nature communications

Article | Article in Press

https://doi.org/10.1038/s41467-026-77202-y

Quantum nonlinearity for optical neural computing

Received: 28 July 2026

Accepted: 12 August 2026

Published online: 27 August 2026

Check for updates

Qingyi Zhou[1,4] ✉, Jungmin Kim[1,4], Yutian Tao[2,4], Guoming Huang[1], Ming Zhou[3], Zewei Shao[1] & Zongfu Yu[1]

The rapid scaling of deep neural networks comes at the cost of unsustainable power consumption. While optical neural networks offer an alternative, their capabilities remain constrained by the lack of efficient optical nonlinearities. To address this, we propose an optical neural computing architecture by embedding quantum emitters in inverse-designed nanophotonic structures. Due to their saturability, quantum emitters exhibit exceptionally strong nonlinearity compared with conventional materials. Using physics-aware training, we numerically demonstrate that the proposed architecture can solve complex tasks, including nonlinear classification and reinforcement learning, within all-optical neural networks. To enable fair comparison across different platforms, we introduce a framework that quantitatively links nonlinearity to a network's expressive power. Analysis shows that our quantum activation operates at nW/μm^2 intensity, which is seven orders of magnitude below the nonlinearity threshold of conventional optical materials. Looking ahead to large language models, we estimate the nonlinearity-limited optical power, which scales sublinearly with model size. Our results indicate that quantum nanophotonics may provide a route toward sustainable AI inference.

In the past decade, the rapid advancement of deep learning has profoundly transformed science and technology. Deep neural networks have achieved state-of-the-art performance across diverse fields, ranging from computer vision[1] and game-playing[2] to protein design[3] and language processing[4]. Such progress has been driven by the continuous scaling of the model size. However, this scaling trend imposes an energy cost toward unsustainable levels[5]. A growing effort has been directed towards finding alternative computing paradigms. In particular, following the pioneering work of Shen et al.[6], optical neural networks (ONNs) have emerged as a promising candidate[7], inspired by the vision that a passive optical device can implement linear transformation with high speed[8] and low energy cost[9]. Specifically, recent works have demonstrated that linear optical matrix operations can be performed with sub-photon energy consumption[10]. Despite these advantages, linear operations are insufficient for deep learning. The expressive power of ONN is severely limited by the lack of efficient optical nonlinearity[9,11]. In conventional materials, optical nonlinearities are often perturbative[12]. As a result, existing all-optical nonlinear activation units demand high optical power and large footprint[13–17], making it difficult to scale up. This nonlinearity bottleneck has remained a long-standing challenge for the optical computing community. Existing works that rely on hybrid opto-electronic architectures[6,18–20] suffer from additional latency and substantial system complexity due to frequent optical-electrical-optical (O/E/O) conversions. More recently, structural nonlinearity schemes have been proposed[21–23], in which the input is encoded not in the optical field but in tunable parameters of a linear structure. However, the connection between such systems and standard deep learning models is often unclear.

To address the nonlinearity bottleneck, we first point out that nonlinear optical phenomena are not intrinsically restricted to high intensities. A single quantum emitter (including an atom, quantum dot, or color center) can be saturated by the absorption of a few photons

[1]Department of Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, WI, USA. [2]The Computer Sciences Department, University of Wisconsin-Madison, Madison, WI, USA. [3]Department of Electrical Engineering, Stanford University, Stanford, CA, USA. [4]These authors contributed equally: Qingyi Zhou, Jungmin Kim, Yutian Tao. ✉e-mail: qzhou75@wisc.edu

per lifetime, leading to extremely strong optical nonlinearity[24]. There have been both theoretical proposals[25,26] and experimental demonstrations[27,28] of using quantum emitter media as activation units, underscoring their potential for realizing strong optical nonlinearities. However, it remains unclear what quantitative benefit such quantum nonlinearity offers for ONNs. Furthermore, to fully utilize the nonlinear functionality of quantum emitters, it is essential to enhance the interaction between light and quantum emitters, which can be achieved using properly designed nanophotonic structures. Recent progress in quantum nanophotonics has enabled deterministic integration of individual emitters with on-chip photonic structures[29–33]. Therefore, we believe it is the right time to address the above-mentioned questions and systematically investigate whether quantum technologies can provide the strong nonlinearity required by optical neural networks.

In this work, we present a theoretical proposal for a low-power optical neural network architecture that exploits the strong nonlinearity of individual quantum emitters. Specifically, we introduce a quantum-enhanced activation unit by embedding quantum emitters into adjoint-optimized nanophotonic structures[34]. A strong nonlinear response can be achieved at an intensity level of nW/μm^2. With full-wave simulations, we verify the performance on nonlinear classification task as well as reinforcement learning tasks, demonstrating functionality beyond linear models. To enable a fair comparison across different physical nonlinearities, we develop a theoretical framework that quantifies the expressive power of an arbitrary activation unit. Unlike existing theoretical analyses that focus on purely linear optics[35–38] or restricted classes of unitary transformations[39], our framework directly links a nonlinear input-output response to the depth-wise growth of ONN expressivity. This allows us to translate a targeted expressive power into the required light intensity, enabling a quantitative assessment of ONN's energy efficiency. Applying this framework, we show that conventional platforms based on the Kerr effect or saturable absorption would require prohibitively high intensities to match the digital baseline. In contrast, the proposed quantum-enhanced activation reduces the required intensity by seven orders of magnitude. Finally, we look into the future by estimating the nonlinearity-limited optical power for optical large language models (LLMs). Our analysis shows that this lower bound, set by the nonlinear activation alone, remains at the watt level for the proposed scheme and scales sublinearly with model size. Taken together, our results indicate that large-scale optical neural computing, powered by quantum-enhanced nonlinearities, could reduce the energy footprint of AI inference, providing a path beyond the limits of electronic hardware.

Results

Overcoming the nonlinearity bottleneck with quantum activations

We consider a generic multi-layer ONN architecture. Each layer consists of a linear transformation $\mathbf{y} = \mathbf{W}^{(i)}\mathbf{x} + \mathbf{b}^{(i)}$, comprising a weight matrix $\mathbf{W}^{(i)}$ and a bias $\mathbf{b}^{(i)}$, followed by a nonlinear activation $f(\cdot)$, mirroring the architecture of a typical multi-layer perceptron (MLP). In machine learning theory, it is well established that the expressive power of a deep neural network grows exponentially with its depth[40–42]. A purely linear network cannot enjoy this benefit since a composition of linear transformations is still linear. As a result, the overall expressive power of an ONN is limited by its nonlinear activation units. We follow the framework developed in refs. 41,42 and introduce a metric for quantifying expressive power. As illustrated in Fig. 1, a closed trajectory of data points serves as the input. The total curvature K of this trajectory provides a robust measure of curve complexity and is monitored as the curve propagates through successive layers. Intuitively, the total curvature measures the degree of folding applied to the data manifold, a capability that is fundamentally impossible with linear transformations. As shown in Fig. 1(a), for a purely linear network, the total curvature remains constant. In contrast, in the presence of nonlinear activations, the curvature increases by a growth factor $r > 1$ after each layer, leading to an exponential growth $\sim r^L$ with depth L[41,42]. The growth factor r is therefore used as a quantitative measure of expressive power (see Supplementary Note 10 for details).

