Photonic Restricted Boltzmann Machine for Content Generation Tasks

Li Luo,* Yisheng Fang,* Wanyi Zhang, and Zhichao Ruan†

Zhejiang Key Laboratory of Micro-nano Quantum Chips and Quantum Control, School of Physics, and State Key Laboratory for Extreme Photonics and Instrumentation, Zhejiang University, Hangzhou 310027, China

(Received 21 August 2025; accepted 4 March 2026; published 31 March 2026)

The restricted Boltzmann machine (RBM) is a neural network based on the Ising model, well known for its ability to learn probability distributions and stochastically generate new content. However, the high computational cost of Gibbs sampling in content generation tasks imposes significant bottlenecks on electronic implementations. Here, we propose a photonic restricted Boltzmann machine (PRBM) that leverages photonic computing to accelerate Gibbs sampling, enabling efficient content generation. By introducing an efficient encoding method, the PRBM eliminates the need for computationally intensive matrix decomposition and reduces the computational complexity of a single dot product in Gibbs sampling from $O(N)$ to $O(1)$. Moreover, with the advantage of photonic computation of non–von Neumann computing architecture, we physically encode the model parameters onto the phases of different wavelength light fields on a spatial light modulator, substantially enhancing computation efficiency for large-scale RBMs. We experimentally validate the photonic-accelerated Gibbs sampling by simulating a two-dimensional Ising model, where the observed phase transition temperature closely matches the theoretical predictions. Beyond physics-inspired tasks, the PRBM demonstrates strong performance in generating and restoring diverse content, including images and temporal sequences. The scalability and reduced training cost of the PRBM framework underscore its potential as a promising pathway for advancing photonic computing in generative artificial intelligence.

DOI: 10.1103/k6f7-gtr7

Subject Areas: Interdisciplinary Physics, Optics

I. INTRODUCTION

The restricted Boltzmann machine (RBM) is a stochastic generative neural network based on the Ising model, characterized by spin interactions solely between the visible and hidden layers [1,2]. RBMs can be applied to a variety of tasks by learning the probability distribution of the training data. It is renowned for the development of deep learning [3–6] and for demonstrating features in classification and recognition tasks, particularly generating new content [6–10]. Furthermore, by combining a RBM with a recurrent neural network, the time-varying RBM enables effective modeling of temporal data [11]. In RBMs, Gibbs sampling plays a crucial role by decomposing high-dimensional problems into multiple conditional distribution problems with Markov chains and sampling sequentially from the conditional distribution [12,13]. However, traditional electronic computing struggles with long chains of Gibbs sampling, especially when dealing with large-scale datasets, which demand enormous computational resources [6,14–16].

Recently, there has been growing interest in developing unconventional computational architectures to implement optical neural networks [17–20]. In addition, to solve NP-hard problems, various Ising machines have been developed for simulating Ising Hamiltonians, based on optical parametric oscillators [21–25], lasers [26–30], polaritons [31–34], trapped ions [35], atomic and photonic condensates [36,37], electronic memorisers [38], superconducting qubits [39–41], and optoelectronic circuits [42–51]. Unlike other Ising machines evolving spin configurations according to nonlinear dynamics to search for the ground state, the spatial photonic Ising machine (SPIM) has been demonstrated reliable for sampling large-scale Ising spin systems at a given temperature, even up to tens of thousands of spins [52]. By leveraging multiple degrees of freedom of light, such as time [53], space [54,55], and wavelength [56], several division multiplexing schemes have been developed, harnessing the advantages of the ultrafast-speed and high-power-efficiency photonic computation [53–56]. In contrast to traditional von Neumann architectures, SPIMs encode both the spin interaction matrix and spin configurations using spatial light modulators (SLMs) [52,53,57–61] or digital

*These authors contributed equally to this work.
†Contact author: zhichao@zju.edu.cn

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

micromirror devices [62,63], with measuring light intensity and instantaneously computing Ising Hamiltonian with Mattis-type interaction [52,57,58,62], antiferromagnetic interaction [59], low-rank interaction [53,60], and four-body interaction [61].

On the other hand, these Ising computation schemes are unable to accelerate the computation for RBMs, due to the discrepancy of the spin distribution and interaction. These Ising machines are designed to simulate the Ising model wherein spins reside within a single layer and interact with each other, whereas RBMs partition spins into distinct visible and hidden layers, with spin interactions existing solely between two layers. In particular, SPIMs simulating such spin interactions in RBMs have to introduce substantial redundant zero-value elements [64] and further require digital computation of an eigenvalue decomposition [53,54] or a Cholesky-like decomposition [56] with the cost of computation complexity $O(N^3)$. Such computation cost becomes significant when the total spin number in the visible and hidden layers is large and is further intensified when accounting for external magnetic fields on each spin. Additionally, SPIMs utilize Metropolis-Hastings sampling evaluating the impact of a single spin flip on the global system, whereas Gibbs sampling evaluates the impact of the locally correlated system on a single spin, which is crucial for accelerating computation in complex high-dimensional problems. Therefore, how to effectively encode the visible and hidden layers and how to implement Gibbs sampling constitute two key challenges in demonstrating the photonic computing acceleration for RBMs.

In this article, we propose a photonic restricted Boltzmann machine (PRBM) that leverages photonic computing to accelerate Gibbs sampling and demonstrate content generation tasks. The PRBM is designed based on wavelength-division multiplexing spatial Ising machines, inheriting the scalability of SPIM architecture. By developing an encoding method, we eliminate the need for the decomposition computation of interaction matrix and reduce the computational complexity of a single dot product in Eq. (3) from $O(N)$ to $O(1)$, thereby significantly reducing computational costs. In addition, with the advantage of photonic computation of non–von Neumann computing architecture, we physically encode the model parameters onto the phases of different wavelength light fields on the SLM, substantially enhancing computation efficiency for large-scale RBMs. We further validate Gibbs sampling process by simulating an analytically solvable two-dimensional Ising model [65] and experimentally observe the phase transition at the critical temperature agreeing well with the exact theoretical value. Moreover, we experimentally demonstrate that the proposed PRBM effectively acquires the capability of generating novel content, with successful implementations in image generation and temporal data processing, as exemplified by music generation. Finally, we prospect the scalability and the training advances of the PRBM framework, suggesting its potential to offer a promising pathway for applications of generative artificial intelligence (AI).

II. PHOTONIC COMPUTING SCHEME OF RESTRICTED BOLTZMANN MACHINE

We consider RBMs with a Hamiltonian $H(\mathbf{v},\mathbf{h}) = -\sum_{ij} v_i W_{ij} h_j - \sum_i a_i v_i - \sum_j b_j h_j$, where v_i and h_j taking values of ± 1 represent the spin configurations in the visible and the hidden layers, comprising N_v and N_h spins, respectively. The spins interact between the two layers solely as W_{ij}, where a_i and b_j are the magnetic field terms for visible-layer spins and hidden-layer spins, respectively [see Fig. 1(a) and Appendix A]. Figure 1(b) shows the proposed PRBM, with a specific encoding method to enable the photonic computing of Gibbs sampling (the detailed experimental setup in Supplemental Material Sec. I [66]). Here, a collimated supercontinuum laser is diffracted by a grating, and different wavelength light focuses on a SLM along the x axis with a cylindrical lens.

