SCIENCE CHINA
Information Sciences

• Supplementary File •

An optoelectronic artificial spiking neuron array with biomimetic spike-temporal pattern for in-sensor visual prediction

Rui Wang[1,2†], Guolei Liu[2†], Dingwei Li[2†], Xiaotao Jing[1], Fanfan Li[2], Zhixian Wu[3], Zhongfang Zhang[2], Huihui Ren[2], Saisai Wang[3], Qi Huang[3], Xiaohua Ma[1], Bowen Zhu[2,3,4*], Min Qiu[2,3], Hong Wang[1*] & Yue Hao[1]

[1]*State Key Laboratory of Wide Band Gap Semiconductor Devices and Integrated Technology, School of Microelectronics, Xidian University, Xi'an 710071, China*
[2]*Key Laboratory of 3D Micro/Nano Fabrication and Characterization of Zhejiang Province, School of Engineering, Westlake University, Hangzhou 310024, China*
[3]*Westlake Institute for Optoelectronics, Hangzhou 311421, China.*
[4]*Westlake Institute for Advanced Study, Hangzhou 310024, China.*

* Corresponding author (email: zhubowen@westlake.edu.cn, hongwang@xidian.edu.cn)
† These authors contributed equally to this work.

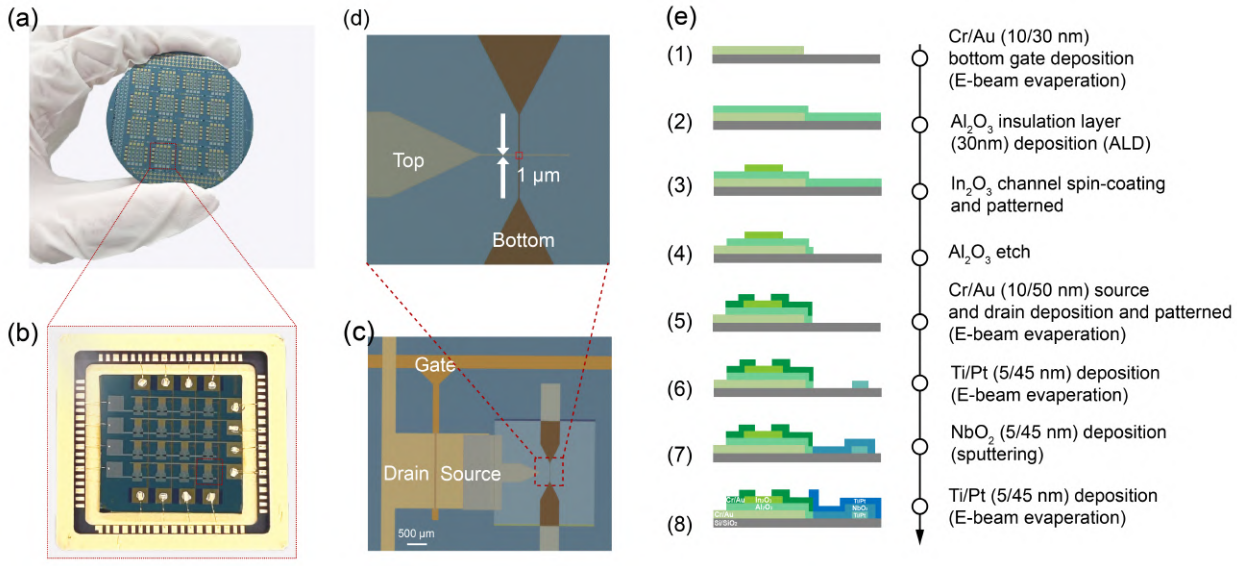

Figure S1 (a) Photograph of 2-inch wafer-scale OAN arrays. (b) QFN72 packaged OAN array. (c-d) Optical image of In_2O_3 synaptic transistor and NbO_x threshold switching device. (e) Fabrication process flow of the array.

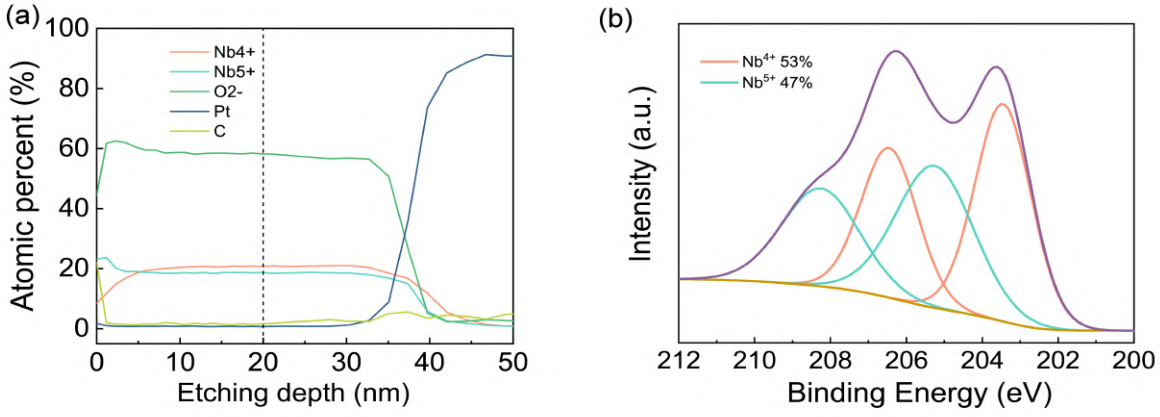

Figure S2 (a) Depth-dependent concentration of various atomic in pristine NbO_x film. (b) XPS peak of the Nb^{4+} and Nb^{5+} at an etching depth of 20 nm for the pristine NbO_x film.

To assess variations in the oxidation state of pristine NbO_x thin films at varying depths, we conducted in-situ X-ray photoelectron spectroscopy (XPS) depth profiling employing Ar^+ sputtering. Through XPS depth profiles, the atomic content of Nb^{4+}, Nb^{5+}, O^{2-}, Pt, and C can be analyzed as a function of etching depth on the Pt electrode. The relative ratios of these elements at different etching depths are illustrated in Figure S2a. The pristine NbO_x film exhibits a uniform distribution of Nb^{4+} and Nb^{5+} throughout its depth, except for the surface region (at an etch depth of 5 nm), which is exposed to air and undergoes oxidation. Additionally, to quantify the oxidation states of Nb, detailed XPS spectra of Nb 3d corresponding to the dashed lines in Figure S2a were extracted (Figure S2b). At an etching depth of 20 nm, Nb^{4+} exhibits a higher concentration than Nb^{5+}. These findings indicate that the Nb:O ratio in the entire film remains at 1:2.2.

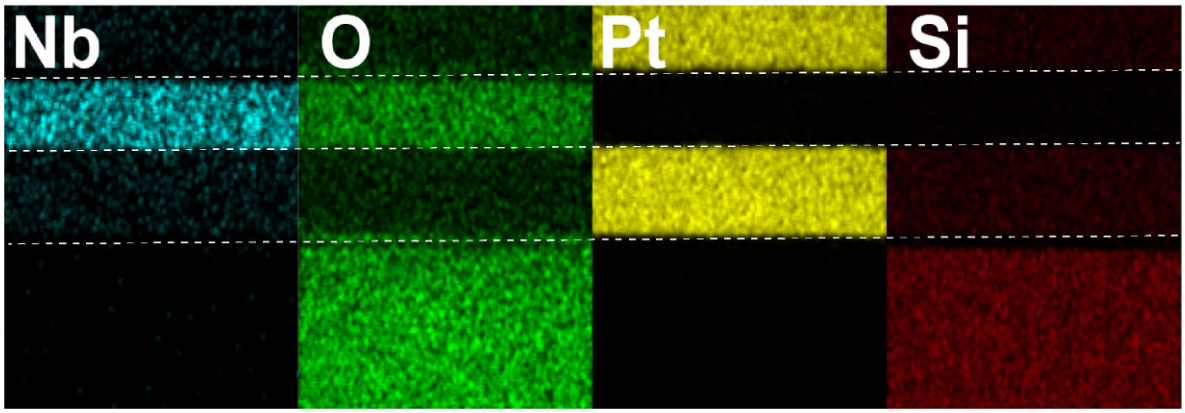

Figure S3 EDS mapping of Nb, O, Pt, and Si elements in the device.

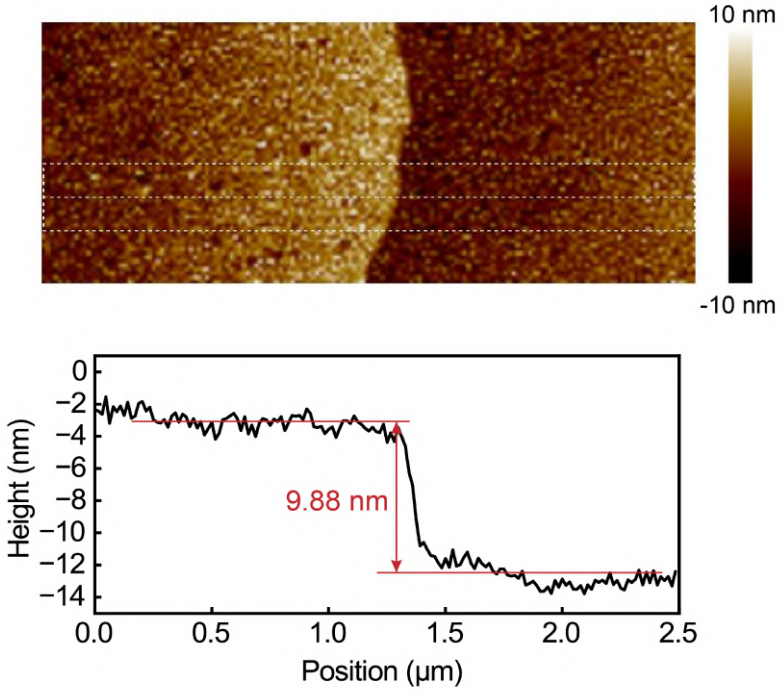

Figure S4 AFM images and height profiles of the channel region of In_2O_3 film.

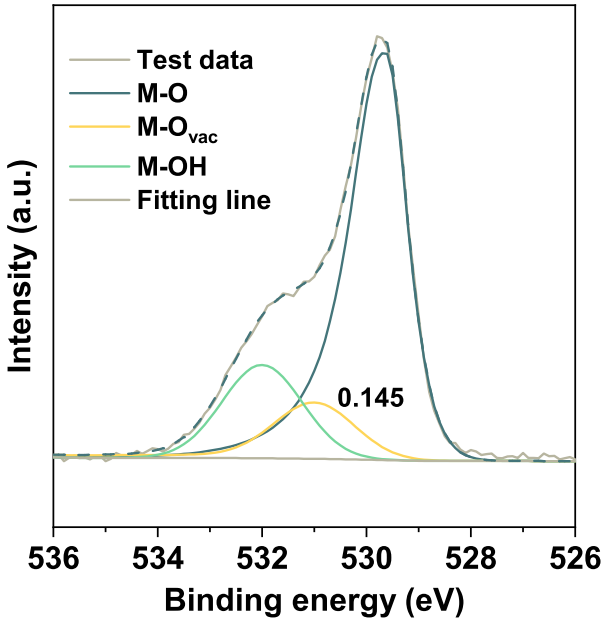

Figure S5 XPS characterization of fabricated In_2O_3 thin films.