Conventional optical materials possess small nonlinear susceptibilities[12]. At realistic light intensities, the response is only weakly nonlinear, which yields a growth factor $r \approx 1$. The expressive power, as shown in Fig. 1(b), shows little increase with depth. Existing all-optical activation units typically require mW/μm^2 laser intensities, together with large footprints[13,14,16,43,44] (see Supplementary Note 1 for a summary of representative designs obtained from literature). Such requirements are incompatible with large-scale ONNs. On the other

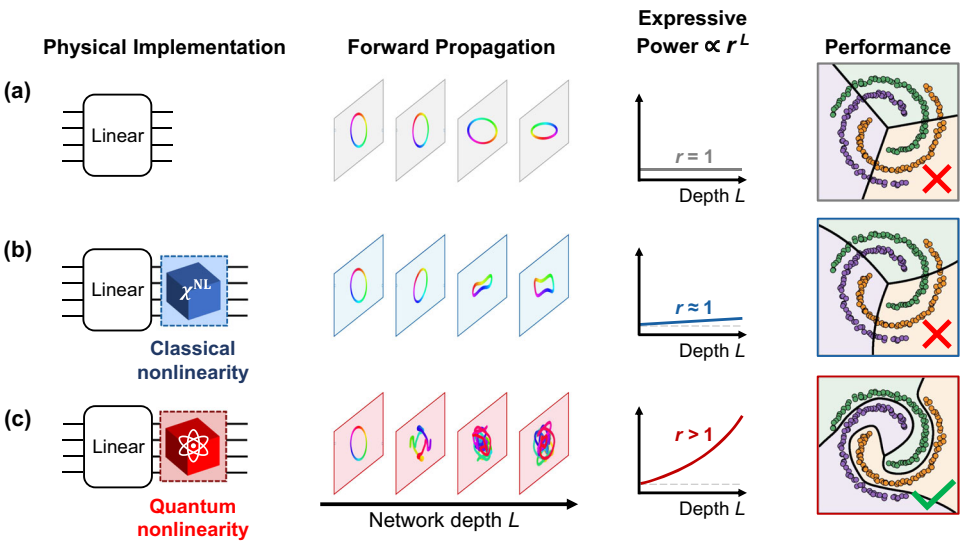

Fig. 1 | Comparison of ONN architectures with different nonlinearities. a A linear ONN, whose expressive power remains constant regardless of depth L. **b** ONN with classical nonlinearity. With weak nonlinearity, the expressive power increases slightly with depth, yet is not enough for complex tasks. **c** ONN with quantum nonlinearity. The expressive power $\propto r^L$ grows exponentially with $r > 1$, and is able to handle complex tasks.

https://doi.org/10.1038/s41467-026-77202-y

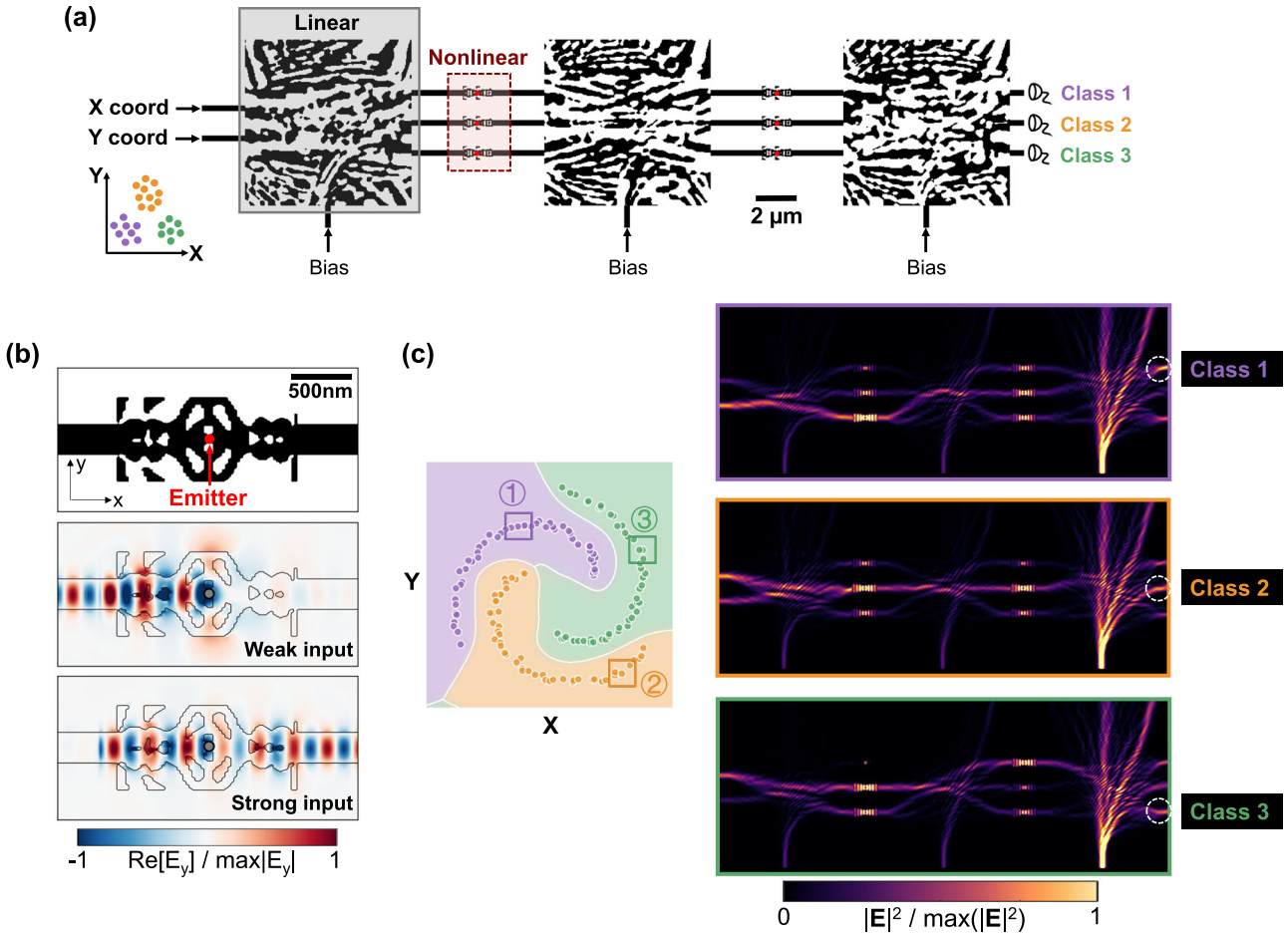

Fig. 2 | Physics-aware training and verification on nonlinear classification. a All-optical neural network design, constructed by stacking nonlinear activation units between linear blocks, which are also designed through adjoint optimization. The 2D coordinates are encoded as input from the left ports, while a separate port at the bottom provides a constant optical bias. The detected intensities are interpreted as classification results. **b** The proposed quantum nonlinear activation unit. A single SiV^- color center is embedded in an inverse-designed GaP-on-diamond structure. Simulated Re [E_y] field distributions (obtained via 3D nonlinear FDFD) illustrate the transition from resonant scattering to saturation, resulting in a nonlinear response at low intensity. **c** Performance verification. Classification results for the spiral dataset are obtained via nonlinear FDFD simulations. The $|E|^2$ intensity distributions of 3 representative inputs (marked with square boxes) are visualized. Different intensity distributions at the output port correspond to different predictions.

hand, it has long been recognized that low-dimensional systems exhibit much stronger optical nonlinearities than bulk media[45–47], owing to enhanced oscillator strength under quantum confinement[48]. In particular, zero-dimensional quantum emitters behave as two-level systems (TLSs) that can be saturated by the absorption of only a few photons per lifetime. This leads to extremely strong nonlinear scattering responses, which have been observed in various waveguide- and cavity-QED platforms[49–53]. These observations suggest that emitter-based nonlinearities could be leveraged to overcome the nonlinearity bottleneck in ONNs (Fig. 1(c)).

