Supplementary Information for

Entanglement engineering of optomechanical systems by reinforcement learning

Li-Li Ye, Christian Arenz, Joseph M. Lukens and Ying-Cheng Lai

Corresponding author: Ying-Cheng Lai (Ying-Cheng.Lai@asu.edu)

CONTENTS

Supplementary Note 1. BACKGROUND RELATED TO OUR WORK

Quantum control. Quantum control [1] is essential to quantum engineering and technology [2–4], where open-loop control [5] has been successfully demonstrated with methods such as gradient-ascent pulse engineering (GRAPE) [6] in spin systems [7], coupled qubits [8], Jaynes-Cummings systems [9], and qubit-cavity lattices [10]. Recently, the open-loop GRAPE algorithm has been extended to feedback GRAPE [11] based on gradient ascent of quantum dynamics for state engineering under strongly stochastic measurement. Open-loop control, however, requires a differentiable model of the quantum dynamics that may not always be available. In realistic situations where such a model is not available, closed-loop feedback control strategies conditioned on experimental measurement outcomes can be applied. Combined with data-driven machine learning, feedback control has been implemented in experiments in a model-free fashion [12–14].

Deep reinforcement learning. In general, RL is a machine-learning paradigm based on a trial-and-error learning process, incorporating traditional optimal control to maximize the accumulated reward. The use of deep neural networks in the learning process leads to deep RL, which explores and exploits the available measurement data to search for a globally optimal policy. In deep RL, many algorithms are available such as deep-Q network (DQN) [12], deep deterministic policy gradient (DDPG) [15], and trust region proximal optimization (TRPO) [16]. A state-of-the-art deep RL algorithm for continuous control is proximal policy optimization (PPO) [17], whose performance can exceed that of TRPO. Incorporating recurrent neural networks [18] into the PPO algorithm leads to improved performance [19]. In recent years, measurement-based feedback control with deep RL has been applied to quantum systems for tasks such as quantum error correction for discrete gates [20], state preparation and stabilization for a single particle [21–24] with an unstable potential [21] or a double-well potential [22], discrimination between entangled states [25] for quantum meteorology, and long-distance entanglement distribution on quantum networks [26]. Experimentally, time scales of the RL action sequences shorter than the coherence time of the underlying quantum system have been realized, rendering feasible real-time deep-RL feedback control [27].

Quantum measurement. In quantum systems, projective measurement can be used to extract the full information about the quantum state but, as a back action, the quantum state will collapse after the measurement [28]. To avoid a complete collapse, one can exploit weak measurements [29, 30], in which the probe is weakly coupled to the system to yield partial information about the quantum state. Examples of weak measurements include continuous monitoring [31] of driven dissipative quantum-optical systems - a basic component of quantum meteorology [32, 33]. A form of weak measurement, the so-called weak continuous measurement (WCM), is fundamental to a broad range of applications. For example, WCM has been used to detect the quadrature operators [34], Wigner [34] and Husimi Q functions [35] with a homodyne apparatus [36], rendering observing both pure [37] and mixed [29] quantum states experimentally feasible. WCM has been experimentally implemented by a weak-field homodyne detector [34, 36, 38] to measure the photon-number statistical distribution over the Fock basis. In another example, WCM has been realized in an atomic spin ensemble [30] via Faraday rotation of an off-resonance probe beam to create and probe nonclassical spin state and dynamics. The concept of WCM has also been used to develop fundamental theories, such as Heisenberg's measurement-disturbance relationship [39] and error-disturbance uncertainty relation [40]. Because of the typical time scales of the quantum dynamics, WCM cannot be regarded as occurring instantaneously [41]. Theoretically, the

impact of WCM on the underlying quantum system can be described by the stochastic master equation [41].

