

1
2
3
4
5
6
7
8
9
10
11
12

Supplementary Information for Hundred-layer photonic deep learning

Supplementary Table S1. Comparisons of deep AI performances..... 2

Supplementary Note 1. The error model of ONN. 3

Supplementary Note 2. Formulations of the single-layer photonic computing. 4

Supplementary Note 3. Error-bounding with single-layer photonic computing..... 7

Supplementary Note 4. Computing capabilities of single-layer photonic chip..... 8

Supplementary Figures (Fig. S1-Fig. S16) 10

Supplementary Table S1. Comparisons of deep AI experiments

Methods	Task	Depth	Cycle time	Energy efficiency	Source
SLiM chip (Optoelectronics)	Image Classification Acc. 85.4% (ImageNet 1000 Classes)	100	0.1 ns @10 GHz	856.85 TOPS/W ^a	This work
	Text generation Loss function value 3.0397 (Wikipedia text)	384 (96×4)			
	Image Generation Loss function value 7.321 (ImageNet dataset)	640 (64×10)			
Nvidia A100 (Digital electronics)	Image Classification Acc. 85.9% (ImageNet 1000 Classes)	100	0.71 ns @1.4 GHz	0.078 TOPS/W	Ref ¹
	Text generation Loss function value 2.955 (Wikipedia text)	384 (96×4)			
	Image Generation (ImageNet dataset) Loss function value 7.278	640 (64×10)			
Free-space Diffractive Networks (Optoelectronics)	Image Classification Acc. 97.0% (MNIST 10 Classes)	5	5 μs	1.5 TOPS/W	Ref ²
On-chip 3-layer ONN (Optoelectronics)	Letter Classification Acc. 89.8% (Letters 4 Classes)	3	0.57 ns	0.071 TOPS/W	Ref ³
Memristor (Analog electronics)	Speech Recognition WER 9.3% (Librispeech)	9	300 ns	12.4 TOPS/W	Ref ⁴

14 *Acc., Accuracy; WER, word error rate

15 **Performances of Nvidia A100 and SLiM were evaluated on the same neural networks.

16 a. Calculation includes the power consumption of modulators, detectors, optical
 17 power sources, transimpedance amplifiers (TIAs), Digital-to-Analog Converters
 18 (DACs), and Analog-to-Digital Converters (ADCs).

19 **Supplementary Note 1. The error model of ONN.**

20 Implementing an ONN layer on traditional photonic computing chips requires multiple
 21 optical propagations and the nonlinear activation. As outlined in the main text, this
 22 process could be mathematically represented as $Y = f(\prod_{i=1}^L W_i X)$, where X , Y , f , and
 23 W_i respectively denotes the input, output, nonlinear function, and propagation layers.

24 When considering error, the distorted output is expressed as $Y + \Delta Y = f\left(\prod_{i=1}^L (W_i + \right.$
 25 $\left. \Delta W_i)(X + \Delta X)\right)$. Our findings demonstrate that the relative error tends to increase with
 26 the layer depth, which could be formulated as

$$27 \quad \text{Var}\left(\frac{\Delta Y}{Y}\right) \sim \sum_{i=1}^L \text{Var}\left(\frac{\Delta W_i}{W_i}\right). \quad (1.1)$$

28 **Proof:** Here we provide a proof utilizing the widely used ReLU nonlinear activation.
 29 First, we introduce an intermediate variable to capture the input to the nonlinear
 30 function: $A = \prod_{i=1}^L W_i X$ leading to $Y = f(A)$. Our aim is to establish that

$$31 \quad \text{Var}\left(\frac{\Delta A}{A}\right) \geq \sum_{i=1}^L \text{Var}\left(\frac{\Delta W_i}{W_i}\right). \quad (1.2)$$

32 To analyze the error propagation, we consider the expansion of the distorted result $A +$
 33 $\Delta A = \prod_{i=1}^L (W_i + \Delta W_i)(X + \Delta X)$. This can be expressed as: $A + \Delta A = A + A\left(\frac{\Delta X}{X} + \right.$
 34 $\left. \sum_i \left(\frac{\Delta W_i}{W_i}\right) + \sum_{i,j} \left(\frac{\Delta W_i \Delta W_j}{W_i W_j}\right) + \dots + \left(\frac{\Delta W_1 \Delta W_2 \dots \Delta W_N}{W_1 W_2 \dots W_N}\right)\right)$, from which it follows that $\Delta A/A =$
 35 $\left(\frac{\Delta X}{X} + \sum_i \left(\frac{\Delta W_i}{W_i}\right) + \sum_{i,j} \left(\frac{\Delta W_i \Delta W_j}{W_i W_j}\right) + \dots + \left(\frac{\Delta W_1 \Delta W_2 \dots \Delta W_N}{W_1 W_2 \dots W_N}\right)\right)$.

36 We first verify the Equation (1.2) for $L = 1$. From the definition of $\frac{\Delta A}{A}$, we have

$$37 \quad \text{Var}\left(\frac{\Delta A}{A}\right) = \text{Var}\left(\frac{\Delta X}{X} + \frac{\Delta W_1}{W_1} + \frac{\Delta X \Delta W_1}{X W_1}\right) = \text{Var}\left(\left(\frac{\Delta X}{X} + 1\right)\left(\frac{\Delta W_1}{W_1} + 1\right)\right). \text{ Following this,}$$

38 we simplify to get: $\text{Var}\left(\frac{\Delta A}{A}\right) = \text{Var}\left(\frac{\Delta X}{X}\right) + \text{Var}\left(\frac{\Delta W_1}{W_1}\right) + \text{Var}\left(\frac{\Delta X}{X}\right) \text{Var}\left(\frac{\Delta W_1}{W_1}\right) \geq$

$$39 \quad \text{Var}\left(\frac{\Delta X}{X}\right) + \text{Var}\left(\frac{\Delta W_1}{W_1}\right) \geq \sum_{i=1}^L \text{Var}\left(\frac{\Delta W_i}{W_i}\right).$$

40 Secondly, assuming Equation (1.2) holds for $L = n$, we can now analyze the case
 41 when $L = n + 1$:

$$42 \quad \text{Var}\left(\frac{\Delta A^{(n+1)}}{A^{(n+1)}}\right) = \text{Var}\left(\frac{\Delta A^{(n)}}{A^{(n)}} + \frac{\Delta W_{n+1}}{W_{n+1}} + \frac{\Delta A^{(n)} \Delta W_{n+1}}{A^{(n)} W_{n+1}}\right) = \text{Var}\left(\left(\frac{\Delta A^{(n)}}{A^{(n)}} + 1\right)\left(\frac{\Delta W_{n+1}}{W_{n+1}} + \right.\right.$$

 43 $\left. 1\right)\right) \geq \text{Var}\left(\frac{\Delta A^{(n)}}{A^{(n)}}\right) + \text{Var}\left(\frac{\Delta W_{n+1}}{W_{n+1}}\right). \text{ By employing mathematical induction, we}$

44 conclude that the Equation (1.2) holds for all integers L .

Next, we apply the Taylor expansion of $Y = f(A)$ around $A = A_0$. Considering that ΔY and ΔA are small, we express $Y = f(A) = f(A_0) + f'(A_0)(A - A_0) + o(A - A_0)$. Neglecting the higher-order terms, we find $\Delta Y = f(A) - f(A_0) = f'(A_0)\Delta A$. This implies $\frac{\Delta Y}{Y_0} = \frac{A_0 f'(A_0)}{f(A_0)} \cdot \frac{\Delta A}{A_0}$. In the non-vanishing region of ReLU nonlinearity, $\frac{A_0 f'(A_0)}{f(A_0)} = 1$. Consequently, we derive $\frac{\Delta Y}{Y_0} = \frac{\Delta A}{A_0}$. Combined with Equation (1.2), this indicates that the Equation (1.1) is validated and illustrates that the variations of the activation terms are proportional to the sum of the variations of the constituent variables.

We numerically validated the error trend by employing 10 samples of 10×10 matrices, with each entry randomly sampled from gaussian distribution. During the validation, a 10×1 random vector is sequentially multiplied with these 10 matrices, the computation is gradually distorted with the noise. For example, to analyze the errors with k layers, the first k layers are distorted with the noise. Finally, the errors are calculated for these 10 consecutive matrix multiplications. This procedure was repeated 10,000 times, and the mean and standard deviation of the errors were computed. As shown in Fig. S15, the mean error increases linearly with the number of noisy layers, corroborating our theoretical analysis.

Supplementary Note 2. Formulations of the single-layer photonic computing.

2.1 Formulation of the multi-wavelength spatial propagation

The optical transmission process can be modeled using matrix-vector multiplication (MVM), represented as $Wx = y$ with $W \in R^{M \times N}$, $x \in R^N$, and $y \in R^M$. To achieve this modeling, it is crucial to shape the propagation beam to correspond with the values represented in matrix W . Denoting the channel matrix as G . For a given spectral input, the propagation process can be described by the equation:

$$W_i = |G_i \Phi_i|^2, 1 \leq i \leq N, \quad (2.1)$$

where $W_i \in R^{M \times 1}$ denotes the i -th column of the target matrix, G_i signifies the propagation channel, and $\Phi_i \in R^{N_{array} \times 1}$ represents the input from the phase-modulated array. It is important to note that since $N_{array} \geq M$, the number of elements N_{array} sets the upper limit for the tunable dimension. To optimize the usage of optical bandwidth, the wavelengths λ_i are closely spaced between 1400 nm and 1600 nm. Given the proximity in wavelength distribution, the columns of the propagation matrix W_i and the input profile of the phase-modulated array Φ_i display significant similarities across different wavelengths, resulting in $G_i \approx G_0$, and $\Phi_i \approx \Phi_0$ for $1 \leq i \leq N$. Consequently, the matrices W_i can be approximated as $W_i \approx W_0$. In essence, Equation (2.1) describes a collection of N dependent first-order linear equations, defined by N_{array} variables. This formulation effectively reduces the DOFs from N_{array} to M .