Device design and verification of nonlinear classification

Motivated by the above considerations, we propose an all-optical activation unit, consisting of a single quantum emitter embedded inside an inverse-designed GaP-on-diamond nanophotonic structure (design region 1.5 × 0.7 μm²), as shown in Fig. 2(a). We utilized adjoint optimization to design a nanophotonic interface that maximizes light-matter interaction and minimizes loss (see Supplementary Note 4 for design details). The emitter is modeled as a TLS with a field-driven dipole moment $d_y \propto \frac{\Omega/\Gamma_0}{1+2(\Omega/\Gamma_0)^2}$, where the Rabi frequency Ω is set by the local electric field[54,55] and $\Gamma_0 = 2\pi \times 94$ MHz is the spontaneous emission rate (parameters are obtained from SiV^- color centers, assuming lifetime-limited linewidth; see Supplementary Note 2 for details). The electric field distributions for two different input intensities are shown in Fig. 2(b), both obtained using full-wave three-dimensional nonlinear finite-difference frequency-domain (FDFD) simulations (see Supplementary Note 3 for details). The nonlinear activation unit is constructed based on a realistic GaP-on-diamond platform (200 nm patterned GaP on a 160 nm diamond film, with an SiO_2 substrate). The passive linear weight blocks are designed based on the variational effective-index method[56,57]. Note that the specific implementation of linear blocks is not central to this work and could be realized with other schemes such as Mach–Zehnder interferometers. The device is engineered to operate in two distinct regimes: in the weak-field limit, the emitter acts as a linear scatterer, while in the strong-field limit, it becomes nearly transparent. In this two-port geometry, the emitter is configured to induce a change in transmission coefficient $|\Delta t| = 1$ via interference, which results in a strong nonlinearity. Using $N > 1$ emitters could in principle achieve a larger transmission change of $|\Delta t| = 2$ (see Supplementary Note 4 for the analysis). From the simulated input-output curve, we extract an effective nonlinear activation function with an ultra-low intensity threshold. To evaluate the expressive power of the obtained activation function, we compute its growth factor $r(I)$, which reaches $r \approx 1.14$ at intensity $I = 0.24$ nW/μm², exceeding the digital baseline. In contrast, conventional silicon- and graphene-based

https://doi.org/10.1038/s41467-026-77202-y

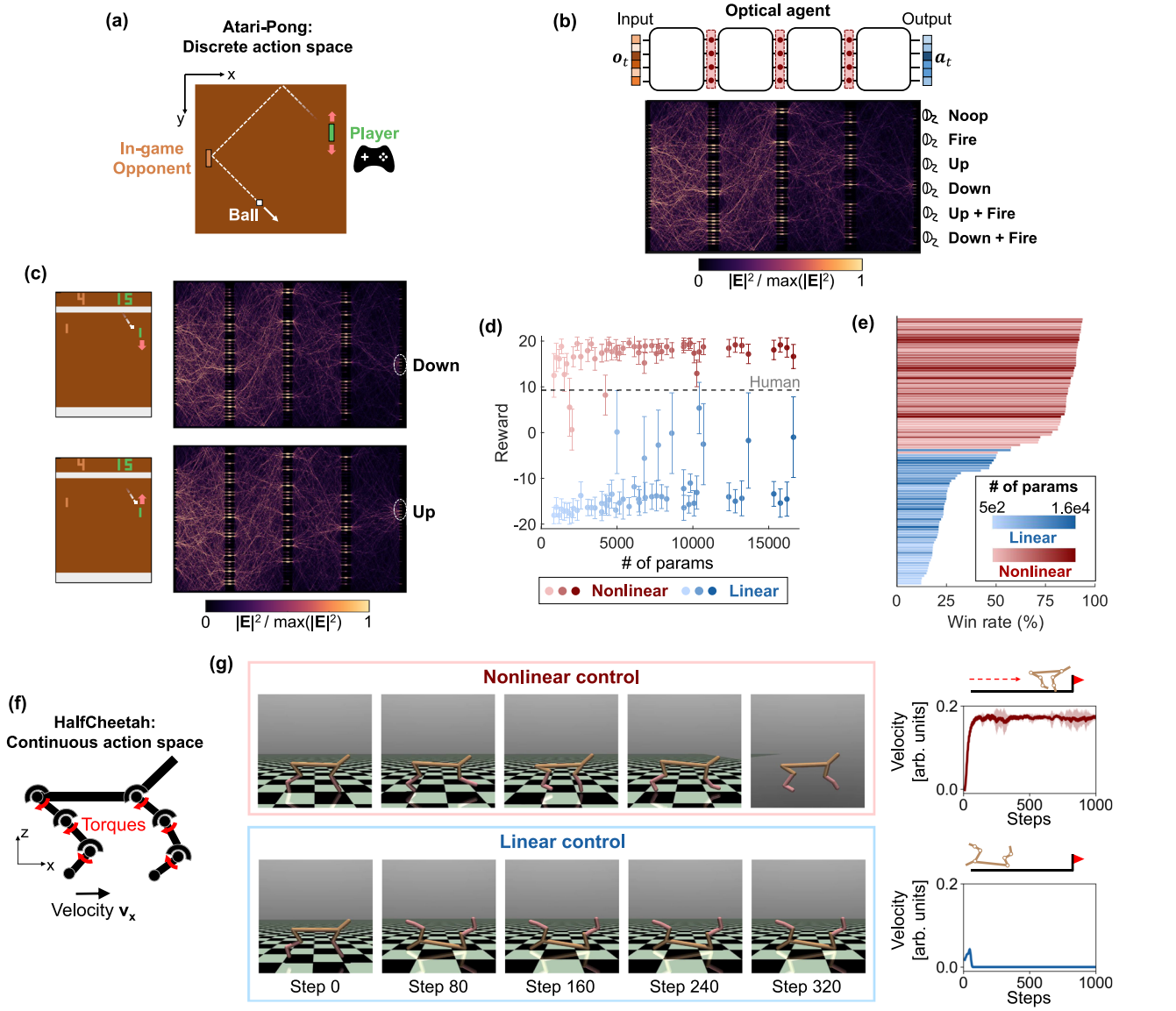

Fig. 3 | Generalizing to reinforcement learning tasks. a Schematic illustration of the Pong environment. The player controls the green paddle and tries to block the ball (white) to win score. **b** The structure of our optical agent, which functions as a policy network. At each time step t, historical game frames are encoded into the input $\mathbf{o}_t$. The network outputs the logits for six discrete actions. **c** Visualization of the learned policy. Light intensity distributions of two key frames are displayed. **d** Final reward versus model size. Linear models (blue) saturate at a low performance ceiling with high variance. In contrast, nonlinear models (red) converge to near-perfect play very quickly. Error bars denote standard deviations obtained over 30 episodes. The human-level performance (reward = 9.3)[2] is visualized using gray dashed line. **e** Performance ranking. Best achieved rewards for 104 trained models (52 linear, 52 nonlinear) are sorted. Nonlinear models consistently outperform linear models. **f** Schematic illustration of the HalfCheetah control task. **g** Snapshots obtained during testing. The nonlinear ONN runs stably, while the linear ONN falls down. The insets plot the corresponding velocity curves, averaged over 10 episodes. The shaded areas visualize the standard deviations.

nonlinearities remain near $r = 1$ under similar operating conditions (see Supplementary Note 11 and Supplementary Fig. 22 for details). We have also analyzed the effect of low quantum efficiency and demonstrate that our method remains robust even when the TLS' quantum efficiency drops to 60% (see Supplementary Note 5 for details). We also characterize the robustness of the activation unit against two different non-idealities: lateral position randomness and spectral disorder (see Supplementary Note 14 for details). These results confirm our key physical intuition: by utilizing saturable quantum emitters embedded in nanophotonic structures, strong optical nonlinearity can be realized at ultra-low optical power.

