nature communications

Article

https://doi.org/10.1038/s41467-026-75808-w

Wavelength-encoded neuromorphic inference enabled by microcavity MoS_2 photodetector arrays

Received: 30 January 2026

Accepted: 8 July 2026

Published online: 18 July 2026

Check for updates

Xiang Chen[1,11], Kangjian Di[2,3,11], Fuhao Yu[3,4,11], Jiashu Li[2], Lixiang Sun[1], Shengdi Chen[2], Silin Chen[2,3], Andy Liu[2], Xiang Peng[5], Shitong Zhu[6], Taotao Li[2,7], Wei Jiang[3,4], Xuping Zhang[3,4], Yi Shi[1], Xinran Wang[2,7,8], Li Gao[9,10] ✉, Zehua Hu[1] ✉ & Ningmu Zou[2,3,7,8] ✉

Biological vision systems tightly couple spectral sensing with temporal integration to extract task-relevant information with minimal data movement. In contrast, conventional optoelectronic vision hardware typically separates photodetection from electronic computation, incurring substantial latency and energy costs. Optical neural networks can alleviate this bottleneck, but many implementations are difficult to scale. Here, we propose a bio-inspired optoelectronic inference architecture based on a Fabry-Perot microcavity-integrated MoS_2 photodetector array, in which sensing, weighting, and accumulation are unified within each pixel. Cavity-engineered wavelength selectivity encodes neural-network weights in the spectral domain, while the finite carrier lifetime of MoS_2 enables analog temporal accumulation without external memory. The system achieves test accuracies of 99.6%, 94.8% and 94.0% on MNIST, CIFAR-10 and the Free Spoken Digit Dataset, respectively. Post-training optical Hessian pruning further reduces optical complexity while maintaining robustness. This architecture provides a compact route toward wavelength-aware in-sensor neuromorphic inference.

Conventional optoelectronic vision systems are predominantly built on a modular architecture in which optical sensing, signal conditioning, and digital computation are physically separated. Visual information is first captured by passive photodetectors and then transferred to electronic processors for feature extraction and classification, resulting in excessive data movement, energy consumption, and latency. Optical and photonic neural networks have been explored to alleviate these bottlenecks by exploiting the parallelism and bandwidth of light[1,2], however, many implementations rely on interferometric meshes[3–5], coherent phase control[6–8], or complex optical routing[9–11], which impose stringent stability requirements and hinder scalability and robustness for compact vision hardware[12,13].

Biological vision systems offer a fundamentally different paradigm, in which sensing and computation are tightly coupled at the physical level. Inspired by this principle, recent bio-inspired optoelectronic devices based on emerging semiconductors have

[1]School of Electronic Science and Engineering, National Laboratory of Solid-State Microstructures, Nanjing University, Nanjing, China. [2]School of Integrated Circuits, Nanjing University, Suzhou, China. [3]Key Laboratory of Intelligent Optical Sensing and Manipulation, Nanjing University, Ministry of Education, Nanjing, China. [4]College of Engineering and Applied Sciences, Nanjing University, Nanjing, China. [5]School of Chemistry and Chemical Engineering, Nanjing University, Nanjing, China. [6]Nanjing Extremo Technology Co. Ltd., ., Nanjing, China. [7]Suzhou Laboratory, Suzhou, China. [8]Interdisciplinary Research Center for Future Intelligent Chips (Chip-X), Nanjing University, Suzhou, China. [9]School of Science, Jiangsu Key Laboratory of Quantum Computing Science and Devices, Nanjing University of Posts and Telecommunications, Nanjing, China. [10]School of Materials Science and Engineering, State Key Laboratory of Flexible Electronics (LoFE) and Institute of Advanced Materials (IAM), Nanjing University of Posts and Telecommunications, Nanjing, China. [11]These authors contributed equally: Xiang Chen, Kangjian Di, Fuhao Yu. ✉e-mail: iamlgao@njupt.edu.cn; zehuahu@nju.edu.cn; nzou@nju.edu.cn

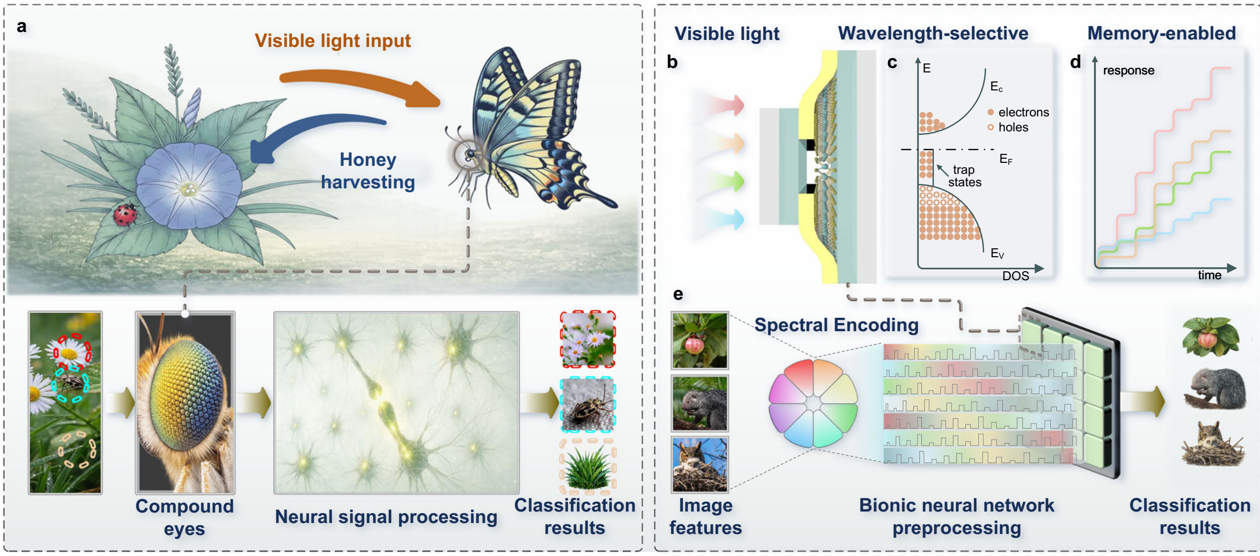

Fig. 1 | Bio-inspired optical neural inference concept and system architecture. **a** Biological vision-inspired perception, where spectral sensing and neural integration enable object recognition. **b** Resonant-cavity-assisted wavelength-selective coupling of visible light into the MoS_2 photodetector, enabling wavelength-dependent responsivity enhancement in the MoS_2 channel for spectral filtering and analog weighting. **c** Schematic of the band structure of the MoS_2 photodetector. DOS, density of states. **d** Memory-enabled, time-domain accumulated photodetector responses enabling optical-domain classification. **e** Bio-inspired optical preprocessing and inference framework using wavelength- and time-encoded visible light and a MoS_2 photodetector array. The representative object image in panel **e** was adapted from an image in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2017 test set. We acknowledge the ImageNet/ILSVRC project; see Russakovsky et al. [43].

demonstrated the direct implementation of perceptual functions within photodetectors themselves[14–20]. In particular, MoS_2 photodetectors with intrinsic memory effects have enabled low-power biomimetic collision detection by exploiting carrier accumulation to emulate looming-sensitive neurons, allowing distance-related perception without external computation[21–23]. Similarly, bioinspired in-sensor visual adaptation has been realized using MoS_2-based devices that dynamically adjust responsivity in response to illumination changes, enabling accurate perception across a wide dynamic range[24,25]. Despite these advances, existing biomimetic photodetector systems predominantly operate along a single intensity dimension. Core perceptual capabilities of biological vision—such as wavelength-dependent sensing and color-aware information processing—remain a nascent frontier at the device level[26,27]. This limitation constrains their applicability to high-dimensional visual inference tasks, where spectral information plays a critical role.