2.2 Realizing full-DOF matrix with dispersive perturbation

To tackle the dimensionality reduction issue and realize arbitrary matrix W , we propose a method that separates different wavelength channels by successively introducing propagation distances d to each propagation channel via random phase perturbation. The phase term is adjusted to $\widehat{\Phi}_{ik} = \Phi_{ik} \cdot e^{jn2\pi\Delta(k)/\lambda_i}$, $1 \leq i \leq N$, $1 \leq k \leq N_{array}$. In this context, n is the effective index of refraction, while d signifies the difference in optical path lengths. The modified governing equation is now designated as Equation 2.2 and formulated as follows

$$W_i = |G_0 \widehat{\Phi}_i|^2, 1 \leq i \leq N. \quad (2.2)$$

The set of equations includes N linearly independent equations, enabling the realization of arbitrary matrices. Subsequently, we evaluate the effects of perturbation levels on the capability to generate arbitrary matrices by examining the comprehensive transmission matrix that links the spatial input phase with the spatial output across different wavelengths. For the analysis, we utilize a spatial array of dimensions 256, with both input and output dimensions fixed at 16, and a propagation distance of 0.5 m. The full transmission matrices for different wavelengths are depicted in Fig. S7a, which illustrates the phase (in the top-left panel) and the amplitude (in the bottom-left panel) of the transmission matrix realized with uniform propagation. It is important to note that phase and amplitude characteristics show considerable similarity across various wavelengths due to their close distribution. When perturbation is introduced, the phase distribution appears random (as seen in the middle column), leading to the loss of correlation among different spatial-spectral channels. The resultant decorrelation has similar effects of complete random perturbations in the right column.

The singularity of the transmission matrices is assessed through singular value decomposition (SVD). The results are visually depicted in Fig. S7b, demonstrating that propagation with dispersive perturbation yields a predominant distribution of eigenvalues. For uniform propagation, a degree of freedom (DOF) of 101 is maintained, with approximately half of the DOFs lost at a threshold of 0.01. The DOFs for dispersive perturbations at distances of 5 μm , 10 μm , 20 μm , and 30 μm are recorded as 225, 237, 252, and 249, respectively, which is comparable to the DOF of 253 observed with entirely random perturbations. Consequently, the perturbation length of 20 μm was selected for chip fabrication.

To validate the efficacy of dispersive perturbation, we conducted experiments on the fabricated chip. Utilizing the array characteristics, we evaluated the DOFs through variations in both input and output dimensions. A 128-element array was chosen for this assessment, which indicates a theoretical upper DOF limit of 128. Results with and without dispersion perturbation are shown in Fig. 2d. The precision was evaluated by varying input dimensions while keeping the product of input and output dimensions constant at 128. Notably, the method that did not incorporate dispersion perturbation exhibited errors exceeding 10% for input dimensions ranging from 6 to 16, whereas the error associated with the dispersive perturbation method remained below 1% for input dimensions ranging from 2 to 16.

2.3 Error decorrelation with nonlinearity.

In order to decorrelate the error from the signals, a zero-gradient nonlinear function is required that prevents error propagation to the next layer, specifically where $\partial f = 0$. However, a zero-gradient function tends to degenerate into a constant function, effectively nullifying the contributions of all subsequent layers. To isolate the error while preserving the nonlinearity, we introduce pulse noise into the gradient function, which is represented as $\delta(x - \Delta f)$, where δ here symbolizes the Kronecker delta function. The solution to this function is given by $f = \int \delta(x - \Delta f)$, resulting in a step activation at the perturbation Δf . During the signal detection phase, this activation can be efficiently performed through analog-to-digital conversion, eliminating the need for specialized nonlinear transformations. As elaborated in Methods and illustrated in Fig. S10, a dedicated controller board has been developed to facilitate parallel activation during the detection phase.

2.4 Advantage of redundancy-free computing.

Arbitrary matrix dimension and entry value. During data transmission across spatial-spectral channels, input data is allocated to the spectral dimension and broadcast spatially, accommodating variable input dimensions based on the number of utilized wavelengths. Our research further illustrates that the achievement of arbitrary DOFs with minimal errors can be realized through the perturbation injection. In contrast, when input is solely encoded within the spatial dimension, the potential to achieve arbitrary matrix entry values is limited, and both input and output dimensions remain fixed after device fabrication.

Error resilience. The resilience to errors originates from two principal physical aspects. First, the implementation of matrices necessitates only a single propagation step, irrespective of the matrix DOFs, effectively preventing the accumulation of errors that frequently arises with increased grid depth in existing Mach-Zehnder Interferometer (MZI) solutions that require multi-layer modulation and multiple propagation steps. Second, the perturbation injection at the detection phases facilitates nonlinearities with signal-error decorrelation, thereby fostering the development of deep error-tolerant physical AI systems and enabling the construction of deep neural networks with hundreds of layers.

Chip ensemble. By utilizing the spectral input dimension, we can segregate the input-loading chip from the matrix chip. This design approach eliminates the constraint of the matrix size to that of a single chip's physical dimensions. Instead, chips can be assembled post-fabrication to form larger matrices, adhering to the chiplet concept. As demonstrated in Fig. 3d, we constructed a 16-chip ensemble that increased the overall DOFs from 256 to 4096. Most importantly, the discrete spectral transmission allows for seamless integration with contemporary wavelength multiplexed data transmission systems, facilitating direct injection and computation of wideband transmitted signals on-chip.

169 Supplementary Note 3. Error-bounding with single-layer photonic computing.

170 In this section, we detail both the analytical framework and experimental procedures
171 for implementing error-tolerant deep ONNs.

172

173 **Analytical framework** We begin by defining the ideal inputs and weights for each
174 layer of the deep neural network, represented as X and W (with $W \in R^{N \times N}$ and $X \in$
175 $R^{N \times 1}$), along with their associated errors ΔX and ΔW . Our objective is to maintain
176 the error rates $\mathcal{E}(X)$ within a specified threshold $\mathcal{E}_0$ irrespective of the depth of the
177 neural network layers. To achieve this, on the SLiM computing, we propose the error-
178 tolerant computing methodology based on two principles. First, the normalized input
179 and weight spaces are uniformly compressed to i discrete levels, $-1, (-1 + 2/(i -$
180 $1)), \dots, (1 - 2/(i - 1)), 1$. The second principle involves defining the output levels
181 through $(i - 1)$ levels, $(-1 + 1/(i - 1)), \dots, 0, \dots, (1 - 1/(i - 1))$. We aim to
182 demonstrate that our single-layer computing framework guarantees bounded errors
183 across the deep layers. The error bounding criterion is encapsulated in two key
184 propositions.

185 **Proposition-1:** For any input error levels $Std(\Delta X) < 44\%$, an effective error
186 tolerance of $\mathcal{E}_0 < 50\%$ is achievable such that $\mathcal{E}((W + \Delta W)(X + \mathcal{E}(X) +$
187 $\Delta X))_D \leq \mathcal{E}_0$, provided $\mathcal{E}(X) \leq \mathcal{E}_0$.

188 Proof: The output variance is aligned to that of the input by normalizing it by a factor
189 of $1/\sqrt{N}$. The deviation between the experimental output of $((W + \Delta W)(X + \mathcal{E}(X) +$
190 $\Delta X))/\sqrt{N}$, and the ideal output $WX/\sqrt{N}$, is defined as $e = \frac{1}{\sqrt{N}}(W\mathcal{E}(X) + W\Delta X +$
191 $\Delta W(X + \mathcal{E}(X)))$. Omitting higher-order terms, each error term can be represented as a
192 random variable:

$$193 \quad e = \frac{1}{\sqrt{N}} \left(\left(\frac{2}{i-1} + 2 \right) / 2 \sum_{j=1}^{\mathcal{E}(X)N} w_j + \sum_{j=1}^N \Delta x_j w_j + \sum_{j=1}^N \Delta w_j x_j \right). \quad (3.1)$$

194 In the above derivation, we assume that the error rate $\mathcal{E}(X)$ correlates with errors in
195 the first $\mathcal{E}(X)N$ entries with error uniformly distributed between $\frac{2}{i-1}$ and 2. We then
196 can derive an upper bound for the variance of e under the assumption of mutual
197 independence of the w , x , Δw , Δx :

$$198 \quad Var(e) \leq \left(\frac{1}{i-1} + 1 \right)^2 \mathcal{E}_0 Var(w) + Var(\Delta x) Var(w) + Var(\Delta w) Var(x). \quad (3.2)$$

199 From the first principle, both w_j and x_j follow uniform distributions, yielding
200 variances of $(i + 1)/3(i - 1)$. The sum of these uniform distributions follows an
201 Irwin-Hall distribution, which approximates a normal distribution for large N with
202 variance $Var(e)$ and zero mean. Thus, we can estimate the output error ratios as
203 follows:

$$\mathcal{E} \left(((W + \Delta W)(X + \mathcal{E}(X) + \Delta X))_D \right) = P(|e| > \frac{2}{i-1}). \quad (3.3)$$

The experimental results in Fig. 2e indicate that $Var(\Delta w)$ remains below 0.01% (Std less than 1%). By integrating Equation (3.3) into the error evaluation $\mathcal{E} \left(((W + \Delta W)(X + \mathcal{E}(X) + \Delta X))_D \right) \leq \mathcal{E}_0$ with reference to the variance specified in Equation (3.2), we deduce that smaller values of i yield enhanced error tolerance. The optimal scenario occurs when the squeezed space comprises only two states as adopted in the experimental evaluation of SLiM, maximizing the upper limit of $\mathcal{E}_0$ at 50%. With $\mathcal{E}_0 = 50\%$ and $i = 2$, the solution of input error variance $Var(\Delta x) = 0.196$, and the corresponding standard deviation $Std(\Delta x) = 0.443$.