To demonstrate the practical advantage of the proposed activation, we benchmark our system on nonlinear classification task that is challenging for models that are weakly nonlinear. We adopt a physics-aware training approach to design our ONN in a physically consistent manner. We first characterize the nonlinear activation unit using nonlinear FDFD simulations to obtain its input-output transmission curve (see Supplementary Fig. 8), which is then used as the activation function in a PyTorch-based ONN model. The linear weight matrices are represented as complex transmission matrices subject to energy-preserving constraints, ensuring that they can be implemented by passive structures. With this differentiable model, the ONN is trained in the digital domain via backpropagation. After training, we map the trained network to concrete photonic structures in a modular fashion. For each layer, the trained complex weight matrix is interpreted as a target transmission matrix between input and output ports. We then run a separate adjoint-based optimization to realize a compact block that implements this target matrix with high fidelity (see

https://doi.org/10.1038/s41467-026-77202-y

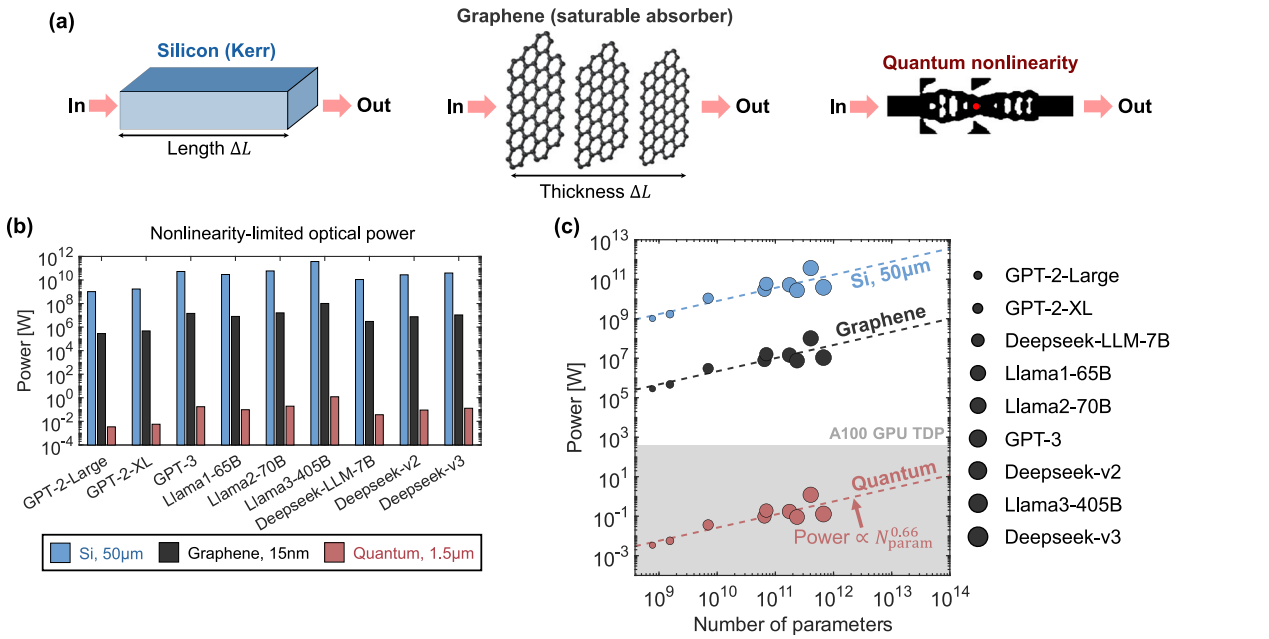

Fig. 4 | Scalability of nonlinearity-limited power requirements. a Schematic illustration of nonlinear platforms: conventional bulk material (silicon, 50 μm length), 2D saturable absorber (stacked graphene, 15 nm thickness), and the proposed quantum activation unit. **b** Histogram showing the estimated nonlinearity-limited optical power of optical LLMs. Conventional materials demand prohibitive power levels, while the proposed scheme is not restricted by nonlinearity. **c** Estimated nonlinearity-limited optical power versus model size N_{param}. The shaded area marks the thermal design power of a single NVIDIA A100 GPU (≈400 W), shown only as a reference. ONNs follow a sublinear scaling $P \propto N_{\text{param}}^{0.66}$, as shown by the dashed lines, indicating that optical computing has a growing advantage as models scale up.

Supplementary Notes 4 and 8 for quantitative metrics). The nonlinear layers are implemented by inserting the quantum activation units between these inverse-designed linear blocks, as illustrated in Fig. 2(a). By dissecting the network into multiple modules that can be designed individually, this approach avoids heavy full-wave simulations of the entire network during training (the training procedure is explained in Supplementary Note 7). We verify the final design using full-wave nonlinear FDFD simulations. The classification result for a three-class spiral dataset is shown in Fig. 2(c). The corresponding light intensities for three representative input points are also visualized, revealing how the network steers optical energy toward different output regions associated with different classes. We provide more examples in Supplementary Note 6 (performance on MNIST and FashionMNIST) and Supplementary Note 8 (performance on nonlinear regression task). These results confirm that the physics-aware training yields physically realizable ONNs. Within realistic optical intensities well below 1 mW/μm^2, conventional materials cannot provide sufficient expressive power. In stark contrast, quantum-enhanced nonlinearity can function below 1 nW/μm^2, enabling the system to solve complex tasks.

Generalizing to intelligent optical agents: reinforcement learning

Having established that quantum-enhanced activations enable nonlinear classification, we next ask whether the same photonic building blocks can support more complex tasks. Reinforcement learning (RL), which has achieved impressive results on game-playing benchmarks[2,58] and robotics[59], provides a natural testbed for our purpose. Note that RL based on photonic spiking network has recently been demonstrated on control benchmarks[60]. To demonstrate the generality of our approach, we evaluate it on two distinct tasks: a discrete game-playing task (Atari Pong) and a continuous control task (HalfCheetah), both provided by the Gymnasium library[61].