Here, inspired by biological visual systems that combine spectral selectivity with temporal integration[28], we propose a bio-inspired optoelectronic inference architecture based on a microcavity-integrated MoS_2 photodetector array. Leveraging the high optoelectronic quality and ease of integration of 2D materials, we embed MoS_2 within a Fabry-Perot microcavity to intrinsically unify sensing, weighting, and accumulation within a single pixel. Specifically, the cavity-engineered wavelength selectivity encodes neural network weights in the spectral domain, while the finite carrier lifetime of MoS_2 enables analog temporal accumulation without external memory. This architecture directly addresses the spectral-processing bottleneck inherent in existing biomimetic optoelectronic systems. The proposed platform is experimentally validated on MNIST, CIFAR-10, and Free Spoken Digit Dataset (FSDD), demonstrating accurate wavelength-encoded optical classification and close agreement with electronic reference models. Furthermore, a post-training optical Hessian pruning strategy is introduced to enhance robustness against optical noise while reducing effective system complexity. Together, these results establish microcavity-enhanced MoS_2 photodetectors as a scalable bio-inspired pathway toward compact, wavelength-aware optical neural inference hardware.

Results

Bio-inspired optical neural inference

Nectar-foraging insects achieve robust flower recognition by jointly exploiting spectral sensing and neural integration in their compound-eye vision. At long range, chromatic contrast enables efficient target localization against foliage, whereas at close range additional cues—such as shape and petal patterns—support fine discrimination among flower structures and surrounding distractors (Fig. 1a). Motivated by this spectral-neural co-processing paradigm, we implement a bio-inspired optical preprocessing and inference framework that operates directly in the optical domain using wavelength- and time-encoded visible light in conjunction with a wavelength-selective MoS_2 photodetector array (Fig. 1e). A resonant optical cavity integrated with the MoS_2 photodetector (Fig. 1b) tailors its spectral responsivity, enabling wavelength-selective detection and thereby supporting spectral encoding of image features across multiple optical channels. In this scheme, feature components are multiplexed onto distinct wavelengths, while the device's wavelength-dependent photoresponse provides an effective analog weighting during optical-to-electrical conversion.

At the device level, the trap-mediated photoresponse dynamics (schematically illustrated in Fig. 1c) endow the MoS_2 photodetector with memory-like behavior. This enables time-domain accumulation of sequential photodetector responses (Fig. 1d), analogous to neuronal integration, which can implement in-sensor multiply-accumulate-type operations and facilitate optical-domain classification with reduced downstream electronic processing.

Fabrication and basic characterization

The microcavity-integrated MoS_2 photodetector is fabricated through a sequence of thin-film deposition, material transfer, and lithographic patterning steps designed to ensure precise cavity definition and high-quality electrical contacts. The overall fabrication flow is summarized

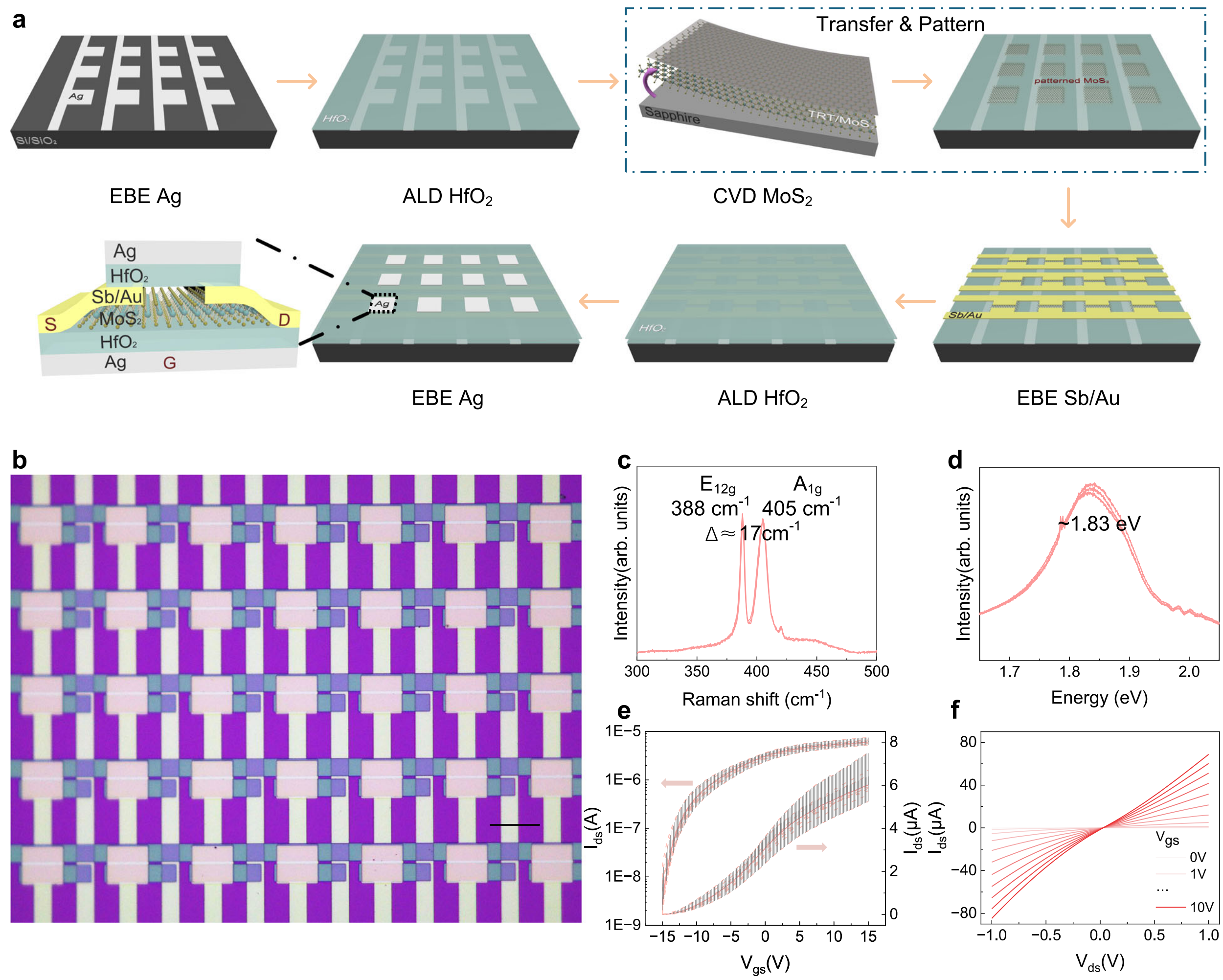

Fig. 2 | Fabrication and characterization of the microcavity-integrated MoS_2 photodetector. **a** Schematic of the fabrication process. **b** Optical microscopy image of the fabricated device, showing the overall morphology and structural details. Scale bar: 30 μm. **c, d** Overlaid Raman **c** and photoluminescence **d** spectra measured from 3 locations. **e** 12 overlaid transfer curves in log (left axis) and linear (right axis) scales. **f** Output characteristics under various V_{gs}.

in Fig. 2a. A 100 nm-thick Ag electrode array was first patterned and deposited on a silicon substrate. This Ag layer serves dual roles simultaneously as the back-gate electrode of the photodetector and as the bottom mirror of the optical microcavity. A 72 nm HfO_2 layer was then deposited by atomic layer deposition, functioning both as the gate dielectric and as a cavity spacer that defines the optical resonance. Single-crystal MoS_2 grown on sapphire by van der Waals epitaxy was subsequently transferred onto the prepared substrate using a wet-transfer process[29–31]. The MoS_2 film was patterned and etched to define the active device channel, after which source and drain electrodes composed of Sb/Au were deposited to form electrical contacts. Finally, a second 72 nm HfO_2 spacer layer and a 40 nm Ag top mirror were deposited, completing the Fabry-Perot microcavity structure integrated with the MoS_2 photodetector. An optical micrograph of a representative device is shown in Fig. 2b.