Proposition-2: For deep neural networks, if the input to the first layer satisfies $\mathcal{E}(X) \leq \mathcal{E}_0$, and if Proposition-1 holds for the subsequent layers (that is, $\mathcal{E} \left(((W + \Delta W)(X + \mathcal{E}(X) + \Delta X))_D \right) \leq \mathcal{E}_0$, when $\mathcal{E}(X) \leq \mathcal{E}_0$), then the error remains constrained below $\mathcal{E}_0$ across all layers.

Proof: Since $\mathcal{E} \left(((W + \Delta W)(X + \mathcal{E}(X) + \Delta X))_D \right)$ represents the output of a given layer with input X , we employ mathematical induction to prove Proposition-2 by successively feeding the output of one layer as the input to the subsequent layer, leveraging the error bounds established in Proposition-1. The derived error bound characterizes the upper limits of error-tolerant photonic computing. In practical implementations, actual errors are often significantly lower, contingent upon the network architecture. Fig. S3b illustrates that in the 10-G experiment, errors were constrained to 16%.

Supplementary Note 4. Computing capabilities of single-layer photonic chip

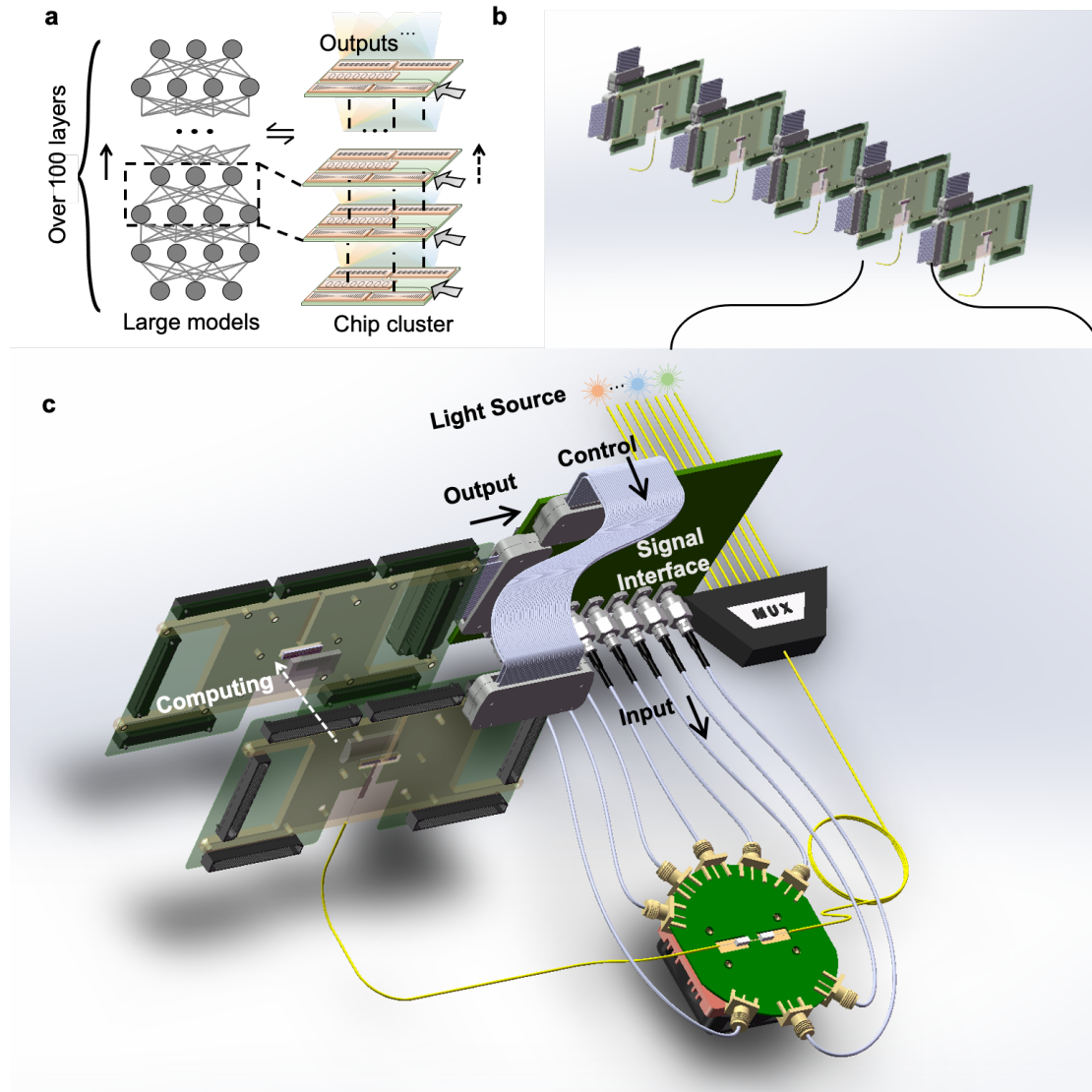
Contemporary digital electronics, which are extensively employed in various industries, are capable of supporting very deep neural network architectures; however, they are often limited in terms of energy efficiency. In contrast, neuromorphic hardware implemented with analog electronics demonstrates significantly enhanced energy efficiency, achieving tens of TOPS/W. Nevertheless, these technologies face constraints in terms of network depth, typically restricted to fewer than 10 layers, and operating frequencies that peak around several MHz, thus limiting their effectiveness on advanced datasets. Although optical neural networks provide exceptional computational parallelism and extremely low latency in the sub-nanosecond range, their inability to incorporate deep architectures restricts their broader applicability. The proposed SLiM chip represents a novel paradigm that establishes the error tolerance in deep ONNs, potential for diverse applications in both data centers and edge computing.

The error-tolerant SLiM computing enhances the degrees of freedom (DOFs) by implementing on-chip perturbation and encoding inputs across distinct wavelength channels, allowing for the rapid realization of diverse error-tolerant matrices for streamlined input processing.

Denoting the modulator number as K_m and output detector number as K_d , the optical propagation realizes a matrix of size $K_m \times K_d$ with total number of multiply-accumulate (MAC) operations being $K_m K_d$, and the total number of operations sums to $2K_m K_d$ ^{2, 5, 6}. The computational capacity is expressed in terms of operations per second (OPS) and is thus defined mathematically as $2K_m K_d F$, with F representing the modulation and detection frequency approaching tens of gigahertz. In this context, in the SLiM experiments of Fig. 4-5, $K_m = K_n = 128$, resulting in a computational capacity of 327.68 Tera OPS when utilizing a 10-GHz processing rate. The overall computing power could increase furthermore through chip ensemble, for example by integrating 16 chips as shown in Fig. 3, to 335.54 Peta OPS. The additional computational energy overhead associated with the error-tolerant system primarily arises from the driving power required for the phase shifter array within the transmission channel. The use of a low-power phase shifter facilitates a significant reduction in energy consumption to 1.7 μ W per π phase shift, achieved through energy-efficient resonant modulation. We also included the other components of the system including the Digital-to-Analog Converters (DACs), Analog-to-Digital Converters (ADCs), optical power sources, detectors, and calculate system energy efficiency. The SLiM chip employs 1-bit converters in high-speed experiments, which simplifies the circuit design and reduces energy consumption compared to higher-bit converters. With current technologies, the power dissipation of 1-bit DAC driving corresponds to the capacitance charge of the modulators, which consumes 6.3 fJ/bit⁷. The ADC process could be simplified to a low-power comparator, which draws 27.3fJ/bit⁸. The transimpedance amplifier consumes around 2.2 mW⁹. The specific power consumptions per channel for each component are as follows: $P_{SLiM} = 1.7 \mu$ W, $P_{DAC} = 63 \mu$ W, $P_{opt} = 200 \mu$ W, $P_{DET} = 250 \mu$ W, $P_{TIA} = 2.2$ mW, and $P_{ADC} = 273 \mu$ W. Denoting the input channel number, output channel number, and working frequency as N , M , and F (10-GHz), the overall system efficiency is calculated as $\eta =$

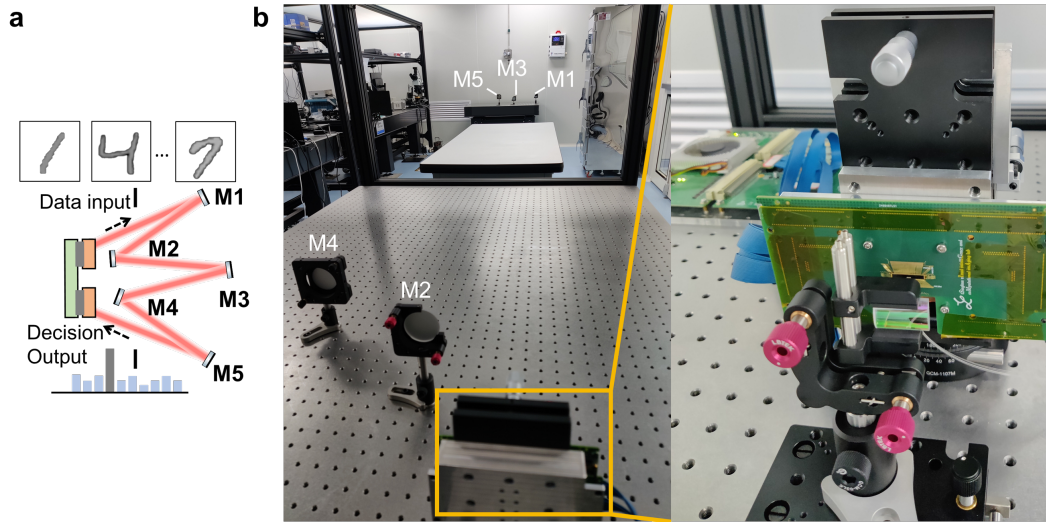
$$2MNF / (N(P_{DAC} + P_{opt}) + M(P_{ADC} + P_{DET} + P_{TIA}) + MNP_{SLiM})$$

. Our findings indicate that in the high-speed experiments, the system efficiency reaches 856.85 TOPS/W while considering the power consumption of modulators, detectors, optical power sources, transimpedance amplifiers (TIAs), Digital-to-Analog Converters (DACs), and Analog-to-Digital Converters (ADCs).

