Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Polarization-coupled complete operators in sliding ferroelectric semiconductors for native in-memory attention computing

Xing Wu

hkxlwu@nju.edu.cn

Nanjing University https://orcid.org/0000-0002-2787-3069

di wang

https://orcid.org/0009-0009-0269-0539

Yukang Liu

Nanjing University

Jiabei Huang

Nanjing University https://orcid.org/0009-0007-1683-237X

Zijing Guo

Nanjing University

Yinuo Zou

Nanjing University

Tianyun Wang

Suzhou University of Science and Technology, China

Liuyang Sun

Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences
https://orcid.org/0000-0001-9896-4382

Xiufang Sun

Nanjing University Hospital

Zhiyong Zhang

Nanjing University

Dongjun Wei

ESCP Business School, Paris

Geliang Yu

Nanjing University https://orcid.org/0000-0002-0169-7781

Hongyi Wu

HKU Business School

Jinlei Zhang

Suzhou University of Science and Technology

Article

Keywords:

Posted Date: September 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10824719/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Abstract

Attention computing, central to Transformer-based generative AI, calls for native in-memory hardware with linear and nonlinear operators intrinsically coupled. Current compute-in-memory (CIM), however, remains tailored to linear algebra, lacking integrated nonlinear operators. Here, we demonstrate complete-operator hardware that intrinsically couples both operator types in situ via ferroelectric polarization. Our device integrates a vertical ferroelectric tunnel junction for linear multiply-accumulate operations and a lateral ferroelectric semiconductor field-effect transistor for exponential functions in sliding-ferroelectric semiconductor ReS_2. Tuning lateral transport length and vertical barrier thickness yields designable exponential operators with tunable sensitivity and operating range. A fabricated (9×3)×9 array realizes the first in-memory hardware execution of full Transformer attention primitives, including intrinsic Softmax, with device-level 600-ps latency and 320-aJ/operation energy. A closed-loop hardware prototype enables real-time text generation, with experimentally calibrated simulations supporting neural-signal-to-language scalability. This work embeds the missing nonlinearity directly into memory, advancing CIM toward complete, native hardware for generative AI.

Introduction

The rapid evolution of generative artificial intelligence (AI) imposes escalating demands on the computational capacity and energy efficiency of attention computing[1,2]. Compute-in-memory (CIM) architectures have emerged as a promising route around von Neumann bottlenecks by performing matrix-vector multiplications (MVM) directly within memory arrays[3-5]. However, current CIM architectures can only support linear operations because memory arrays are governed by Kirchhoff's law[6-13]. Hardware support for essential nonlinear operators, such as exponential and logarithmic functions, remains absent in these architectures. This gap severely restricts the ability of CIM hardware to support full attention computing in generative AI[2,14-18]. Therefore, an in-situ hardware realization of a complete operator architecture (COA) is urgently needed. Such a framework should natively co-realize linear and nonlinear operators within a single unit cell.

To achieve COA, a robust nonlinear process alone is insufficient, as strong and stable coupling between linear and nonlinear operators is also essential. Nonlinear responses originating from defect migration can be exploited for feature extraction, but the inherent instability of defect-mediated mechanisms precludes their reliable use as general nonlinear operators[19,20]. Moreover, these nonlinear responses are difficult to physically register with linear conductance states, resulting in weak coupling and hindering the realization of COA. Other approaches utilize peripheral circuits to implement HardSigmoid activations, compensating for the lack of nonlinear functionality[2,13,21]. While enabling attention computing, these approaches are still limited to specific activation functions. Ferroelectric polarization is nonvolatile and electrically switchable, making it well suited for programmable operator coupling in ferroelectric semiconductors[22]. Driven by these constraints, we use ferroelectric polarization as this shared state variable to control both in-plane transport and out-of-plane tunneling in a ferroelectric

semiconductor. In this device, the vertical ferroelectric tunnel junction (FTJ) performs linear MVM through polarization-modulated quantum tunneling, consistent with Kirchhoff's law. Simultaneously, the same remanent polarization regulates lateral transport in a ferroelectric semiconductor field-effect transistor (FeS-FET), providing a parallel output branch for in-situ linear-nonlinear operator coupling. This COA, with intrinsic linear and nonlinear functionalities, offers a fundamental hardware building block for unified in-memory attention computing in generative AI.

Here, we fabricate the first COA device integrating a vertical FTJ and a lateral FeS-FET in sliding-ferroelectric ReS_2. The device exhibits 10^5-cycle endurance and nonvolatile retention exceeding one year. It supports in situ linear operators with up to 128 FeS-FET and 64 FTJ conductance states, alongside nonlinear operators for exponential computing. By tuning the lateral transport length and vertical barrier thickness, we establish a designable exponential-operator space with structurally controlled sensitivity and operating range. A (9×3)×9 COA array demonstrates the first in-memory hardware execution of Transformer attention primitives, including intrinsic Softmax. The array exhibits picosecond-scale device latency and attojoule-scale operation energy. This physical integration reduces the operator-level overhead of ADC-intensive peripheral circuits and external nonlinear modules. Experimentally calibrated simulations further support its scalability toward neural-signal-to-language generation from mouse local field potential (LFP) inputs. This transition from conventional linear accelerators to native complete-operator hardware establishes a device-level design paradigm for next-generation generative AI hardware.

Polarization-coupled operators for complete in-memory computing

To realize this device-level paradigm, linear and nonlinear operators must be physically coupled in situ within a unit cell, rather than implemented as functionally separated computational stages. As conceptually illustrated in Fig. 1a, crossbar arrays of FTJ devices naturally realize linear matrix operations through out-of-plane quantum tunneling modulated by ferroelectric polarization. Simultaneously, nonlinear transformations can be mapped onto the in-plane transport characteristics of the FeS-FET, which are also determined by the same ferroelectric polarization node. Therefore, this polarization-coupled ferroelectric semiconductor architecture provides the physical building blocks for complete linear and nonlinear operations, giving rise to a COA.

Guided by this design principle, ReS_2, a semiconductor with high carrier mobility and sliding ferroelectricity, is employed as both the ferroelectric tunnel barrier and semiconductor channel to realize complete operators (Fig. 1b). At each array site of the COA, a vertical FTJ and a lateral FeS-FET are integrated and share a common ferroelectric polarization state. In this geometry, this shared polarization state simultaneously modulates the vertical tunneling barrier and the lateral carrier density. Crucially, the FTJ resistance is exponentially sensitive to the polarization-modulated tunneling barrier, whereas the FeS-FET conductance scales proportionally with the polarization-induced carrier density[23–26]. This intrinsic asymmetry between the two responses underlies the nonlinear cross-terminal mapping. Because both the vertical and lateral transport mechanisms are governed by this single physical variable,

linear and nonlinear operators are intrinsically coupled within a single unit cell. As a result, complete operator computing can be realized directly in unit-cell-level COA hardware.