We divide the SLM into three distinct regions along the y axis. We take the example of the generation process as illustrated in Fig. 1(c). When sampling from the visible layer $\mathbf{v}$ to the hidden layer $\mathbf{h}$, for each hidden-layer spin h_j, the visible-layer spin v_i with the spin interaction W_{ij} and under the magnetic field b_j is encoded at the jth wavelength, and the phase modulation along the y axis of the SLM is given as

$$\begin{aligned}
&\text{Region I:}\ \{\phi_j^{m,n}\}^{\mathrm{I}} = h_j \frac{\pi}{2},\\
&\text{Region II:}\ \{\phi_{i,j}^{m,n}\}^{\mathrm{II}} = v_i \frac{\pi}{2} + (-1)^{m+n}\alpha_{ij},\\
&\text{Region III:}\ \{\phi_j^{m,n}\}^{\mathrm{III}} = s \frac{\pi}{2} + (-1)^{m+n}\beta_j. \qquad (1)
\end{aligned}$$

Here, an auxiliary spin $s = 1$ is used for magnetic field b_j in region III. We encode each spin by a macropixel which consists of $N_x \times N_y$ pixels, where m and n are the pixel indices within each spin, $1 \le m \le N_x$, and $1 \le n \le N_y$.

In order to implement the spin interaction and the external magnetic field, we propose the gauge transform [57,59]: In regions II and III, the phase modulations have additional contributions from a checkerboard modulation, corresponding to rotating each spin by the angle $\alpha_{ij} = \arccos(W_{ij}/L)$ and $\beta_j = \arccos(b_j/L)$, which are equivalent spins $v'_{ij} = v_i W_{ij}$ and $s'_j = s b_j$, respectively. The normalization parameter L is the maximum absolute value of the interactions and external magnetic fields.

The key physical process of the proposed PRBM is contributed by focusing the modulated light beam with a Fourier lens and measuring the light intensity at the center. The coherent interference leads to the intensity contribution for each wavelength is derived as $I_j(0,0) = [(A_0 C)^2/(\lambda f)^2](h_j + b_j/L + \sum_{i=1}^{N_v} W_{ij} v_i/L)^2$, where A_0 is the

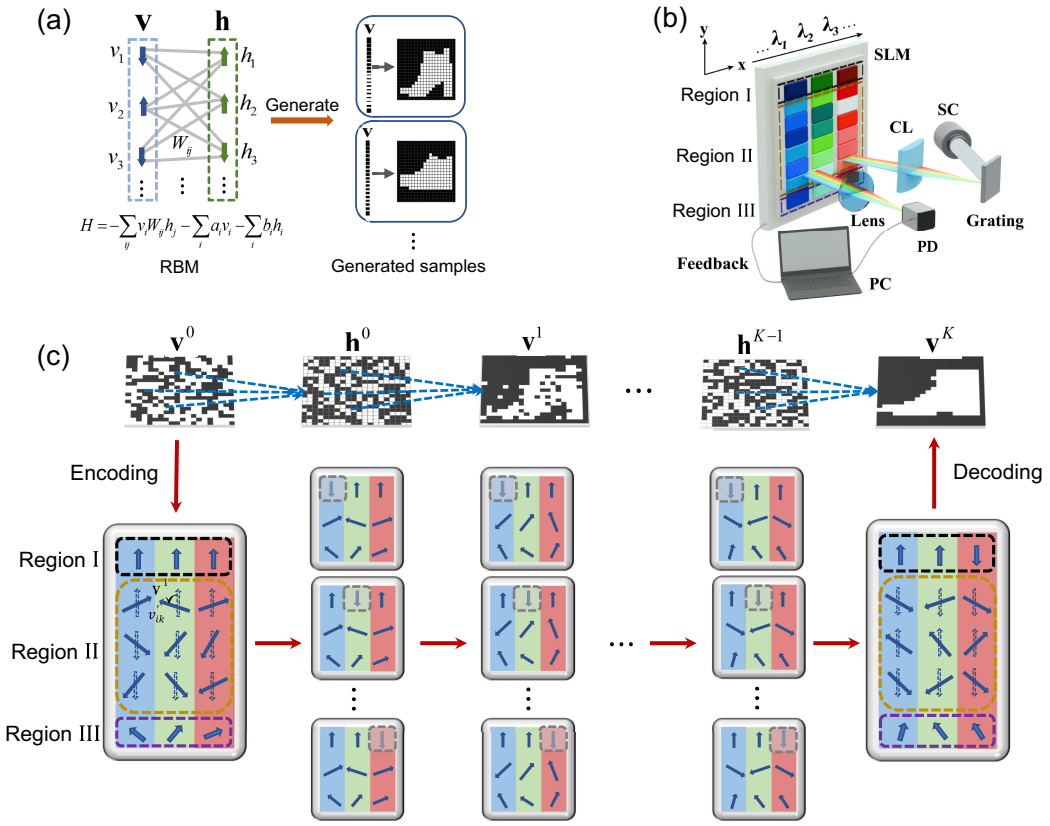


FIG. 1. Schematic of photonic restricted Boltzmann machine with Gibbs sampling. (a) The RBM consists of visible and hidden layers of spins $\mathbf{v}$ and $\mathbf{h}$, with spin interactions W_{ij} solely between two layers and under external magnetic fields a_i and b_j, respectively. For content generation tasks, RBM is trained by maximizing the log-likelihood function and performed to generate new content, by Gibbs sampling. (b) Schematic of the PRBM architecture. In this setup, light with different wavelengths is diffracted and focuses on a phase-only SLM along the x direction, while the pixels in the y direction are coherently illuminated by incident light of the same wavelength. The spins are encoded with phase modulation on the SLM at three different regions (circuited by dashed lines). SC, supercontinuum laser; CL, cylindrical lens; PD, photodetector. (c) Experimental process of content generation by PRBM. Initial random visible spin configuration is encoded onto region II. By performing iterative photonic Gibbs sampling between the visible and hidden layers, a generated content is decoded from the final visible layer.

incident light amplitude, f is the focal length, and C represents the effective pixel area occupied by a single spin. Given the light incoherence among different wavelengths, we have a total intensity via summing the contribution by each wavelength light and define a normalized total intensity $\tilde{I}$ (detailed in Supplemental Material Sec. II [66]) as

$$\tilde{I} = \sum_{j=1}^{N_h} \left(h_j + b_j/L + \sum_{i=1}^{N_v} W_{ij} v_i / L \right)^2. \tag{2}$$

The cross terms arising from the expansion of these squared terms serve as the optical analog of the interaction energy in the Hamiltonian, laying the foundation for the subsequent sampling process.