Supplementary Note 2. QUANTUM OPTOMECHANICAL SYSTEM

The standard Hamiltonian of a quantum optomechanical system in the rotating frame of the laser is given by [42, 43]

$$\begin{aligned}\tilde{H} = &-\hbar\Delta\hat{a}^\dagger\hat{a} + \hbar\omega_m\hat{b}^\dagger\hat{b} + \hbar g_0(\hat{b}^\dagger + \hat{b})\hat{a}^\dagger\hat{a} \\ &+ \hbar(\alpha_L\hat{a}^\dagger + \alpha_L^*\hat{a}),\end{aligned} \tag{S1}$$

where $\hat{a}, \hat{b}$ ($\hat{a}^\dagger, \hat{b}^\dagger$) are the annihilation and creation operators of the optical cavity and mechanical mode, respectively. The frequency detuning is $\Delta \equiv \omega_L - \omega_c$, where ω_L is the frequency of the driven laser and ω_c is the intrinsic frequency of the cavity. The nonlinear coupling g_0 between the single cavity and mechanical mode arises from the frequency dispersion relationship with respect to the displacement $\hat{q}$ of the mechanical mode. The complex amplitude of the driven electromagnetic field is denoted as α_L. A detailed description of how the Hamiltonian is derived is as follows.

Consider a single optical cavity and a mechanical mode (with a movable mirror). The resonant frequency of the cavity mode is controlled by the displacement of the movable end-mirror $\omega_c(\hat{q})$ or the length of the cavity, which can be expanded to the first order about the intrinsic frequency $\omega_c(\hat{q} = 0)$ of the cavity, leading to the following nonlinear coupling term:

$$\begin{aligned}\hat{H}_0 &= \hbar\omega_c(\hat{q})\hat{a}^\dagger\hat{a} + \hbar\omega_m\hat{b}^\dagger\hat{b} \\ &= \hbar(\omega_c + (\partial\omega_c(q)/\partial q)\hat{q})\hat{a}^\dagger\hat{a} + \hbar\omega_m\hat{b}^\dagger\hat{b} \\ &= \hbar\omega_c\hat{a}^\dagger\hat{a} + \hbar\omega_m\hat{b}^\dagger\hat{b} + \hbar g_0\hat{a}^\dagger\hat{a}(\hat{b}^\dagger + \hat{b}),\end{aligned} \tag{S2}$$

where $g_0 \equiv (\partial\omega_c(q)/\partial q)q_{\text{zpf}}$ is the single-photon optomechanical coupling strength and the position operator of the mechanical mode is $\hat{q} \equiv (\hat{b} + \hat{b}^\dagger)q_{\text{zpf}}$ with $q_{\text{zpf}} = \sqrt{\hbar/(2m\omega_m)}$ being the mechanical zero-point fluctuations. The radiation pressure force is acted on the mechanical resonator by the photon number operator multiplying the displacement operator $\hat{q}$.

The Hamiltonian $\hat{H} = \hat{H}_0 + \hat{H}_{driven}$ in the rotating frame is defined as [44]:

$$\tilde{H} = \hat{U}^\dagger\hat{H}\hat{U} - \hat{A} \tag{S3}$$

with $\hat{U} \equiv \exp(-i\omega_L\hat{a}^\dagger\hat{a}\ t)$ and $\hat{A} \equiv \hbar\omega_L\hat{a}^\dagger\hat{a}$. Using the following identities:

$$\begin{aligned}&\exp(i\omega_L\hat{a}^\dagger\hat{a}\ t)\hat{a}\exp(-i\omega_L\hat{a}^\dagger\hat{a}\ t) = \hat{a}\exp(-i\omega_L t), \\ &\exp(i\omega_L\hat{a}^\dagger\hat{a}\ t)\hat{a}^\dagger\exp(-i\omega_L\hat{a}^\dagger\hat{a}\ t) = \hat{a}^\dagger\exp(i\omega_L t),\end{aligned} \tag{S4}$$

we have

$$\hat{U}^\dagger\hat{a}^\dagger\hat{a}\hat{U} = \hat{a}^\dagger\hat{a}.$$

In the rotating frame, with the detuning $\Delta \equiv \omega_L - \omega_c$, we then have

$$\tilde{H}_0 = -\hbar\Delta\hat{a}^\dagger\hat{a} + \hbar\omega_m\hat{b}^\dagger\hat{b} + \hbar g_0\hat{a}^\dagger\hat{a}(\hat{b}^\dagger + \hat{b}). \tag{S5}$$