To quantitatively validate this coupling, we calculated the resistive mappings between the two branches using density functional theory (DFT). As shown in Fig. 1c, the calculated FTJ and FeS-FET resistance states form well-defined exponential cross-terminal mappings (Supplementary Fig. 1 and Supplementary Note. 1)[23–26]. This theoretical result confirms that the intrinsic transport asymmetry gives rise to an exponential cross-terminal function. Because both tunneling and transport processes are regulated by a single polarization state, exponential computing can be directly realized by controlling the ferroelectric polarization using electrical pulses. Moreover, a family of exponential operators can be systematically tuned by varying the lateral transport length and vertical barrier thickness (Fig. 1c). This provides a deterministic physical foundation for designable complete-operator computing.

Mechanically exfoliated ReS_2 flakes with a non-centrosymmetric Td phase are prepared, as shown in Fig. 1d. ReS_2 is a typical layered semiconductor known for its high carrier mobility and large on/off switching ratio. Complementary Raman spectroscopy and atomic force microscopy confirm the crystallographic phase and thickness uniformity of the ReS_2 flakes (Supplementary Figs. 1 and 2). Using piezoresponse force microscopy (PFM), we characterize ferroelectric properties in various ReS_2 samples, including twisted bilayers with a twist angle of 0.6°, non-twisted bilayers, and few-layer ReS_2[27–31]. All samples show clear ferroelectric hysteresis and domain structures, confirming the general sliding ferroelectricity of ReS_2 (Fig. 1e,f and Extended Data Fig. 1). The sliding ferroelectricity observed in non-twisted ReS_2 layers of varying thicknesses allows compatibility with both tunneling and transport processes, without introducing defects or requiring complex stacking procedures. These results thus position sliding-ferroelectric ReS_2 as a physically consistent platform for constructing a polarization-coupled COA capable of realizing complete operators.

Experimental realization of polarization-coupled complete operators

The COA is experimentally realized using a dual-channel ferroelectric semiconductor unit cell, where a lateral FeS-FET and a vertical FTJ are co-integrated within the same ReS_2 ferroelectric medium (Extended Data Fig. 2). Because the vertical FTJ and lateral FeS-FET share the same ferroelectric polarization state, the two channels can be programmed by the same ultrafast voltage pulse[16,32–34]. Figure 2a shows the R-V_p hysteresis loops of the FTJ and FeS-FET measured under identical voltage pulses with a duration (t_d) of 600 ps. The resulting loops exhibit synchronous switching behavior in response to the applied pulses and demonstrate continuously tunable resistance states. A + 7 V pulse, corresponding to downward ferroelectric polarization, simultaneously programs the low-resistance states (LRS) by increasing tunneling probability and accumulating channel carriers. Conversely, a −7 V pulse, corresponding to upward ferroelectric polarization, switches both channels to high-resistance states (HRS). After programming, these continuous conductance states can be used as linear weights.

Under read bias, Kirchhoff's law enables independent linear multiply–accumulate operations along either transport direction.

Continuous-variable resistive states can be obtained by adjusting the pulse amplitude. Specifically, 128 distinct and stable resistive states in the FeS-FET and 64 stable resistive states in the FTJ are achieved without degradation over 10^3 s (Fig. 2b,c). Long-term retention measurements further confirm that the programmed resistive states can be retained for more than one year (Supplementary Fig. 3). The stability and modulation tolerance of these resistive states originate from ferroelectric polarization, a distinctive property that is absent in nonlinear responses governed by defects and thermal effects[19,20]. The two channels show synchronous, stable switching at a low voltage of ± 1.5 V with a duration of 20 ns (Fig. 2d), persisting even under ultrafast 600 ps pulses (Fig. 2e). This robust synchronization is further confirmed at ± 3 V across various pulse durations from 100 ns down to 600 ps (Extended Data Fig. 3). Remarkably, the switching energy of the COA is as low as 320 aJ, highlighting its strong potential for implementing energy-efficient in-memory linear operators in integrated circuits. Crucially, since a single pulse simultaneously modulates both channels, this ultralow-energy programming writes the linear weights while setting the paired FTJ/FeS-FET states used for the subsequent cross-terminal nonlinear operation.

Building on this intrinsic coupling, nonlinear operators can be realized through cross-terminal operations, as shown in Fig. 2f. Under varying programming pulse amplitudes applied through the FTJ terminal, the paired resistive states of the FTJ and FeS-FET form deterministic nonlinear mappings characterized by exponential functions. Exponential computing is achieved with an ultralow energy consumption of 5.12 fJ, an ultrafast device-level latency of 600 ps, and a calibrated mapping accuracy of 99.99%, consistent with the theoretical prediction in Fig. 1c. Thus, complete operator hardware is realized for the first time through polarization-mediated coupling between tunneling and lateral transport. This COA, which features deterministic device-level coupling rather than relying on external synchronization or peripheral circuitry, provides a crucial physical foundation for native in-memory Transformer attention hardware.

Geometric designability of exponential operators

For COA hardware, practical attention computation requires exponential operators to be designable across different response sensitivities and operating ranges. We therefore use two geometric tuning routes: lateral transport length to control local sensitivity and vertical barrier thickness to expand the global operating range. First, at a fixed ReS_2 thickness, increasing the channel spacing of the FeS-FET systematically modulates its resistive states, while leaving the HRS and LRS of the FTJ nearly unchanged (Fig. 3a,b). The lateral FeS-FET resistance contains a transport-length-dependent channel component, allowing spacing tuning to selectively rescale the FeS-FET output within a fixed FTJ weight window. This geometry-defined rescaling therefore enhances the nonlinear gain of the exponential function and improves its response sensitivity to FTJ weight variations (Fig. 3c).