The Atari Pong environment is illustrated schematically in Fig. 3(a). In Pong, an agent controls the right paddle against the in-game opponent. An episode ends when a player reaches 21 points, and the reward is defined as the final score difference. As shown in Fig. 3(b), the ONN acts as a policy network: at each time step t, a stack of F recent frames is encoded into an observation $\mathbf{o}_t$, which serves as the input (see Supplementary Fig. 17 for details). The output intensities (divided into six regions) are interpreted as logits for six discrete actions. An action a_t is sampled from this distribution and sent back to the environment, which then advances to the next time step. The policy parameters are trained using a standard proximal policy optimization (PPO) algorithm[62], based on the same physics-aware framework described above (see Supplementary Note 9 for details). We note that while the nonlinear activation mechanism has been verified with 3D simulation in Fig. 2, the present Atari Pong demonstration relies on 2D simulations due to limited computing resources. The optical intensity distributions for two representative game frames are shown in Fig. 3(c). Different spatial configurations of the ball and paddles lead to distinct activation patterns and different intensity hotspots at the output ports. We then systematically benchmark the performance of linear versus nonlinear ONNs. We train 104 models, sweeping across network width H, number of hidden layers L, and the number of input frames F. All models are trained for an identical number of iterations (see Supplementary Note 9 for details). The results are summarized in Fig. 3(d), where the final reward is plotted against the number of parameters. Fig. 3(e) further ranks all the trained models based on their final rewards. Linear models saturate at a low performance ceiling regardless of model size and exhibit high variance, indicating that the learned strategies cannot win reliably. In contrast, ONNs equipped with quantum activations achieve much higher rewards as they scale up, converging to near-perfect play.

We further evaluate our ONN on the MuJoCo HalfCheetah control benchmark, as illustrated in Fig. 3(f). With a continuous action space, HalfCheetah is much more challenging than Pong. At each time step, the optical agent receives a 17-dimensional observation (joint positions and velocities, see Supplementary Fig. 18) and outputs a 6-dimensional

action vector that specifies torques applied at the six hinge joints. We adopt a similar ONN backbone as in Pong, with one key difference: the output uses balanced detection to support negative action values (see Supplementary Note 9 for details). Training is performed using the standard soft actor-critic (SAC) algorithm[63]. The corresponding snapshots collected during testing are shown in Fig. 3(g). With quantum activation, the ONN learns to run smoothly, whereas a linear ONN fails to acquire a viable control policy and falls down early in the episode. The insets in Fig. 3(g) plot the averaged velocity v_x over 10 episodes, highlighting the stability enabled by optical nonlinearity. The above results confirm that strong nonlinearity is essential for enabling complex capabilities in deep ONNs.

Nonlinearity-limited power requirements for large-scale networks

Having established the importance of strong nonlinearity at the device level, we now examine how the choice of nonlinearity constrains the optical power of large-scale systems. We consider only the optical power needed to drive the nonlinear activations, and ask what this nonlinearity-limited power would be for present-day LLMs. As a quantitative baseline, we note that in standard digital MLPs, common activation functions typically increase the total curvature by $r_{\text{digital}} \simeq 1.045 \sim 1.095$ per layer (see Supplementary Note 10 for details). We therefore ask: for a given physical nonlinearity, what is the minimum optical intensity $I_{\min}$ required to match this digital baseline? Using our established framework, we compute the expressive power $r(I)$ for three representative platforms: Kerr nonlinearity in a 50 μm long silicon waveguide[64,65], saturable absorption in stacked graphene layers[45,66] (15 nm total thickness), and our proposed quantum activation unit (see Supplementary Note 11 for details). We find that maintaining the target expressive power in silicon requires intensities exceeding 72.6 W/μm^2, while graphene requires approximately 0.02 W/μm^2. In contrast, the quantum activation unit achieves the same baseline at merely ~ 0.24 nW/μm^2. This represents an efficiency improvement of roughly 8.3×10^7 times relative to graphene and 3.0×10^{11} times relative to silicon. This quantitative comparison reveals the inadequacy of conventional nonlinear materials and shows that the quantum nonlinearity proposed here can overcome this limitation.

Given these intensity thresholds, we next estimate the nonlinearity-limited optical power for LLM-scale models. For a standard decoder-only transformer architecture with context length L_{seq}, embedding dimension d_{model}, and L transformer layers[4], we estimate the optical input dimension per layer as $3L_{\text{seq}}d_{\text{model}}$, accounting for the parallel projection of query, key, and value matrices. Assuming each optical neuron occupies an effective cross-sectional area $A \approx 0.1\mu\text{m}^2$, corresponding to an on-chip waveguide mode,[6] and is driven at $I_{\min}$, the nonlinearity-limited optical power is estimated as

$$P \approx I_{\min} A \cdot (3L_{\text{seq}}d_{\text{model}}) \cdot L. \quad (1)$$

This value should be understood as a lower bound set by the nonlinear activation alone and does not represent system-level power consumption. Using the architectural parameters of representative LLMs ranging from GPT-2 to DeepSeek-V3[67–74] (see Supplementary Note 12 for details), we evaluate Eq. (1) and summarize the results in Fig. 4(b). When using conventional nonlinearities based on silicon or graphene, the required optical power quickly reaches prohibitive levels (exceeding 10^8 W for the largest models). However, the proposed quantum architecture keeps this nonlinearity-limited power below 1.2W across all investigated models. Finally, we analyze how this power requirement scales with model size. Fig. 4(c) plots the estimated optical power against the total number of trainable parameters, N_{param}. We also indicate the thermal design power of a single high-end GPU (NVIDIA A100, $\approx$ 400W) as a shaded area. We include this only as a rough reference for scale. We do not use it to claim that an optical system would consume less total power than a GPU. For ONNs the data points follow a sublinear scaling law, $P \propto N_{\text{param}}^{0.66}$. This behavior stems from the geometric nature of the network: the parameter count grows with the volume of the network ($\sim L \cdot d_{\text{model}}^2$), whereas the required optical power scales with the number of inputs ($\sim L \cdot d_{\text{model}} \cdot L_{\text{seq}}$). Such a distinction leads to a scaling exponent smaller than one, consistent with known results[75]. We would like to point out that the dimensionality of current integrated photonic circuits[76,77] remains far below LLM scale, so a near-term large-scale optical LLM is more plausibly realized on free-space platforms (e.g., diffractive networks). The same sublinear scaling applies to such platforms, since the exponent comes from the network geometry and is platform-independent. In contrast, the power consumption of electronic processors typically scales linearly with N_{param}. The nonlinearity-limited optical power therefore scales more favorably with model size, although translating this into a system-level energy advantage would require addressing the additional overheads. Overall, these results suggest that quantum nonlinearity can substantially relax the optical-power bottleneck that conventional nonlinear materials impose at large scale. By shifting to quantum nonlinearities, it should be possible to construct optical deep learning models within a feasible power budget.

Discussion

In summary, we have presented a theoretical framework to address the nonlinearity bottleneck in optical computing. At the device level, by integrating quantum emitters with inverse-designed nanophotonic structures, strong nonlinearity can be realized at intensities below 1 nW/μm^2. This enables complex functionalities ranging from nonlinear classification to reinforcement learning, presenting a clear performance gap over linear ONNs. Moreover, we have developed a general theoretical framework to quantify the expressive power of arbitrary nonlinear physical systems, which in turn allows us to determine the light intensity requirements. At the scale of large language models, we estimate the nonlinearity-limited optical power and find that it stays at the watt level, with a favorable sublinear scaling in model size. Together, these results suggest that the lack of nonlinearity is not a fundamental limit, but rather a challenge that could be overcome with quantum technologies.