The results show that the O 1s peak of the In_2O_3 film is divided into three small peaks centered at 529.6 ± 0.1 eV, 530.9 ± 0.1 eV, and 531.8 ± 0.1 eV. According to the peak separation results, the blue line centered at 529.6 eV ± 0.1 eV represents the metal-oxygen bond (M-O), the yellow line centered at 530.9 ± 0.1 eV represents the peak of oxygen vacancies (M-O_{vac}), and the green line centered at 531.8 ± 0.1 eV represents the peak of residual hydroxide in the film or impurity water adsorbed on the surface (M-OH). It is well known that as the annealing temperature increases, the oxygen vacancy defect increases, the number of carriers increases, and the resistance decreases. When the In_2O_3 film is annealed at 300°C for 1 hour, the ratio of oxygen vacancy reaches 0.145, and the resistivity matches NbO_x well.

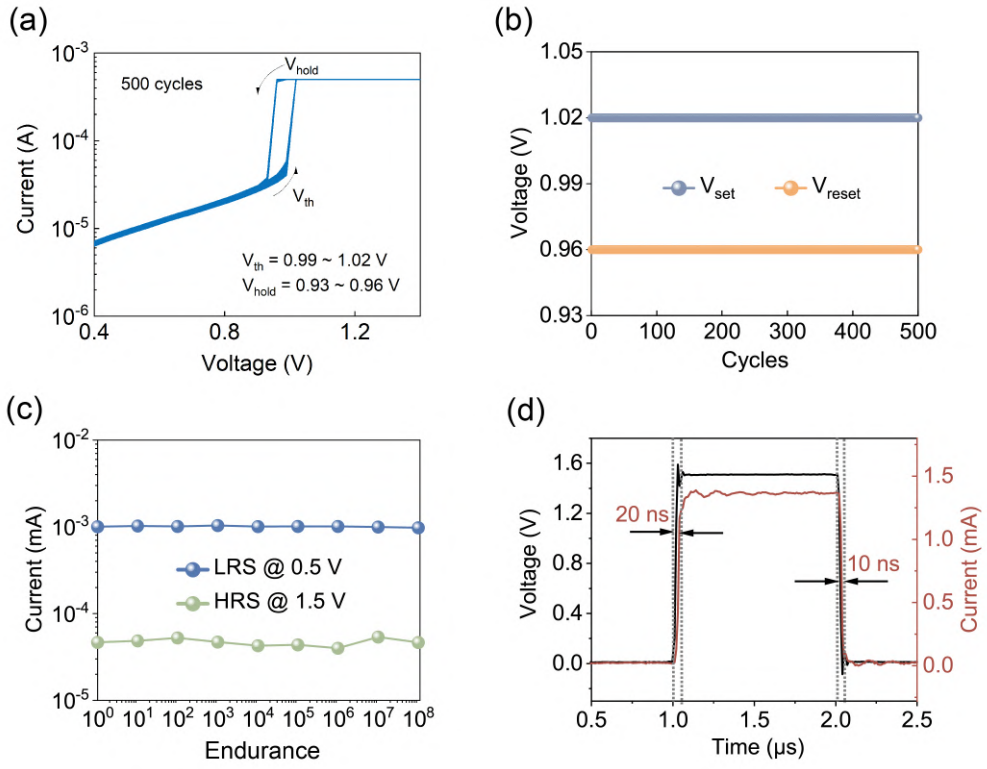

Figure S6 (a) I-V behavior of the NbO_x TS device repeated for 500 cycles. (b) Distributions of V_{th} and V_{hold} values of the NbO_x device in 500 repeated cycles, showing excellent stability. (c) The endurance of the NbO_x device for 10^8 cycles. (d) Real-time switching response triggered by an input pulse. Response time from on- to off-state and from off- to on-state is, respectively, 20 ns and 10 ns.

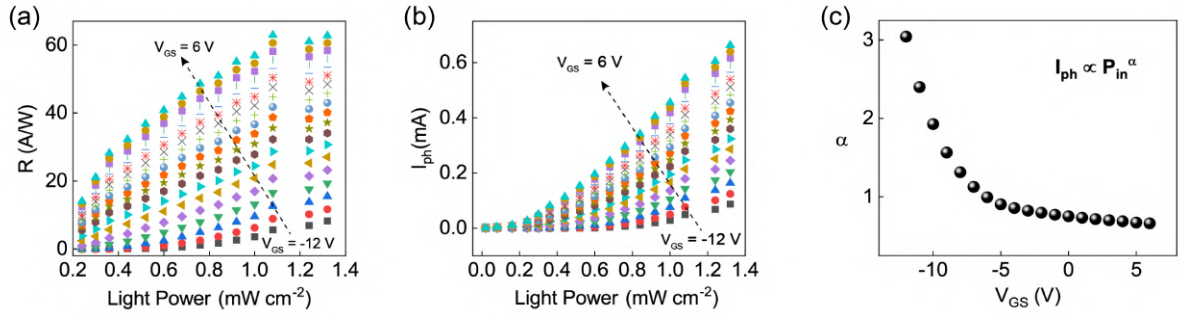

Figure S7 (a) Relationship between P_{in} and I_{ph} under different gate voltages. (b) Extracted α as a function of the V_{GS}. (c) Extracted photoresponsivity R ($R = I_{ph}/P_{in}$) as a function of V_{GS} and P_{in}.

We quantitatively evaluate the photoresponsivity (R) of the device by $I_{ph}/(P_{in} \times A)$, and the active area of the device is considered to be equal to the channel area (A = channel width (W) $\times$ channel length (L)). The W/L of the In_2O_3 synaptic transistor is 10/400 μm. As a result, Figure S7a shows a high photoresponsivity (60 A/W) with UV power (P_{in}) at a 365 nm wavelength. The I_{ph} is defined as $I_{ph} = I_{illumination} - I_{dark}$, where Iillumination and Idark are I_D under light illumination and dark at drain-source voltage (V_{DS}) of 2 V (Figure S7b). Figure S7c shows the relationship between P_{in} and I_{ph} ($I_{ph} \propto P_{in}^{\alpha}$); the α is the exponent factor. As V_{GS} varies from -12 to 6 V, the relationship between P_{in} and I_{ph} changes from nonlinear $\alpha = 3$ to $\alpha = 0.5$. The non-linearity response to light facilitates neural network implementation. [1]

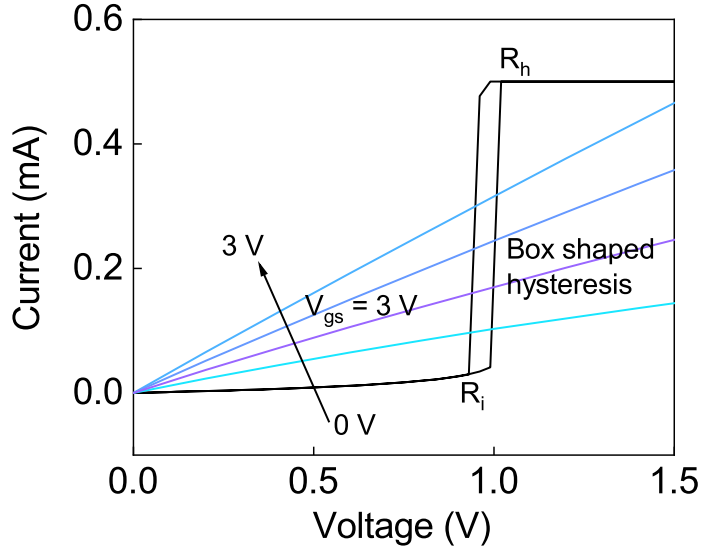

Figure S8 The current-voltage (I-V) characteristics of the NbO_x TS and the output characteristics of the In_2O_3 PT.

In addition, to ensure the successful integration of neurons and synapses, synaptic transistor selection is critical because the current levels between the memristor and the synaptic transistor must be consistent to achieve OAN design. Figure S8 shows the I-V characteristics of the NbO_x TS and the output characteristics of the In_2O_3 PT. Sustained spikes can be achieved in the range of $R_m < R_{ch} < R_i$, where R_{ch} represents the channel resistance of the In_2O_3 transistor, and R_i and R_m represent the insulation resistance and metal resistance of NbO_x, respectively. Because when the voltage reaches the threshold of NbO_x, R_i is converted to R_h, the voltage divider of NbO_x is reduced, and NbO_x is converted from R_h to R_i again. The continuous state switching completes the neuron oscillation characteristics. As shown in Figure S8, when V_{gs} is 3 V and V_{ds} is 0-3 V, the current levels of the two devices match each other.

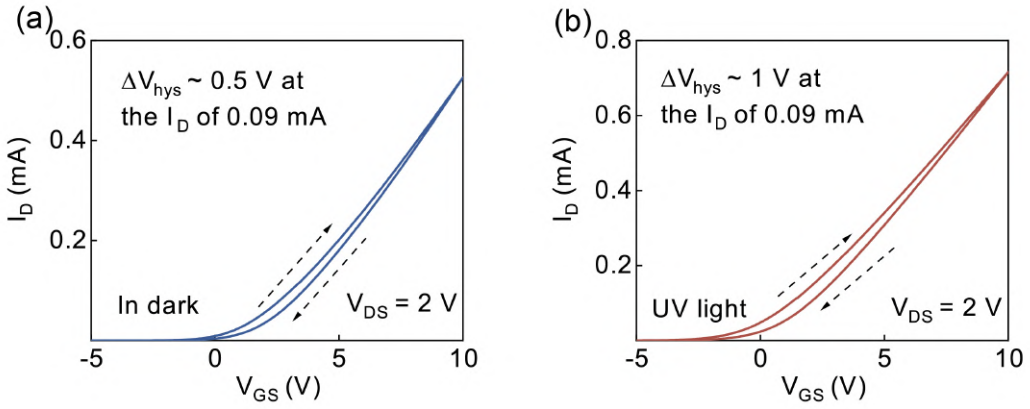

Figure S9 Transfer curves of In_2O_3 synaptic transistors under darkness (a) and UV irradiation (b). Under UV irradiation, the device exhibits negative V_{TH} and an increased voltage hysteresis window (ΔV_{hys}).

Figure S9a-b illustrates the transfer curves (I-V) under darkness and UV irradiation. Analysis of these results reveals a notable negative threshold voltage (V_{TH}) and an expanded voltage hysteresis window in the I-V curve. This phenomenon is attributable to an increase in the concentration of ionized oxygen vacancies and the density of charge trapping. [2] Such behavior engenders a persistent photoconductive effect, facilitating the generation of a memory spike current within the sensing array. Furthermore, there is a certain degree of coupling between the visual information captured at time T and $T+\tau$. The observed clockwise hysteresis loop corroborates the occurrence of charge-trapping and de-trapping processes. The hysteresis voltage window (ΔV_{hys}) is defined as the difference in the gate voltage (V_{GS}), yielding $\Delta V_{hys} = \sim$1 V at the I_D of 0.09 mA. The trap charge density (N_t) is estimated to be approximately $1.7\times10^{12} cm^{-2}$, calculated as $N_t = (\Delta V_{hys} \times C_{ox})/q$, where C_{ox} represents the oxide capacitance between the channel and the local bottom gate, measured at $170 \times 10^{-9} F/cm^2$, and q denotes the electron charge.