Now we perform Gibbs sampling in two steps through the measurement and feedback scheme. In the first step, we set all hidden-layer spins as $h_j = 1$ and encode them with visible-layer spins onto the SLM following Eq. (1). Thus, the normalized intensity $\tilde{I}_0$ at the center position of the back focal plane is obtained, and the Hamiltonian is computed as $H_0 = -L\tilde{I}_0 = -2\sum_{j=1}^{N_h}\sum_{i=1}^{N_v} v_i W_{ij} - 2\sum_{j=1}^{N_h} b_j - Q$, where Q is a constant (Supplemental Material Sec. II [66]). In the second step, every spin in the hidden layer is flipped independently, as indicated by the gray shadowed box in Fig. 1(c). For the kth spin, we set $h_k = -1$ and measure the intensity corresponding to the Hamiltonian $H_k = -2\sum_{j\neq k}^{N_h}\sum_{i=1}^{N_v} v_i W_{ij} - 2\sum_{j\neq k}^{N_h} b_j + 2\sum_{i=1}^{N_v} v_i W_{ik} + 2b_k - Q$. Therefore, by measuring the difference of intensity in two steps, we can compute ΔH_k as

$$\Delta H_k = (H_0 - H_k)/2 = -2\sum_{i=1}^{N_v} v_i W_{ik} - 2b_k. \tag{3}$$

Based on the feedback scheme, we update the spin configuration with Gibbs sampling at temperature T, by calculating the probability of the kth spin being $+1$ as $P(h_k = 1|\mathbf{v}) = 1/[1 + \exp(\Delta H_k/T)]$.

We note that the proposed encoding method for Gibbs sampling does not require the decomposition of the spin interaction matrix [53,54,56] and can be generalized for sparse spin interactions (see more details in Appendix B). When sampling from the hidden layer to the visible layer, we encode the visible-layer spin v_i and the external magnetic field a_i in region I and III, respectively, and the equivalent spin $h'_{ji} = h_j W_{ij}$ in region II. In the process of generating an image, we begin with a randomized visible-layer spin configuration encoded onto region II. By performing iterative photonic computing of Gibbs sampling between the visible and hidden layers, a generated content is decoded from the final visible layer as shown in Fig. 1(c).

III. VALIDATION OF PRBM BY OBSERVING PHASE TRANSITION

We first validate the proposed PRBM with Gibbs sampling by observing the phase transition in the two-dimensional Ising model $H = -J\sum_{\langle ij\rangle}\sigma_i\sigma_j$, where each spin interacts only with its four nearest neighbors of the strength J. Analogous to a RBM, we transform the spins at the four corner positions of the square into visible layer [red circular dots in Fig. 2(a)] and hidden layer [blue triangles in Fig. 2(a)], respectively. Through this transformation, we study the phase transition and compare the experimental results with the analytical critical temperature [65]. We experimentally conduct photonic computing of the Gibbs sampling on a 10×10 lattice at 14 different temperatures. In order to calculate the phase diagram, we generate 20 samples at each temperature, with each sample obtained after ten iterations from a random initial spin configuration, with each complete iteration consisting of Gibbs sampling from the visible layer to obtain the hidden layer, and vice versa. Figure 2(b) displays the obtained spin configurations after ten iterations at four different temperatures $T = 6.1J$, $3.1J$, $1.9J$, and $1.1J$. At high temperature, such as $T = 6.1J$, the spin configuration exhibits disordered, while, as the temperature decreases, clusters of aligned spins begin to emerge and the spin configuration becomes increasingly ordered.

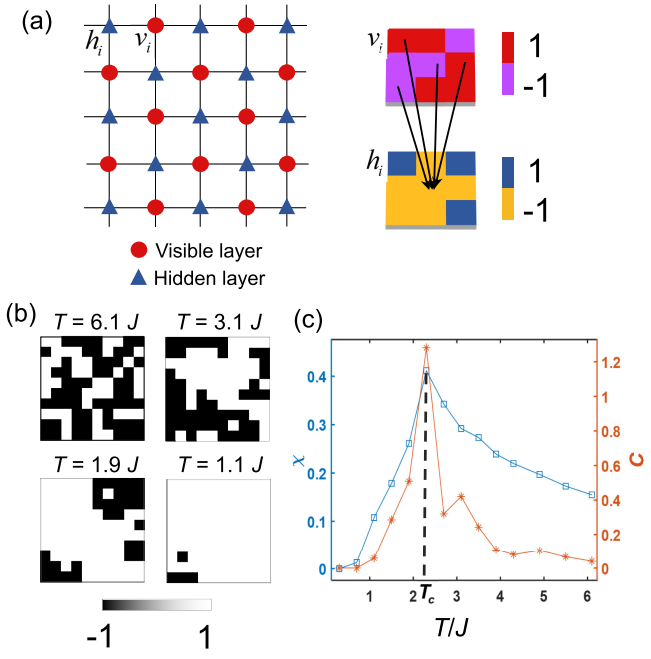


FIG. 2. Validation of PRBM by observing phase transition in two-dimensional Ising model. (a) Left: the spins in the 2D Ising model are divided into visible and hidden layers, with each spin interacting only with its nearest-neighboring spins. Right: the visible-layer spins are color coded as red and purple, while blue and yellow correspond to spins in the hidden layer. (b) The equilibrium states sampling from four different temperatures by Gibbs sampling through photonic computing. (c) The variation of the order parameter susceptibility and heat capacity with temperature, and the phase transition temperature is $T_c = 2.3J$.

To precisely evaluate the phase transition temperature, we calculate the heat capacity $C = [(\langle H^2\rangle - \langle H\rangle^2)/NT^2]$ and susceptibility $\chi = \sum_{i=1}^{N}[(1 - m_i^2)/NT]$, where m_i represents the statistical average of a spin and N is the number of spins. As shown in Fig. 2(c), the experimental results show that the phase transition temperature is $T_c = 2.3J$, which agrees well with the theoretical predictions $T_c = \{2J/[\ln(1+\sqrt{2})]\} \approx 2.27J$ in the infinite 2D model [65]. We also employ the annealing method [67] with Gibbs sampling to search the ground state (Supplemental Material Sec. III [66]). All these results demonstrate that, despite limited noise and aberrations, the proposed photonic computing of Gibbs sampling effectively simulates the spin interaction between the visible and hidden layers for RBMs.

IV. IMAGE CONTENT GENERATION AND RESTORATION BY PRBM

Next, we demonstrate the capability of the photonic restricted Boltzmann machine for content generation. In image generation, we utilize the Fashion MNIST dataset [68] and MNIST digit images [69] as the visible layer spins to train the RBM, where each pixel in an image corresponds to a spin. Before training, we compress the original images with 28×28 spins to a resolution of 14×14 (yielding 196 spins) and binarize the images. During both the training and generation processes, Gibbs sampling is conducted at a fixed temperature $T = 1$. Furthermore, we train the model in the absence of a magnetic field and separately train it on the fashion items "boot," "pants," and the digit "0." Following the Hinton recipe for training RBM [70], we initialize the interaction weights with small random values drawn from a zero-mean Gaussian distribution, and the training method is detailed in Supplemental Material Sec. IV [66].