A broadly designable exponential-operator family requires not only local sensitivity tuning but also global span control. Vertical barrier-thickness tuning provides this global control by directly determining the polarization-modulated tunneling path and barrier profile in the FTJ. It expands the FTJ weight window, while the FeS-FET response varies only weakly in comparison because the FTJ resistance is exponentially sensitive to the vertical tunneling barrier. As shown in Fig. 3d,e, thickness tuning modifies the paired FeS-FET and FTJ states, with the dominant change appearing as a broadened FTJ weight range. Representative pulse measurements further confirm that paired FTJ/FeS-FET switching is preserved across the thickness series, maintaining the cross-terminal correspondence required for extracting thickness-defined exponential functions (Extended Data Fig. 4). Importantly, as the barrier thickness increases beyond the direct-tunneling limit, the exponential dependence is still retained under field-assisted vertical transport (Supplementary Note. 1)[33,34]. The vertical FTJ resistance therefore preserves an exponential cross-terminal mapping with the FeS-FET response, allowing thickness-defined exponential functions to be extracted beyond the thin-barrier limit (Fig. 3f). Thus, spacing tuning primarily controls nonlinear gain and response sensitivity, whereas thickness tuning expands the global operating range. Together, these two tuning routes convert the fixed cross-terminal exponential function into a designable family of exponential operators. Ultimately, this geometric customizability yields exponential operators with tailored nonlinear gains and weight ranges for array-level attention execution.

COA array enables native in-memory attention execution

Transformer attention requires linear MVM and nonlinear Softmax normalization, which are typically separated between memory arrays and digital peripherals. By leveraging the polarization-coupled linear FTJ and nonlinear FeS-FET responses within each cell, our COA array directly overcomes this obstacle. We implemented a (9×3)×9 COA array as the core attention-computing hardware (Fig. 4a and Extended Data Fig. 5). This array demonstrates the first in-memory hardware execution of Transformer attention primitives. It retains spatially uniform 600-ps writing and 10^5-cycle write-read endurance across its FTJ and FeS-FET cells, supporting repeated fast updates for hardware Softmax and in situ training (Extended Data Fig. 6).

In this array, linear MVM—the core of query, key, and value (QKV) projection — is performed in situ by the continuous-variable conductance states of the vertical FTJs. Input word tokens are encoded as voltage vectors applied to the rows, and the weighted charge accumulation along the columns directly computes the context-dependent dot products (Supplementary Figs. 4 and 5). Crucially, the nonlinear Softmax step —the exponential operation for attention normalization—is generated by the laterally coupled FeS-FET. Its channel current responds exponentially to the very same polarization state that defines the FTJ conductance. The resulting exponential currents are summed and normalized in the hardware current domain, realizing a hardware Softmax primitive without digital conversion in the core attention path. Under simultaneous activation of one to nine COA cells, the cross-terminal mapping error remains low and bounded, showing no scaling with the number of active cells (Extended Data Fig. 7). This behavior directly enables error-resilient, in-array probability normalization. The resulting attention-weighted

outputs are subsequently read as column currents, completing a Transformer attention block within the memory fabric.

Remarkably, this array supports in situ hardware training, with rapid convergence of the cross-entropy loss. The sentence-reasonableness score also progressively increases toward unity, indicating improved semantic consistency in the generated action-object pairs (Fig. 4b). The COA array delivers substantial operator-level performance gains by reducing the major overheads of conventional architectures (Supplementary Fig. 6a). Estimated latency and energy analyses based on measured device parameters indicate picosecond latency and attojoule-scale energy per operation, outperforming analog CIM with its peripheral ADC bottleneck by one to two orders of magnitude in both metrics[3–18]. The advantage is even more pronounced for the nonlinear Softmax function. Here, the intrinsic exponential response arises directly from polarization-coupled tunneling and transport under ferroelectric switching. This establishes an efficiency gap that originates from the physical integration of operators, not merely circuit-level optimization, whereas CIM implementations rely on digital arithmetic or look-up tables for nonlinear functions (Extended Data Fig. 8).

To validate functional completeness, we integrated the COA array with a lightweight FPGA controller and a display interface to form a closed hardware inference prototype (Supplementary Figs. 7 and 8)[35]. All core computations—QKV transformations and in situ Softmax normalization—are executed within the COA array; the FPGA only handles token sequencing and interface timing. When prompted with the input phrase “I am hungry”, the system computes attention-based inference and generates the continuation “need pizza now”, displaying it in real time (Fig. 4c). Additional examples across emotional contexts confirm the robustness of this hardware generation (Extended Data Fig. 9). This demonstration shows that the COA architecture sustains a complete attention-driven generative loop within memory, transitioning CIM from an accelerator of isolated linear layers to native complete-operator inference hardware.

Building on the hardware validation of the COA attention module, we further evaluated whether the experimentally calibrated COA operator set can scale from the small-array hardware prototype to a simulated neural-signal-to-language task. To this end, we constructed a 1024×1024 COA-array simulation using measured device parameters and applied it to neural-signal-to-language generation[36–39]. In this framework, multichannel mouse LFP traces from four affective states, Calm, Hungry, Excited and Fearful, are projected onto the simulated COA array for emotion-conditioned language generation (Fig. 4d and Supplementary Fig. 9)[40]. Linear MVMs extract spatiotemporal features, while the intrinsic exponential response performs nonlinear compression and probabilistic encoding. This process yields a compact emotion state vector E, which encodes both the discrete affective category and its continuous intensity. The simulated system achieves accurate emotion-state discrimination across the four LFP classes (Fig. 4e). The vector **E** then drives a COA-based Transformer decoder, producing graded language outputs that preserve categorical semantics while varying with affective intensity (Fig. 4f and Supplementary Fig. 10). Additional text and speech generation results are provided in Extended Data

Fig. 10. Estimated latency and energy analyses for this simulated pipeline are summarized in Supplementary Fig. 6b and benchmarked against conventional CIM implementations. These simulations indicate that the complete-operator paradigm can scale beyond isolated hardware functions toward complex sensory-to-language tasks, providing a physics-based foundation for efficient real-world generative AI.

Conclusion

In conclusion, we have demonstrated a complete-operator architecture (COA) that overcomes the long-standing nonlinearity barrier in compute-in-memory systems. By exploiting polarization-mediated coupling in the sliding ferroelectric semiconductor ReS_2, we monolithically integrate linear and nonlinear operators within a single four-terminal device (FTJ/FeS-FET). This intrinsic fusion enables Transformer attention primitives, including linear matrix operations and hardware Softmax, to be executed natively in memory with picosecond-scale latency (600 ps) and attojoule-level energy efficiency (320 aJ). Validated from unit cells to a functional array and extended through experimentally calibrated large-scale simulation, the COA supports scalable complete-operator computing toward complex, cross-modal generative tasks. Beyond accelerating isolated layers, this work establishes a new hardware paradigm in which computational completeness is embedded at the physical level of materials and devices, paving the way for energy-efficient complete-operator AI hardware at the edge.