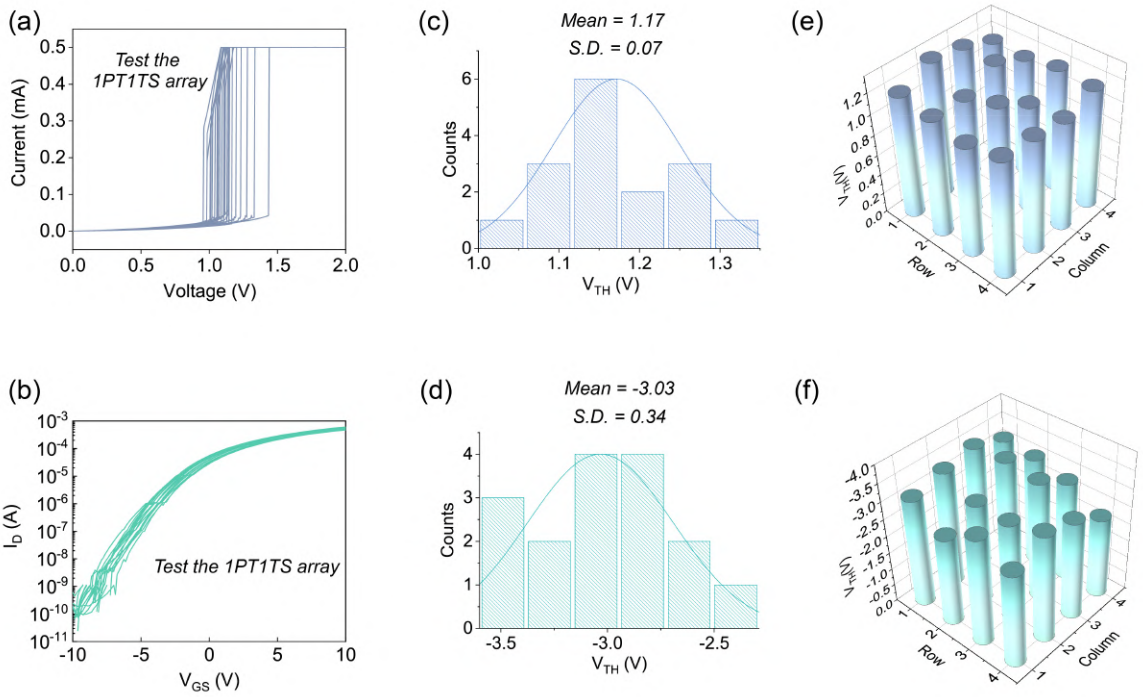

Figure S10 (a, b) I-V characteristics of 16 NbO_x and In_2O_3 devices in the array, respectively. (c, d) The statistical analysis of V_{TH} in 16 NbO_x and In_2O_3 devices, respectively. (e, f) The V_{TH} distribution of NbO_x and In_2O_3 devices in the array, respectively. All 16 devices exhibit robust performance.

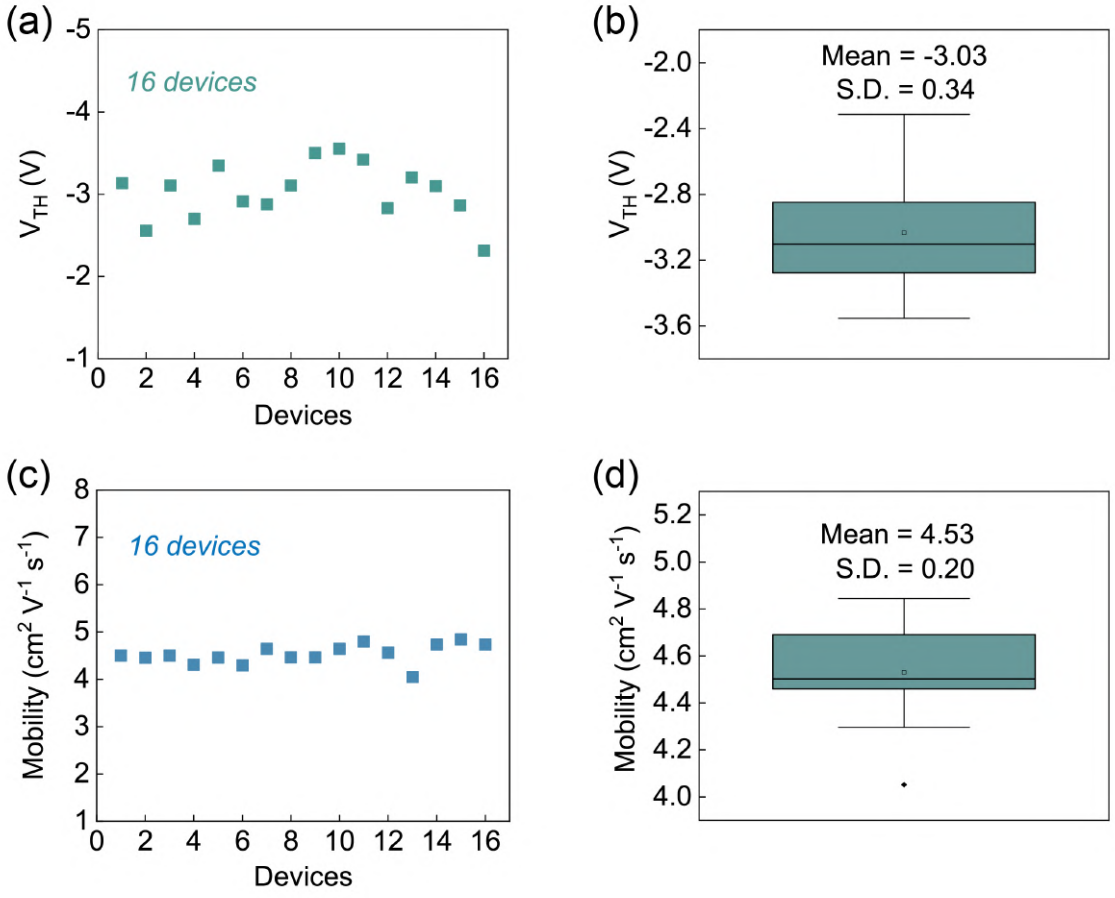

Figure S11 (a-b) Threshold voltage of the 16 In_2O_3 phototransistors with a mean of -3.04 V and standard deviations of 0.34 V, respectively. (c-d) Mobility of the 16 In_2O_3 phototransistors with a mean of 4.53 $cm^2V^{-1}s^{-1}$ and standard deviations of 0.20 $cm^2V^{-1}s^{-1}$, respectively.

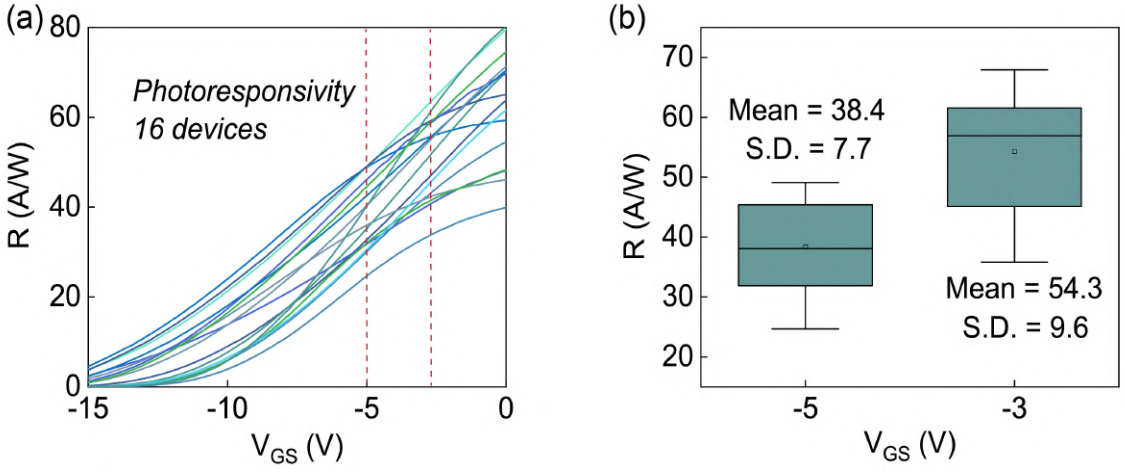

Figure S12 (a) Photoresponsivity (R) of 16 In_2O_3 phototransistors as a function of V_{GS} under light power of 1.32 $mWcm^{-2}$. (b) Statistical analysis of R when V_{GS} is -5 and -3 V (corresponding to the red line in Figure S12a).

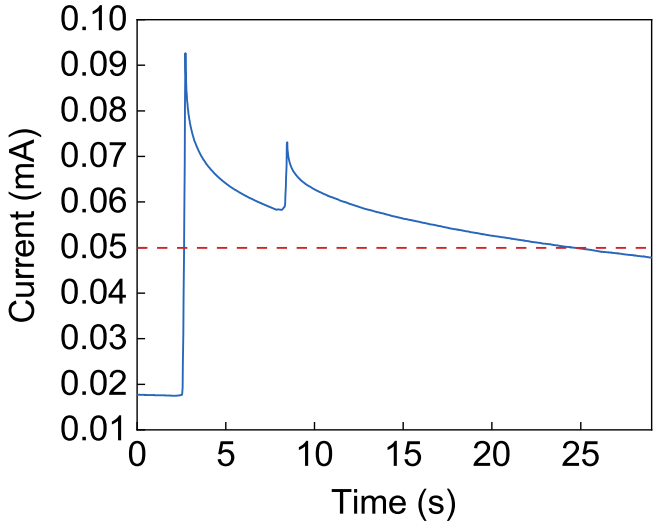

Figure S13 Persistent photoconductivity effect of the In_2O_3 transistor. The red line represents the boundary condition of neuron oscillation when V_{in} is 2 V.

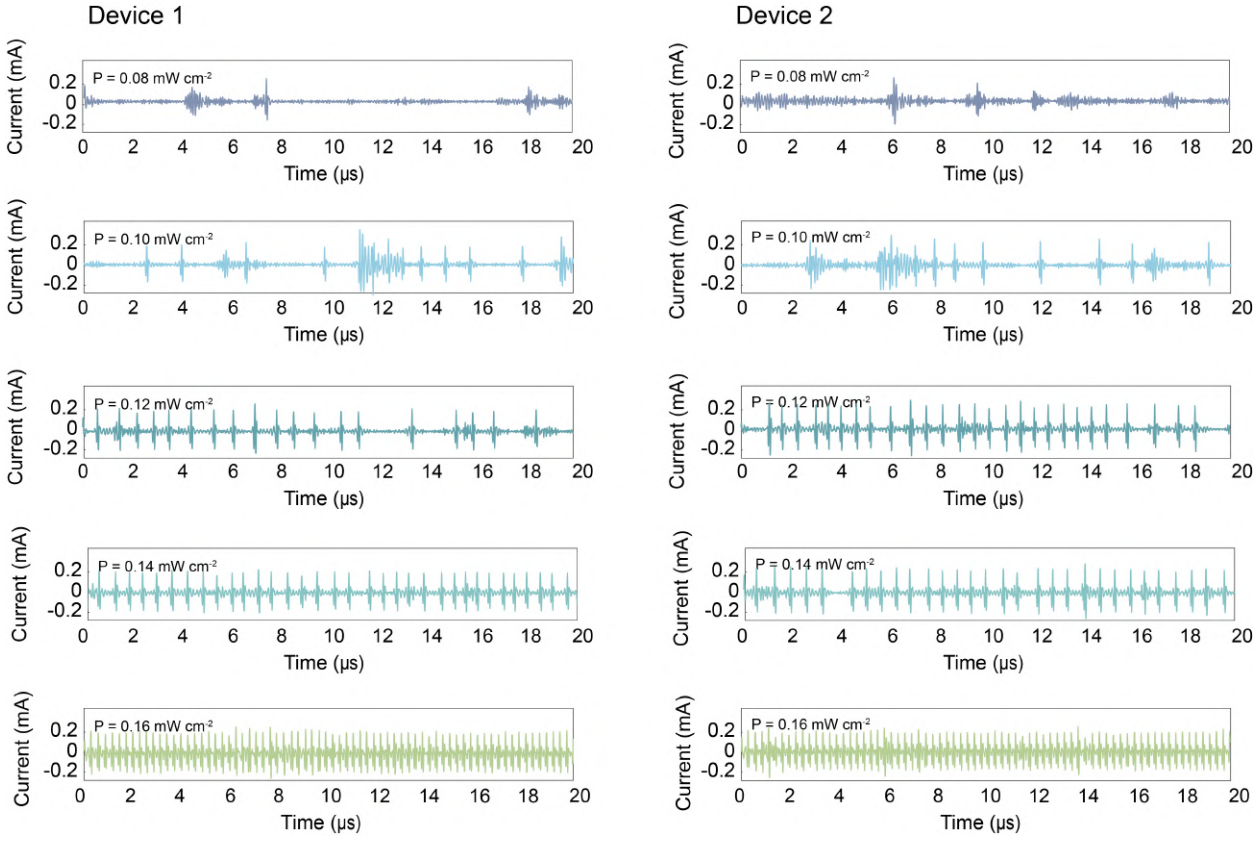

Figure S14 The current spiking response with different light powers.