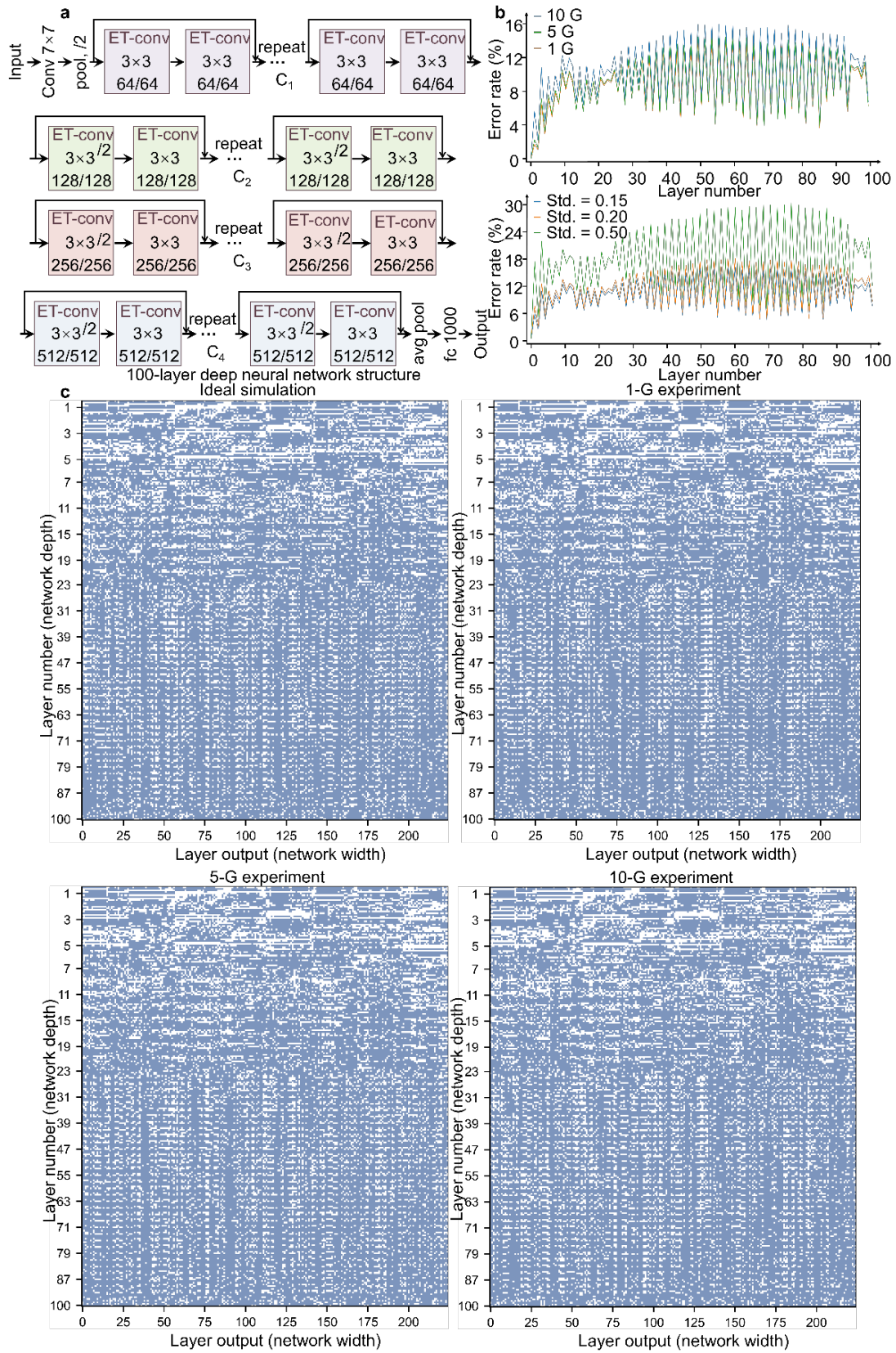


276

277 **Supplementary Figure S1. Deep neural networks with single-layer photonic chips (related to**
 278 **networks presented in Figs. 4-5).** **a**, The schematic illustration of deep learning with single-layer
 279 photonic chip. **b**, Corresponding system designs of (a). **c**, Experiment system for large models in
 280 Fig. 4 and Fig. 5. Eight-channel light sources (Light Source) with ITU DWDM channels are
 281 multiplexed into the input loading chip, which was driven by 10 GHz signals from the arbitrary
 282 waveform generators (Input). The computation was facilitated between single-layer chips with
 283 cylindrical lenses collimating the light from and to the chip, following which the high-speed results
 284 were captured with an oscilloscope (Output). The weights were configured with the controller board
 285 (Control) and the inter-chip distance was set to 0.5 m.
 286

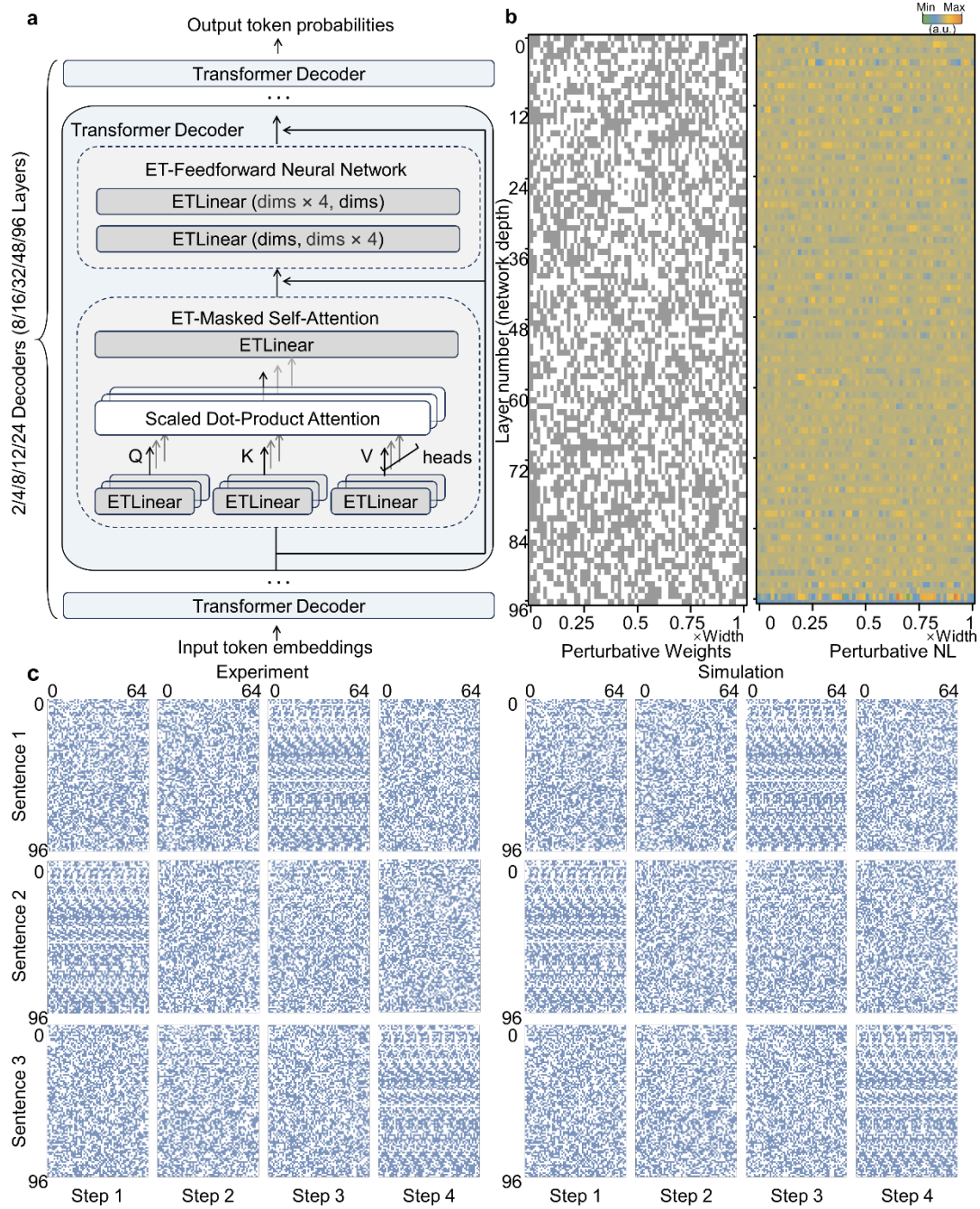


Supplementary Figure S2. Schematics (a) and Experimental setup (b) for the evaluation of spatial depth limit with single-layer photonic computing (related to networks presented in Fig. 3). M1 to M5 represent five mirrors utilized to increase spatial separation by approximately sixfold, allowing for propagation distances ranging from 0.01 m to 45 m.