Declarations

Acknowledgements

This work was supported by the NSF of Jiangsu Province (No. BK20243062) and the National Natural Science Foundation of China (Nos. 12274210, 62227820 and 62531009). Partial support is from the Fundamental Research Funds for the Central Universities (0210/14380228).

Author Contributions: D.W. and Y.K.L designed the experimental setup, conducted the experiments, analyzed the data, and co-wrote the manuscript. J.B.H prepared the samples and analyzed the data. Z.J.G fabricated the devices and provided the electrodes. Y.N.Z performed the behavioral experiments and data analyses and polished the manuscript. T.Y.W conducted the experiments. L.Y.S provided the mouse LFP signals. X.F.S performed the mouse maintenance and electrode fixation. Z.Y.Z performed the conductance calculations. D.J.W performed AI calculation. G.L.Y provides the experimental equipment. H.Y.W analyzed the data and performed AI computations. J.L.Z suggested and supported the electrical measurements of the device, analyzed the data, and reviews the manuscript. X.L.W conceived the project, analyzed the data, and co-wrote the manuscript. D.W, Y.K.L, and J.B.H contributed equally to this work.

Funding: This work was supported by NSF of Jiangsu Province (No. BK20243062), National Natural Science Foundation of China (Nos. 12474032, 62227820, and 62531009), andFundamental Research

Funds for the Central Universities (0210/14380228).

Competing interests The authors declare no competing interests.

Correspondence and requests for materials should be addressed to Xinglong Wu,

References

1. Vaswani A et al (2017) Curran Associates,. Attention is all you need. In Proc. 31st International Conference on Neural Information Processing Systems, NIPS' 17 6000–6010
2. Leroux N, Manea PP, Sudarshan C et al (2025) Analog in-memory computing attention mechanism for fast and energy-efficient large language models. Nat Comput Sci 5:813–824
3. Xia Q, Yang JJ (2019) Memristive crossbar arrays for brain-inspired computing. Nat Mater 18:309–323
4. Singh A, Le Gallo M, Vasilopoulos A et al (2025) The design of analogue in-memory computing tiles. Nat Electron 8:1156–1169
5. Huang Y, Ando T, Sebastian A et al (2024) Memristor-based hardware accelerators for artificial intelligence. Nat Rev Electr Eng 1:286–299
6. Yao P, Wu H, Gao B et al (2020) Fully hardware-implemented memristor convolutional neural network. Nature 577:641–646
7. Aguirre F, Sebastian A, Le Gallo M et al (2024) Hardware implementation of memristor-based artificial neural networks. Nat Commun 15:1974
8. Rasch MJ, Carta F, Fagbohungbe O et al (2024) Fast and robust analog in-memory deep neural network training. Nat Commun 15:7133
9. Ning H, Yu Z, Zhang Q et al (2023) An in-memory computing architecture based on a duplex two-dimensional material structure for in situ machine learning. Nat Nanotechnol 18:493–500
10. Sebastian A, Le Gallo M, Khaddam-Aljameh R et al (2020) Memory devices and applications for in-memory computing. Nat Nanotechnol 15:529–544
11. Lanza M, Pazos S, Aguirre F et al (2025) The growing memristor industry. Nature 640:613–622
12. Zhou J, Li W, Chen Y et al (2025) Optoelectronic polymer memristors with dynamic control for power-efficient in-sensor edge computing. Light Sci Appl 14:309
13. Choi Y, Jeong S, Jeong H et al (2025) Advanced AI computing enabled by 2D material-based neuromorphic devices. npj Unconv Comput 2:8
14. Jain S, Li S, Zheng H et al (2025) Heterogeneous integration of 2D memristor arrays and silicon selectors for compute-in-memory hardware in convolutional neural networks. Nat Commun 16:2719
15. Hong H, Du Z, Jiang M et al (2025) Memristor-based adaptive analog-to-digital conversion for efficient and accurate compute-in-memory. Nat Commun 16:9749

16. Schranghamer TF, Pannone A, Kumar JM et al (2025) Large-scale crossbar arrays based on three-terminal MoS_2 memtransistors. Nat Commun 16:9518
17. Du N, Polian I, Bengel C et al (2025) Mixed-mode in-memory computing: towards high-performance logic processing in a memristive crossbar array. Commun Eng 4:163
18. Martemucci M, Rummens F, Malot Y et al (2025) A ferroelectric–memristor memory for both training and inference. Nat Electron 8:921–933
19. Zolfagharinejad M, Büchel J, Cassola L et al (2025) Analogue speech recognition based on physical computing. Nature 645:886–892
20. Li S, Li B, Feng X et al (2021) Electron-beam-irradiated rhenium disulfide memristors with low variability for neuromorphic computing. npj 2D Mater Appl 5:1
21. Yang J, Mao R, Jiang M et al (2025) Efficient nonlinear function approximation in analog resistive crossbars for recurrent neural networks. Nat Commun 16:1136
22. Liu Y, Tang W, Zeng J et al (2025) Ferroelectric-based neuromorphic memory devices for bio-inspired computing. Nat Rev Electr Eng 2:773–787
23. Pantel D, Alexe M (2010) Electroresistance effects in ferroelectric tunnel barriers. Phys Rev B 82:134105
24. Maksymovych P et al (2009) Polarization control of electron tunneling into ferroelectric surfaces. Science 324:1421–1425
25. Si M, Saha AK, Gao S et al (2019) A ferroelectric semiconductor field-effect transistor. Nat Electron 2:580–586
26. Wang S, Liu L, Gan L et al (2021) Two-dimensional ferroelectric channel transistors integrating ultra-fast memory and neural computing. Nat Commun 12:53
27. Wan Y, Hu T, Mao X et al (2022) Room-Temperature Ferroelectricity in 1T'-ReS_2 Multilayers. Phys Rev Lett 128:067601
28. Zhang Q, Fan A, Wang Y et al (2025) Emerging frontiers in two-dimensional sliding ferroelectrics. npj 2D Mater Appl 9:76
29. Zhong T-T, Wu M-H (2020) Research progress of two-dimensional interlayer-sliding ferroelectricity. Acta Phys Sin 69:217707
30. HuangFu C, Zhou Y, Ke C et al (2024) Out-of-plane ferroelectricity in two-dimensional 1T-MoS_2 above room temperature. ACS Nano 18:14708–14715
31. Cui C, Hu W-J, Yan X et al (2018) Intercorrelated in-plane and out-of-plane ferroelectricity in ultrathin two-dimensional layered semiconductor In_2Se_3. Nano Lett 18:1253–1258
32. Kaiser MAA, Adebunmi KO, You H et al (2025) Multi-functional polymorphic memory based on 2D ferroelectric tunnel junctions. npj 2D Mater Appl 9:85
33. Berdan R, Marukame T, Ota K et al (2020) Low-power linear computation using nonlinear ferroelectric tunnel junction memristors. Nat Electron 3:259–266