To generate an image by the PRBM, we first randomly initialize the state vector $\mathbf{v}^0$ and subsequently obtain $\mathbf{v}^K$ as the terminal generated image after $K = 15$ iterations by

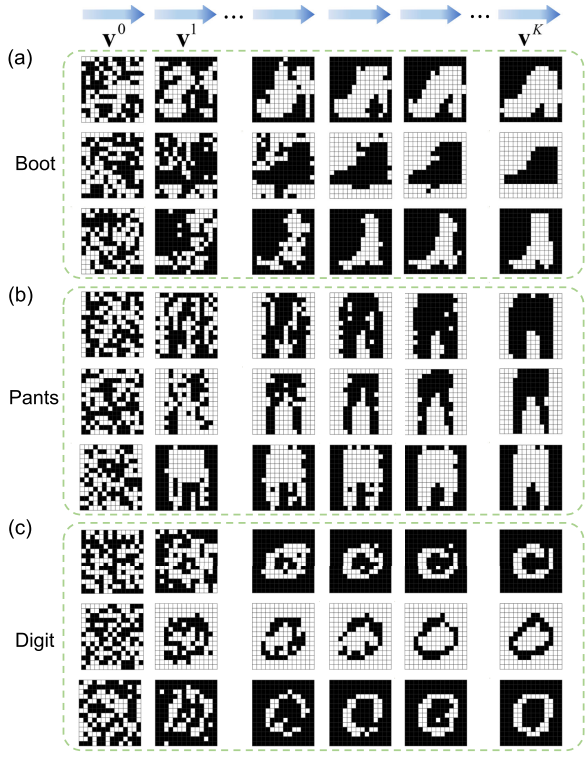


FIG. 3. Image content generation with PRBM. (a)–(c) The processes of generating three independent images for each of "boot," "pants," and the digit "0," respectively.

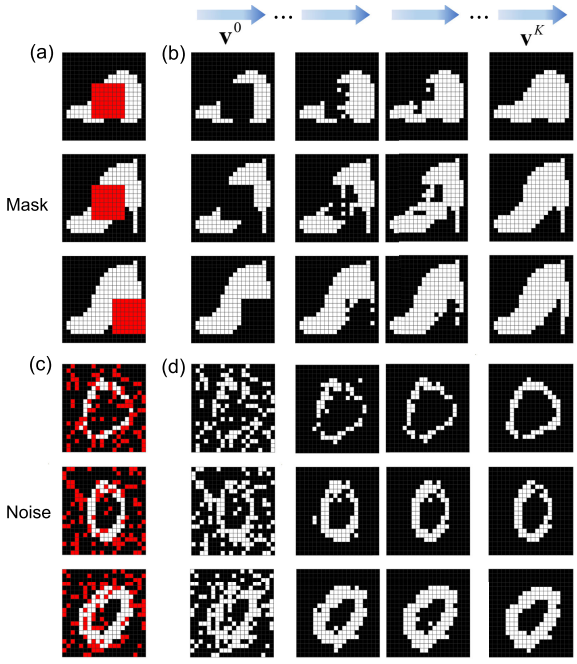


FIG. 4. Image restoration with PRBM. All the initial images are excluded in the training dataset and distorted. (a),(b) The processes of image restoration for boot in test dataset are demonstrated, and a portion of the image is occluded (marked by red color). (c),(d) The processes of image restoration for digit 0 in the test dataset are demonstrated, and random noise (marked by red color) is added to digit images.

performing photonic computing of Gibbs sampling. With the stabilized spin configuration in the visible layer, Fig. 3 illustrates three distinct image generation processes for each experimental item. Additional experimental results on image generation are provided in Supplemental Material Sec. VII [66]. The diversity observed across these generated images demonstrates the PRBM's capacity to produce novel content with significant variability. To investigate the impact of the number of iterations K, we evaluate the Kullback-Leibler divergence as K varies for Fig. 3(c), which consistently exhibits a clear convergence to a steady-state value when the iteration count exceeds $K = 10$ [Fig. S7(a) in Supplemental Material [66]].

We further validate photonic computing of Gibbs sampling by evaluating whether the trained model exhibits overfitting. Here, we use the PRBM to perform a restoring-image test, where the initial images are excluded in the training dataset and distorted. We compress images of the fashion item boot and the digit 0 to 20×20 spins and train the model. For the image restoration experiments, we mask three different boot images randomly selected from the test dataset, as indicated by the red areas in Fig. 4(a). Figure 4(b) represents the initial input images $\mathbf{v}^0$ (the first column), where spins in the masked regions are set to a value of -1. After $K = 15$ iterations, the three restored images $\mathbf{v}^K$ are obtained, as shown in the last column in Fig. 4(b).

As another example, we also restore three different 0 images from the test dataset by adding random noise, as indicated by the red marks in Fig. 4(c). Before performing image restoration, we set the spins in the red-marked regions to the negative values of their respective original values, as is shown in the first column $\mathbf{v}^0$ in Fig. 4(d). After 15 iterations, the final restored images $\mathbf{v}^K$ are presented in the last column in Fig. 4(d). More other experimental results about image restoration are shown in Supplemental Material Sec. VII [66]. All these results demonstrate that the image generation capabilities of the PRBM are not attributable to overfitting and also validate the proposed photonic computing of Gibbs sampling. To investigate the impact of the number of iterations K, we evaluate the Hamming distance as K varies for Fig. 4(a), which shows convergence to a steady-state value when the iteration count exceeds $K = 10$ [Fig. S7(b) in Supplemental Material [66]].

V. TEMPORAL CONTENT GENERATION BY PRBM

To demonstrate time-varying RBM systems, we generate pieces of piano music by using a recurrent neural network-restricted Boltzmann machine (RNN-RBM) [11], which forms a sequential step through each RBM cell [the green box in Fig. 5(a)] along the temporal dimension. Different to the ordinary RBM, the magnetic field coefficients in

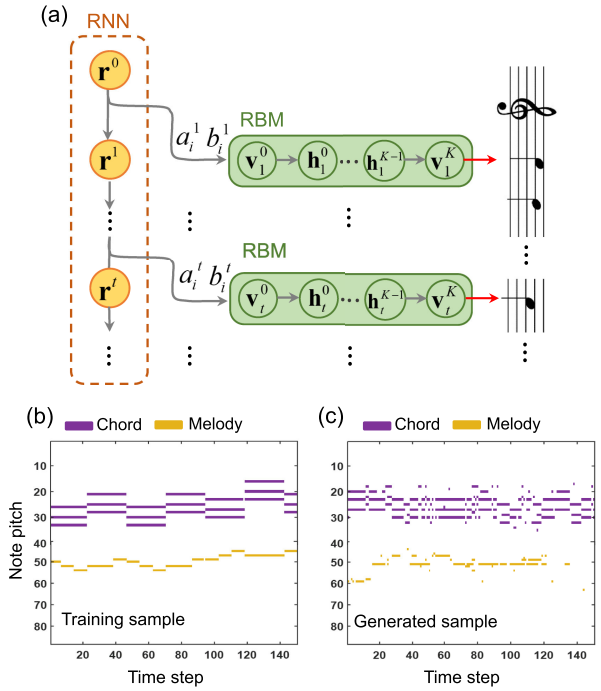


FIG. 5. Temporal content generation with PRBM. (a) The structure of RNN-RBM. A green box denotes a RBM cell, and the visible-layer spin configuration obtained after several iterations is mapped to the piano keys to be pressed at the current time step. (b) A music training sample. (c) The generated music piece.