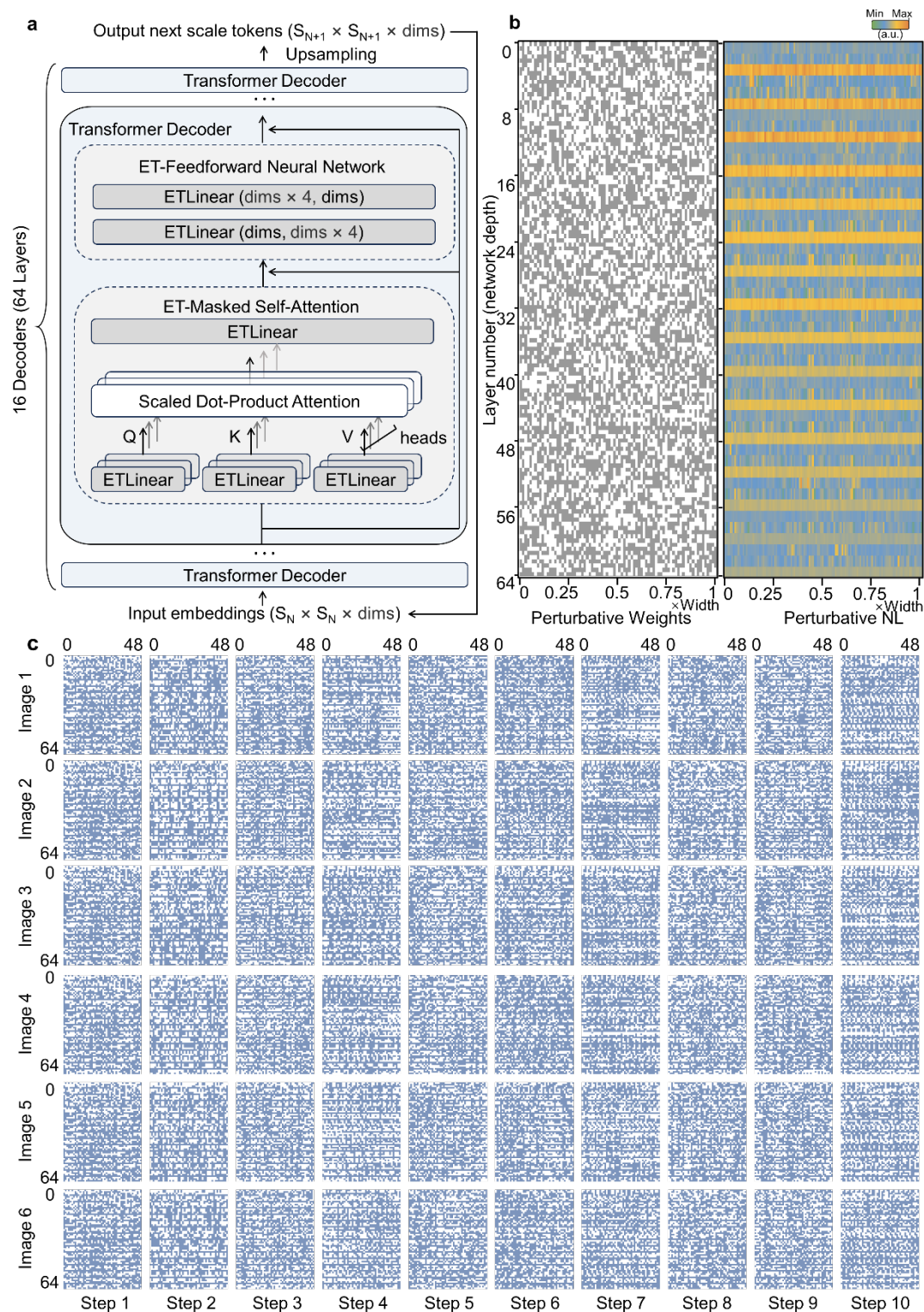
Supplementary Figure S3. The performance of the error-tolerant 100-layer deep photonic neural network, along with the corresponding experimental results on the ImageNet-1000 dataset. a, The network architecture is illustrated, which is built upon four main convolutional

blocks, concluding with a fully connected layer. Each block comprises sequential residual convolution modules, with their respective quantities labeled as C1, C2, C3, and C4. A series of models were developed featuring different layers for experimental purposes: an 18-layer model (C1/C2/C3/C4 = 2/2/2/2), a 34-layer model (C1/C2/C3/C4 = 3/4/6/3), and a comprehensive 100-layer model (C1/C2/C3/C4 = 3/8/35/3). **b**, The error rates associated to the experiments carried out using the 100-layer error-tolerant model. The upper section showcases results from tests conducted at frequencies of 1 G, 5 G, and 10 G, yielding error rates of 9.04%, 9.22%, and 10.34%, respectively. The lower section details the outcomes under conditions of Gaussian noise, where standard deviations of 0.15, 0.20, and 0.50 corresponded to error rates of 10.48%, 11.56%, and 18.78%. **c**, The experimental results for the first sample drawn from the ImageNet validation set as processed by the 100-layer deep neural network. The outputs from the four convolutional blocks were 200,704, 100,352, 501,706, and 25,088. The data for the error-tolerant activations from each layer are organized into four blocks, each with designated configurations and output quantities. The downsampling ratio is set at 110, with the arrangement of rows in the four blocks being 8/4/2/1, creating a structure where each row contains 225 columns. In this schematic representation, activated neurons are indicated by white grids (1s) and inactive neurons by blue grids (-1s). ET-conv refers to error-tolerant convolution.



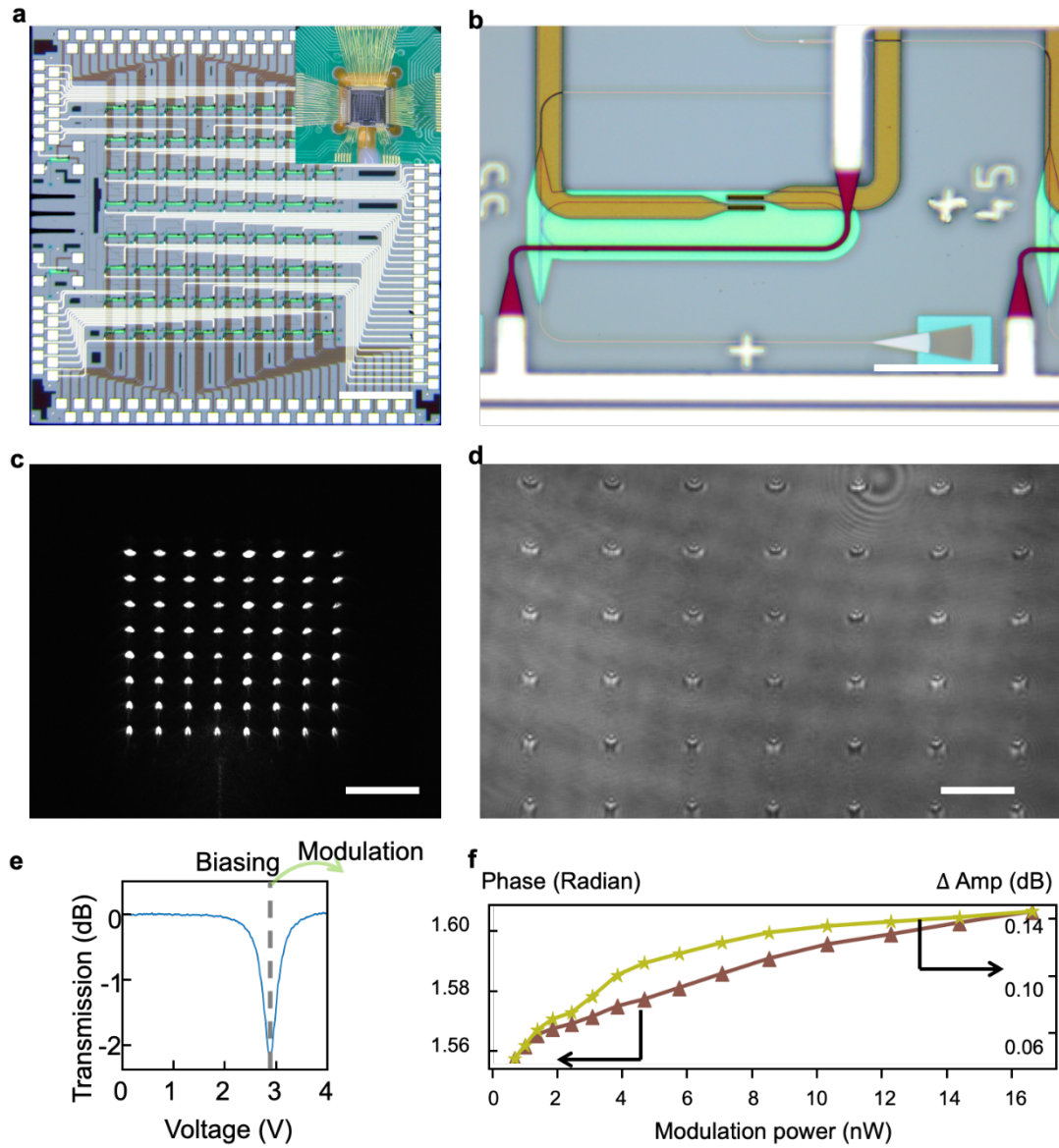
Supplementary Figure S4. Error-tolerant ONNs as a large language model. **a**, The structure of error-tolerant large ONN, consisting of cascaded blocks of error-tolerant decoders. Each decoder integrates a self-attention mechanism and a feedforward neural network, comprising four error-tolerant linear layers (ETLinear). The analysis includes four configurations with 2, 4, 8, 12, and 24 decoders. **b**, The network models that have successfully converged, featuring 96 layers. The white and grey grids symbolize activations of 1s and -1s, respectively, with each layer's width adjusted to 64 for clarity in visualization. **c**, This panel compares experimental results with simulations of a three-sentence generation task. It illustrates activations from 24 transformer blocks within the 96-layer GPT-2 neural network, where each transformer block consists of 4 layers. The outputs for the first layer (QKV) are formatted as $(L \times 3 \times 1024)$, while the second through fourth layers output $(L \times 1024)$, $(L \times 4096)$, and $(L \times 1024)$ respectively. To enhance figure visualization, nearest-neighbor

downsampling was used to reduce each layer’s output to a 64-dimensional representation, resulting in each row containing 64 columns that span 96 layers. The initial input token lengths for the three sentences are 6, 8, and 5, increasing by one token with each subsequent step. Again, the white and blue grids represent on and off neuron activations. A.u., arbitrary unit.