Despite these advances, transforming our theoretical proposal into large-scale hardware is still facing several practical challenges. A key trade-off exists between intensity threshold and operation bandwidth: the high sensitivity is inherently related to the emitter's long radiative lifetime. For a bare SiV^- center, the intrinsic response bandwidth lies in the sub-GHz range, below the 10–50 GHz modulation rates typically used in optical computing. We point out that this limit is not fundamental, since the response speed can be increased through Purcell enhancement in optimized photonic structures[78]. Our inverse-designed activation units already exhibit an emergent Purcell factor $F_P \approx 2.74$ (Supplementary Note 13), and Purcell-enhanced linewidth as large as $2\pi \times 4.6$ GHz has been demonstrated in photonic crystal cavity[79], offering a realistic route to GHz-scale operation. Another challenge is the inhomogeneity of solid-state emitters. Since the activation requires each emitter to be resonant with the optical signal, the inhomogeneous broadening of transition frequencies becomes an obstacle as the system scales up. To tackle this issue, solutions such as DC Stark tuning have been demonstrated to tune the resonance of individual emitters[80], which provides a route to align each emitter independently. Regarding integration, while deterministic placement of emitters remains difficult, recent advances in fabrication techniques offer promising solutions for large-scale integration[32,81,82]. Finally, solid-state quantum emitters often require cryogenic operation to suppress dephasing and to approach lifetime-limited linewidths[83,84], which introduces an additional power overhead for cooling.

Looking forward, this work shows that in order to unlock the full potential of optical computing, we should exploit the strong

nonlinearity provided by quantum emitters rather than conventional bulk materials. By combining inverse-designed nanophotonics with modern deep learning theory, we provide a path toward low-power optical neural computing, in which the lack of strong nonlinearity is no longer the limiting factor. Realizing this vision will ultimately pave the way toward sustainable, next-generation artificial intelligence.

Methods

Nonlinear electromagnetic simulations

We model the quantum emitter as an ideal two-level system with ground state $|g\rangle$ and excited state $|e\rangle$. The emitter has a transition frequency ω_0 and a transition dipole moment $\mathbf{d}_0$. In this work we neglect pure dephasing. When the emitter is driven by a monochromatic field, the field-driven dipole moment at steady state can be derived as

$$\mathbf{d} = \frac{\Omega(-\Delta + i\Gamma_0/2)}{\Delta^2 + (\Gamma_0/2)^2 + \Omega^2/2}\mathbf{d}_0, \tag{2}$$

where $\Omega = d_0 E_{\mathrm{inc}}/\hbar$ is the Rabi frequency, $\Delta = \omega_{\mathrm{d}} - \omega_0$ is the detuning, and Γ_0 is the radiative decay rate. The device simulations are performed under resonant excitation with $\Delta = 0$ (analysis regarding the robustness to spectral disorder can be found in Supplementary Note 14). The dipole moment leads to a current source $\mathbf{J}(\mathbf{r}) = -i\omega\mathbf{d}\cdot\delta(\mathbf{r} - \mathbf{r}_0)$. This current source, which depends nonlinearly on the local electric field, is included in a three-dimensional finite-difference frequency-domain model. Specifically, after spatial discretization, Maxwell's equations take the form

$$\mathbf{F}(\mathbf{x}) \equiv \mathbf{A}\mathbf{x} - \mathbf{b}_0 - \mathbf{b}_{\mathrm{TLS}}(\mathbf{x}) = \mathbf{0}, \tag{3}$$

where $\mathbf{x}$ represents the electric field distribution, $\mathbf{A}$ is the discretized Maxwell operator, $\mathbf{b}_0$ is the external source, and $\mathbf{b}_{\mathrm{TLS}}$ corresponds to the dipole source. The above nonlinear equation can be solved using the Newton–Raphson method. At each iteration, the equation is linearized around the current solution, and the resulting Jacobian system is solved for the field update. Details of the derivation, including the treatment of the dipole's primary radiation and the full Jacobian matrix, are provided in Supplementary Notes 2 and 3.

Platform and inverse design

The nonlinear activation unit is based on a single SiV^- color center. At cryogenic temperature, we focus on its zero-phonon line at $\lambda_0 = 737.134$ nm. The platform consists of a 200-nm-thick patterned GaP layer on a 160-nm-thick unetched diamond film and a SiO_2 substrate. The refractive indices of GaP, diamond, and SiO_2 are 3.22, 2.40, and 1.45, respectively. The input light enters through a GaP strip waveguide (260 nm wide and 200 nm thick) via its fundamental TE mode. A single SiV^- center is placed at the center of the design region, 30 nm below the diamond surface, with its dipole oriented along the y direction.

We optimize the in-plane GaP pattern within a $1.5 \times 0.7\ \mu m^2$ design region using the adjoint method. The optimization runs for 500 iterations using Adam. A conic filter (radius 100 nm) is also applied to control the minimum feature size. The objective function and material parameterization are explained in Supplementary Note 4. We also design the passive linear blocks on the same platform to realize the weight matrices. We use a 2.5D variational effective-index method to reduce the computational cost. Details of the effective-index model and adjoint optimization are provided in Supplementary Note 4.

Physics-aware neural-network training

We use the response curve obtained from nonlinear FDFD simulations as the activation function in the neural network. We first train the neural network in the digital domain using PyTorch. All complex weight matrices are constrained to be isometric, so that they can be implemented using passive photonic structures. After training, each matrix is treated as a target transmission matrix and translated into a linear optical block through adjoint optimization. The optimized linear blocks are then concatenated with the nonlinear activation units to obtain the final structure.

For the reinforcement learning tasks, the Atari Pong task is trained using proximal policy optimization (PPO), while the HalfCheetah task is trained using soft actor–critic (SAC). For each task, the linear and nonlinear models use the same network architecture and training settings, with the activation disabled in the linear baseline. Further details on network training and evaluation are provided in Supplementary Notes 7 and 9.

Quantification of expressive power

In order to evaluate the expressive power of a given neural network, we follow the approach of Raghu et al.[42] and use a circular trajectory as the network input,

$$\mathbf{x}(\theta) = \sqrt{N_{\mathrm{dim}}}[\mathbf{u}\cos(\theta) + \mathbf{v}\sin(\theta)], \qquad \theta \in [0, 2\pi], \tag{4}$$

where N_{dim} is the input dimension, and $\mathbf{u}$ and $\mathbf{v}$ are random orthonormal unit vectors. Specifically, we propagate this trajectory through successive network layers and calculate its total curvature K after each layer. A linear transformation does not increase the total curvature, whereas a nonlinear activation progressively deforms the trajectory, which can lead to an exponential increase in total curvature with network depth. The growth factor r can be extracted by fitting the layer-dependent curvature to $K \propto r^L$, where L denotes the network depth. Thus, a larger r indicates stronger expressive power. For optical activations, we repeat this procedure at different input intensities to obtain r as a function of intensity, $r(I)$. Details related to the trajectory sampling, network dimensions, and curve fitting are provided in Supplementary Notes 10 and 11.

Data availability

The data supporting the findings of this study are publicly available together with the associated code in the GitHub repository at https://github.com/zhouqingyi616/quantum-nonlinearity-for-optical-neural-computing. The same files are permanently archived on Zenodo at https://doi.org/10.5281/zenodo.21762053.

Code availability

The custom code used to generate and analyze the results reported in this study is publicly available together with the supporting data in the GitHub repository at https://github.com/zhouqingyi616/quantum-nonlinearity-for-optical-neural-computing. The same files are permanently archived on Zenodo at https://doi.org/10.5281/zenodo.21762053.