The structural and optoelectronic characteristics of the devices are evaluated by material and electrical measurements, as shown in Fig. 2c–f. Raman and photoluminescence measurements collected from three spatially separated locations across a 2-inch wafer exhibit nearly identical spectral features (Fig. 2c, d), indicating high material quality and excellent wafer-scale uniformity of the transferred MoS_2. This uniformity is essential for reproducible device behavior across large-area photodetector arrays. Electrical characterization further confirms the high quality of the fabricated devices. The MoS_2 field-effect transistors display an on-off current ratio of approximately 10^4 and an output current of ~6 μA at $V_d = 0.1$ V (Fig. 2e). The output characteristics show a linear dependence on drain voltage over the measured range (Fig. 2f), demonstrating ohmic contact behavior[32]. This contact performance is attributed to the low-temperature deposition of the semimetal Sb, which effectively suppresses Fermi-level pinning and preserves the intrinsic crystal structure of MoS_2. Together, these results establish a reliable electrical and material foundation for subsequent optoelectronic and computing experiments.

Optoelectronic performance

The resonant wavelength λ of a Fabry-Perot cavity is related to its physical length L by the condition: $2nL = m\lambda, m = 1, 2, 3, \ldots$, where n denotes the refractive index of the medium filling the cavity. In this work, we adopt the fundamental mode ($m = 1$) to realize the simplest possible cavity configuration. The spectral sharpness of the resonance is quantified by the finesse $\mathscr{F}$, defined as $\mathscr{F} = \frac{\pi (R_1R_2)^{1/4}}{1-(R_1R_2)^{1/2}}$, where R_1 and R_2 are the reflectance of the two mirrors. The full width at half maximum (FWHM) of the resonance peak, $\Delta\lambda$, is then given by $\Delta\lambda = \frac{\lambda^2}{2nL\mathscr{F}}$, which underpins the design of our optical classifier part.

The optoelectronic response of the microcavity-integrated MoS_2 photodetector is systematically evaluated using a home-built micro-

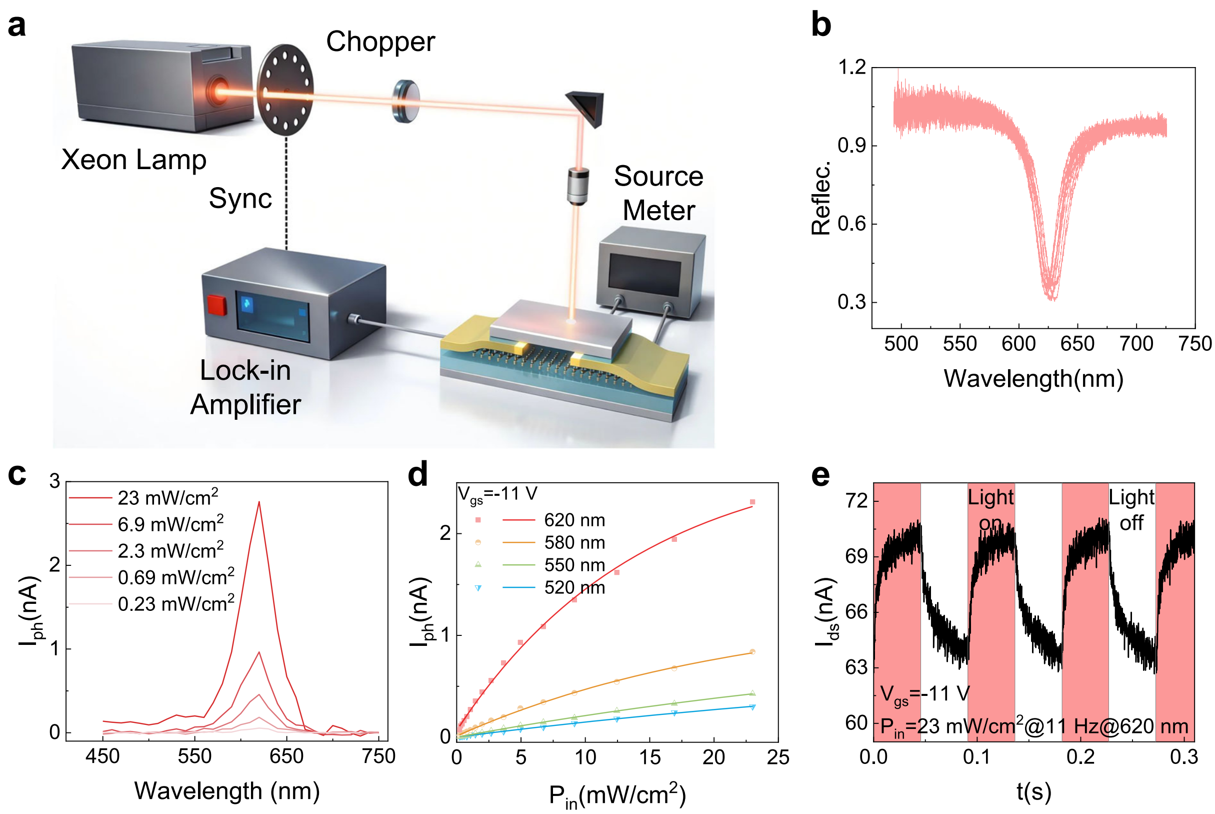

Fig. 3 | Optical characterization of the microcavity-integrated MoS_2 photodetector. a Schematic of the home-made micro-optoelectronic measurement setup. **b** Overlaid reflection spectra from 12 devices. **c** Wavelength-dependent photocurrent under different incident powers. **d** Power-dependent photocurrent for varying incident wavelengths. **e** The time-resolved photoresponse under 620 nm illumination at $V_{gs} = -11$ V, and $V_d = 0.1$ V.

optoelectronic measurement setup, schematically illustrated in Fig. 3a. This setup enables simultaneous control and characterization of optical excitation conditions and electrical readout, allowing detailed assessment of the wavelength-dependent and dynamic photoresponse of the device.

The micro-reflection spectra collected from 12 nominally identical devices reveal a highly uniform cavity resonance behavior centered at approximately 625 nm with a FWHM of ~25 nm, corresponding to a cavity quality factor of ~30 (Fig. 3b). The small variation in resonance position and linewidth across devices indicates excellent control over cavity thickness and material uniformity, which is essential for scalable array-level operation. The optical resonance directly translates into wavelength-selective photodetection. As shown in Fig. 3c, the photocurrent reaches its maximum near ~620 nm, closely matching the cavity resonance wavelength. Away from resonance, the photocurrent decreases, confirming that the spectral responsivity of the device is governed by microcavity-enhanced absorption. This wavelength selectivity provides a physical mechanism for encoding spectral information into electrical signals. Power-dependent photocurrent measurements further elucidate the device behavior under resonant and off-resonant excitation. As the incident wavelength approaches the cavity resonance, the photocurrent response evolves from linear to sub-linear with increasing optical power. This behavior is attributed to the more rapid saturation of interface states or defect-related traps under enhanced resonant excitation, where the optical field intensity inside the cavity is maximized[23,33].

In addition to its spectral response, the device exhibits fast transient characteristics. Time-resolved measurements reveal rise and fall times of 9.87 and 14.12 ms, respectively, under resonant illumination (Fig. 3e). These response times are sufficiently short to support time-multiplexed operation while remaining long enough to enable temporal accumulation of sequential optical pulses. The coexistence of wavelength selectivity and intrinsic temporal integration establishes the microcavity-integrated MoS_2 photodetector as a suitable building block for wavelength-encoded optical analog computation.