The quantized electromagnetic field can be written as

$$\hat{H}_{driven} = \hbar \left[\alpha_L \exp(-i\omega_L t)\hat{a}^\dagger + \alpha_L^* \exp(i\omega_L t)\hat{a} \right]. \tag{S6}$$

Through the unitary transformation, we obtain

$$\hat{U}^\dagger \hat{H}_{driven} \hat{U} = \hbar\alpha_L \hat{a}^\dagger + \hbar\alpha_L^* \hat{a}. \tag{S7}$$

Finally, the total Hamiltonian driven by the electromagnetic field in the rotating frame is given by

$$\begin{aligned} \tilde{H} &= \hat{U}^\dagger(\hat{H}_0 + \hat{H}_{driven})\hat{U} - \hat{A} \\ &= \tilde{H}_0 + \hat{U}^\dagger \hat{H}_{driven} \hat{U} \\ &= -\hbar\Delta \hat{a}^\dagger \hat{a} + \hbar\omega_m \hat{b}^\dagger \hat{b} + \hbar g_0 \hat{a}^\dagger \hat{a}(\hat{b}^\dagger + \hat{b}) \\ &\quad + \hbar(\alpha_L \hat{a}^\dagger + \alpha_L^* \hat{a}). \end{aligned} \tag{S8}$$

Supplementary Note 3. QUANTUM STOCHASTIC MASTER EQUATION

The starting point is von Neumann equation, which governs the unitary evolution of the density matrix and is given by

$$\dot{\rho} = \frac{1}{i\hbar}[\hat{H}, \rho] \equiv \mathcal{L}\rho, \tag{S9}$$

where $\mathcal{L}$ is the Liouvillian superoperator. Equation (S9) can be derived from the Schrödinger equation and its conjugate:

$$\begin{aligned} i\hbar \frac{\partial}{\partial t}|\psi\rangle &= \hat{H}|\psi\rangle, \\ -i\hbar \frac{\partial}{\partial t}\langle\psi| &= \langle\psi|\hat{H}, \end{aligned} \tag{S10}$$

with Hermitian Hamiltonian $\hat{H}^\dagger = \hat{H}$. Since the density matrix is defined as a mixture of quantum states, $\rho = \sum_i P_i |\psi_i\rangle\langle\psi_i|$ with $\sum_i P_i = 1$, we have

$$\begin{aligned} i\hbar\dot{\rho} &= \sum_i P_i (i\hbar|\dot{\psi_i}\rangle)\langle\psi_i| - \sum_i P_i |\psi_i\rangle(-i\hbar\langle\dot{\psi_i}|) \\ &= \sum_i P_i \hat{H}|\psi_i\rangle\langle\psi_i| - \sum_i P_i |\psi_i\rangle\langle\psi_i|\hat{H} \\ &= \hat{H}\rho - \rho\hat{H} = [\hat{H}, \rho], \end{aligned} \tag{S11}$$

where $\partial\rho/\partial t \equiv \dot{\rho}$ and $\partial|\psi\rangle/\partial t \equiv |\dot{\psi}\rangle$.