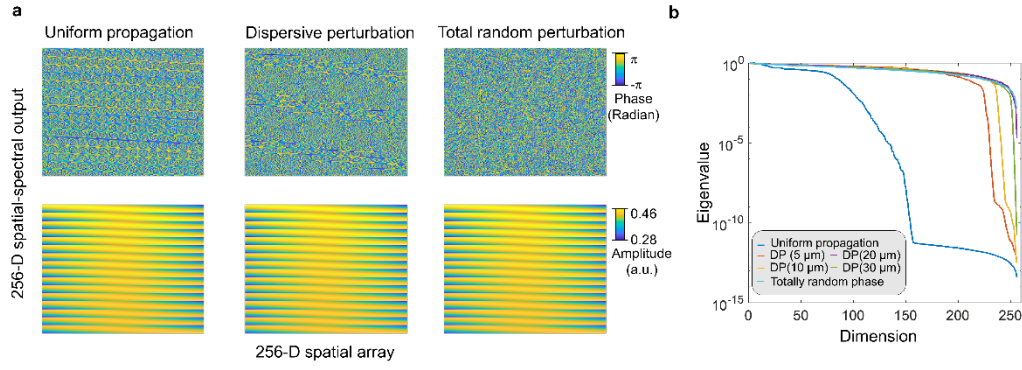
Supplementary Figure S5. Error-tolerant architectures for conditional image generation. a, The structure of networks composed of cascaded blocks of error-tolerant decoders. Each decoder

integrates a self-attention mechanism along with a feedforward neural network, featuring four fault-tolerant linear layers (ETLinear). Operating at a dimension of 1024, the generative process unfolds over ten steps, with each step propagating through 16 decoders. The outputs from each step are upsampled following a scaling order of 1, 4, 9, 16, 25, 36, 64, 100, 169, and 256, serving as the input tokens for the subsequent step. **b**, The converged network models consisting of 64 layers. The white and grey grids in the network signify activations of 1s and -1s, respectively. For enhanced visualization, the width of each layer has been adjusted to 160 (divides to two rows, each with 80 columns). **c**, The experimental results from the generation of six images. The activations from 16 transformer blocks within the 64-layer neural network are displayed, with each transformer block consisting of 4 layers. The first (QKV) layer outputs ($P^2 \times 3 \times 1024$), while the second, third, and fourth layers produce outputs of ($P^2 \times 1024$), ($P^2 \times 4096$), and ($P^2 \times 1024$), respectively. Nearest-neighbor downsampling has been utilized to rescale each layer output to 48 dimensions, resulting in each row containing 48 columns across 64 layers. The input token lengths for the six images begin at 1, incrementing through the sequences of 1, 4, 9, 16, 25, 36, 64, 100, 169, and 256. Again, the white and blue grids denote on and off neuron activations. A.u., arbitrary unit.

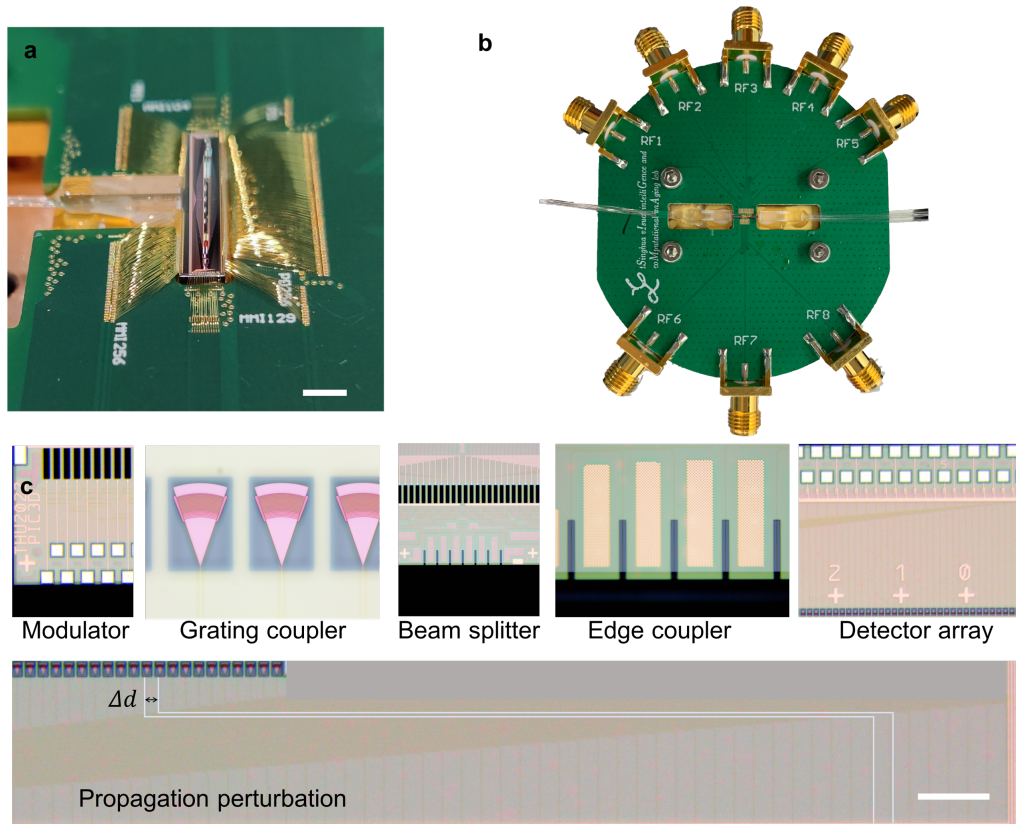


352

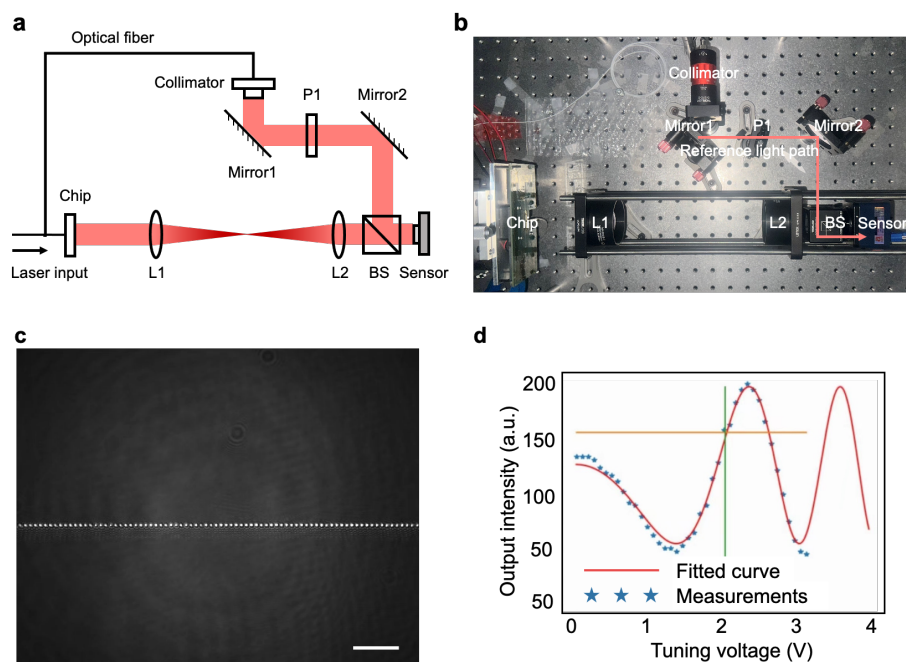
353 **Supplementary Figure S6. High-efficiency phase modulation chip.** **a**, The overall design of the
 354 low-power modulation chip featuring an 8×8 array, with an inset highlighting the complete
 355 packaging. **b**, The resonant modulation unit, which is optimized for operation in the over-coupling
 356 regime. **c**, The emission characteristics of the array. **d**, The observed pattern during the phase
 357 measurement of the array. **e**, The process of biasing and modulating the resonant phase modulator
 358 is detailed. **f**, The modulation characteristics in response to varying modulation power level. Scale
 359 bars are set at $500 \mu\text{m}$ (a), $50 \mu\text{m}$ (b), $500 \mu\text{m}$ (c), and $200 \mu\text{m}$ (d).



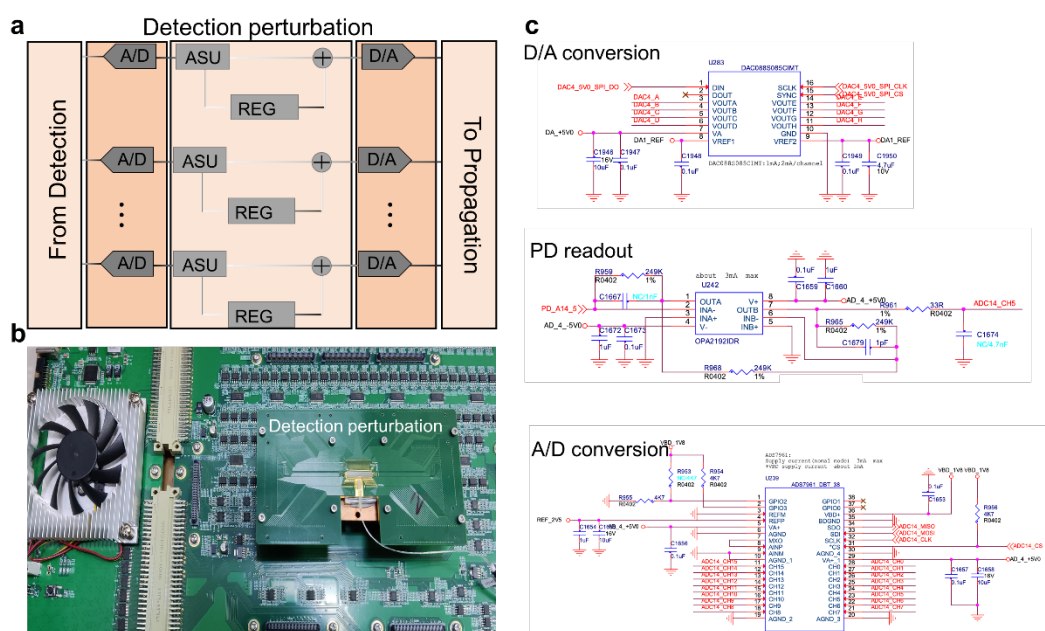
Supplementary Figure S7. Channel matrices analysis with dispersive propagation. **a**, Channel matrices illustrating spatial-spectral propagation dynamics, compared under conditions with and without perturbation injections. **b**, Distribution of eigenvalues.



