/PLUS" ?result "tran") 4e-06 8e-06 " ") * 0.9)	-308.9p



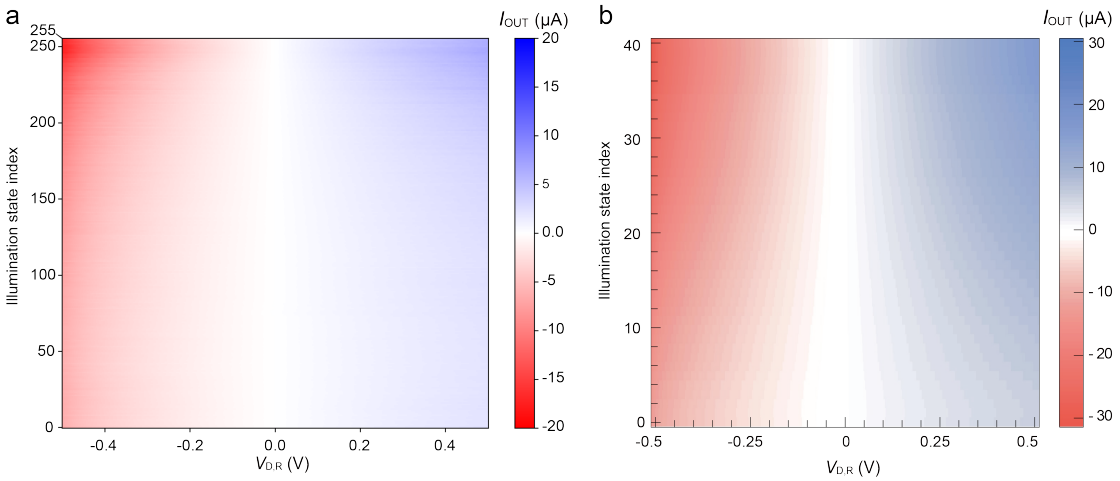
Supplementary Fig. 21. Simulation setup and results for evaluating the energy consumption of the drain MUXs.



Supplementary Fig. 22. Basis for ADC energy consumption. The figure summarizes the energy consumed per conversion by ADCs with different precisions reported at the IEEE International Solid-State Circuits Conference (ISSCC) and the IEEE Symposium on VLSI Technology and Circuits (VLSI) from 1997 to 2026. The green dashed line indicates representative ADC performance from the latest ISSCC/VLSI reporting cycle. Based on this trend, the typical energy consumption per conversion used in our analysis is 1 pJ for a 5.36-bit ADC, 6.24 pJ for an 8-bit ADC, and 100 pJ for a 12-bit ADC. [This figure is adopted from B. Murmann, *ADC Performance Survey*, Online available: https://github.com/bmurmann/ADC-survey.]



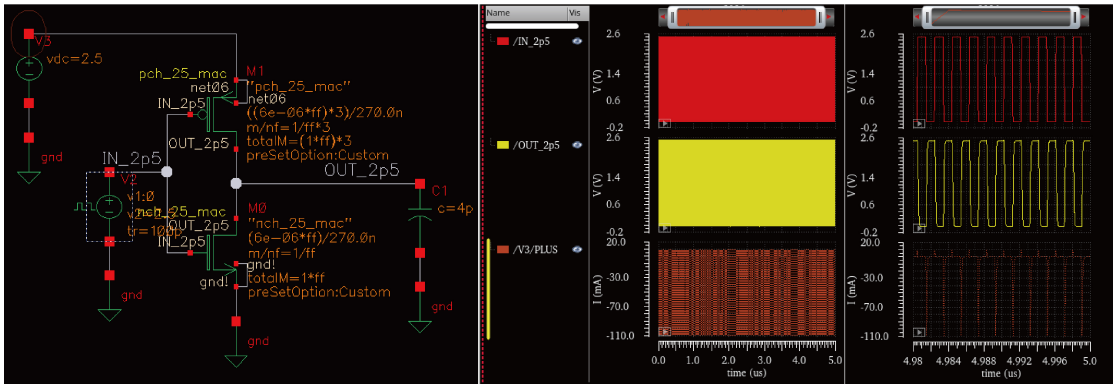
Supplementary Fig. 23. Simulation setup and results for evaluating the I/O data-movement energy required to transfer the 5.36-bit raw-pixel data of a 16 × 16 patch.