RNN-RBM are dynamically updated at each time step, whereas the interactions parameters remain invariant. In the experimental implementation of piano music generation with the PRBM, as the time step t increases, we need only to modify the rotation β_k in region III of the SLM by utilizing the gauge transform. We use the piano dataset Nottingham [71] as the training dataset. Within the PRBM framework, a total of 88 piano keys ranging from A0 to C8 are mapped onto 88 spins v_i in the visible layer of the RBM, where v_i assumes the value of $+1$ to represent the pressing of the ith piano key and -1 otherwise.

Figure 5(a) demonstrates the music generation process. The gray arrows indicate the connections between the elements. The red arrows signify that the spins $\mathbf{v}_t^K$ in each RBM cell correspond to the specific piano keys requiring activation at the current time step. In the experiments, the proposed PRBM executes the music generation composed of 150 time steps, where each time step stands for a quaver. The architecture of each RBM cell within this temporal sequence incorporates 88 spins in the visible layer, corresponding to the full range of piano keys, and 96 spins in the hidden layer to capture the underlying musical patterns and dependencies.

Figure 5(b) presents one of the pieces in training dataset. Following the training process, the rhythmic patterns and stylistic characteristics of the music are encoded within the coefficient parameters. Subsequently, we generate novel melodies exhibiting similar rhythmic and stylistic properties. For each time step t, $\mathbf{v}_t^0$ is randomly initialized as the spins in the visible layer of the RBM. The generated music $\mathbf{v}_t^K$ is produced after 20 iterations, as shown in the tth column in Fig. 5(c). The $\mathbf{v}_t^K$ is then utilized for updating the RNN elements $\mathbf{r}^t$ in the current time step. The generation processes are conducted repeatedly, and, after 150 time steps, a piano music piece is generated as shown in Fig. 5(c). The generated musical composition demonstrates considerable similarity to the training examples in terms of rhythmic structure and stylistic attributes, thereby validating the efficacy of the proposed PRBM for temporal content generation. The detailed statistical analysis of the note pitch distributions is illustrated as Fig. S6 in Supplemental Material Sec. X [66], and the Pearson correlation coefficient between the generated sample and training dataset is approximately 0.80.

VI. DISCUSSION AND OUTLOOK

In conclusion, we propose a photonic restricted Boltzmann machine with a novel encoding method to accelerate Gibbs sampling processes. In contrast to SPIMs, the proposed encoding method eliminates the decomposition computation of the interaction matrix and the magnetic field, thereby significantly improving computing efficiency. We verify the PRBM and the photonic computing of Gibbs sampling by observing the phase transition in the 2D lattice Ising model. Furthermore, we experimentally demonstrate the capability of the PRBM to generate novel content by generating and restoring images. To further illustrate the advantages of the encoding method, we demonstrate that the PRBM can implement time-varying spin interactions for temporal content generation, such as music, where nondecomposition computation is particularly crucial for efficient processing.

The PRBM can be scaled to larger spin sizes by extending the wavelength range and increasing the number of SLM pixels. Within our spatial encoding scheme, the scale of the system is primarily limited by the pixel size of the SLM, as each spin requires at least 2×2 pixels for modulation. In the current experiments, we have demonstrated a system with 400 spins, while the theoretical maximum number of spins by the present setup is $N_{\max} = 1080/2 = 540$. Correspondingly, the maximum number of encodable parameters can reach approximately $N_{\max}^2 \approx 2.9 \times 10^5$. As emerging lithium niobate photonic-crystal modulators enable pixel sizes approaching the submicrometer scale [72], it would be possible to encode on the order of 10^{10} parameters, allowing the realization of models with $N = 10^5$ spins within a 100 mm $\times$ 100 mm area. Leveraging a non–von Neumann photonic computing architecture, both the interaction matrix and the magnetic fields are directly encoded on the SLM. This approach significantly reduces memory requirements and eliminates

the data-transfer bottleneck between the CPU/GPU and memory inherent to conventional electronic computing, leading to substantial improvements in computational efficiency, particularly for large-scale RBMs. Although a camera is used in the experiment setup, the PRBM fundamentally requires only a single-pixel photodetector (PD). Commercial PDs can operate at speed of tens of gigahertz with low energy consumption. A detailed comparison of PRBM using different modulators and PDs is provided in Sec. VIII and Table S2 in Supplemental Material [66].

We also evaluate the computational capacity of the proposed system. For each photonic Gibbs-sampling operation, the computational load is described by Eq. (3), where each conditional dot product requires N effective multiply-accumulate (MAC) operations, with N denoting the number of spins. In our experimental implementation, the latency for updating each hidden spin is given by $T_{\text{spin}} = 2(T_{\text{SLM}} + T_{\text{PD}} + T_{\text{ADC}}) + T_{\text{CTL}}$, where T_{SLM}, T_{PD}, T_{ADC}, and T_{CTL} are the SLM modulation time, the PD readout time, the analog-to-digital conversion (ADC) time, and the control time required to compute the Gibbs update rule, respectively. Thus, the time required per effective MAC operation is T_{spin}/N for the PRBM. Notably, as the PRBM requires only a single-pixel PD, T_{PD} can be reduced to below 1 ns using commercially available PDs. With the rapid development of advanced SLM technologies capable of gigahertz modulation rates and low power consumption [72–74], T_{SLM} can also approach the nanosecond scale. In the current implementation, the Gibbs factor is computed electronically, and both T_{ADC} and T_{CTL} on the order of nanoseconds have already been demonstrated in optoelectronic neural networks [17]. In the long term, an all-optical PRBM architecture could be realized by exploiting optical bistability in nonlinear materials, analogous to all-optical neural networks [19]. As a performance benchmark, we compare the estimated time and energy consumption per MAC operation as a function of the number of spins in Supplemental Material [66] Fig. S4 and Sec. VIII, for both digital and photonic RBMs implemented on various hardware platforms. As the spin size N scales up, the time and the energy consumption per effective MAC operation scale inversely with N, indicating that the computational throughput and energy efficiency of the proposed PRBM can potentially outperform conventional electronic computing architectures.