Figure S15 The three distinct devices exhibit varying degrees of spike memory behavior in response to different optical powers.

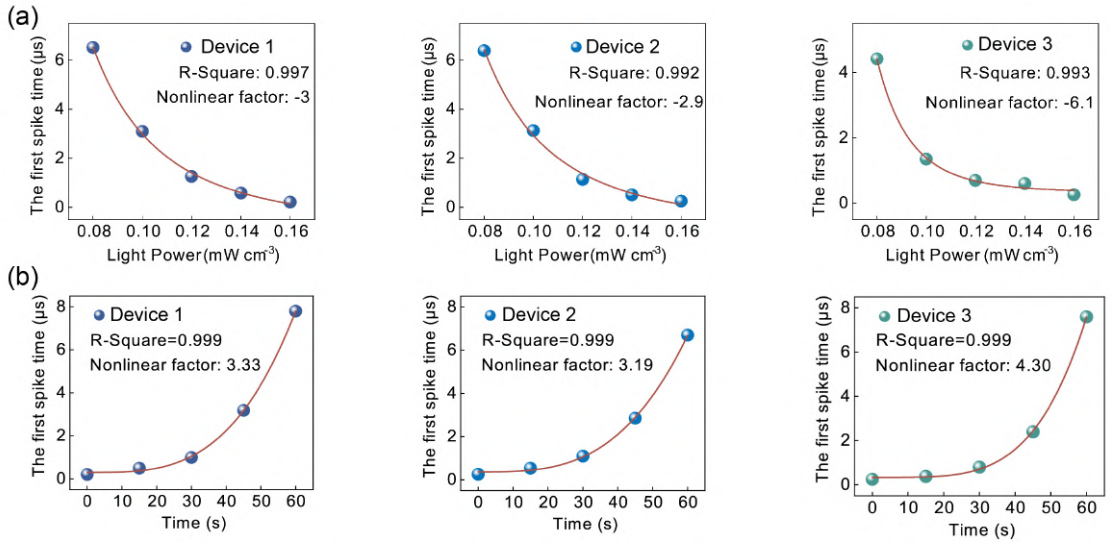

Figure S16 The nonlinearity sensing (a) and memory (b) analysis.

The nonlinearity of light response and storage of the array can be characterized by the formula $T = a + b * p^c$, where a and b are constants, c represents the nonlinear factor, p denotes optical power, and T signifies the first-spike time. This nonlinearity facilitates the RSTI equation to achieve in-sensor prediction.

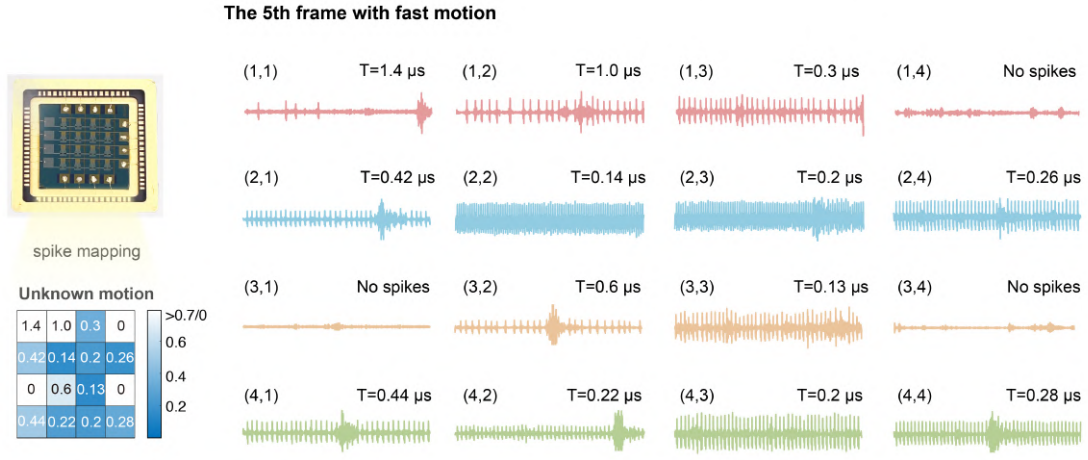

Figure S17 The 5th frame spiking current with fast motion.

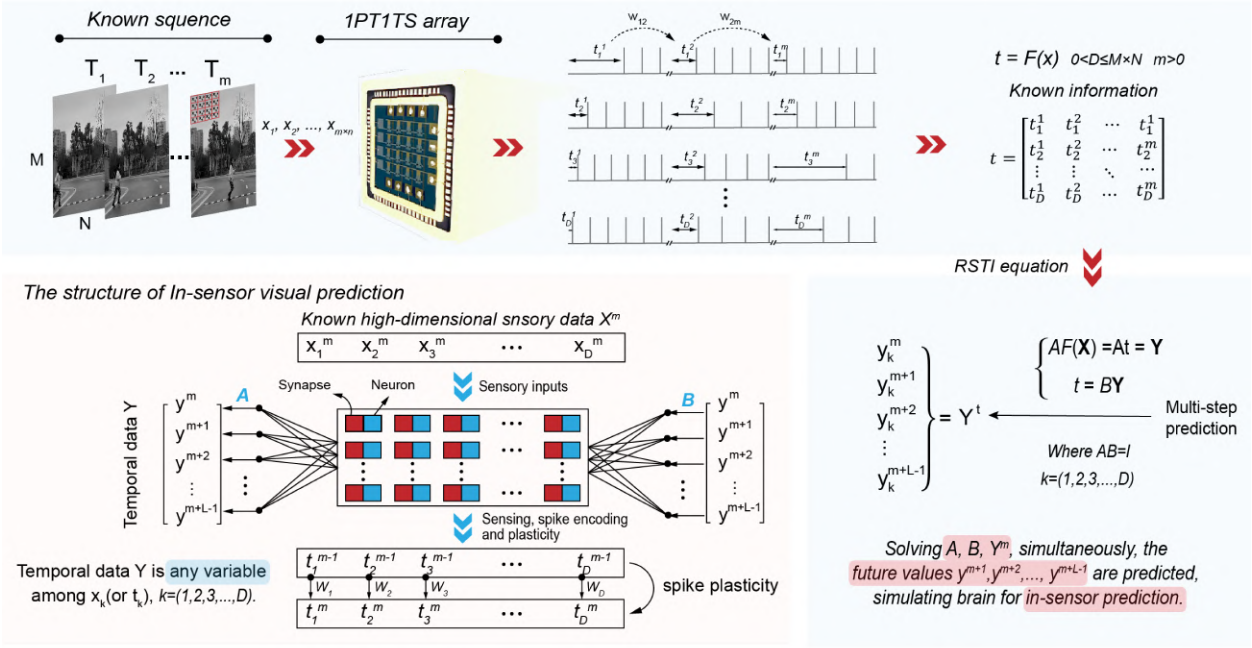

Figure S18 The theory of in-sensor prediction.

The OAN array with a 1PT1TS structure detects action trajectories, $X^m = (x_1^m, x_2^m, \cdots, x_D^m)$, with D and m variables and performs nonlinear spike encoding and memory behavior to obtain $T^m = (T^m + \gamma T^{m-\tau}) = (t_1^m, t_2^m, \cdots, t_D^m)$, where $T = F(X)$ forms a RSTI equation. F represents the nonlinear operation of the array. Among them, γ represents the memory factor, coupling the response at time $m - \tau$ to time m to express the coupling weight w. Subsequently, utilizing the delay-embedding theory, we construct a delayed vector, $Y^m = (y^{m+1}, y^{m+2}, \cdots, y^{m+L-1})$, for any target variable to be predicted $(e.g., Y^m = t_k^m = x_k^m, k = 1, 2, \cdots, D)$. Both the primary and conjugate forms (two weight matrices A and B) are concurrently solved to forecast future information of the target variable Y, particularly for limited time series, facilitating the transformation of information flows from $X \to F(X) \to T \to Y$.

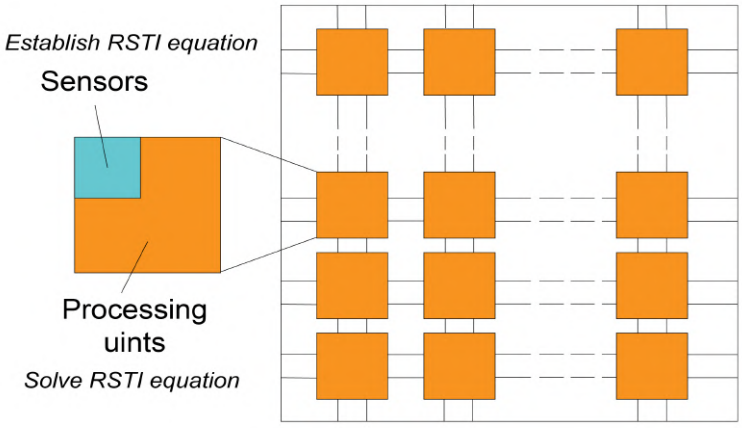

Figure S19 In-sensor prediction architecture. Image sensing and processing are fused in the sensor itself.

We experimentally verified that the RSTI equation established based on OAN only requires 4 frames and simple equation solutions to complete future predictions. As shown in Figure S19, the sensing unit is used to sense and establish the RSTI spatiotemporal equation, and the processing unit is used to solve the RSTI equation in real time and output future prediction results.

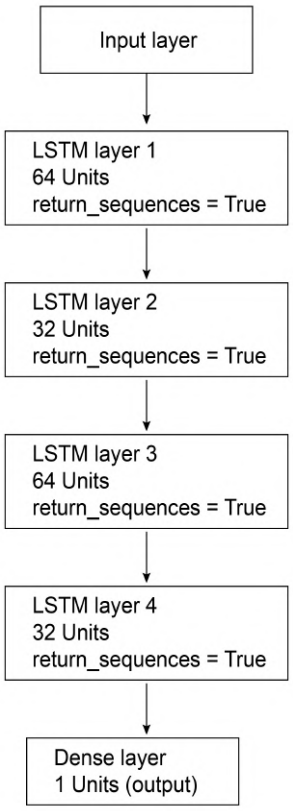

Figure S20 The diagram of the LSTM structure.

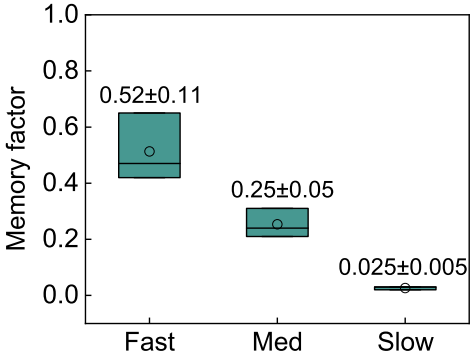

Figure S21 Memory factors at different speeds.