Experimental setup for optical classifier

Based on the intrinsic Fabry-Perot cavity effect of MoS_2 photodetector, we construct the optical classification platform shown in Fig. 4 and evaluated it on three datasets (MNIST[34,35], CIFAR-10[36], and FSDD[37]), as conceptually illustrated in Fig. 4a. The input samples are first processed by the electronic feature-extraction backbone to generate a compact feature representation (feature map / feature vector). The detailed architecture of this feature-extraction network is provided in Supplementary Note 1 and Fig. S1. Briefly, the feature extractor is a compact CNN with hierarchical feature aggregation, producing a 512-dimensional feature vector that is subsequently processed by the wavelength-encoded optical classifier. The extracted features are then encoded into a time-multiplexed sequence of optical pulses generated by a supercontinuum laser, with the pulse intensities carrying the feature amplitudes. These pulses are subsequently injected into the optical system to execute the final classification stage (Fig. 4b).

A broadband supercontinuum source provides dense wavelength channels that enable wavelength-encoded weighting in hardware. After collimation by the fiber collimator (FC), the beam is routed by a lens and beam splitter (BS) to both an optical spectrum analyzer (OSA) for spectral monitoring and the MoS_2 photodetector array for computation (Fig. 4b). The measured spectrum (Fig. 4c) confirms a wide wavelength span with sufficiently uniform spectral power, allowing flexible allocation of discrete wavelength channels. Each pixel is engineered to exhibit a distinct cavity resonance wavelength and tailored FWHM (Fig. 4d), yielding a pixel-specific

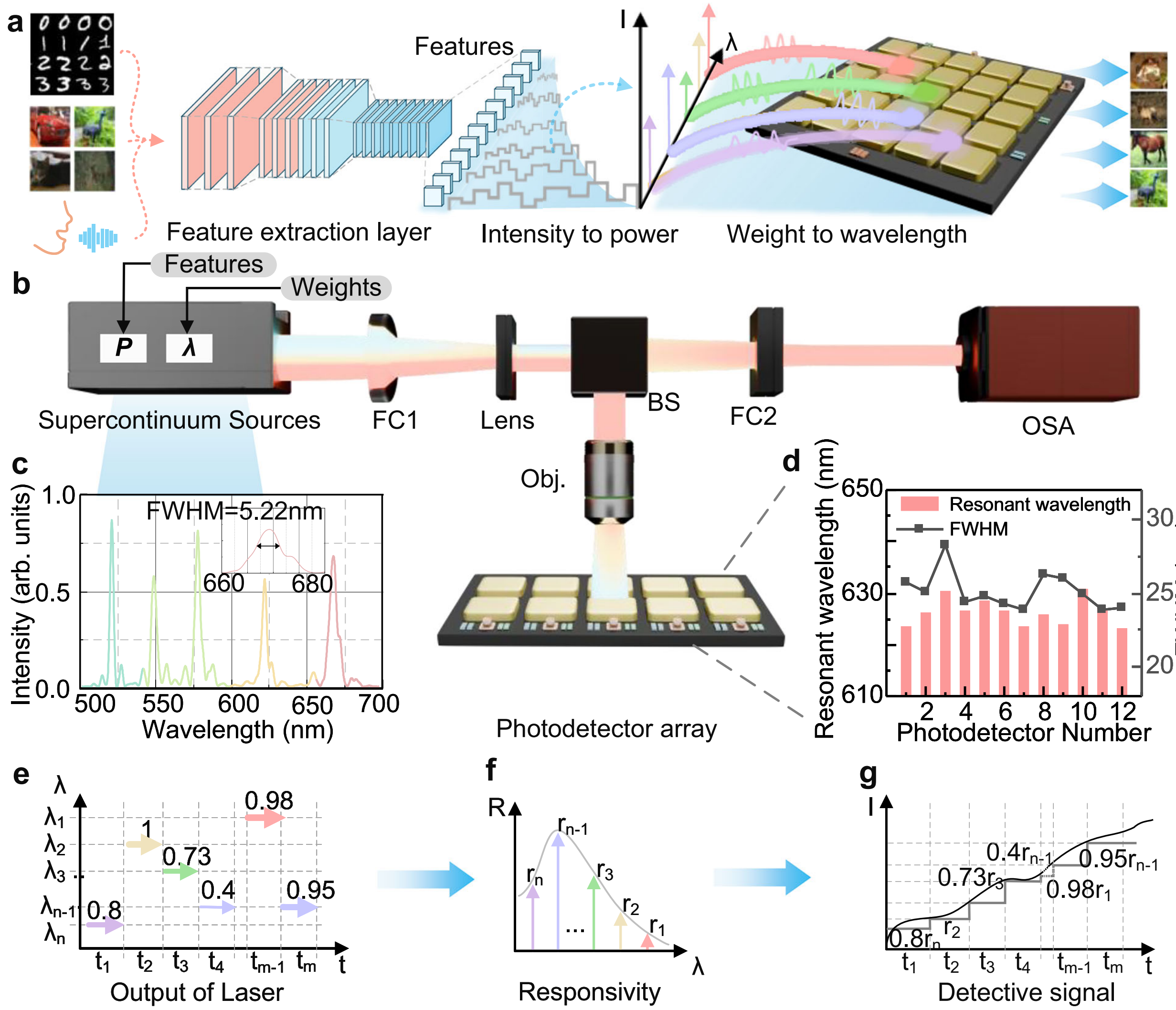

Fig. 4 | Experimental setup for the optical classifier. **a** Schematic diagram of the classification task based on the MoS_2 photodetector. **b** System setup diagram, including the supercontinuum laser source, fiber collimator (FC), lens, beam splitter (BS), optical spectrum analyzer (OSA), objective lens (Obj.), and the MoS_2 photodetector array. **c** OSA measurement of the output spectrum from the supercontinuum laser. **d** The resonance wavelengths and the FWHM of 10 photodetectors in the MoS_2 photodetector array. **e** Example of the time–wavelength multiplexed pulse sequence used in inference, where pulse amplitude encodes the input feature a_j and the assigned wavelength encodes the corresponding weight channel $\lambda_{k,j}$. **f** Representative wavelength-dependent responsivity $R_k(\lambda)$ of a pixel, where sampling at $\lambda_{k,j}$ implements the effective weight $\widetilde{w}_{k,j}$. **g** Measured integrated readout of the photodetector in response to the sequence in **e**, demonstrating wavelength-selective multiply-accumulate through intrinsic temporal accumulation. The representative object image in panel **a** was adapted from an image in the MNIST and CIFAR-10 datasets. MNIST was created by Yann LeCun and Corinna Cortes; CIFAR-10 was created by Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. See refs. [35,36].

responsivity spectrum $R_k(\lambda)$(Fig. 4f). During inference, the input feature vector is mapped to a time-multiplexed pulse sequence in which the pulse amplitude encodes the feature component a_j, while the trained weight is encoded by assigning a wavelength channel $\lambda_{k,j}$ to each time slot (Fig. 4e).

Thus, the two operands of the optical multiply–accumulate operation (MAC) are encoded in different physical dimensions: the feature value is carried by pulse amplitude, whereas the weight is selected by the wavelength-dependent responsivity of the microcavity photodetector.