34. Luo Z, Wang Z, Guan Z et al (2022) High-precision and linear weight updates by subnanosecond pulses in ferroelectric tunnel junction for neuro-inspired computing. Nat Commun 13:699
35. Xu YX et al (2024) Ferroelectric FET-based context-switching FPGA enabling dynamic reconfiguration for adaptive deep learning machines. Sci Adv 10:eadk1525
36. Liu Z, Mei J, Tang J et al (2025) A memristor-based adaptive neuromorphic decoder for brain–computer interfaces. Nat Electron 8:362–372
37. Wairagkar M, Card NS, Singer-Clark T et al (2025) An instantaneous voice-synthesis neuroprosthesis. Nature 644:145–152
38. Anderson AJ, Fernandino L, Gross WL et al (2025) Neural decoding of autobiographical mental image features with a general semantic model. Nat Commun 16:11208
39. Adikari A, Alahakoon D, Pallewela N et al (2025) Reconstructing impaired language using generative AI for people with aphasia. Sci Rep 15:40877
40. Jin S, Li JZ, Jiang LY et al (2025) Instant noninvasive near-infrared deep brain stimulation using optoelectronic nanoparticles without genetic modification. Sci Adv 11, eadt4771

Methods

Fabrication of twisted bilayer ReS_2

Monolayer and few-layer ReS_2 flakes (thickness 1.5–3.0 nm) were mechanically exfoliated from bulk crystals onto polydimethylsiloxane (PDMS) substrates. Flakes with uniform thickness and well-defined crystalline domains were selected using optical microscopy and atomic force microscopy (AFM). Twisted bilayer structures were assembled via a polymer-assisted dry-transfer method: the first ReS_2 layer was picked up using a PPC/PDMS stamp, and the second layer was aligned with a high-precision rotational stage (0.1° angular resolution). Twist angles between 1° and 8° were reproducibly achieved and calibrated by combining rotational stage readings with crystallographic orientation from the anisotropic optical contrast of ReS_2. During stacking, the release temperature was maintained between 20–120 °C to ensure clean van der Waals contact without sliding. After assembly, samples were annealed at 150 °C for 5–10 min to remove residual polymer and prevent interlayer drift. The twisted structures were transferred onto predefined electrode arrays for device fabrication[27-29].

Device fabrication

Devices were fabricated using standard dry transfer technology and electron-beam lithography (EBL), following previously established protocols. A 3×3 array design was used, with each unit containing 3×9 coupled arrays. ReS_2 and graphene flakes were exfoliated onto an oxidized Si wafer (285 nm SiO_2). For reference devices with gates, the bottom gate for each array unit was defined via EBL and Ti/Au (5/20 nm) deposition and the upper electrodes are made of graphene. ReS_2 and graphene flakes with smooth, clean surfaces and appropriate thickness were selected, then transferred onto the pre-fabricated bottom gates using dry transfer with a PPC film. The coupled array geometry was defined by EBL, with additional

etching to shape the structures. Finally, 3×9 electrode arrays contacting with graphene were created through EBL, followed by deposition of Ti/Au (5/60 nm).

Electrical characterization

Electrical measurements in this study were performed using a combination of precision source meters and high-speed pulse generators. DC and quasi-static I–V characteristics were obtained using three source-measure units: the Keithley 2400, Keithley 2410, and the Keithley 4200 parameter analyzer. The instrument was selected according to the required voltage range and measurement resolution for voltage sweeping and current readout. High-resistance and low-current measurements were conducted with a Keithley 6517B to ensure accurate characterization in the high-resistance state. High-speed switching and transient responses were evaluated using a picosecond pulse generator (Tektronix PSPL10300B), which delivered voltage pulses to the device through a 50-Ω coaxial line. Following each pulse, the device resistance was read at a small sensing voltage using the source meter to monitor resistance changes[23-26,32-34].

Hardware implementation

The hardware architecture developed in this study consists of a (9×3)×9 ferroelectric-coupled device array integrated with an FPGA control platform based on the Intel/Altera Cyclone IV EP4CE10. Row/column addressing, pulse scheduling, and readout timing are generated by on-chip FPGA logic, including a pulse-control module, row/column decoding module, ADC sampling module, and data-processing module. Write and erase operations are performed by applying voltage pulses with predefined amplitude and width through an external pulse-driving unit controlled by the FPGA. During inference, the FPGA applies a low readout voltage and acquires the drain-current signals of each column through a 12-bit high-speed ADC (AD9230). Pulse generation and amplitude tuning are handled by a custom driver circuit, while unselected cells remain floating to prevent unintended programming. All current signals are converted through a transimpedance amplifier (TI OPA657) before entering the ADC, and subsequently quantized and buffered by the FPGA[2,35].

A lightweight generative-model module implemented on the FPGA executes normalization, state updates, and token selection according to a predefined computation graph, outputting the corresponding generated token after each inference cycle. All timing-critical parameters — including pulse width, addressing window, readout delay, and voltage amplitude—are implemented as fixed hardware timing sequences on the EP4CE10 to ensure strict cycle-to-cycle reproducibility.

Software architecture

The software framework for generative tasks runs in the host-side control environment of the FPGA, handling instruction scheduling, data management, and interaction with the hardware array. The system is implemented using a hybrid Python (3.10) and C workflow: Python serves as the high-level interface for task configuration, data preprocessing, and model-parameter initialization, whereas low-level

communication and state management are executed through a C-based serial protocol. The structure of the lightweight generative model—including the embedding table, linear-mapping matrices, and normalization rules—is constructed offline via Python scripts, which produce the corresponding quantized-weight files. These quantized parameters are then converted into pulse-sequence formats compatible with the array and transmitted to the FPGA.

During inference, the software sends trigger instructions to the FPGA over USB and receives, at each inference step, the measured conductance values, intermediate computational states, and the probability distribution of the current token. All returned data are processed by a lightweight post-processing module in Python, which performs normalization, thresholding, and template-based syntactic matching, yielding the final textual output[1-2,19].