Supplementary Fig. 24. Extension of the experimentally measured mapping from 41 to 256 optical input states. a, measured light-$V_{D,R}$-I_{OUT} mapping using 256 optical input states. **b,** measured light-$V_{D,R}$-I_{OUT} mapping using 41 optical input states, as used in Fig. 4b. The programming conditions were kept the same in both measurements. For each programmed optical input state, the $V_{D,R}$ was swept from -0.5 to 0.5 V using 1001 voltage points, and the I_{OUT} was recorded. The denser sampling in **a** shows that the experimentally characterized optical input resolution can be extended from 41 levels to 256 levels, providing a higher-resolution lookup table for physical tokenization.

$$E_{mv} = VDD \int_{t_0}^{t_0+4\mu s} I_{VDD}(t)dt$$

	Name/Signal/Expr	Value
1	(2.5 * integ(i("/V3/PLUS" ?result "tran") 1e-06 5e-06 " "))	-57.44n



Supplementary Fig. 25. Simulation setup and results for evaluating the I/O data-movement energy required to transfer the 8-bit raw-pixel data of a 16 × 16 patch.

Supplementary Note 6. Scalability and Engineering Considerations.

Here, we outline practical engineering considerations in the present physical tokenizer prototype and highlight optimization opportunities in future scaling. Monolayer MoS_2 in this context offers attractive attributes, including analog-tunability, high photosensitivity, potential for large-area integration, and compatibility with heterogeneous/monolithic 3D integration. These characteristics motivate a practical pathway to advance the demonstrated physical tokenizer concept toward more engineering-ready vision systems.

Speed. Speed is a key factor for real-time edge deployment. Here we discuss what currently limits the readout/settling speed in our prototype, and the headroom for faster programming/refresh in future implementations.

Readout speed. The ~4 μs setup time reported for the current prototype (Supplementary Fig. 19) is mainly limited by an effective Resistor-Capacitance settling process. The dominant capacitive contributions include parasitics from board-level routing and wiring, bonding pads, MUX input/output capacitances, and cabling. The corresponding resistive components include the finite on-resistance of the MUX switches and the on-state resistance of the MoS_2 transistors. Together, these elements form a low-pass network that determines the settling behavior of the readout path. In future implementations, board- and cable-related parasitics and large bonding-pad capacitances can be substantially reduced. Meanwhile, MUX parasitics can be further minimized through tighter integration and shorter interconnects. Therefore, there is significant headroom to shorten the setup time in scaled implementations.

Programming speed. In the present proof-of-concept experiments, the effective write latency is mainly bounded by the timing granularity and coordination overhead of the optical/electrical programming setup rather than by an intrinsic device-speed limit. Specifically, optoelectronic characterizations rely on LabVIEW-coordinated optical shutter/SMU control, and the tokenization demonstrations coordinate the DMD projection and FPGA-driven biases through Python. These instrument/control loops typically operate at millisecond-scale timing resolution, which is relatively slower than the switching dynamics of floating-gate devices. Recent reports have demonstrated ultrafast operation of 2D-material floating-gate transistor arrays[3], suggesting substantial headroom for faster programming and refresh when the programming drivers, timing control, and peripheral circuitry are co-optimized in a tightly integrated hardware platform.

Precision. We consider two practical precision aspects relevant to scaled photosensitive memory arrays: programming precision, which governs how accurately the incident image can be written into the array as nonvolatile pixel states, and computation

precision, which governs the accuracy of the analog dot-product during patch embedding. These issues tend to become more pronounced at the array level due to device variations and array/interconnect non-idealities.

Programming precision. The effective input-image precision is determined by the number of reliably distinguishable and retainable nonvolatile states under array operation. In the present prototype, we demonstrate 41 programmable states with good retention, which is sufficient to validate the physical tokenizer concept. For scaled arrays, improving programming precision is closely tied to pixel-to-pixel isolation and the suppression of array-level leakage during write and hold operations. From an engineering perspective, introducing an access transistor per pixel (e.g., a 1T-1FG) can strengthen isolation and improve state controllability, enabling finer and more reliable image programming.

Computation precision. Patch-embedding accuracy depends on operating the devices within a defined near-linear readout window, such that the current is approximately proportional to the applied read bias under fixed read conditions. We adopt Sb contacts to obtain near-ohmic behavior and explicitly select a small-signal readout window, validating linear and symmetric I-V characteristics within this range (Supplementary Fig. 6d). For scaled implementations, computation precision can be further impacted by device/interconnect variations and line-resistance-induced voltage drops, motivating additional improvements via device engineering (e.g., contact optimization), circuit-level calibration, and more compact integration.

Array scaling and peripheral overhead. Recent progress in wafer-scale 2D semiconductor growth and wafer-level integration with improving yield and device uniformity[3] suggests a realistic route to scale physical tokenizer beyond the current prototype. Key considerations for future scaling include lowering defect density and improving grain-boundary control to reduce device-to-device variability, and achieving wafer-scale reproducibility for dielectric deposition and low-resistance contacts without degrading the ultra-thin 2D channel.