For pixel k, the instantaneous photocurrent is governed by wavelength-selective transduction, and can be written as:

$$i_k(t) = \int R_k(\lambda)P(\lambda,t)d\lambda \approx P_0 \sum_{j=1}^{M} R_k(\lambda_{k,j})a_j p(t-t_j) \quad (1)$$

where $p(t)$ is the pulse shape and t_j denotes the j-th time slot. Device memory (finite carrier lifetime) further performs temporal accumulation, so that the integrated readout at t_{read} approximates an analog dot product

$$y_k(t_{read}) \propto \sum_{j=1}^{M} \underbrace{R_k(\lambda_{k,j})}_{\widetilde{w}_{k,j}} a_j \quad (2)$$

with the effective weight $\widetilde{w}_{k,j}$ implemented by responsivity sampling at the assigned wavelength. The resulting integrated photocurrent (Fig. 4g), determined jointly by the programmed time-wavelength sequence (Fig. 4e) and the measured responsivity spectrum (Fig. 4f), forms the class-dependent output used for decision making. The usable resonance range of a single pixel and its responses at quantized wavelength channels are summarized in Supplementary Fig. 7. A full

https://doi.org/10.1038/s41467-026-75808-w

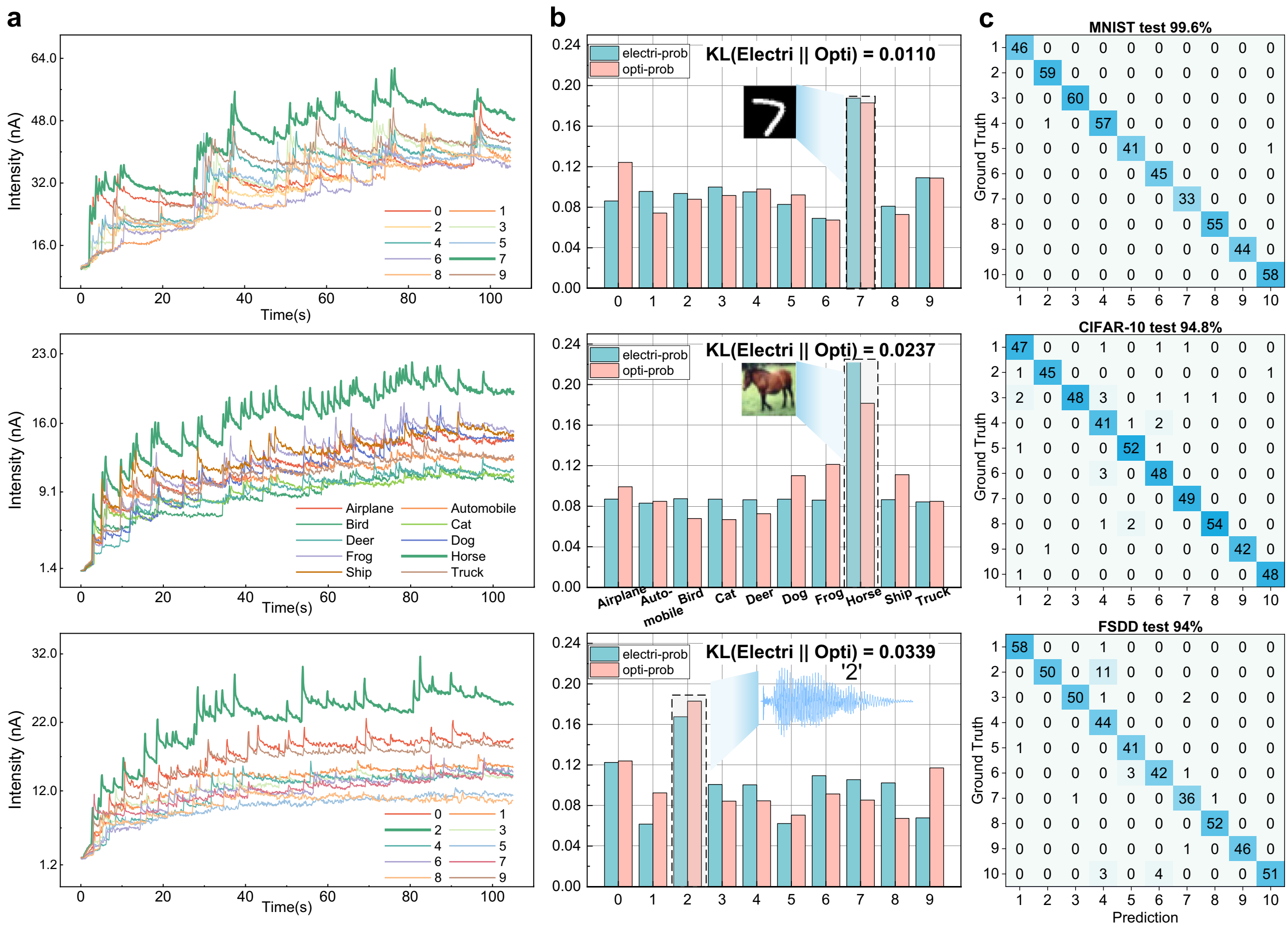

Fig. 5 | Results of the MoS_2 photodetector optical classifier on three 10-category classification datasets (MNIST, CIFAR-10, FSDD). The ten output channels correspond to the ten candidate classes rather than ten individual test samples: **a** System responses for different datasets. **b** Output distribution at the final layer: simulation vs. Optical Classifier. **c** Confusion matrix of the Optical Classifier. MNIST / CIFAR-10 / FSDD test accuracy = 99.6% / 94.8% / 94%.

derivation including the temporal kernel and calibration strategy is provided in Supplementary Note 2.

The end-to-end inference capability of the MoS_2 photodetector based optical classifier is evaluated on three representative datasets: MNIST, CIFAR-10, and FSDD. The trained classifier weight vectors for MNIST, CIFAR-10, and FSDD are visualized in Supplementary Fig. 4–6, respectively. The overall system-level performance is summarized in Fig. 5. In this architecture, features extracted by an electronic backbone are encoded into the amplitudes of a time-multiplexed supercontinuum pulse sequence, while the trained non-negative classifier weights are mapped onto wavelength channels through the wavelength-selective responsivity of the MoS_2 photodetector array. Because the responsivity-based effective weight is inherently non-negative, the final optical classification layer was trained under a non-negative weight constraint and without bias. The affine normalization maps feature values into non-negative optical pulse amplitudes, rather than encoding negative weights. Owing to the intrinsic device memory arising from the finite carrier lifetime, each photodetector pixel performs analog MAC over the entire pulse sequence and outputs a class-dependent integrated photocurrent that serves as the decision logit.

The temporal evolution of the photocurrent during inference reveals clear accumulation behavior. As shown in Fig. 5a, the measured responses of all ten output channels exhibit progressive integration as the feature pulses are sequentially applied. For all three datasets, the channel corresponding to the ground-truth class consistently develops the largest final integrated response, enabling a straightforward argmax decision in the optical-electrical domain. Small residual fluctuations and trial-to-trial variations are observed, which are mainly attributed to dynamic intensity noise from the supercontinuum source as well as device- and coupling-level nonidealities. Nevertheless, these effects do not compromise class separability under the demonstrated operating conditions. The fidelity of the optical classifier relative to the electronic reference network is quantified by comparing final-layer output probability distributions. As illustrated in Fig. 5b, the probability distributions obtained from the optical classifier (opti-prob) closely match those from electronic simulation (electri-prob) after normalization and softmax. Representative examples include MNIST (digit 7), CIFAR-10 (horse), and FSDD (digit 2). The corresponding Kullback-Leibler (KL) divergence values are 0.0110 for MNIST, 0.0237 for CIFAR-10, and 0.0339 for FSDD, indicating high-fidelity implementation of the fully connected classification layer through wavelength-selective weighting and temporal accumulation[38]. The slightly higher divergence observed for FSDD is consistent with increased sensitivity to source intensity fluctuations and hardware mismatch when output probabilities are more narrowly distributed across competing classes.