It is worth noting that the proposed photonic Gibbs-sampling approach can significantly improve computational efficiency and reduce resource consumption during training. Since the RBM training relies on Gibbs sampling to estimate the gradient of the log-likelihood function, it is inherently a backpropagation-free method [2]. Consequently, PRBMs equipped with photonic Gibbs sampling naturally avoid backpropagation-based training, which is known to involve various trade-offs in many physical neural network implementations [75]. Moreover, the PRBM enables efficient scaling to larger neural networks through accelerated photonic Gibbs sampling. For example, state-of-the-art generative AI models require tremendous computational resources for backpropagation-based training. Transformer-based language models implemented on digital computers typically demand computational operations on the order of $6MBS$, where M, B, and S are the number of model parameters, the training batch size, and the iteration number, respectively [76]. In comparison, the total number of digital computational operations required to train RBMs is of the same order of magnitude, approximately $9MBS$, where $M \approx N^2/4$, N is the total spin number (Supplemental Material Sec. V [66]). Owing to the accelerated Gibbs sampling enabled by the proposed PRBM, the complexity can be reduced to $3MBS + 3NBS$. As a result, photonic computing acceleration allows the training efficiency of RBMs to outperform that of transformer-based models, highlighting a promising pathway toward unconventional computational architectures for generative AI. As shown in Supplemental Material Table S1 [66], the PRBM reduces the training time by approximately 2 orders of magnitude compared to an NVIDIA H100 when training a GPT-3 model with the same number of parameters and training data.

These advantageous features of PRBM may inspire the development of efficient photonic generation models extending beyond RBMs to modern deep generative architectures, such as diffusion models or transformer. Given that the proposed photonic computing scheme exhibits strong compatibility with diverse data types, including both sequential and nonsequential data, as well as binary and continuous variables [4,70], PRBMs are expected to achieve high practical efficiency in learning complex probability distributions and generating sophisticated content. This capability further suggests their potential applicability to large-scale generative tasks, including language modeling.

ACKNOWLEDGMENTS

The authors acknowledge funding through the National Natural Science Foundation of China (124B2085, 12474393, and 12174340) and the National Key Research and Development Program of China (2022YFA1405200).

DATA AVAILABILITY

The data that support the findings of this article are openly available [77].

APPENDIX A: RESTRICTED BOLTZMANN MACHINE

At temperature T, the probability of a spin configuration follows the Boltzmann-Gibbs distribution $P(\mathbf{v}, \mathbf{h}) \propto \exp[-H(\mathbf{v}, \mathbf{h})/T]$, where H is the energy of that state (here, we take the convention of the Boltzmann constant

$k_B = 1$). As shown in Fig. 1(a), the training set is binarized as the spins in the visible layer, and the training objective of RBMs is to maximize the log-likelihood function $\ell(\boldsymbol{\theta}) = \langle \ln P(\mathbf{v}, \mathbf{h}) \rangle_{\text{data}}$ with the parameters $\boldsymbol{\theta} = \{W_{ij}, a_i, b_i\}$, at temperature $T = 1$. Gibbs sampling is utilized to compute the partial derivatives of $\ell(\boldsymbol{\theta})$ with respect to each parameter, which enables the update of parameters. The probability of the kth spin being $+1$ in the hidden layer is $P(h_k = 1|\mathbf{v}) = P(h_k = 1, \mathbf{v})/P(\mathbf{v}) = 1/\{1 + \exp[-2(b_k + \sum_i v_i W_{ik})]\}$ (Supplemental Material Sec. VI [66]), while the probability of the kth spin in the visible layer being $+1$ is calculated as $P(v_k = 1|\mathbf{h}) = 1/\{1 + \exp[-2(a_k + \sum_i W_{ki} h_i)]\}$. In the generation process, Gibbs sampling is employed to generate novel content of the same type based on the trained parameters.

APPENDIX B: GENERAL SPARSE INTERACTION MATRIX

The proposed encoding method is also applicable to Ising models with complex sparse interactions. During Gibbs sampling, the spin σ_i is encoded in region I of the SLM, while the spins connected to σ_i are encoded in region II. Therefore, this photonic computing system can be extended to simulate general Ising models with sparse interactions. Especially for combinatorial optimization problems, such as the graph partitioning problem with arbitrary sparsity, the system is expected to exhibit unique performance without requiring matrix decomposition.

[1] P. Smolensky, *Chapter 6: Information processing in dynamical systems: Foundations of harmony theory*, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, edited by D. E. Rumelhart and J. L. McClelland (MIT Press, 1986), pp. 194–281.

[2] G. E. Hinton, *Training products of experts by minimizing contrastive divergence*, Neural Comput. **14**, 1771 (2002).

[3] R. Salakhutdinov and G. Hinton, *Deep Boltzmann machines*, in *Artificial Intelligence and Statistics* (PMLR, Florida, USA, 2009), pp. 448–455.

[4] G. E. Hinton and R. R. Salakhutdinov, *Reducing the dimensionality of data with neural networks*, Science **313**, 504 (2006).

[5] M. N. Bojnordi and E. Ipek, *Memristive Boltzmann machine: A hardware accelerator for combinatorial optimization and deep learning*, in *Proceedings of the 2016 IEEE International Symposium on High Performance Computer Architecture (HPCA)* (IEEE, New York, 2016), pp. 1–13.

[6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, 2016).

[7] H. Larochelle and Y. Bengio, *Classification using discriminative restricted Boltzmann machines*, in *Proceedings of the 25th International Conference on Machine Learning* (ACM, Helsinki, Finland, 2008), pp. 536–543.

[8] J. Fernandez de Cossio Diaz, S. Cocco, and R. Monasson, *Disentangling representations in restricted Boltzmann machines without adversaries*, Phys. Rev. X **13**, 021003 (2023).

[9] S. Patel, P. Canoza, and S. Salahuddin, *Logically synthesized and hardware-accelerated restricted Boltzmann machines for combinatorial optimization and integer factorization*, Nat. Electron. **5**, 92 (2022).

[10] X. Li, C. Wan, R. Zhang, M. Zhao, S. Xiong, D. Kong, X. Luo, B. He, S. Liu, J. Xia *et al.*, *Restricted Boltzmann machines implemented by spin–orbit torque magnetic tunnel junctions*, Nano Lett. **24**, 5420 (2024).

[11] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, *Modeling temporal dependencies in high-dimensional sequences: application to polyphonic music generation and transcription*, in *Proceedings of the 29th International Conference on Machine Learning* (Omnipress, Edinburgh, UK, 2012), pp. 1881–1888.

[12] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques* (MIT Press, Cambridge, MA, 2009).

[13] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, *An introduction to MCMC for machine learning*, Mach. Learn. **50**, 5 (2003).

[14] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, *A learning algorithm for Boltzmann machines*, Cogn. Sci. **9**, 147 (1985).

[15] P. M. Long and R. Servedio, *Restricted Boltzmann machines are hard to approximately evaluate or simulate*, in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (Citeseer, 2010), pp. 703–710.

[16] S. Niazi, S. Chowdhury, N. A. Aadit, M. Mohseni, Y. Qin, and K. Y. Camsari, *Training deep Boltzmann networks with sparse Ising machines*, Nat. Electron. **7**, 610 (2024).