References

1. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, Vol. 25 (2012).
2. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
3. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
4. Vaswani, A. et al. Attention is all you need. *Advances in neural information processing systems***30** (2017).
5. Patterson, D. et al. Carbon emissions and large neural network training. *arXiv preprint arXiv*: http://arxiv.org/abs/2104.10350 (2021).
6. Shen, Y. et al. Deep learning with coherent nanophotonic circuits. *Nat. Photonics* **11**, 441–446 (2017).
7. Lin, X. et al. All-optical machine learning using diffractive deep neural networks. *Science* **361**, 1004–1008 (2018).

8. Shekhar, S. et al. Roadmapping the next generation of silicon photonics. *Nat. Commun.* **15**, 751 (2024).
9. Wetzstein, G. et al. Inference in artificial intelligence with deep optics and photonics. *Nature* **588**, 39–47 (2020).
10. Wang, T. et al. An optical neural network using less than 1 photon per multiplication. *Nat. Commun.* **13**, 123 (2022).
11. Shi, W., Huang, Z., Fu, T. & Chen, H. Review of nonlinear activation functions in optical neural networks. *Adv. Photonics* **7**, 064004–064004 (2025).
12. Boyd, R. W., Gaeta, A. L. & Giese, E. Nonlinear optics. In *Springer Handbook of Atomic, Molecular, and Optical Physics*, 1097–1110 (Springer, 2008).
13. Feldmann, J., Youngblood, N., Wright, C. D., Bhaskaran, H. & Pernice, W. H. All-optical spiking neurosynaptic networks with self-learning capabilities. *Nature* **569**, 208–214 (2019).
14. Wu, B., Li, H., Tong, W., Dong, J. & Zhang, X. Low-threshold all-optical nonlinear activation function based on a ge/si hybrid structure in a microring resonator. *Optical Mater. Express* **12**, 970–980 (2022).
15. Wu, T., Li, Y., Ge, L. & Feng, L. Field-programmable photonic nonlinearity. *Nat. Photonics* **19**, 725–732 (2025).
16. Jha, A., Huang, C. & Prucnal, P. R. Reconfigurable all-optical nonlinear activation functions for neuromorphic photonics. *Opt. Lett.* **45**, 4819–4822 (2020).
17. Yanagimoto, R. et al. Programmable on-chip nonlinear photonics. *Nature* 1–8 (2025).
18. Williamson, I. A. et al. Reprogrammable electro-optic nonlinear activation functions for optical neural networks. *IEEE J. Sel. Top. Quantum Electron.* **26**, 1–12 (2019).
19. Pour Fard, M. M. et al. Experimental realization of arbitrary activation functions for optical neural networks. *Opt. Express* **28**, 12138–12148 (2020).
20. Hu, Y. et al. Integrated lithium niobate photonic computing circuit based on efficient and high-speed electro-optic conversion. *Nat. Commun.* **16**, 8178 (2025).
21. Yildirim, M., Dinc, N. U., Oguz, I., Psaltis, D. & Moser, C. Nonlinear processing with linear optics. *Nat. Photonics* **18**, 1076–1082 (2024).
22. Xia, F. et al. Nonlinear optical encoding enabled by recurrent linear scattering. *Nat. Photonics* **18**, 1067–1075 (2024).
23. Wanjura, C. C. & Marquardt, F. Fully nonlinear neuromorphic computing with linear wave scattering. *Nat. Phys.* **20**, 1434–1440 (2024).
24. Lodahl, P., Mahmoodian, S. & Stobbe, S. Interfacing single photons and single quantum dots with photonic nanostructures. *Rev. Mod. Phys.* **87**, 347–400 (2015).
25. Zhu, C., Wang, T., McMahon, P. L. & Soh, D. Quantum optical neural networks using atom-cavity interactions to provide all-optical nonlinearity. *arXiv preprint arXiv*: https://arxiv.org/abs/2511.06167 (2025).
26. Canora, R., Xu, X., Niu, Z., Alaeian, H. & Du, S. Engineering nonlinear activation functions for all-optical neural networks via quantum interference. *arXiv preprint arXiv*: http://arxiv.org/abs/2504.04009 (2025).
27. Zuo, Y. et al. All-optical neural network with nonlinear activation functions. *Optica* **6**, 1132–1137 (2019).
28. Ryou, A. et al. Free-space optical neural network based on thermal atomic nonlinearity. *Photonics Res.* **9**, B128–B134 (2021).
29. Ohno, K. et al. Engineering shallow spins in diamond with nitrogen delta-doping. *Appl. Phys. Lett.* **101**, 082413 (2012).
30. Chen, Y.-C. et al. Laser writing of individual nitrogen-vacancy defects in diamond with near-unity yield. *Optica* **6**, 662–667 (2019).
31. Day, A. M., Dietz, J. R., Sutula, M., Yeh, M. & Hu, E. L. Laser writing of spin defects in nanophotonic cavities. *Nat. Mater.* **22**, 696–702 (2023).
32. Yama, N. S., Wu, C.-C., Hatami, F. & Fu, K.-M. C. A scalable gallium-phosphide-on-diamond spin-photon interface. *arXiv preprint arXiv*: https://arxiv.org/abs/2601.04733 (2026).
33. Schröder, T. et al. Scalable focused ion beam creation of nearly lifetime-limited single quantum emitters in diamond nanostructures. *Nat. Commun.* **8**, 15376 (2017).
34. Lalau-Keraly, C. M., Bhargava, S., Miller, O. D. & Yablonovitch, E. Adjoint shape optimization applied to electromagnetic design. *Opt. Express* **21**, 21693–21701 (2013).
35. Kulce, O., Mengu, D., Rivenson, Y. & Ozcan, A. All-optical information-processing capacity of diffractive surfaces. *Light Sci. Appl.* **10**, 25 (2021).
36. Miller, D. A. Why optics needs thickness. *Science* **379**, 41–45 (2023).
37. Li, Y. & Monticone, F. The spatial complexity of optical computing: toward space-efficient design. *Nat. Commun.* **16**, 8588 (2025).
38. Onodera, T. et al. Arbitrary control over multimode wave propagation for machine learning. *Nat. Phys.* 1–8 (2025).
39. Yu, S., Piao, X. & Park, N. Nonlinear unitary circuits for photonic neural networks. *ACS Photonics* **12**, 6737–6744 (2025).
40. Montúfar, G., Pascanu, R., Cho, K. & Bengio, Y. On the number of linear regions of deep neural networks. *Advances in Neural Information Processing Systems*, Vol. 27 (2014).
41. Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J. & Ganguli, S. Exponential expressivity in deep neural networks through transient chaos. *Advances in Neural Information Processing Systems*, Vol. 29 (2016).
42. Raghu, M., Poole, B., Kleinberg, J., Ganguli, S. & Sohl-Dickstein, J. On the expressive power of deep neural networks. in *International Conference on Machine Learning*, 2847–2854 (PMLR, 2017).
43. Shi, Y. et al. Nonlinear germanium-silicon photodiode for activation and monitoring in photonic neuromorphic networks. *Nat. Commun.* **13**, 6048 (2022).
44. Li, G. H. et al. All-optical ultrafast ReLU function for energy-efficient nanophotonic deep learning. *Nanophotonics* **12**, 847–855 (2023).
45. Bao, Q. et al. Atomic-layer graphene as a saturable absorber for ultrafast pulsed lasers. *Adv. Funct. Mater.* **19**, 3077–3083 (2009).
46. Shi, J. et al. 3r mos2 with broken inversion symmetry: a promising ultrathin nonlinear optical device. *Adv. Mater.* **29**, 1701486 (2017).
47. Liu, B. et al. Giant second harmonic generation in two-dimensional tellurene with synthesis and thickness engineering. *Appl.Phys. Rev.* **12**, 011414 (2025).
48. Hanamura, E. Rapid radiative decay and enhanced optical nonlinearity of excitons in a quantum well. *Phys. Rev. B* **38**, 1228 (1988).
49. Javadi, A. et al. Single-photon non-linear optics with a quantum dot in a waveguide. *Nat. Commun.* **6**, 8655 (2015).
50. Volz, J., Scheucher, M., Junge, C. & Rauschenbeutel, A. Nonlinear π phase shift for single fibre-guided photons interacting with a single resonator-enhanced atom. *Nat. Photonics* **8**, 965–970 (2014).
51. Shomroni, I. et al. All-optical routing of single photons by a one-atom switch controlled by a single photon. *Science* **345**, 903–906 (2014).
52. Hacker, B., Welte, S., Rempe, G. & Ritter, S. A photon–photon quantum gate based on a single atom in an optical resonator. *Nature* **536**, 193–196 (2016).
53. Lukin, D. M. et al. 4H-silicon-carbide-on-insulator for integrated quantum and nonlinear photonics. *Nat. Photonics* **14**, 330–334 (2020).
54. Zhou, Q., Gangaraj, S., Zhou, M. & Yu, Z. Simulating quantum emitters in arbitrary photonic environments using FDTD: beyond the semi-classical regime. *arXiv preprint arXiv*: https://arxiv.org/abs/2410.16118 (2024).
55. Wang, H. & Fan, S. Lorentz–Drude dipoles in the radiative limit and their modeling in finite-difference time-domain methods. *Ann. Phys.* **537**, e00156 (2025).