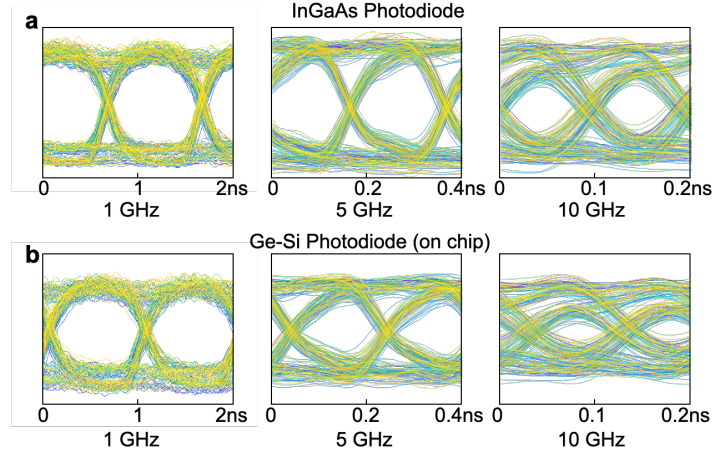
Supplementary Figure S8. The details of the fabricated chips. **a**, The packaged error-tolerant chip is presented. **b**, The packaged chip designed for input loading. High-speed signal is injected via eight SMA connectors. **c**, The on-chip photonic structures and the propagation perturbation. Scalebar, 3 mm (a) and 200 μm (c).



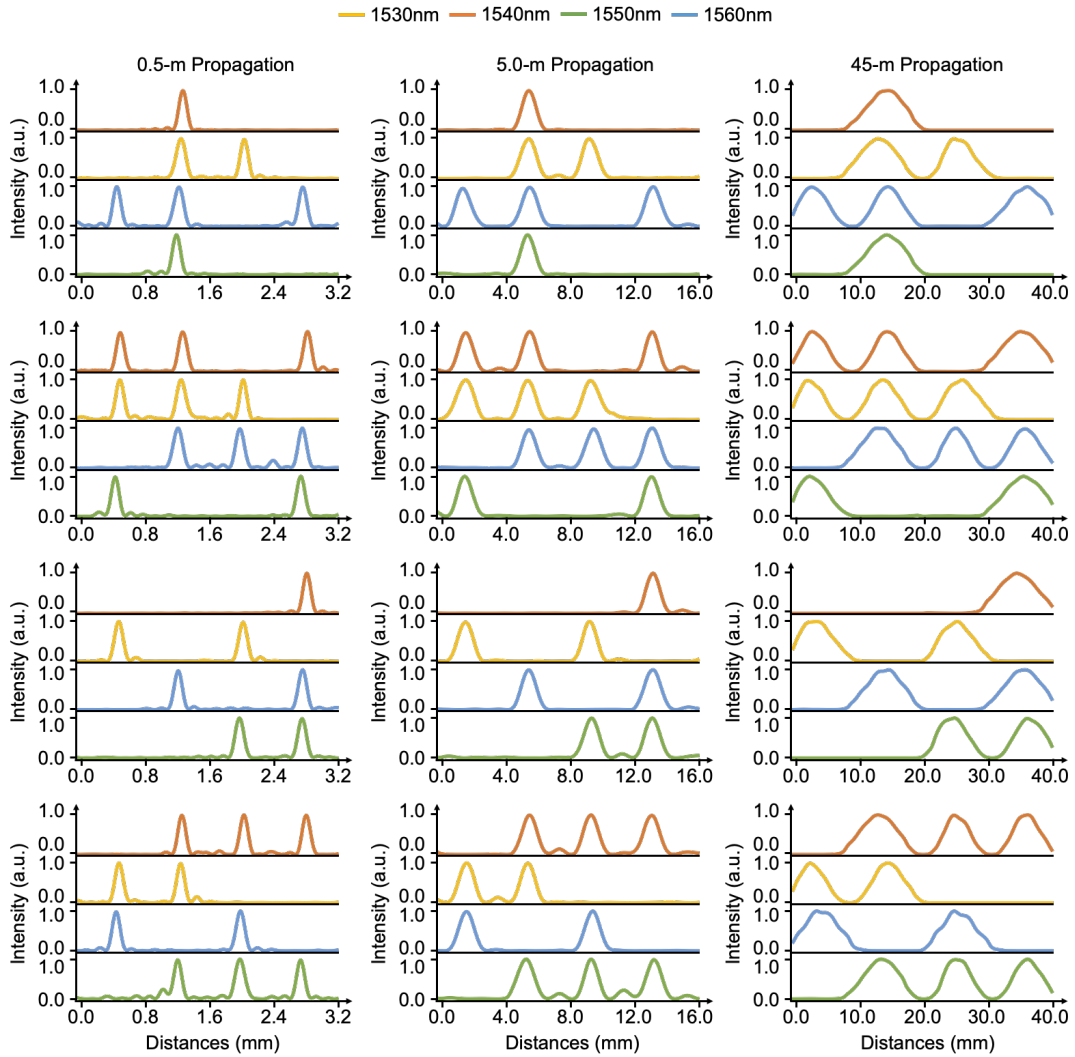
Supplementary Figure S9. Characterization of the single-layer chip. **a**, Schematic representation of the calibration holographic systems. **b**, The experimental setup utilized for testing. **c**, Infrared interference image captured during the experiment. **d**, Sample curve fitting results for a specific modulator. Scale bar: 400 μm .



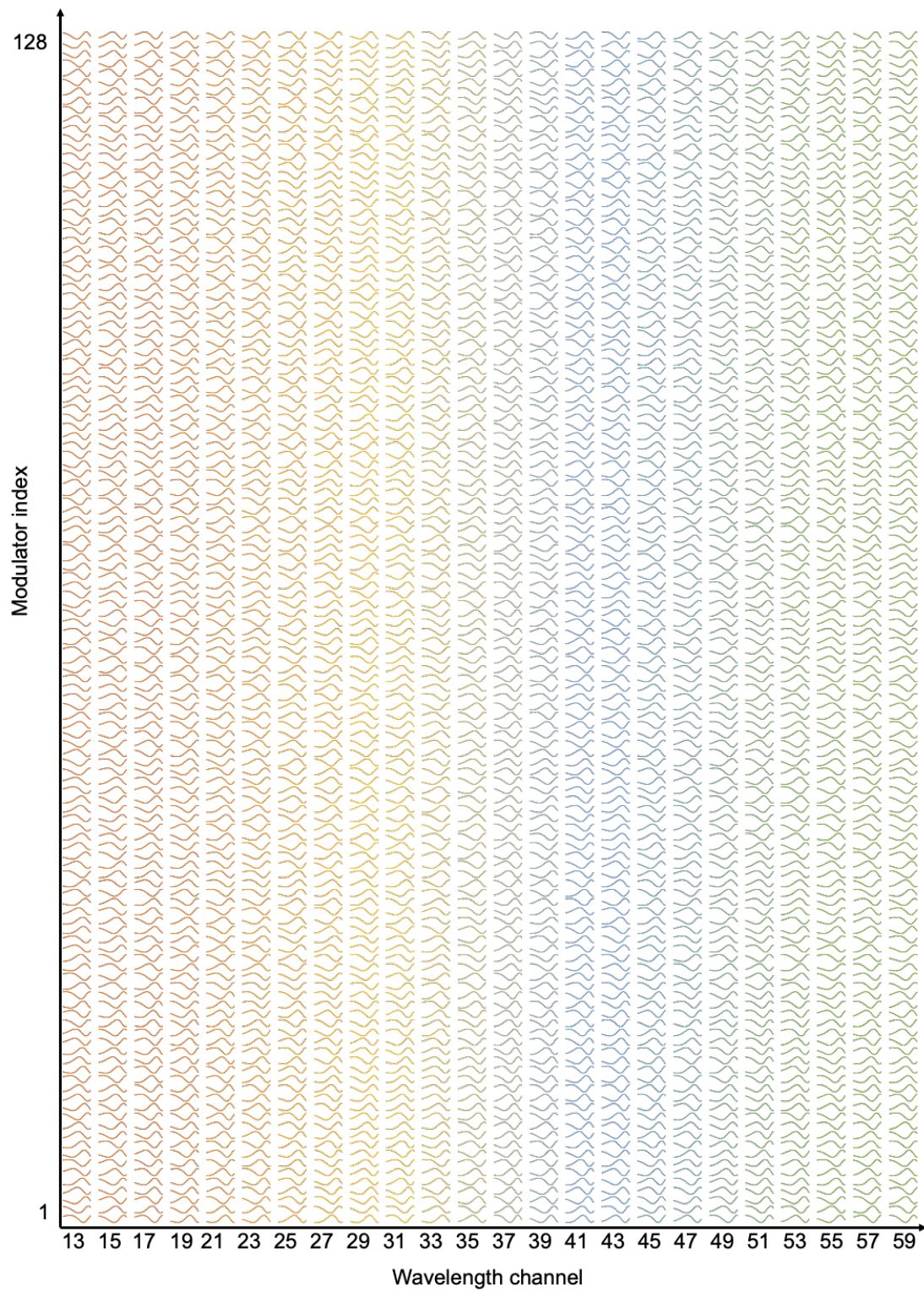
Supplementary Figure S10. Controller board design and integration for detection perturbation. **a**, The diagram illustrating the design of the controller. The signals flow from the detection part, and proceed to the propagation on next chip after detection perturbation. **b**, The packaged controller board in conjunction with the error-tolerant single-layer chip. **c**, Schematic representations of the circuitry details in (a).