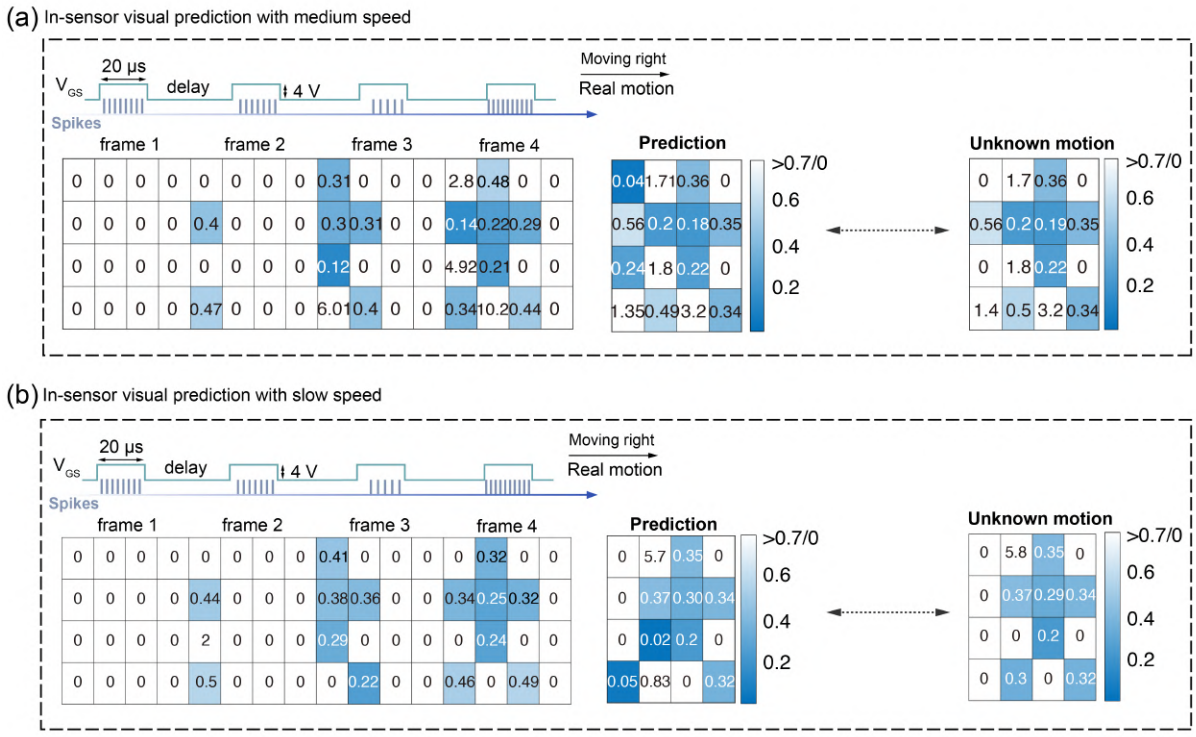

Figure S22 The array senses 16 motion features with delays of 30 s (a, medium speed) and 60 s (b, slow speed), respectively. Digital "0" means no spike is generated.

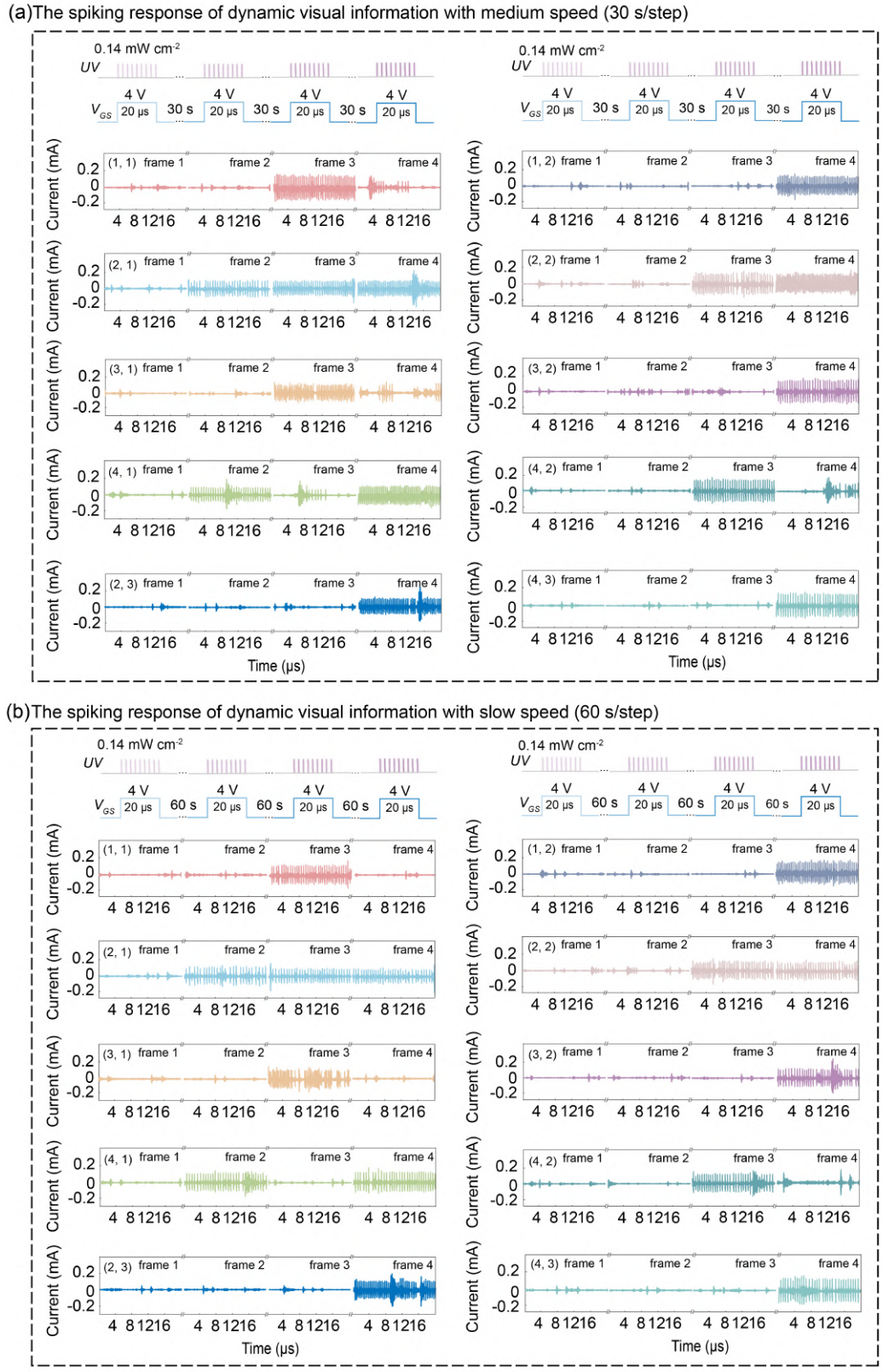

Figure S23 The spiking response to 4 × 4 motion targets with medium (a) and slow (b) speed.

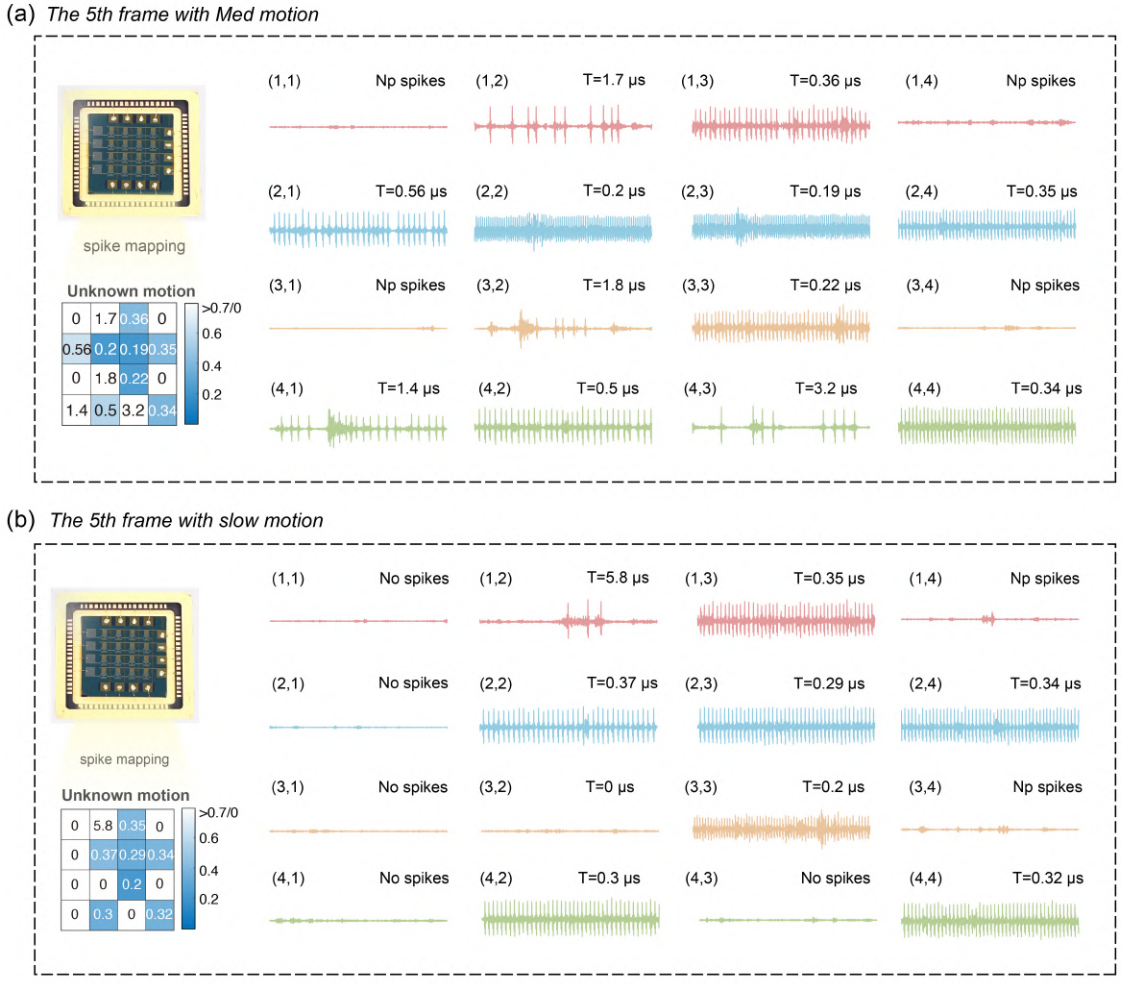

Figure S24 The 5th frame spiking currents with medium (a) and slow (b) speeds were tested on the array.

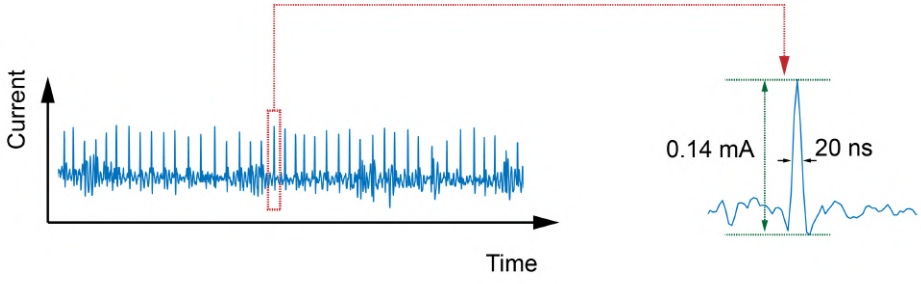

Figure S25 The spike current generated from light information.