[17] G. Wetzstein, A. Ozcan, S. Gigan, S. Fan, D. Englund, M. Soljačić, C. Denz, D. A. B. Miller, and D. Psaltis, *Inference in artificial intelligence with deep optics and photonics*, Nature (London) **588**, 39 (2020).

[18] P. L. McMahon, *The physics of optical computing*, Nat. Rev. Phys. **5**, 717 (2023).

[19] M. Matuszewski, A. Prystupiuk, and A. Opala, *Role of all-optical neural networks*, Phys. Rev. Appl. **21**, 014028 (2024).

[20] R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D. Englund, *Large-scale optical neural networks based on photoelectric multiplication*, Phys. Rev. X **9**, 021032 (2019).

[21] P. L. McMahon, A. Marandi, Y. Haribara, R. Hamerly, C. Langrock, S. Tamate, T. Inagaki, H. Takesue, S. Utsunomiya, K. Aihara *et al.*, *A fully programmable 100-spin coherent Ising machine with all-to-all connections*, Science **354**, 614 (2016).

[22] T. Inagaki, Y. Haribara, K. Igarashi, T. Sonobe, S. Tamate, T. Honjo, A. Marandi, P. L. McMahon, T. Umeki, K. Enbutsu *et al.*, *A coherent Ising machine for 2000-node optimization problems*, Science **354**, 603 (2016).

[23] T. Inagaki, K. Inaba, R. Hamerly, K. Inoue, Y. Yamamoto, and H. Takesue, *Large-scale Ising spin network based on degenerate optical parametric oscillators*, Nat. Photonics **10**, 415 (2016).

[24] F. Böhm, G. Verschaffelt, and G. Van der Sande, *A poor man's coherent Ising machine based on opto-electronic feedback systems for solving optimization problems*, Nat. Commun. **10**, 1 (2019).
[25] A. Marandi, Z. Wang, K. Takata, R. L. Byer, and Y. Yamamoto, *Network of time-multiplexed optical parametric oscillators as a coherent Ising machine*, Nat. Photonics **8**, 937 (2014).
[26] S. Utsunomiya, K. Takata, and Y. Yamamoto, *Mapping of Ising models onto injection-locked laser systems*, Opt. Express **19**, 18091 (2011).
[27] M. Babaeian, D. T. Nguyen, V. Demir, M. Akbulut, P.-A. Blanche, Y. Kaneda, S. Guha, M. A. Neifeld, and N. Peyghambarian, *A single shot coherent Ising machine based on a network of injection-locked multicore fiber lasers*, Nat. Commun. **10**, 1 (2019).
[28] C. Tradonsky, I. Gershenzon, V. Pal, R. Chriki, A. A. Friesem, O. Raz, and N. Davidson, *Rapid laser solver for the phase retrieval problem*, Sci. Adv. **5**, eaax4530 (2019).
[29] M. Parto, W. Hayenga, A. Marandi, D. N. Christodoulides, and M. Khajavikhan, *Realizing spin Hamiltonians in nanoscale active photonic lattices*, Nat. Mater. **19**, 725 (2020).
[30] M. Honari-Latifpour and M.-A. Miri, *Mapping the XY Hamiltonian onto a network of coupled lasers*, Phys. Rev. Res. **2**, 043335 (2020).
[31] K. P. Kalinin, A. Amo, J. Bloch, and N. G. Berloff, *Polaritonic XY-Ising machine*, Nanophotonics **9**, 4127 (2020).
[32] N. G. Berloff, M. Silva, K. Kalinin, A. Askitopoulos, J. D. Töpfer, P. Cilibrizzi, W. Langbein, and P. G. Lagoudakis, *Realizing the classical XY Hamiltonian in polariton simulators*, Nat. Mater. **16**, 1120 (2017).
[33] K. P. Kalinin and N. G. Berloff, *Simulating Ising and n-state planar Potts models and external fields with nonequilibrium condensates*, Phys. Rev. Lett. **121**, 235302 (2018).
[34] W. Verstraelen and Timothy C. Liew, *Optical resonators constitute a universal spin simulator*, Phys. Rev. B **111**, 014307 (2025).
[35] K. Kim, M.-S. Chang, S. Korenblit, R. Islam, E. E. Edwards, J. K. Freericks, G.-D. Lin, L.-M. Duan, and C. Monroe, *Quantum simulation of frustrated Ising spins with trapped ions*, Nature (London) **465**, 590 (2010).
[36] J. Struck, M. Weinberg, C. Ölschläger, P. Windpassinger, J. Simonet, K. Sengstock, R. Höppner, P. Hauke, A. Eckardt, M. Lewenstein *et al.*, *Engineering Ising-XY spin-models in a triangular lattice using tunable artificial gauge fields*, Nat. Phys. **9**, 738 (2013).
[37] M. Vretenar, B. Kassenberg, S. Bissesar, C. Toebes, and J. Klaers, *Controllable Josephson junction for photon Bose-Einstein condensates*, Phys. Rev. Res. **3**, 023167 (2021).
[38] F. Cai, S. Kumar, T. Van Vaerenbergh, X. Sheng, R. Liu, C. Li, Z. Liu, M. Foltin, S. Yu, Q. Xia, J. J. Yang, R. Beausoleil, W. D. Lu, and J. P. Strachan, *Power-efficient combinatorial optimization using intrinsic noise in memristor Hopfield neural networks*, Nat. Electron. **3**, 409 (2020).
[39] M. W. Johnson, M. H. S. Amin, S. Gildert, T. Lanting, F. Hamze, N. Dickson, R. Harris, A. J. Berkley, J. Johansson, P. Bunyk *et al.*, *Quantum annealing with manufactured spins*, Nature (London) **473**, 194 (2011).
[40] S. Boixo, T. F. Rønnow, S. V. Isakov, Z. Wang, D. Wecker, D. A. Lidar, J. M. Martinis, and M. Troyer, *Evidence for quantum annealing with more than one hundred qubits*, Nat. Phys. **10**, 218 (2014).
[41] A. D. King, J. Carrasquilla, J. Raymond, I. Ozfidan, E. Andriyash, A. Berkley, M. Reis, T. Lanting, R. Harris, F. Altomare *et al.*, *Observation of topological phenomena in a programmable lattice of 1,800 qubits*, Nature (London) **560**, 456 (2018).
[42] C. Roques-Carmes, Y. Shen, C. Zanoci, M. Prabhu, F. Atieh, L. Jing, T. Dubček, C. Mao, M. R. Johnson, V. Čeperić *et al.*, *Heuristic recurrent algorithms for photonic Ising machines*, Nat. Commun. **11**, 249 (2020).
[43] M. Prabhu, C. Roques-Carmes, Y. Shen, N. Harris, L. Jing, J. Carolan, R. Hamerly, T. Baehr-Jones, M. Hochberg, V. Čeperić *et al.*, *Accelerating recurrent Ising machines in photonic integrated circuits*, Optica **7**, 551 (2020).
[44] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund *et al.*, *Deep learning with coherent nanophotonic circuits*, Nat. Photonics **11**, 441 (2017).
[45] K. Wu, J. G. De Abajo, C. Soci, P. P. Shum, and N. I. Zheludev, *An optical fiber network oracle for NP-complete problems*, Light **3**, e147 (2014).
[46] Y. Okawachi, M. Yu, J. K. Jang, X. Ji, Y. Zhao, B. Y. Kim, M. Lipson, and A. L. Gaeta, *Demonstration of chip-based coupled degenerate optical parametric oscillators for realizing a nanophotonic spin-glass*, Nat. Commun. **11**, 1 (2020).
[47] S. Hua, E. Divita, S. Yu, B. Peng, C. Roques-Carmes, Z. Su, Z. Chen, Y. Bai, J. Zou, Y. Zhu *et al.*, *An integrated large-scale photonic accelerator with ultralow latency*, Nature (London) **640**, 361 (2025).
[48] J. Ouyang, S. Liu, Z. Yang, W. Wang, X. Feng, Y. Li, and Y. Huang, *16-channel photonic solver for optimization problems on a silicon chip*, Chip **4**, 100117 (2025).
[49] B. Wu, W. Zhang, S. Zhang, H. Zhou, Z. Ruan, M. Li, D. Huang, J. Dong, and X. Zhang, *A monolithically integrated optical Ising machine*, Nat. Commun. **16**, 4296 (2025).
[50] Z. Yang, C. Li, Y. Ran, Y. Li, X. Feng, K. Cui, F. Liu, H. Sun, W. Zhang, Y. Ye *et al.*, *SUANPAN: Scalable photonic linear vector machine*, Light **15**, 16 (2026).
[51] Z. Hu, Y. Ren, Y. Meng, T. Wang, Y. Jiang, M. Wei, Y. Xiao, Z. Li, and M. Li, *Programmable optoelectronic Ising machine for optimization of real-world problems*, Light **15**, 6 (2026).
[52] D. Pierangeli, G. Marcucci, and C. Conti, *Large-scale photonic Ising machine by spatial light modulation*, Phys. Rev. Lett. **122**, 213902 (2019).
[53] H. Yamashita, K. I. Okubo, S. Shimomura, Y. Ogura, J. Tanida, and H. Suzuki, *Low-rank combinatorial optimization and statistical learning by spatial photonic Ising machine*, Phys. Rev. Lett. **131**, 063801 (2023).
[54] D. Veraldi, D. Pierangeli, S. Gentilini, M. C. Strinati, J. Sakellariou, J. S. Cummins, A. Kamaletdinov, M. Syed, R. Z. Wang, N. G. Berloff *et al.*, *Fully programmable spatial photonic Ising machine by focal plane division*, Phys. Rev. Lett. **134**, 063802 (2025).
[55] S. Shimomura, J. Tanida, and Y. Ogura, *Parallel spatial photonic Ising machine using spatial multiplexing for accelerating combinatorial optimization*, Opt. Lett. **50**, 6674 (2025).