Classification performance across the full test sets is further summarized by the confusion matrices shown in Fig. 5c. All three matrices exhibit strong diagonal dominance, confirming robust class discrimination by the optical classifier. The achieved test accuracies are 99.6% for MNIST, 94.8% for CIFAR-10, and 94% for FSDD. The near-ceiling performance on MNIST reflects the higher intrinsic separability of handwritten digit data, while the remaining errors in CIFAR-10 and FSDD are primarily

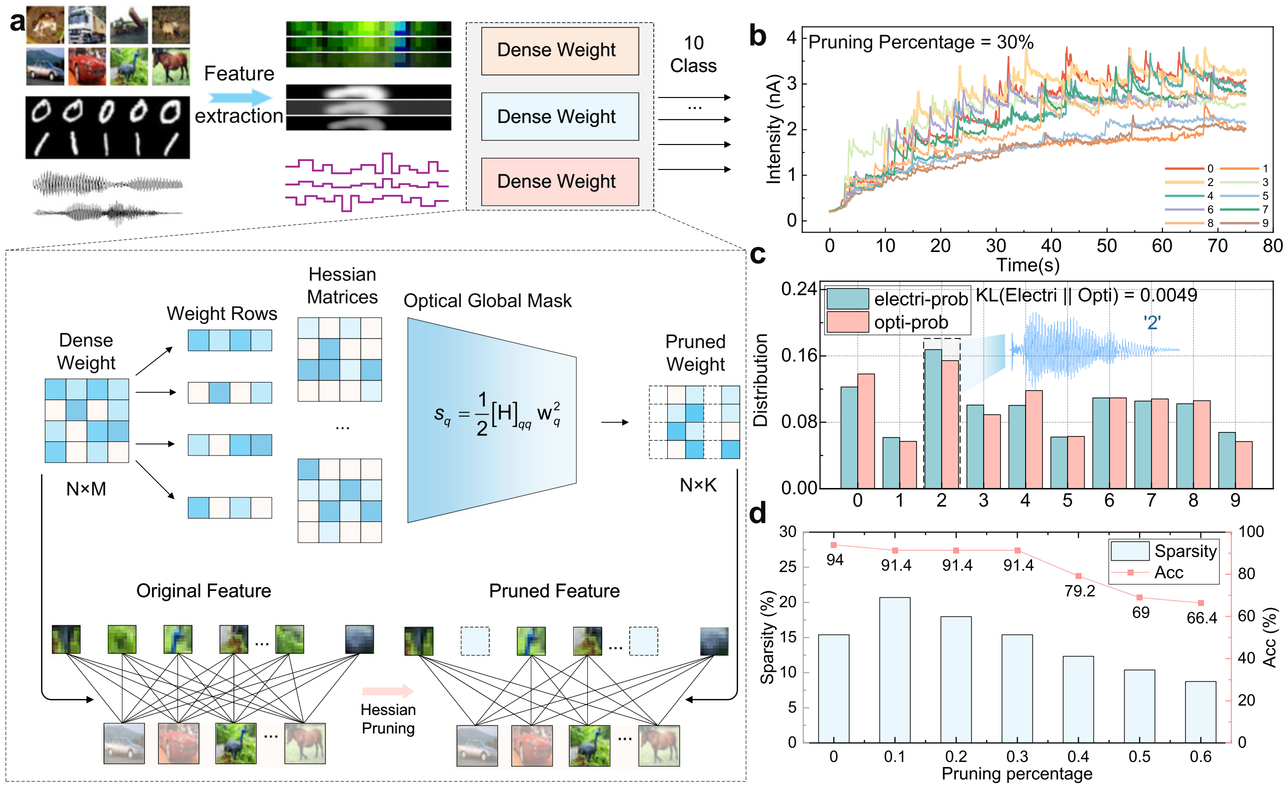

Fig. 6 | Optical Hessian pruning for robust and lightweight MoS_2 photodetector optical classification (FSDD). **a** Schematic of the proposed post-training optical Hessian pruning workflow. **b** Measured temporal responses of the MoS_2 photodetector classifier to different digit classes under Hessian pruning with $\alpha = 0.3$. The response associated with the ground-truth class 2 exhibits the highest integrated output. **c** Final-layer output probability distribution comparison between the electronic reference model and the optical classifier after Hessian pruning. **d** Test accuracy of the optical classifier as a function of pruning ratio.

associated with inter-class similarity and hardware-induced distribution perturbations.

To further evaluate the role of hardware calibration and microcavity-induced wavelength selectivity, we perform additional control experiments on the FSDD task. First, using the same 10-device MoS_2 photodetector array without pixel-wise calibration, the classification accuracy decreases to 76.4% (Supplementary Fig. 8), whereas calibration based on the measured responsivity spectra restores the accuracy to 94.0%. This result indicates that device-to-device spectral variation mainly introduces deterministic weight mismatch, which can be compensated by calibration. Second, a control MoS_2 photodetector array without microcavities shows a much lower classification accuracy of 16.4% (Supplementary Fig. 9). Together with the non-cavity reflection spectra in Supplementary Fig. 13, this control experiment confirms that the microcavity-induced wavelength-selective responsivity is essential for implementing the optical classifier.

Optical Hessian pruning

To improve the robustness of the optical classifier against dynamic intensity noise from the laser source, a post-training compression strategy based on Hessian pruning is employed. This approach operates on a fully trained network and selectively removes parameters that contribute weakly to the loss function. By leveraging second-order information from the Hessian of the loss with respect to model parameters, this method quantifies parameter importance and identifies low-curvature directions in the loss landscape as pruning candidates with minimal impact on classification accuracy[39–41].

In the adopted architecture, features extracted by the backbone network are flattened into a one-dimensional vector and processed by a fully connected layer, whose parameters form a two-dimensional weight matrix. In the optical classifier, this fully connected layer operation is physically implemented through wavelength-encoded computation: feature values are mapped to the intensities of time-multiplexed optical pulses, while fully connected layer weights are encoded via wavelength-channel assignment and realized through the wavelength-dependent responsivity of the MoS_2 photodetector array[42]. Under this encoding scheme, certain feature dimensions are particularly sensitive to laser intensity fluctuations and wavelength-channel mismatch. Hessian-based pruning is therefore applied to suppress feature dimensions that are both curvature-insensitive and noise-amplifying, thereby enhancing robustness to optical nonidealities.

The pruning workflow is illustrated in Fig. 6a. To estimate how much the loss changes when a parameter is removed, we expand the training loss $L(\mathbf{w})$ around the trained weights $\mathbf{w}$ using a second-order Taylor approximation:

$$L(\mathbf{w} + \Delta\mathbf{w}) \approx L(\mathbf{w}) + \nabla L(\mathbf{w})^T \Delta\mathbf{w} + \frac{\Delta\omega^T \boldsymbol{H} \Delta\mathbf{w}}{2} \tag{3}$$

Near a converged solution, the first-order term is typically small ($\nabla L(\mathbf{w}) \approx 0$), so the loss increase is dominated by the curvature captured by the Hessian matrix $\boldsymbol{H} = \nabla^2 L(\mathbf{w})$.

Pruning the q-th weight means setting $\mathbf{w}_q \to 0$, which corresponds to the perturbation:

$$\Delta\mathbf{w} = -\mathbf{w}_q e_q \tag{4}$$

Substituting this perturbation into the second-order term and using the common diagonal approximation (keeping only $\boldsymbol{H}_{qq}$) gives the Hessian pruning importance score:

$$s_q \triangleq \Delta L_q \approx \frac{\boldsymbol{H}_{qq}\mathbf{w}_q^2}{2} \tag{5}$$

Hessian-based importance scores are computed for each column of the weight matrix, corresponding to individual feature dimensions. Columns with low importance are removed, yielding a lightweight model characterized by a pruning coefficient α. If the original number of weights is NK, the number of retained weights after pruning is αNK. Implementation details are provided in Supplementary Note 3.

Figure 6b–d summarizes the results obtained on the FSDD dataset. An example inference for digit 2 with $\alpha = 0.3$ is shown in Fig. 6b. Each temporal response corresponds to the accumulated photocurrent of a candidate class, with the ground-truth class 2 maintaining the largest integrated output. Figure 6c compares the final-layer probability distributions before and after pruning, showing improved agreement between the optical classifier and the electronic reference model, as reflected by a reduced KL divergence. This improvement is attributed to reduced sensitivity to system noise, although modest error amplification may occur for certain competing classes.

The dependence of test accuracy on pruning ratio is presented in Fig. 6d. Accuracy decreases gradually with increasing pruning ratio, consistent with trends observed in electronic neural networks. Analysis of sparsity reveals two regimes: sparsity increases from 15.39% to 20.7% as pruning rises to 10%, due to the removal of feature columns with predominantly non-zero weights, and then decreases to 8.75% at 60% pruning as remaining columns become denser. By removing entire feature dimensions rather than individual weights, this column-wise pruning strategy reduces the number of active wavelength channels, mitigating cross-talk and intensity fluctuation effects in the MoS_2 photodetector array while maintaining reliable optical inference.