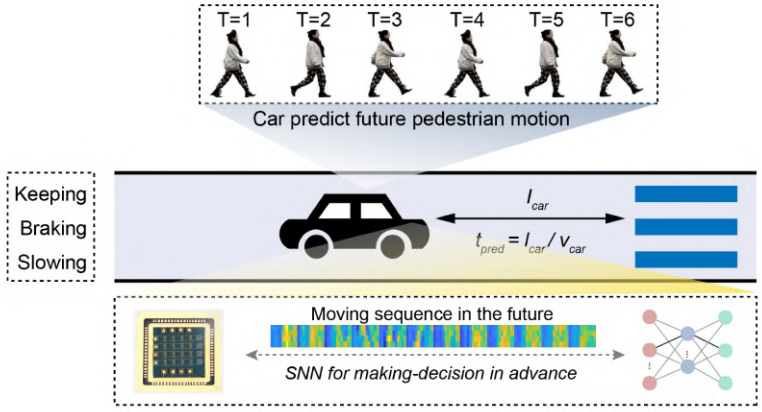

Figure S26 The schematic diagram of pedestrian obstacle avoidance.

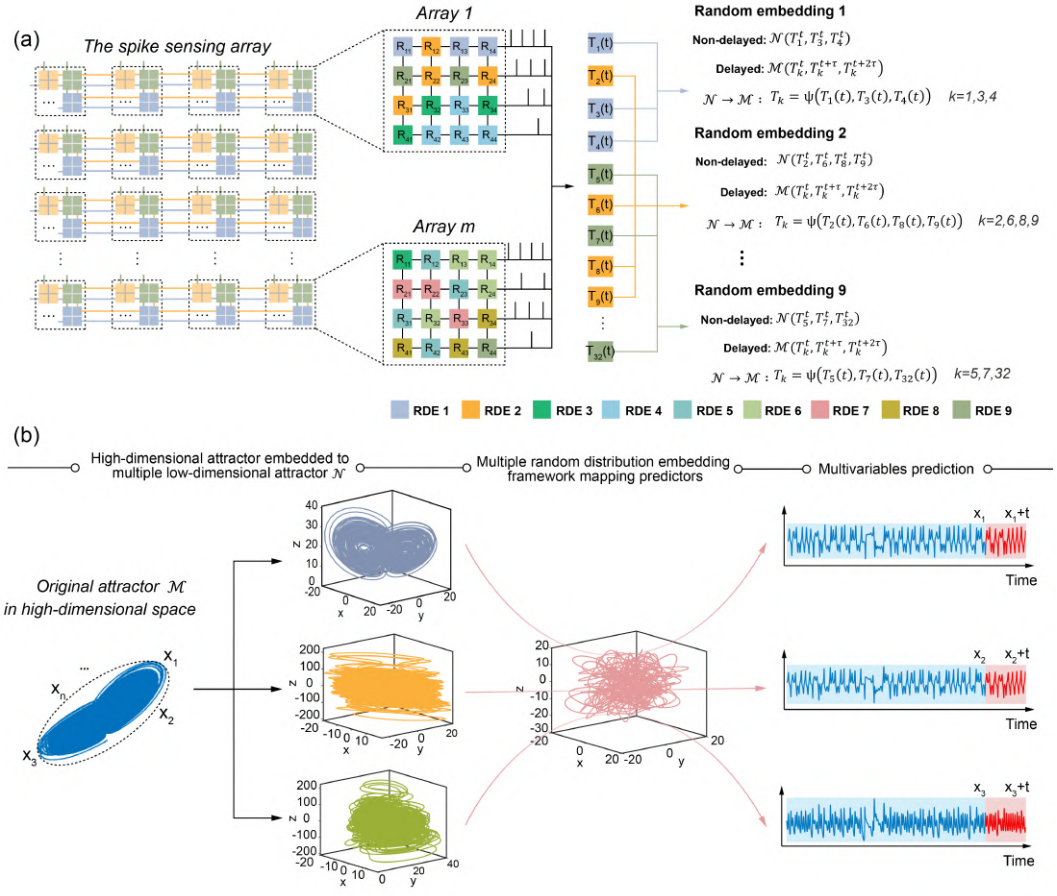

Figure S27 (a) Schematic diagram of deploying random distribution embedding (RDE) framework in the array for parallel prediction. (b) The general principle of the RDE framework.

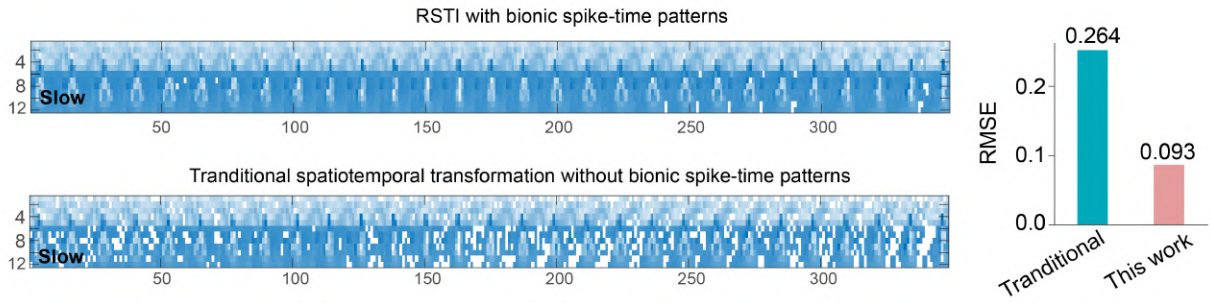

Figure S28 The comparison of RSTI with bionic spike-time patterns and traditional STI equation without bionic spike-time patterns.

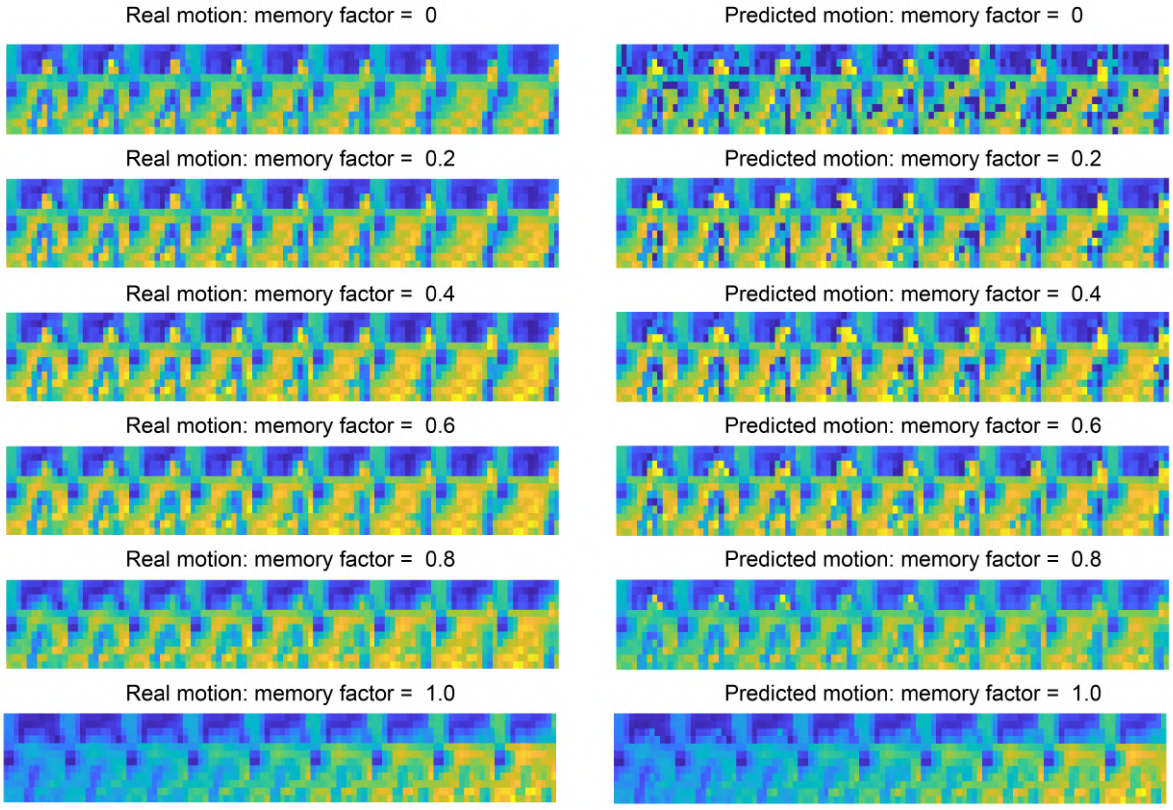

Figure S29 The motion characteristics and prediction results of 9 frames under different memory factors.

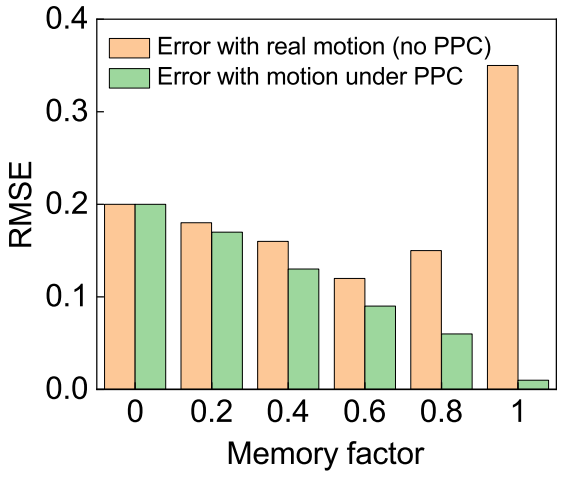

Figure S30 The RMSE values of predictions under different memory factors.

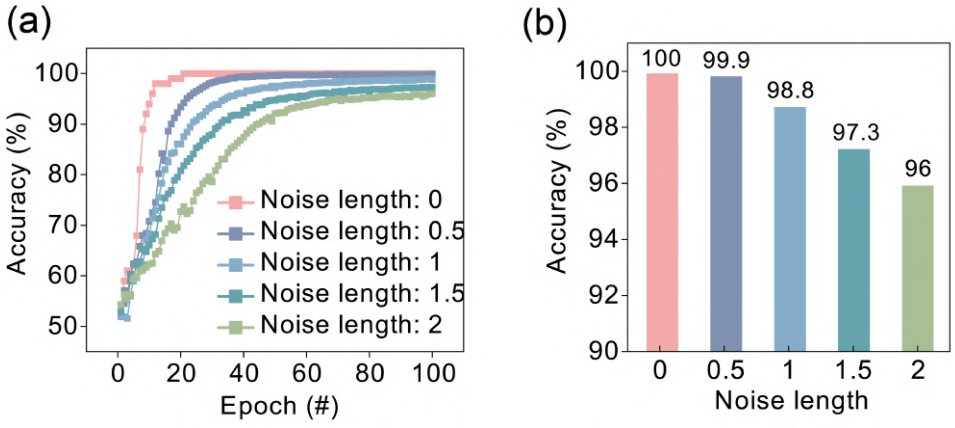

Figure S31 (a) Test iteration accuracy with different noise lengths. (b) The accuracy histogram under different noise. After 100 epochs, when the noise length is 0, 0.5, 1, 1.5, and 2, the IAs decision accuracy is 100%, 99.9%, 98.8%, 97.3%, and 96% respectively.

Before prediction, we map the signal to the (0, 1) interval, then we generate a random number of (0, 1) and multiply it by the noise length. When the noise length is 0.5, 1, and 2, random noise with intervals of (0, 0.5), (0, 1), and (0, 2) will be generated, respectively. Therefore, noise length values of 0.5, 1, and 2 represent 50%, 100%, and 200% noise, respectively.