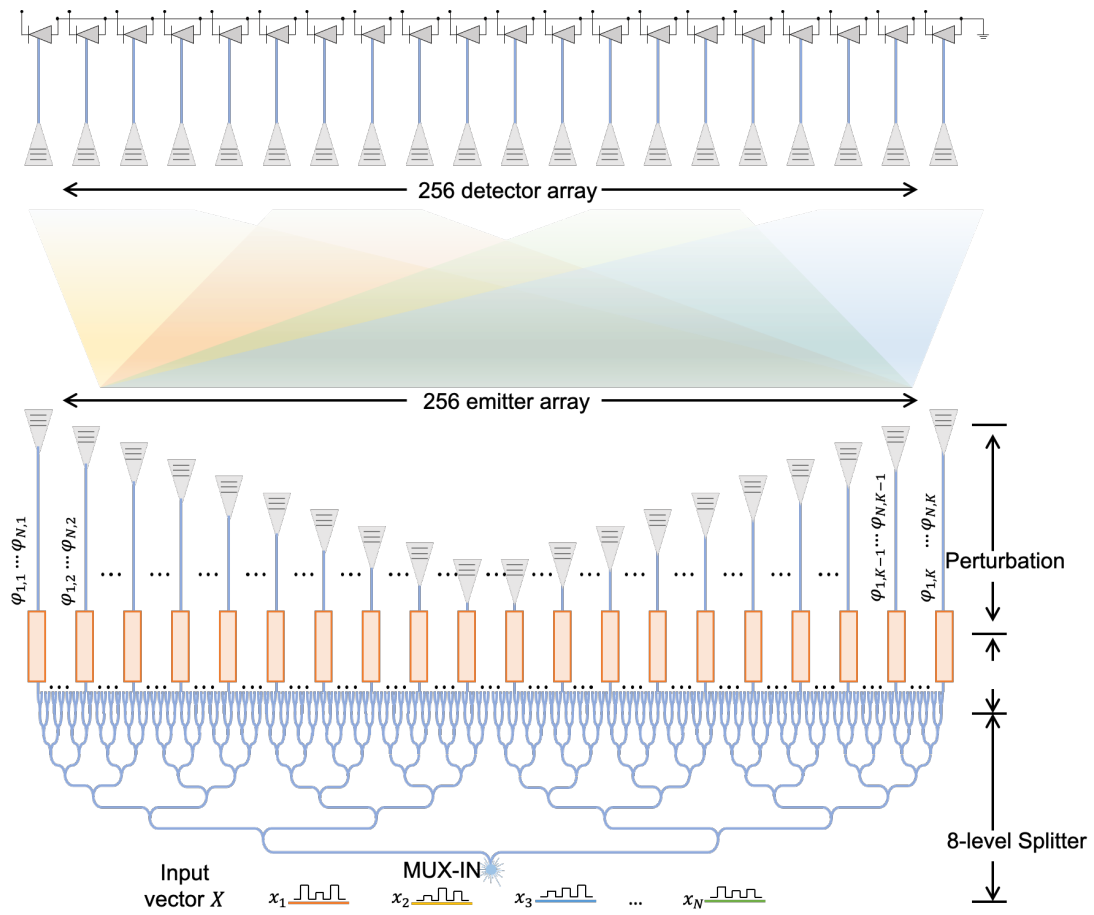
Supplementary Figure S11. The eye-diagram measurements under high-speed scenarios.



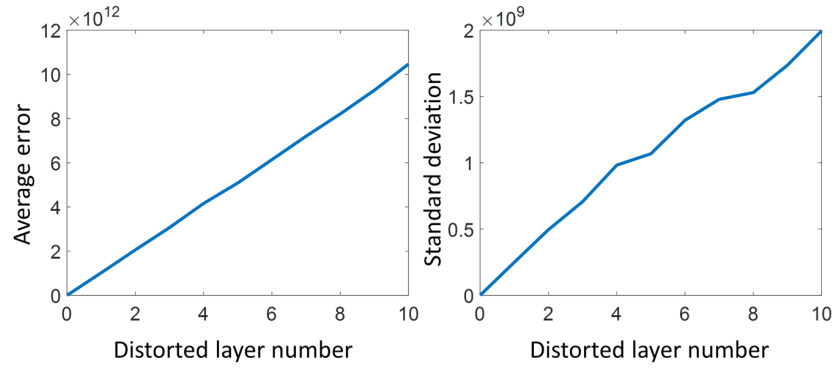
Supplementary Figure S12. The measured matrices at propagation distances of 0.5 m, 5.0 m, and 45 m. The data are organized in four rows, each representing composite vectors: $[e_3, e_1 + e_2 + e_3 + e_4, e_2, e_3]$; $[e_1 + e_2 + e_4, e_1 + e_2 + e_3, e_2 + e_3, e_1 + e_3 + e_4]$; $[e_2, e_3, e_2 + e_4, e_1 + e_3 + e_4]$; $[e_2 + e_3, e_1 + e_2 + e_4, e_1 + e_3 + e_4, e_1 + e_4]$



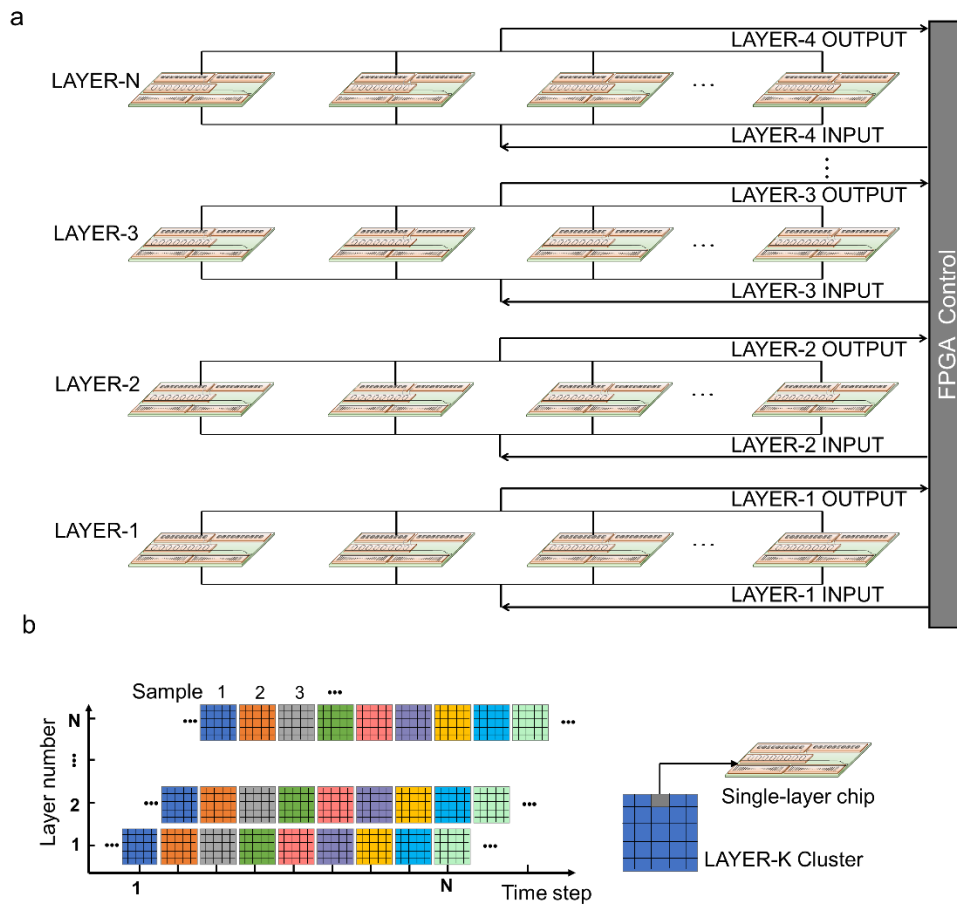
Supplementary Figure S13. Calibration of modulation characteristics across 24 wavelengths (ITU channels C13 - C59), which were utilized for the experiments depicted in Figs. 3 to 5.



Supplementary Figure S14. Circuit diagram of the SLiM chip and schematics of computing.



Supplementary Figure S15. Analysis of the error accumulation with respect to layers. Left panel: average error, right panel: standard deviation.



Supplementary Figure S16. Parallel processing with SLiM chip cluster. a, The N layer neural network could be realized by N SLiM cluster. Each SLiM chip in one cluster parallelly implements the computations of one layer. The chips are scheduled by a controller that allocates the input data and collect the outputs. b, The streamline of the cluster processing, the N cluster can process N samples concurrently.

References

1. Choquette, J., Gandhi, W., Giroux, O., Stam, N. & Krashinsky, R. Nvidia a100 tensor core gpu: Performance and innovation. *IEEE Micro* **41**, 29-35 (2021).
2. Zhou, T. et al. Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. *Nature Photonics* **15**, 367-373 (2021).
3. Ashtiani, F., Geers, A. J. & Aflatouni, F. An on-chip photonic deep neural network for image classification. *Nature*, 1-6 (2022).
4. Ambrogio, S. et al. An analog-AI chip for energy-efficient speech recognition and transcription. *Nature* **620**, 768-775 (2023).
5. Xu, Z., Zhou, T., Ma, M., Deng, C., Dai, Q. & Fang, L. Large-scale photonic chiplet Taichi empowers 160-TOPS/W artificial general intelligence. *Science* **384**, 202-209 (2024).
6. Miscuglio, M. et al. Massively parallel amplitude-only Fourier neural network. *Optica* **7**, 1812-1819 (2020).
7. Yuan, Y. et al. A 5× 200 Gbps microring modulator silicon chip empowered by two-segment Z-shape junctions. *Nature Communications* **15**, 918 (2024).
8. Wang, J., Ye, F. & Ren, J. A Three-Stage Comparator with High Speed and Low Power. In: *2021 IEEE 14th International Conference on ASIC (ASICON)* (2021).
9. Kromer, C. et al. A low-power 20-GHz 52-dB/spl Omega/transimpedance amplifier in 80-nm CMOS. *IEEE Journal of Solid-State Circuits* **39**, 885-894 (2004).