Discussion

This work demonstrates a microcavity-integrated MoS_2 photodetector platform that enables wavelength-aware, memory-enabled optoelectronic inference through device-level physical mechanisms. By embedding spectral selectivity and temporal accumulation within individual photodetector pixels, sensing, weighting, and accumulation are intrinsically unified, eliminating the need for interferometric routing, coherent phase control, or external electronic summation. Compared with conventional optical and optoelectronic neural architectures, this approach offers improved robustness, architectural simplicity, and scalability for compact vision hardware. The wavelength-time multiplexed inference scheme leverages microcavity-defined responsivity for passive spectral weighting, while intrinsic carrier dynamics in MoS_2 provide reliable temporal accumulation of optical inputs. Close agreement between optical hardware outputs and electronic reference models across multiple datasets confirms that device-level physics can be quantitatively integrated into functional machine-learning pipelines without sacrificing inference fidelity. Furthermore, the introduction of optical Hessian pruning highlights the importance of algorithm-hardware co-design in physical neural systems. By suppressing noise-sensitive feature dimensions, pruning improves robustness to optical nonidealities while reducing effective system complexity and optical resource usage.

The latency of the present time-multiplexed experiment is mainly limited by the laboratory supercontinuum-AOTF modulation setup rather than by the intrinsic wavelength-selective detection mechanism. Future implementations using compact high-speed wavelength-tunable sources or integrated modulators could shorten the pulse-sequence duration and reduce inference latency. In the present time-multiplexed implementation, the number of distinguishable wavelength channels is governed by the usable responsivity range and the linewidth of the wavelength-selective source, rather than by the cavity FWHM alone. Narrower-linewidth tunable sources could therefore enable finer responsivity sampling and higher effective weight precision without changing the photodetector structure.

In conclusion, microcavity-enhanced MoS_2 photodetectors emerge as a scalable and energy-efficient platform for wavelength-programmable optoelectronic inference. More broadly, this work illustrates how resonant nanophotonics and semiconductor carrier dynamics can transform photodetectors from passive sensors into compact computational primitives for edge and in-sensor intelligence.

Methods

Device fabrication

The device fabrication process began with the definition of the bottom gate/mirror via laser direct writing (LDW) using an ultraviolet maskless lithography system (TuoTuo Technology, TTT-07-UV Litho-S+) on a substrate consisting of 275 nm SiO_2 thermally grown on heavily doped silicon. This was followed by electron-beam evaporation (EBE) of a 10/100 nm Ti/Ag bilayer and subsequent lift-off. A 72 nm HfO_2 dielectric layer was then deposited by thermal atomic layer deposition (ALD) using tetrakis (dimethylamido) hafnium (TDMAHf) and H_2O as precursors.

Next, a single-crystal MoS_2 film—synthesized on a custom-designed C/A-plane sapphire substrate (as reported in Tao et al.[29])—was transferred onto the target substrate using a polymer-assisted method. Specifically, polymethyl methacrylate (PMMA, 950 K, A4) was spin-coated onto the MoS_2/sapphire substrate at 2000 rpm and annealed at 150 °C. A thermal release tape (TRT) was then laminated onto the PMMA surface and used to mechanically delaminate the PMMA/MoS_2 stack from the sapphire substrate in a KOH etching solution. The resulting TRT/PMMA/MoS_2 assembly was subsequently transferred onto the pre-patterned device substrate inside a nitrogen-purged glovebox and baked at 120 °C to release the tape. The PMMA support layer was finally removed by immersion in acetone.

The MoS_2 active channel was then patterned via LDW and etched into an array of discrete blocks using reactive ion etching. Source and drain contacts were formed by patterning 15/30 nm Sb/Au electrodes via LDW, followed by EBE and lift-off. To enable high-quality HfO_2 deposition directly on MoS_2, a 1 nm Al seed layer was first deposited and allowed to undergo natural oxidation. Subsequently, an additional 72 nm of HfO_2 was deposited using the same ALD process described earlier. Finally, a 40 nm Ag top mirror was defined by LDW, deposited by EBE, and lifted off using the same procedure employed for the bottom mirror, thereby completing the Fabry-Perot microcavity structure. Material characterization of the as-grown MoS_2 film is provided in Supplementary Fig. 2.

Device characterization

Optoelectronic characterization of the microcavity-enhanced MoS_2 photodetector was performed using a scanning photoelectric test system (Metatest, MStarter 200), which operates under a base vacuum pressure of approximately 10^{-3} Pa. A Keithley 2612B source meter was used to apply gate and drain biases and to measure the resulting drain current. The setup is integrated with a xenon arc lamp, enabling standard optoelectronic measurements and facilitating the acquisition of wavelength-dependent photocurrent responses.

For optical classification experiments, the system was further equipped with a supercontinuum laser coupled to an acousto-optic tunable filter. This configuration enables rapid wavelength switching and precise control of incident light intensity.

Optical Hessian pruning

Hessian pruning was applied post training because the optical hardware performs inference with fixed, pre-trained weights. Let $L(\mathbf{w})$ denote the training objective and $\boldsymbol{H} = \nabla^2 L(\mathbf{w})$ the Hessian. Around a converged solution, a second-order Taylor approximation indicates that removing the q-th parameter ($\mathbf{w}_q \to 0$) leads to an approximate loss increase $s_q = \frac{1}{2}\boldsymbol{H}_{qq}\mathbf{w}_q^2$ under the common diagonal curvature approximation. In our implementation, these curvature-weighted saliency scores are computed and aggregated per feature dimension (i.e., per column of the fully connected weight matrix), enabling column-wise structured pruning that directly corresponds to disabling wavelength-encoded feature channels in the MoS_2 photodetector array. Columns with low importance are removed, yielding a lightweight model characterized by a pruning coefficient α. If the original number of weights is NK, the number of retained weights after pruning is αNK. Implementation details are provided in Supplementary Note 3.

Data availability

The MNIST data used in this study are available from the MNIST database[34,35] [http://yann.lecun.com/exdb/mnist]. The CIFAR-10 data used in this study are available from the official CIFAR-10 dataset[36] [http://www.cs.toronto.edu/~kriz/cifar.html]. The FSDD[37] data used in this study are available from Zenodo [https://zenodo.org/records/1342401]. Source data are provided with this paper.

Code availability

The source code of this study is available in the repository of the [https://github.com/zouningmu/WE_Neuromorphic_Inference].