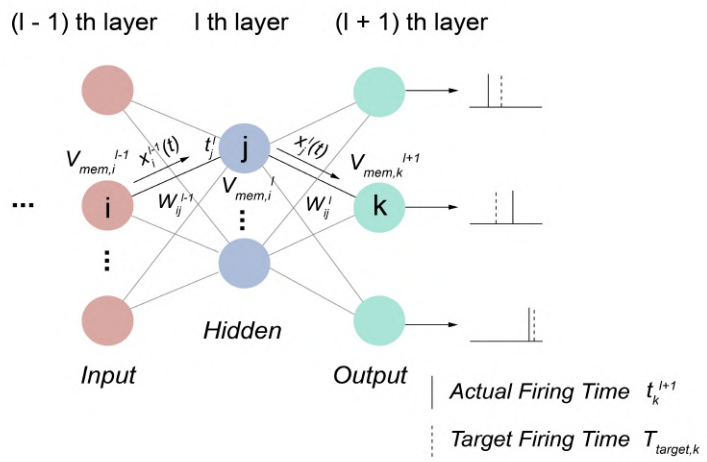

Figure S32 The schematic diagram of a spiking neuron network with first-spike time encoding.

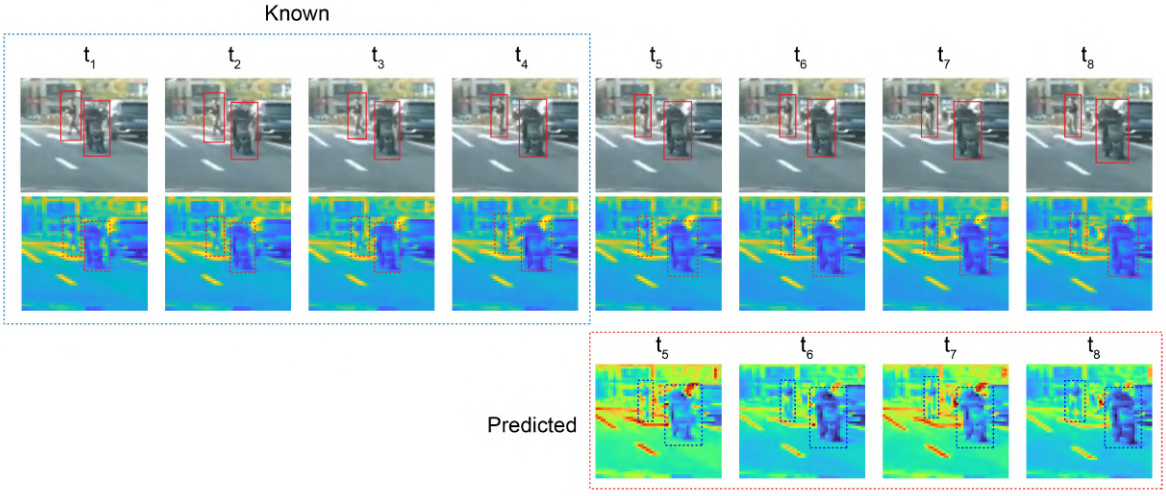

Figure S33 Visual processing capabilities in unstructured environments. Based on the motion data of the previous $t_1 \sim t_4$ frames, the results of the subsequent $t_5 \sim t_8$ frames are predicted.

Appendix A The mechanistic explanation of persistent photoconductivity-spiking generation.

The persistent photoconductivity (PPC) effect of the transistor is the main factor causing the persistent spike behavior. The mechanism of the PPC effect of oxide transistors is explained as follows: illumination with $\lambda < 550$ nm is thought to ionize the deep, neutral V_O states to shallow donor states (V_O^{2+}). The outward relaxation of bonds surrounding the V_O sites then creates an energy barrier (∼0.3 eV) against neutralization of V_O^{2+} sites, thus keeping the material in a state of high conductivity. [3, 4] Consequently, the light-induced negative V_{th} shift, and thus ΔI_{photo}, can be approximately described by assuming that PPC effect simply raises the flat-band Fermi level (EF0) because of the increased electron doping concentration of the film. [5] Figure S13 shows the PPC effect of the In_2O_3 transistor under light stimulation with an intensity of 0.12 mW and a wavelength of 420 nm, respectively. Furthermore, to ensure the generation of spike current, the In_2O_3 transistor current must match that of the NbO_x device (Figure S8). In Figure S8, R_i is 20 kΩ, so $R_{ch} < 20$ kΩ. When V_{in} is 2 V, the current of the artificial pulse neuron satisfies $I > (2 - V_{NbO_x})/2 \times 10^4 A$. Consider a boundary condition ($R_{ch} = R_i$), then I > 0.05 mA can ensure the generation of spike current. As can be seen from Figure S13, the neuronal oscillation conditions are met for a period of time, resulting in the generation of continuous spike currents.

Appendix B RC and LSTM model implementation.

For the RC model, when feedback from the output to the reservoir is absent, the time evolution of the neuronal states in the reservoir is described as follows:

$$r^t = f(W^{in}X^t + Wr^{t-1}), \tag{B1}$$

where t denotes the discrete time, r^t is the state vector of the reservoir units, X^t is the input vector, W^{in} is the weight matrix for the input-reservoir connections, and W is the weight matrix for the recurrent connections in the reservoir. The function f_k among $f = (f_1, f_2, \cdots, f_n)$ represents the kth elementwise activation function of the reservoir units, which is typically a sigmoid-type activation function. Equation (B2) represents a non-autonomous dynamical system forced by the external input X^t. The output is often given by a linear combination of the neuronal states in the reservoir as follows:

$$Y^t = W^{out}r^t, \tag{B2}$$

where Y^t is the output vector and W^{out} is the weight matrix in the readout. In supervised learning, this weight matrix is trained to minimize the difference between the network output and the desired output for a certain time period. Here, we consider a special form of RC by combining the neuronal states and output from Eqs. (B1) and (B2) as:

$$Y^t = W^{out}f(W^{in}X^t + WY^{t-1}), \tag{B3}$$

In RC, all W^{in} and W are randomly given and fixed, and only W^{out} as unknown variables is trained to minimize the difference between the network output and the desired output, with the known time series (X^t, Y^t).

For LSTM, we adopt the following structure (Figure S20). The batch size is 32, the optimizer is adam, the root mean square error calculation function is used, the learning rate is adaptive, and the number of training times is 50.

Furthermore, humans can react quickly to objects in the real world and make behavioral judgments, which stems from the inherent nonlinear pulse coding and long-term short-term memory characteristics of neurons and synapses. Nonlinear pulse coding and short-term pulse memory behaviors are widely present in the primary cortex of the brain and are called spike timing patterns. This helps to efficiently process information in time and space before it is transmitted to the brain, and ultimately achieves accurate predictions within the primary cortex. [6] Although current neuromorphic computing has established neuron and synapse models, it is still based on the traditional neural network framework (including RC and LSTM networks), which requires a large amount of data to be collected in advance and long-term iterative training, which is incompatible with the biological nervous system. Current research uses different material systems to simulate the basic properties of neurons and synapses, but still does not reveal how these properties give artificial neurons a more efficient information processing mode. In this work, we found that the basic properties of neurons and synapses can establish an efficient space-time equation, which helps neuromorphic chips achieve efficient information processing comparable to that of humans. Our research may play an important role in further simulating the human brain efficiently and promoting neuromorphic research.

Appendix C Compared with RC and LSTM.

Compared with RC and LSTM models, the RSTI equation has a key feature that it utilizes the spatiotemporal characteristics of high-dimensional data. Due to the conversion from high-dimensional spatial information to temporal information, RSTI can achieve multi-step ahead prediction even with only short-term data. In RC and LSTM models, a single time series is usually used to train a large number of parameters, and they will encounter overfitting problems when there is only a short time length. Furthermore, the models obtained through training often contain random factors, and the degrees of freedom implicit in the neural network may not match the actual unknown system. For RSTI, it constructs dynamic features based on real-time sensor data, is completely data-driven, and is highly robust to unknown time series. Finally, RSTI utilizes the inherent dynamics of the device array to form random, non-fixed weights, resulting in far fewer parameters. While in-sensor RC calculations employ a similar approach, they utilize the inherent dynamics of individual devices and lack a simultaneous high-dimensional variable reservoir, making them less capable of capturing the spatiotemporal characteristics of short-term variables.

Appendix D Computational complexity comparison.

The long short-term memory network (LSTM) stands as a renowned artificial neural network extensively employed in time series processing. [7] It was specifically selected as the benchmark method for evaluating time efficiency. In the context of the sensory array with N inputs, the array can spontaneously perform the capacity for nonlinear spiking encoding and memory retention. Since row and column address selection are required to read the output, the computational complexity is $O(N)$. A dropout scheme with n iterations is deployed in solving the weight matrix $B_{D\times L}$. In each iteration, $k(k < D)$ variables are randomly selected, and the computational complexity of solving A, B and Y is $O(2k(2/3L^3+2L^2))$, because the computational complexity of solving an indefinite linear equation is $O(2/3L^3 + 2L^2)$ if there are L coefficients to be solved in the equations. Therefore, the total time cost of n iterations is $O(2kn(2/3L^3 + 2L^2))$. Finally, the total computational complexity of RSTI is $O(2kn(2/3L^3+2L^2)+N)$. Notably, for the overall training of n_1 iterations, the total time complexity is $O(N) + n_1 \cdot O(n_2 \cdot N \cdot D) = O(N + n_1 \cdot n_2 \cdot N \cdot D)$.

Because RSTI can implement prediction with limited data, the L $\leqslant$ D, L $\leqslant$ n, L $\leqslant n_1$, and L $\leqslant n_2$. Therefore, time complexity is shown as follows:

$$TC(RSTI) = O(2kn(2/3L^3 + 2L^2) + N) \approx O(L \cdot n \cdot k \cdot L \cdot L), \tag{D1}$$

$$TC(LSTM) = O(L(N + n_1 n_2 ND + s_1^3)) \approx O(L \cdot n_1 \cdot n_2 \cdot N \cdot D + L \cdot s_1^3), \tag{D2}$$

The iteration numbers, denoted as n_1, n_2, and n, pertain to the same computational complexity order within the aforementioned algorithms, where D > k. Additionally, L represents the step size of prediction, characterizing a short sequence consistently smaller than the training size N or the number of iterations n_2 in LSTM, i.e., L $\leqslant$ N, L $\leqslant n_2$. Therefore, $O(L \cdot n \cdot k \cdot L \cdot L) \leqslant O(L \cdot n_1 \cdot n_2 \cdot N \cdot D + L \cdot s_1^3)$, which means the complexity of RSTI is smaller than LSTM.

Appendix E The process of implementing parallel prediction.

The random distribution embedding (RDE) framework is used to parallel in-sensor prediction. [8] Figure S27 shows the schematic diagram of the RDE framework. The RDE operates as follows. Consider a high-dimensional time series data denoted as T_i^t, where i = 1, 2, $\cdots$, D, and construct a delay attractor $M(T_k^t, T_k^{t+\tau}, T_k^{t+2\tau})$ as future time data for a single variable T_i^t. Utilizing the general embedding theorem, establish a non-delayed attractor $N(T_i^t, T_j^t, T_s^t)$, where (T_i^t, T_j^t, T_s^t) are randomly chosen observation time series from x_k^t and are employed for prediction. Here, M represents the delayed attractor, while N signifies the non-delayed attractor. The spatiotemporal transformation maps the non-delayed attractor to a delayed one, denoted as $\phi : N \to M$. For each index tuple (i, j, s), a mapping relationship can be constructed, yielding T_k^t as a predictor variable of the target variable, as shown in equation (E1).

$$T_k = \psi(T_i(t), T_j(t), T_s(t)), \tag{E1}$$

where ψ is a component of Ψ. As a theoretical explanation, the dimension of the non-delayed attractor is selected as three; based on the embedding theory, the prediction dimension of the general attractor can be larger.