Figures

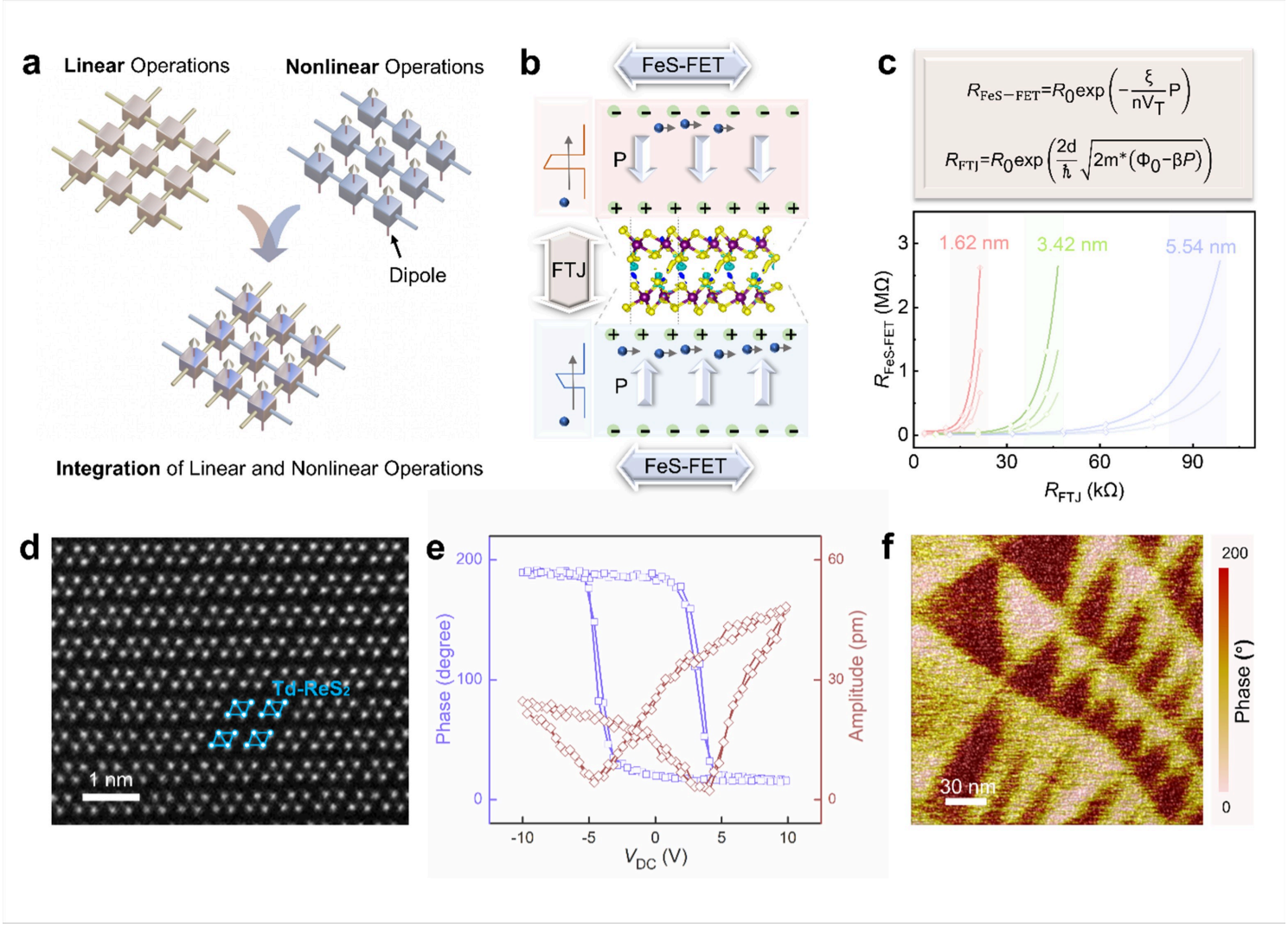


Figure 1

Mechanism of COA based on ferroelectric semiconductor ReS_2. a, In situ linear and nonlinear operations forming the complete operator. The COA is constructed by coupling linear and nonlinear operators along different transport directions through the manipulation of ferroelectric dipoles, overcoming the operator incompleteness commonly found in CIM systems. **b,** A schematic of a COA device utilizing the same ferroelectric semiconductor ReS_2 channel. In the vertical direction, the ferroelectric polarization modulates quantum tunneling in the FTJ, forming a linear operator. In the lateral direction, the ferroelectric polarization also controls the carrier density in the FeS-FET. The paired resistive states of the FeS-FET and FTJ form an exponential cross-terminal mapping, demonstrating a write-programmed and read-extracted nonlinear operator. **c,** Theoretical nonlinear operator for exponential computation under different ReS_2 thicknesses and lateral channel spacings, derived from polarization-dependent quantum tunneling and transport models. **d,** Cs-TEM image of non-centrosymmetric Td-phase ReS_2. **e,** PFM hysteresis loops in bilayer ReS_2 with a twist angle of 0.6°. **f,** Ferroelectric domain in bilayer ReS_2 with a twist angle of 0.6°.