References

1. Lin, X. et al. All-optical machine learning using diffractive deep neural networks. *Science* **361**, 1004–1008 (2018).
2. Xue, Z. et al. Fully forward mode training for optical neural networks. *Nature* **632**, 280–286 (2024).
3. Ahmed, S. R. et al. Universal photonic artificial intelligence acceleration. *Nature* **640**, 368–374 (2025).
4. Hua, S. et al. An integrated large-scale photonic accelerator with ultralow latency. *Nature* **640**, 361–367 (2025).
5. Shen, Y. et al. Deep learning with coherent nanophotonic circuits. *Nat. Photonics* **11**, 441–446 (2017).
6. Feldmann, J. et al. Parallel convolutional processing using an integrated photonic tensor core. *Nature* **589**, 52–58 (2021).
7. Li, G. H. Y. et al. Deep learning with photonic neural cellular automata. *Light Sci. Appl.* **13**, 283 (2024).
8. Qin, J. et al. All-optical Fourier neural network using partially coherent light. *Chip* **4**, 100140 (2025).
9. Huang, C. et al. A silicon photonic–electronic neural network for fibre nonlinearity compensation. *Nat. Electron.* **4**, 837–844 (2021).
10. Xu, X. et al. 11 TOPS photonic convolutional accelerator for optical neural networks. *Nature* **589**, 44–51 (2021).
11. Yu, F. et al. Time-wavelength multiplexed photonic neural network accelerator for distributed acoustic sensing systems. *Adv. Photonics* **7**, 026008 (2025).
12. Fu, T. et al. Optical neural networks: Progress and challenges. *Light Sci. Appl.* **13**, 263 (2024).
13. Xu, D., Ma, Y., Jin, G. & Cao, L. Intelligent photonics: A disruptive technology to shape the present and redefine the future. *Engineering* **46**, 186–213 (2025).
14. Choi, C. et al. Human eye-inspired soft optoelectronic device using high-density MoS2-graphene curved image sensor array. *Nat. Commun.* **8**, 1664 (2017).
15. Jang, H. et al. An atomically thin optoelectronic machine vision processor. *Adv. Mater.* **32**, 2002431 (2020).
16. Wang, Q. H., Kalantar-Zadeh, K., Kis, A., Coleman, J. N. & Strano, M. S. Electronics and optoelectronics of two-dimensional transition metal dichalcogenides. *Nat. Nanotechnol.* **7**, 699–712 (2012).
17. Wang, S. et al. Tellurium nanowire retinal nanoprosthesis improves vision in models of blindness. *Science* **388**, eadu2987 (2025).
18. Hou, X. et al. High-performance Ga_2 O_3 in-memory DUV photodetectors by interface charge reservoir design for multifunctional applications. *Adv. Mater.* **37**, 2506179 (2025).
19. Wang, H. et al. High-sensitivity and fast-response solar-blind photodetectors via band offset engineering for motion tracking. *Nat. Commun.* **16**, 8175 (2025).
20. Tong, L. et al. 2D materials–based homogeneous transistor-memory architecture for neuromorphic hardware. *Science* **373**, 1353–1358 (2021).
21. Huang, P.-Y. et al. Neuro-inspired optical sensor array for high-accuracy static image recognition and dynamic trace extraction. *Nat. Commun.* **14**, 6736 (2023).
22. Jayachandran, D. et al. A low-power biomimetic collision detector based on an in-memory molybdenum disulfide photodetector. *Nat. Electron.* **3**, 646–655 (2020).
23. Liao, F. et al. Bioinspired in-sensor visual adaptation for accurate perception. *Nat. Electron.* **5**, 84–91 (2022).
24. Chen, J. et al. Optoelectronic graded neurons for bioinspired in-sensor motion perception. *Nat. Nanotechnol.* **18**, 882–888 (2023).
25. Zhang, D. et al. Spectral kernel machines with electrically tunable photodetectors. *Science* **390**, eady6571 (2025).
26. Yang, J. et al. Cavity-enhanced near-infrared organic photodetectors based on a conjugated polymer containing [1,2,5]selenadiazolo[3,4- c]pyridine. *Chem. Mater.* **33**, 5147–5155 (2021).
27. Wang, J. et al. Organic cavity photodetectors based on nanometer-thick active layers for tunable monochromatic spectral response. *ACS Photonics* **6**, 1393–1399 (2019).
28. Mennel, L. et al. Ultrafast machine vision with 2D material neural network image sensors. *Nature* **579**, 62–66 (2020).
29. Li, T. et al. Epitaxial growth of wafer-scale molybdenum disulfide semiconductor single crystals on sapphire. *Nat. Nanotechnol.* **16**, 1201–1207 (2021).
30. Zou, X. et al. Robust epitaxy of single-crystal transition-metal dichalcogenides on lanthanum-passivated sapphire. *Science* **390**, eaea0849 (2025).
31. Sun, Q. et al. Designable excitonic effects in van der Waals artificial crystals with exponentially growing thickness. *Nat. Commun.* **16**, 2712 (2025).
32. Li, W. et al. Approaching the quantum limit in two-dimensional semiconductor contacts. *Nature* **613**, 274–279 (2023).
33. Rizzo, D. J. et al. Molecular plasmonic cavities. *Nano Lett.* **25**, 14043–14050 (2025).
34. Deng, L. The MNIST database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Process. Mag.* **29**, 141–142 (2012).
35. LeCun, Y. et al. Backpropagation applied to handwritten zip code recognition. *Neural Comput* **1**, 541–551 (1989).
36. Krizhevsky, A. Learning multiple layers of features from tiny images. in (2009).
37. Jackson, Z. et al. Jakobovski/free-spoken-digit-dataset: v18. Zenodo https://doi.org/10.5281/ZENODO.1342401 (2018).
38. Cover, T. M. & Thomas, J. A. *Elements of Information Theory*. (Wiley, 2001). https://doi.org/10.1002/0471200611.
39. Hassibi, B. & Stork, D. Second order derivatives for network pruning: Optimal Brain Surgeon. in *Advances in Neural Information Processing Systems* (eds Hanson, S., Cowan, J. & Giles, C.) vol. 5 (Morgan-Kaufmann, 1992).
40. Singh, S. P. & Alistarh, D. WoodFisher: Efficient Second-Order Approximation for Neural Network Compression. in *Advances in Neural Information Processing Systems* (eds Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F. & Lin, H.) vol. 33 18098–18109 (Curran Associates, Inc., 2020).

41. Frantar, E. & Alistarh, D. Optimal brain compression: A framework for accurate post-training quantization and pruning. in *Advances in Neural Information Processing Systems* (eds Koyejo, S. et al.) vol. 35 4475–4488 (Curran Associates, Inc., 2022).
42. Ning, H. et al. An index-free sparse neural network using two-dimensional semiconductor ferroelectric field-effect transistors. *Nat. Electron.* **8**, 222–234 (2025).
43. Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).

Acknowledgements
The authors thank Jiangsu Innovation Teams and the AI & AI for Science Project of Nanjing University for their support.

Author contributions
N.Z., Z.H., X.C., and K.D. jointly conceived the idea. X.C., X.P., and L.S. contributed to the photodetector fabrication. T.L. and S.Z. performed the CVD growth of MoS_2. K.D., F.Y., S.C.(Silin Chen), and A.L. assisted with the theory. K.D., X.C., J.L., L.S., and S.C.(Shengdi Chen) performed the experiments. K.D. and X.C. visualized the data. Y.S., X.W., X.Z., Z.H., L.G., and N.Z. supervised and coordinated all the work. K.D., Z.H., X.C., and N.Z. wrote the original manuscript. X.C., X.Z., X.W., L.G., W.J., Z.H., and N.Z. revised and edited the manuscript. All authors contributed to the discussions.

Funding
Z.H. acknowledges support by the National Key Research and Development Program of China (Grant No. 2022YFA1204301), the National Natural Science Foundation of China (Grant Nos. 62522405 and 62405132), and the Natural Science Foundation of Jiangsu Province (Grant No. BK20241227). N.Z. acknowledges support by the Fundamental Research Funds for the Central Universities (Grant No. 2024300421) and the National Natural Science Foundation of China (Grant No. 62341408). L.G. and X.W. acknowledge support by the National Natural Science Foundation of China (Grant No. U24A20295) and the Fundamental Research Funds for the Central Universities (Grant No. 4830-14380006). L.G. acknowledges support by the National Natural Science Foundation of China (Grant No. 62288102), the National Key Research and Development Program of China (Grant No. 2021YFA1202904), the Major Science and Technology Special Project of Jiangsu Province (Grant No. BG2024017), and the Basic Research Program of Jiangsu Province (Grant No. BK20243057). W.J. acknowledges support by the National Natural Science Foundation of China (Grant Nos. 62175103 and 61775094). Y.Z. acknowledges support by the Fundamental Research Funds for the Central Universities (Grant No. 0213-14380265).

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-75808-w.

Correspondence and requests for materials should be addressed to Li Gao, Zehua Hu or Ningmu Zou.

Peer review information *Nature Communications* thanks Peng Huang, Gunuk Wang and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.