Appendix F Comparison with conventional CMOS active pixel sensors.

In table F1, we compare size, power consumption, pixel composition, and functionality with CMOS active drive sensors. For power consumption, Mendis et al. [9] reported the CMOS active pixel image sensors with a power dissipation of 10^7 nJ / fps/ pix. Park et al. [10] presented a CMOS active pixel array for real-time edge image detection with a power dissipation of 0.72 nJ / fps/ pix. Gottardi et al. [11] reported a CMOS vision sensor with 14.95 nJ / fps / pix for event detection. In this work, the optoelectronic artificial neuron integrates In_2O_3 transistors and NbO_x threshold switching devices with a power consumption of only 5.6 pJ / fps / pix. In addition, studies have shown that by using spikes to drive brain-like neuronal computing, more energy-efficient and powerful machine intelligence can be achieved. [12] However, encoding the voltages into spike signals still consumes additional pJ $\sim$ fJ power consumption and 4 $\sim$ 6 additional transistors and capacitors. [13] For area efficiency, CMOS active drive array usually consumes 3 transistors per pixel for readout, selection, and reset. In this work, the optoelectronic artificial neuron consists of only a transistor and a simple two-terminal threshold switch memristors. Memristors can be integrated into transistor source/drain electrodes, so the integration density of 1T1R is limited by the size of the transistor ($> 8F^2$), which is better than the three transistors in the active pixel arrays.

Table F1 Comparison with conventional CMOS active pixel sensors

Ref	Image processing solutions	Array size	Pixel composition	Function	Sensor power
[9]	Active pixel sensors	128 × 128	A sensor and three transistors	Imaging	107 nJ / fps/ pix
[10]	Active pixel sensors	320 × 320	A sensor and row/column drivers	Edge Detection	0.72 nJ / fps/ pix
[11]	Active pixel sensors	500 × 500	A sensor and three transistors	Event Detection	14.95 nJ / fps / pix
This work	Optoelectronic spike neurons	4 × 4	1PT-1TS	In-sensor prediction	5.6 pJ / fps / pix

Appendix G Evaluate the cumulative timing error in the spike-time pattern.

The PPC effect of spike-time pattern may introduce cumulative errors in continuous-time tasks, which is a very important issue. In this paper, we introduce a parameter called the memory factor (γ) to represent the impact of PPC effect. The γ indicates the strength of the continuous spike signals after the light is removed. This makes the optical image at time m retain the peak current characteristics at time $m-\tau$, that is, $T^m = T^m + \gamma T^{m-\tau}$, causing cumulative time error. When $\gamma = 0$, it means that when the next sampling moment arrives, the current at the previous moment has been completely consumed (that is, there is no PPC effect); when $\gamma = 1$, it means that when the next sampling moment arrives, the current at the previous moment has not decayed (that is, the PPC effect is the strongest). Therefore, the cumulative time error caused by the PPC effect can be indirectly evaluated by evaluating the value of the γ. Figure S29 shows the motion characteristics and prediction results of 9 frames under different memory factors. The results show that as the memory factor increases, the prediction effect becomes better, but the motion characteristics gradually disappear. When the memory factor is greater than 0.8, the motion characteristics gradually disappear as the number of frames increases. Figure S30 shows the RMSE values of predictions under different memory factors. The results show that as the memory factor increases, the prediction error relative to the real image without the PPC effect first decreases and then increases; the prediction error relative to the image with the corresponding PPC effect decreases monotonically. This indicates that the PPC effect does indeed enhance the spatiotemporal prediction ability of neurons. However, when the PPC effect is too strong, the characteristics of the original motion almost disappear. Even though the prediction ability is improved, the results deviate significantly from the real motion. When the memory factor is 0, 0.2, 0.4, 0.6, 0.8, and 1, the RMSEs compared with the motion images without PPC are 0.2, 0.18, 0.15, 0.12, 0.17, and 0.35, respectively; and the RMSEs compared with the motion images with corresponding PPC are 0.2, 0.17, 0.13, 0.09, 0.06, and 0.01, respectively. This shows that when the memory factor of the PPC effect is 0.6, it can not only retain the obvious original motion characteristics, but also improve the spatiotemporal prediction ability of the optoelectronic neuron array.

Appendix H Spiking neural network training process.

In this network, the predicted pedestrian motion trajectory is encoded using the first-spike time as the input of the SNN. As shown in equation (H1), the firing time of input neurons is inversely proportional to the input value (I_i) of each neuron.

$$t_i^{input} = [(I_{max} - I_i)/I_{max}T_{max}], \tag{H1}$$

Figure S32 depicts a schematic diagram of an SNN with first-spike time encoding. The firing time of the jth neuron in the lth layer is defined as t_j^l. $x_j^l(t)$ represents the output spikes generated by the jth neuron of the lth layer at time t in the form of a voltage pulse. Input pulses are multiplied by weights and integrated by the non-leaky IF model. When the membrane voltage ($v_{mem,j}^l(t)$) reaches the neuron threshold (v_{th}^l), the neuron fires and generates a spike ($x_j^l(t) = 1$) in the next layer, as shown in equation (H2). Then, the firing time of neuron, t_j^l, is set to t, the time when the membrane reaches the threshold.

$$if \quad l = 1, x_j^l(t) = \begin{cases} 1 & (if \quad t = t_j^l), \\ 0 & (o.w.)), \end{cases} \quad else: x_j^l(t) = \begin{cases} 1 & (if \quad v_{mem,j}^l(t) > v_{th}^l), \\ 0 & (o.w.)), \end{cases} \tag{H2}$$

When the jth neuron of the lth layer is fired and the membrane cumulative function $S_j^l(t)$ is zero at time t, as shown in equation (H3).

$$S_j^l(t) = \begin{cases} 1 & (if \quad t \geqslant t_j^l), \\ 0 & (o.w.)), \end{cases} \tag{H3}$$

The membrane voltage of output neuron j can be calculated by multiplying the cumulative input and weights, as shown in equation (H4).

$$V_{mem,k}^{l+1}(t-1) = V_{mem,k}^{l+1}(t-1) + \sum_{i=j}^{N^l} x_j^l(t) w_{jk}^l = \sum_{i=j}^{N^l} S_j^l(t), w_{jk}^l, \tag{H4}$$

In SNN, the output value of neuron k is expressed as the firing time t_k^o. The error function of training is shown in equation (H5), which makes the output neuron as close to the target discharge time as possible ($T_{target,k}$).

$$L = 1/2(\sum T_{target,k} - t_k^o)/T_{max}, \qquad \delta_k^o = (T_{target,k} - t_k^o)/T_{max}, \tag{H5}$$

Notably, first-spike coding is also consistent with the fact that only sparse coding can explain the fast reaction times observed in the brain, such as visual processing. Researchers are working to develop new computing strategies using these basic spike encoding methods [14], such as deploying them inside sensors to build efficient neuromorphic systems. However, how the human brain uses the first-spike time coding to achieve short-term and rapid visual predictions remains unclear. In this work, we find that first-spike time encoding has nonlinear characteristics, which, combined with short-term memory capacity, can simulate the human brain to achieve efficient visual spatiotemporal prediction. What′s more, combining RSTI with SNN can achieve efficient pedestrian obstacle avoidance.

Appendix I More Information.

Table I1 Benchmarking of emerging machine vision systems based on in-sensor computing

Ref	Pixel	Capacity	Sensing	Training-free	Transient prediction	Sensor-level prediction	Sensor power (pJ)
[15]	1T1R	Static prediction	Resistance	×	×	×	N/A
[16]	1PT	Motion perception	Discrete current	×	×	×	9×10^4
[17]	1PT	Motion perception	Discrete current	×	×	×	7×10^5
[18]	1R	Motion perception	Discrete current	×	✓	×	3×10^4
[19]	1T	Motion perception	Discrete current	×	✓	×	9
[20]	1TS1R	Motion perception	Spike freq.	×	✓	×	15.3
This work	1PT1TS	Motion, location, prediction	First-spike time	✓	✓	✓	5.6

References

1 Zhou Y, Fu J, Wan T, et al. A 2T2R1C vision cell with 140 dB dynamic range and event-driven characteristics for in-sensor spiking neural network, In: Proceedings of IEEE International Electron Devices Meeting (IEDM), San Francisco, 2022
2 Jeon S, Ahn S E, Song I, et al. Gated three-terminal device architecture to eliminate persistent photoconductivity in oxide semiconductor photosensor arrays. Nat Mater 2012, 11: 301
3 Lany S, Zunger A. Anion vacancies as a source of persistent photoconductivity in II?VI and chalcopyrite semiconductors. Phys Rev B 2025, 72: 035215
4 Janotti A, Van de Walle C G. Oxygen vacancies in ZnO. Appl Phys Lett 2025, 87: 122102
5 Jeon S, Ahn SE, Song I, et al. Gated three-terminal device architecture to eliminate persistent photoconductivity in oxide semiconductor photosensor arrays. Nat Mater 2012, 11: 301–305
6 Fuster J M, Alexander G E. Neuron Activity Related to Short-Term Memory. Science 1971, 173: 652
7 Rao A, Plank P, Wild A, et al. A Long Short-Term Memory for AI Applications in Spike-based Neuromorphic Hardware. Nat Mach Intell 2022, 4: 467
8 Ma H, Leng S, Aihara K, et al. Randomly distributed embedding making short-term high-dimensional data predictable. Proc Natl Acad Sci U S A. 2018, 115: E9994
9 Mendis S K, Kemeny S E, Gee R C, et al. CMOS active pixel image sensors for highly integrated imaging systems. IEEE J Solid-State Circuits, 1997, 32: 187–197
10 Park M J, Kim H J. A Real-Time Edge-Detection CMOS Image Sensor for Machine Vision Applications. IEEE Sens J, 2023, 23: 9254–9261
11 Gottardi M, Parmesan L, Tosato P, et al. A 500 × 500 Pixel Image Sensor with Multiple Regions of Interest for Center of Mass-Based Event Detection. IEEE Sens J, 2024, 24: 32043–32052
12 Roy K, Jaiswal A, Panda P. Towards spike-based machine intelligence with neuromorphic computing. Nature, 2019, 575: 607–617
13 Sourikopoulos I, Hedayat S, Loyez C, et al. A 4-fJ/Spike Artificial Neuron in 65 nm CMOS Technology. Front Neurosci, 2017, 11.
14 Su X, Zhang B, Liang C, et al. Integrating Image Perception and Time-to-First-Spike Coding in MoS_2 Phototransistors for Spiking Neural Network, Adv Funct Mater 2024, 34: 2315323.
15 Wang Z, Li C, Lin P, et al. In situ training of feed-forward and recurrent convolutional memristor networks, Nat Mach Intell 2019, 1: 434.
16 Chen J, Zhou Z, Kim B J, et al. Optoelectronic graded neurons for bioinspired in-sensor motion perception. Nat Nanotechnol 2023, 18: 882.
17 Pan X, Shi J, Wang P, et al. Parallel perception of visual motion using light-tunable memory matrix. Sci Adv 2023, 9: eadi4083.
18 Tan H, van Dijken S. Dynamic machine vision with retinomorphic photomemristor-reservoir computing. Nat Commun 2023, 14: 2169.
19 Gao C, Liu D, Xu C, et al. Toward grouped-reservoir computing: organic neuromorphic vertical transistor with distributed reservoir states for efficient recognition and prediction. Nat Commun 2024, 15: 740.
20 Song H, Lee M G, Kim G, et al. A Bioinspired Stretchable Sensory-Neuromorphic System. Adv Mater 2024, 36: e2309708.