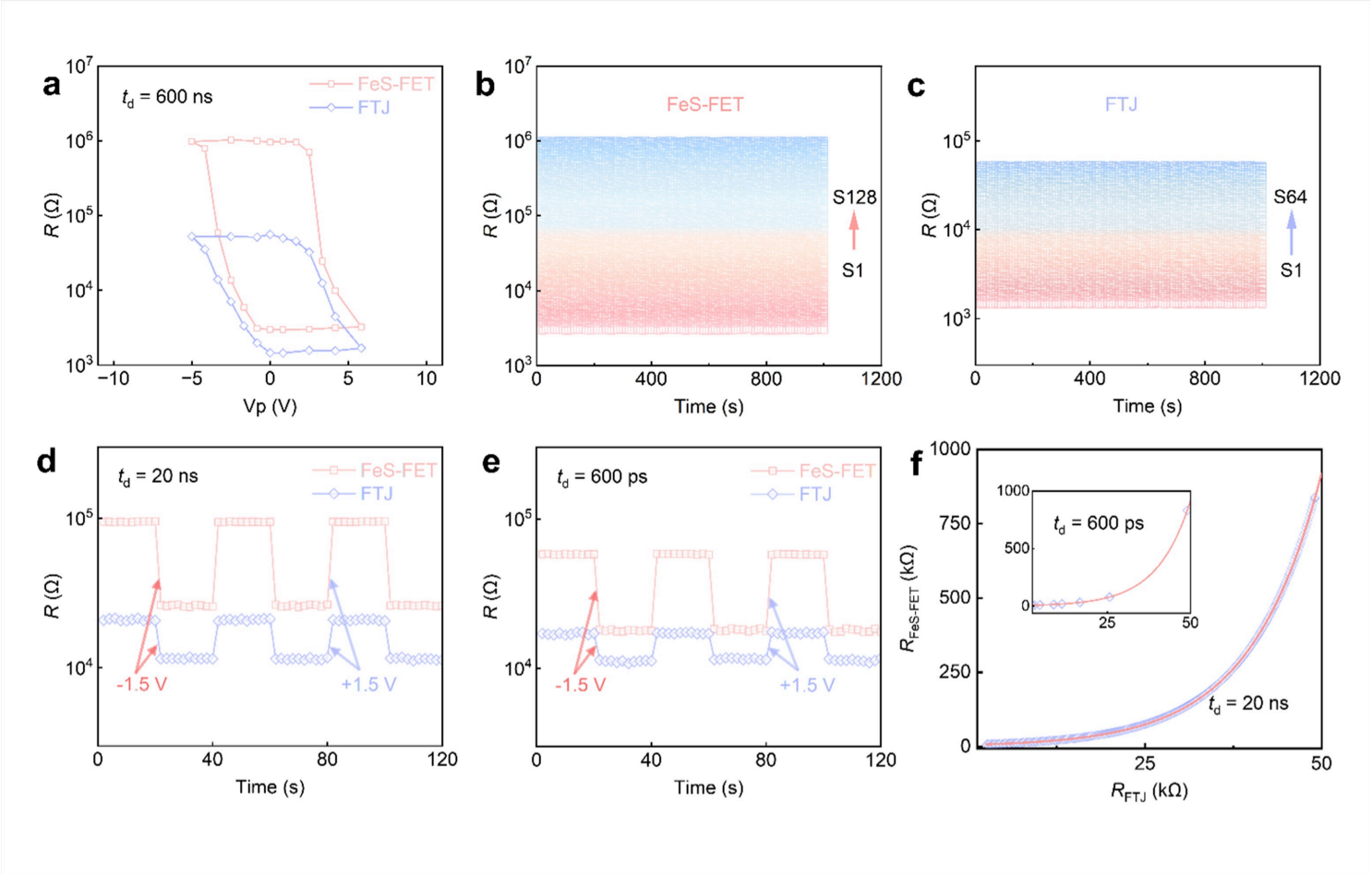


Figure 2

COA based on ferroelectric semiconductor ReS_2 with exponential computing. a, *R-V*p hysteresis loops of the lateral FeS-FET (red) and vertical FTJ (blue), measured at 0.01 V with a pulse duration of t_d = 600 ps. **b-c,** Nonvolatile retention of the FeS-FET (**b**) and FTJ (**c**) ports, showing 128 and 64 resistive states,

respectively. **d-e,** Resistance evolution under repeated ±1.5 V polarization pulses with 20 ns (**d**) and 600 ps (**e**). **f,** Exponential computing can be operated by ±1.5 V pulses with t_d = 20 ns. The inset shows the same operation achieved with t_d = 600 ps, showing ultrafast speed and ultralow energy consuming.