[56] L. Luo, Z. Mi, J. Huang, and Z. Ruan, *Wavelength-division multiplexing optical Ising simulator enabling fully programmable spin couplings and external magnetic fields*, Sci. Adv. **9**, eadg6238 (2023).
[57] Y. Fang, J. Huang, and Z. Ruan, *Experimental observation of phase transitions in spatial photonic Ising machine*, Phys. Rev. Lett. **127**, 043902 (2021).
[58] D. Pierangeli, G. Marcucci, D. Brunner, and C. Conti, *Noise-enhanced spatial-photonic Ising machine*, Nanophotonics **9**, 4109 (2020).
[59] J. Huang, Y. Fang, and Z. Ruan, *Antiferromagnetic spatial photonic Ising machine through optoelectronic correlation computing*, Commun. Phys. **4**, 1 (2021).
[60] W. Sun, W. Zhang, Y. Liu, Q. Liu, and Z. He, *Quadrature photonic spatial Ising machine*, Opt. Lett. **47**, 1498 (2022).
[61] S. Kumar, Z. Li, T. Bu, C. Qu, and Y. Huang, *Observation of distinct phase transitions in a nonlinear optical Ising machine*, Commun. Phys. **6**, 31 (2023).
[62] M. Leonetti, E. Hörmann, L. Leuzzi, G. Parisi, and G. Ruocco, *Optical computation of a spin glass dynamics with tunable complexity*, Proc. Natl. Acad. Sci. U.S.A. **118**, e2015207118 (2021).
[63] S.-T. Yu, M.-G. He, S. Fang, Y. Deng, and Z.-S. Yuan, *Spatial optical simulator for classical statistical models*, Phys. Rev. Lett. **133**, 237101 (2024).
[64] F. Böhm, D. Alonso-Urquijo, G. Verschaffelt, and G. Van der Sande, *Noise-injected analog Ising machines enable ultrafast statistical sampling and machine learning*, Nat. Commun. **13**, 5847 (2022).
[65] L. Onsager, *Crystal statistics. I. A two-dimensional model with an order-disorder transition*, Phys. Rev. **65**, 117 (1944).
[66] See Supplemental Material at http://link.aps.org/supplemental/10.1103/k6f7-gtr7 for additional details.
[67] S. Kirkpatrick, C. D. Gelatt Jr, and M. P. Vecchi, *Optimization by simulated annealing*, Science **220**, 671 (1983).
[68] H. Xiao, K. Rasul, and R. Vollgraf, *Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms*, arXiv:1708.07747.
[69] Y. LeCun, C. Cortes, and C. J. Burges, *The MNIST database of handwritten digits*, https://yann.lecun.org/exdb/mnist/.
[70] G. E. Hinton, *A practical guide to training restricted Boltzmann machines*, in *Neural Networks: Tricks of the Trade: Second Edition* (Springer, New York, 2012), pp. 599–619.
[71] *Nottingham dataset*, https://github.com/jukedeck/nottingham-dataset.git.
[72] M. Li, J. Ling, Y. He, U. A. Javid, S. Xue, and Q. Lin, *Lithium niobate photonic-crystal electro-optic modulator*, Nat. Commun. **11**, 4123 (2020).
[73] A. Smolyaninov, A. El Amili, F. Vallini, S. Pappert, and Y. Fainman, *Programmable plasmonic phase modulation of free-space wavefronts at gigahertz rates*, Nat. Photonics **13**, 431 (2019).
[74] C. L. Panuski, I. Christen, M. Minkov, C. J. Brabec, S. Trajtenberg-Mills, A. D. Griffiths, J. J. McKendry, G. L. Leake, D. J. Coleman, C. Tran *et al.*, *A full degree-of-freedom spatiotemporal light modulator*, Nat. Photonics **16**, 834 (2022).
[75] A. Momeni, B. Rahmani, B. Scellier, L. G. Wright, P. L. McMahon, C. C. Wanjura, Y. Li, A. Skalli, N. G. Berloff, T. Onodera *et al.*, *Training of physical neural networks*, Nature (London) **645**, 53 (2025).
[76] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, *Scaling laws for neural language models*, arXiv:2001.08361.
[77] https://github.com/loli145/PRBM.