56. Hammer, M. & Ivanova, O. V. Effective index approximations of photonic crystal slabs: a 2-to-1-d assessment. *Opt. Quantum Electron.* **41**, 267–283 (2009).
57. Nikkhah, V. et al. Inverse-designed low-index-contrast structures on a silicon photonics platform for vector–matrix multiplication. *Nat. Photonics* **18**, 501–508 (2024).
58. Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
59. Hwangbo, J. et al. Learning agile and dynamic motor skills for legged robots. *Sci. Robot.* **4**, eaau5872 (2019).
60. Xiang, S. et al. Nonlinear photonic neuromorphic chips for spiking reinforcement learning. *Optica* **13**, 457–468 (2025).
61. Towers, M. et al. Gymnasium: a standard interface for reinforcement learning environments. *arXiv preprint arXiv*: https://arxiv.org/abs/2407.17032 (2024).
62. Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv*: https://arxiv.org/abs/1707.06347 (2017).
63. Haarnoja, T., Zhou, A., Abbeel, P. & Levine, S. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. in *International Conference on Machine Learning*, 1861–1870 (PMLR, 2018).
64. Dinu, M., Quochi, F. & Garcia, H. Third-order nonlinearities in silicon at telecom wavelengths. *Appl. Phys. Lett.* **82**, 2954–2956 (2003).
65. Dulkeith, E., Vlasov, Y. A., Chen, X., Panoiu, N. C. & Osgood Jr, R. M. Self-phase-modulation in submicron silicon-on-insulator photonic wires. *Opt. Express* **14**, 5524–5534 (2006).
66. Lau, K. Y., Liu, X. & Qiu, J. A comparison for saturable absorbers: carbon nanotube versus graphene. *Adv. Photonics Res.* **3**, 2200023 (2022).
67. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).
68. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
69. Touvron, H. et al. Llama: open and efficient foundation language models. *arXiv preprint arXiv*: https://arxiv.org/abs/2302.13971 (2023).
70. Touvron, H. et al. Llama 2: open foundation and fine-tuned chat models. *arXiv preprint arXiv*: https://arxiv.org/abs/2307.09288 (2023).
71. Dubey, A. et al. The Llama 3 herd of models. *arXiv e-prints*. https://arxiv.org/abs/2407.21783 (2024).
72. Bi, X. et al. DeepSeek LLM: scaling open-source language models with longtermism. *arXiv preprint arXiv*: https://arxiv.org/abs/2401.02954 (2024).
73. Liu, A. et al. Deepseek-v2: a strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv*: https://arxiv.org/abs/2405.04434 (2024).
74. Liu, A. et al. Deepseek-v3 technical report. *arXiv preprint arXiv*: https://arxiv.org/abs/2412.19437 (2024).
75. Anderson, M., Ma, S.-Y., Wang, T., Wright, L. & McMahon, P. Optical transformers. *Transactions on Machine Learning Research* (2023).
76. Hua, S. et al. An integrated large-scale photonic accelerator with ultralow latency. *Nature* **640**, 361–367 (2025).
77. Ahmed, S. R., Baghdadi, R., Bernadskiy, M. et al. Universal photonic artificial intelligence acceleration. *Nature* **640**, 368–374 (2025).
78. Englund, D. et al. Controlling the spontaneous emission rate of single quantum dots in a two-dimensional photonic crystal. *Phys. Rev. Lett.* **95**, 013904 (2005).
79. Evans, R. E. et al. Photon-mediated interactions between quantum emitters in a diamond nanocavity. *Science* **362**, 662–665 (2018).
80. Laucht, A. et al. Mutual coupling of two semiconductor quantum dots via an optical nanocavity. *Phys. Rev. B–Condens. Matter Mater. Phys.* **82**, 075305 (2010).
81. Chen, Y.-C. et al. Laser writing of coherent colour centres in diamond. *Nat. Photonics* **11**, 77–80 (2017).
82. Laferrière, P. et al. Unity yield of deterministically positioned quantum dot single photon sources. *Sci. Rep.* **12**, 6376 (2022).
83. Sipahigil, A. et al. Indistinguishable photons from separated silicon-vacancy centers in diamond. *Phys. Rev. Lett.* **113**, 113602 (2014).
84. Rogers, L. J. et al. Multiple intrinsically identical single-photon emitters in the solid state. *Nat. Commun.* **5**, 4739 (2014).

Acknowledgements

The authors would like to thank Dr. Erfan Khoram, Dr. Zhicheng Wu, Prof. Jennifer. T. Choy, and Prof. E. Sifakis for insightful discussions.

Author contributions

Q.Z., J.K., and Z.Y. conceived the project and developed the theoretical framework. Q.Z. and Y.T. developed the nonlinear simulation solver. Q.Z. and J.K. developed the inverse-design framework. Q.Z., G.H., and Z.S. performed the training for the supervised learning and reinforcement learning tasks. Q.Z., J.K., Y.T., M.Z., and Z.Y. wrote the manuscript, with input from all authors. Z.Y. supervised the project.

Funding

Q.Z. and Z.Y. disclose support for the research of this work from the National Science Foundation, Grant No. 2016136, for the QLCI center Hybrid Quantum Architectures and Networks.

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/s41467-026-77202-y.

Correspondence and requests for materials should be addressed to Qingyi Zhou.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.