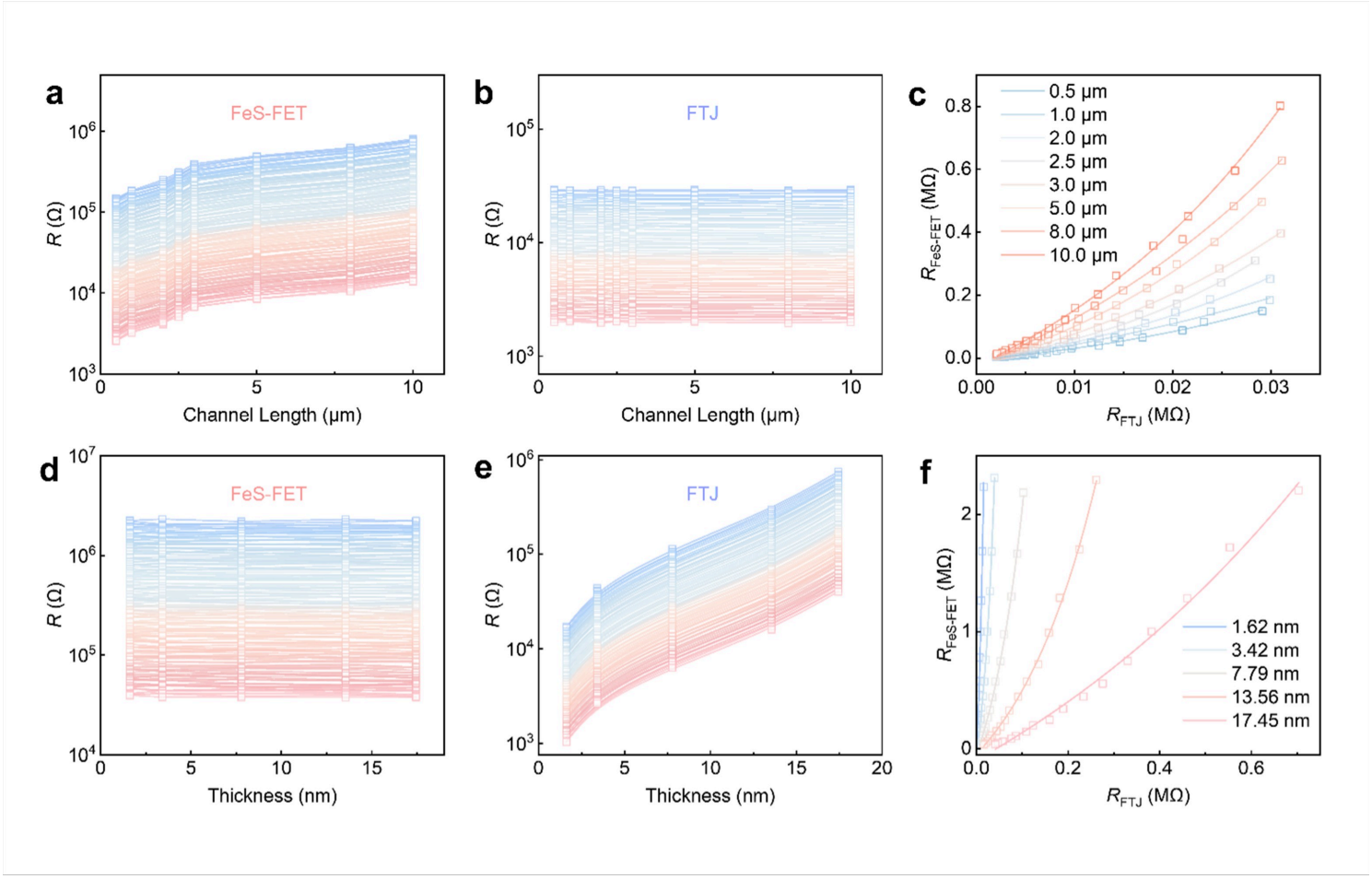


Figure 3

Designable cross-terminal exponential functions enabled by lateral-transport and vertical-barrier tuning. a-b, Resistive states measured of the FeS-FET (**a**) and FTJ (**b**) measured as a function of channel spacing at a fixed at a fixed ReS_2 thickness. **c,** Exponential functions extracted from the paired FeS-FET and FTJ states with different channel spacings. **d-e,** Thickness-dependent resistive states of the FeS-FET (**d**) and FTJ (**e**) obtained by tuning the ReS_2 thickness at a fixed channel spacing. **f,** Exponential functions extracted from the paired FeS-FET and FTJ states at different ReS_2 thicknesses.

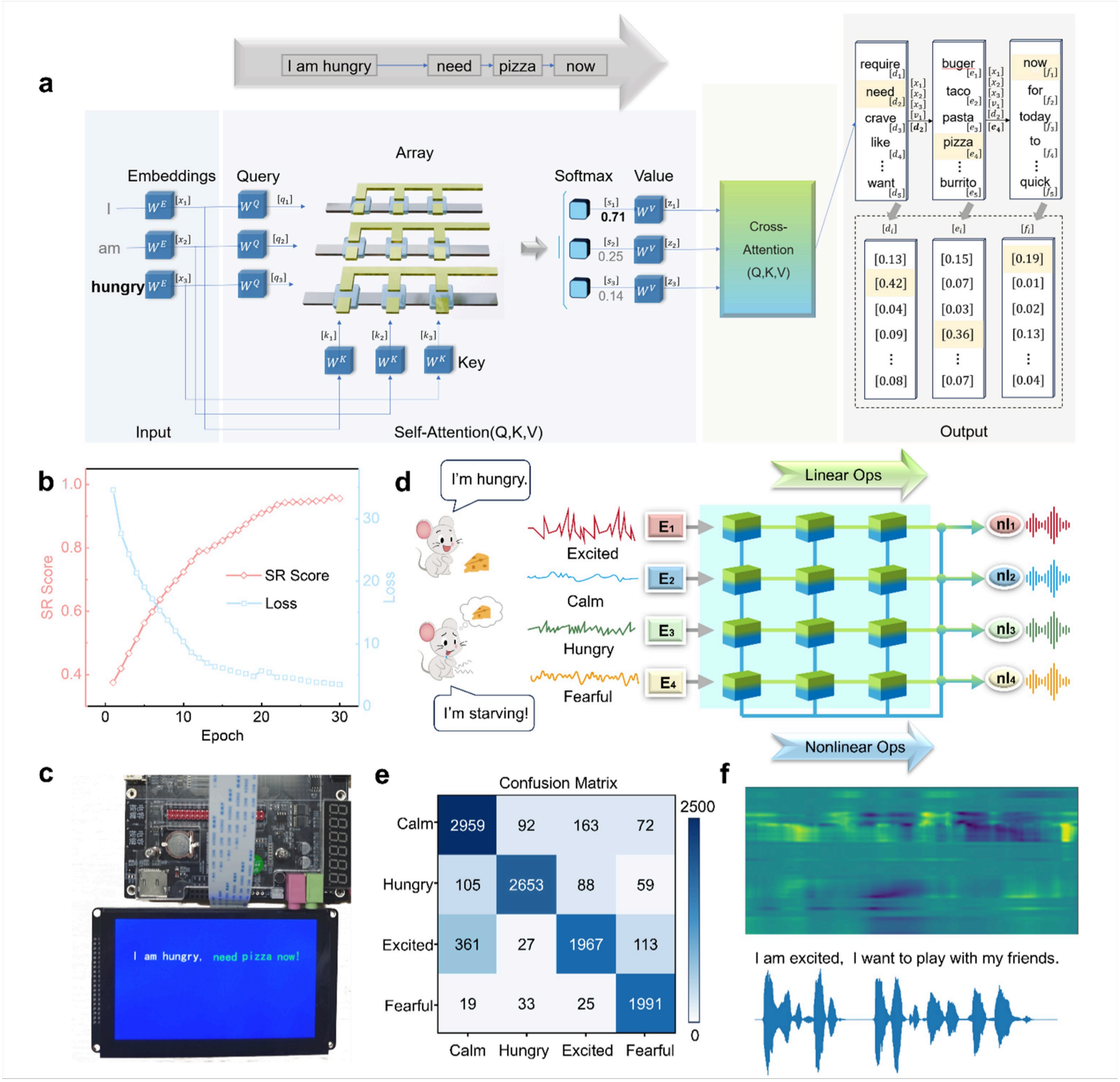


Figure 4

Hardware implementation of a simplified generative task using the COA array and large-scale simulation of COA-enabled LFP-to-language generative modeling. a, Workflow of the COA-based self-attention module executing a minimal text-generation task. Input tokens ("I am hungry") are embedded and mapped to query, key, and value vectors, which are processed by the COA array to perform matrix multiply–accumulate operations and in situ Softmax normalization. The resulting probability distribution drives a cross-attention decoding step, yielding next-token candidates and their associated likelihoods. **b,** Training trajectory of the task, showing rapid loss convergence and a gradual increase in the sentence-

reasonableness (SR) score, indicating improved semantic consistency during generation. **c,**Photograph of the FPGA-controlled hardware prototype executing the generation task in real time, displaying the output sentence “I am hungry, need pizza now!” on the liquid crystal display module. **d,** Schematic of the COA-based emotion-to-language generative framework. Multichannel LFP signals corresponding to four emotional states (excited, calm, hungry, fearful) are projected onto the COA array, where coupled linear and nonlinear operators jointly perform feature extraction and probabilistic semantic decoding. **e**, Confusion matrix, showing accurate emotion-state discrimination across the four LFP classes. **f**, Representative spectrogram and generated output for the excited emotion state. The system produces an expressive sentence (“I am excited, I want to play with my friends”).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ExtendedDataFig.docx
- WangDNatElectronicsSupInf2026.docx