At the same time, vertical integration offers a promising pathway for manufacturable scaling. With increasingly CMOS-compatible process constraints for 2D materials and demonstrated heterogeneous[4] and monolithic 3D integration[5], a vertically integrated layout can reduce interconnect lengths and enable more compact placement of peripheral circuits beneath the sensor array. Continued maturation of wafer-level process modules, including transfer and alignment, contamination control, and BEOL-compatible thermal budgets, is expected to further improve scalability.

To assess the practical impact of scaling physical tokenizer to a typical setting (224×224 input with 16×16 patch), we examine the interconnect IR drop and the associated peripheral-area overhead.

IR drop. To bound the IR drop induced bias error in a scaled physical tokenizer, we set a conservative voltage-drop budget of < 1 mV along the routing path. This budget is motivated by an 8-bit analog drive range of ~0.3 V (one least significant bit ≈ 0.3 V/255 ≈ 1.18 mV). We therefore allocate 0.5 mV to the source-side line and 0.5 mV to the drain-side line, and assume a typical per-cell read current of ~1 μA.

For the source-side line, we consider a worst-case simultaneous-activation condition within a 16×16 patch. Specifically, assuming each source line is shared by the 16 cells along one patch, the maximum current on a source line is ~16 μA. The corresponding resistance constraint is 31.25 Ω (0.5 mV/16 μA). With representative 65-nm CMOS routing, a 12-μm-wide single-layer line yields ~27.8 Ω, satisfying this requirement. Using multiple metal layers in parallel further relaxes the required width proportionally while maintaining a similar effective line resistance.

For the drain-side line, the typical current per line during readout is ~1 μA, giving a much looser requirement of 500 Ω (0.5 mV/1 μA), which can be met with extremely modest routing width (e.g., ~125 nm), well below the 5 μm reserved width. Therefore, under the considered operating conditions, IR drop is not expected to be a dominant limitation when scaling physical tokenizer to a 224×224 configuration.

Peripheral overhead. We estimate the peripheral area overhead for scaling to a 224×224 array with 16×16 patch using representative 65-nm CMOS technology[6,7] under conservative assumptions.

For the drain driver, we conservatively assume 256 independent 8-bit DACs for a 16×16 patch. The DAC[6] capacitor area is 3000 μm^2, the DAC output buffer is assumed to be slightly larger than the capacitor and is set to 4000 μm^2, and control logic is estimated as 1000 μm^2 per channel. This gives (3000+4000+1000)×256≈2.05 mm^2. In addition, we include MUXs cost of 1250 μm^2 per line across 224 lines, i.e., 1250×224≈0.28 mm^2. Therefore, the drain-driver area is approximately 2.33 mm^2 .

For the source-side readout, we use a representative ADC[7] macro area of 37200 μm^2, plus an I-V and bias circuitry of 1000 μm^2, and MUXs of 1250 μm^2 per line across 224 lines (≈0.28 mm^2). This yields a source-side area of ≈0.037+0.001+0.28≈0.32 mm^2.

The total estimated peripheral area is therefore ≈2.33+0.32≈2.65 mm^2. For comparison, the photosensitive memory array area under a 25-μm pixel pitch is $(224\times25\ \mu m)^2$≈31.36 mm^2. Thus, under these representative assumptions, the periphery is not expected to dominate chip area for a 224×224 physical tokenizer implementation.

Reference

1 Lopez-Sanchez, O., Lembke, D., Kayci, M., Radenovic, A. & Kis, A. Ultrasensitive photodetectors based on monolayer MoS2. *Nat. Nanotechnol.* **8**, 497-501, doi:10.1038/nnano.2013.100 (2013).

2 Li, W. *et al.* Approaching the quantum limit in two-dimensional semiconductor contacts. *Nature* **613**, 274-279, doi:10.1038/s41586-022-05431-4 (2023).

3 Liu, C. *et al.* A full-featured 2D flash chip enabled by system integration. *Nature* **646**, 1081-1088, doi:10.1038/s41586-025-09621-8 (2025).

4 Zhu, K. *et al.* Hybrid 2D-CMOS microchips for memristive applications. *Nature* **618**, 57-62, doi:10.1038/s41586-023-05973-1 (2023).

5 Lu, D. *et al.* Monolithic three-dimensional tier-by-tier integration via van der Waals lamination. *Nature* **630**, 340-345, doi:10.1038/s41586-024-07406-z (2024).

6 Chen, L. *et al.* A 0.7-V 0.6-μW 100-kS/s low-power SAR ADC with statistical estimation-based noise reduction. *IEEE J. Solid-State Circuit* **52**, 1388-1398, doi:10.1109/JSSC.2017.2656138 (2017).

7 Lee, S., Kang, H. & Lee, M. in *2024 IEEE International Solid-State Circuits Conference (ISSCC).* 184-186.