The dynamics of a quantum system interacting with the vacuum bath under the continuous measurement of the observable $\hat{c}$ are described by the general stochastic master equation(SME) [1, 41, 44]:

$$d\rho = \frac{1}{i\hbar}[\hat{H}, \rho]dt + \mathcal{L}_{env}\ \rho dt + \mathcal{D}(\hat{c})\rho dt + \mathcal{H}(\hat{c})\rho dW, \tag{S12}$$

where $\mathcal{L}_{env}\ \rho$ is the interaction between the system and vacuum bath, which is given by

$$\mathcal{L}_{env}\ \rho = \kappa\mathcal{D}(\hat{a})\rho + \gamma\mathcal{D}(\hat{b})\rho, \tag{S13}$$

and dW corresponds to a Wiener process with a Gaussian distribution. Concretely, both the cavity and the oscillator modes are coupled to the vacuum bath with the coupling strengths κ and γ, respectively, where the bath is at the absolute zero temperature. In Eq. (S13), the symbols $\mathcal{D}$ and $\mathcal{H}$ denote the Lindblad and measurement superoperators, respectively, which are given by

$$\mathcal{D}(\hat{c})\rho \equiv \hat{c}\rho\hat{c}^\dagger - \frac{1}{2}(\hat{c}^\dagger\hat{c}\rho + \rho\hat{c}^\dagger\hat{c}), \tag{S14}$$

$$\mathcal{H}(\hat{c})\rho \equiv \hat{c}\rho + \rho\hat{c}^\dagger - \langle\hat{c} + \hat{c}^\dagger\rangle\rho. \tag{S15}$$

The actions described by the two superoperators can drive the quantum state into an eigenstate of the observable $\hat{c}$ to some degree. Pertinent to this process is WCM [41]. To understand WCM, we begin with the von Neumann measurement.

The set of eigenstates of an observable forms an orthonormal basis in the Hilbert space: $\{|n\rangle : n = 1, ..., n_{\text{max}}\}$. Any pure quantum state can be completely expanded as $|\psi\rangle = \sum_n c_n|n\rangle$ with the probability distribution $|c_n|^2$ over the basis $\{|n\rangle\}$. The von Neuman measurement, after which the quantum state will be completely projected onto one of the eigenstates of the observable, gives complete information about the collapsed quantum state. More specifically, the measurement can be described by a set of projection operators $\{P_n = |n\rangle\langle n|\}$ based on the orthonormal basis of the observable. If the initial state is $\rho = |\psi\rangle\langle\psi|$, the probability of obtaining the nth eigenvalue will be $\text{Tr}[P_n\rho P_n]$ with the final state given by

$$\rho_f = \frac{P_n\rho P_n}{\text{Tr}[P_n\rho P_n]} = |n\rangle\langle n|. \tag{S16}$$

While von Neumann measurement provides complete information for the collapsed quantum state after being measured since the state has collapsed to an eigenstate of the observable after the projective measurement, it is not the only kind of measurement. Other methods can reduce the uncertainty of the observable but often fail to remove all of it. Such measurements can extract only partial information about the quantum system.

In principle, we can choose a set of m_{max} operators Ω_m with the restriction

$$\sum_{m=1}^{m_{\text{max}}} \Omega_m^\dagger\Omega_m = I,$$

where the number m_{max} of elements can be larger than the dimension of the Hilbert space which they act in. A measurement with N possible outcomes can be designed for

$$\rho_f = \frac{\Omega_m\rho\Omega_m^\dagger}{\text{Tr}[\Omega_m\rho\Omega_m^\dagger]}, \tag{S17}$$

with the probability $\text{Tr}[\Omega_m\rho\Omega_m^\dagger]$. For example, the probability of the observation in the range $[a, b]$ is given by

$$P(m \in [a, b]) = \sum_{m=a}^{b} \text{Tr}[\Omega_m\rho\Omega_m^\dagger] = \text{Tr}[\sum_{m=a}^{b} \Omega_m\rho\Omega_m^\dagger]. \tag{S18}$$

The measurement, associated with a positive operator $M = \sum_{m=a}^{b} \Omega_m^\dagger \Omega_m$ with every subset in the range $m \in [1, m_{\max}]$, is called a positive operator-valued measure (POVM).

POVMs can describe weak measurements, where only partial information is extracted from the measurement by the Gaussian weighted sum over all eigenstates of the observable:

$$\Omega_m = \frac{1}{\mathcal{N}} \sum_n e^{-k(n-m)^2/4} |n\rangle\langle n|, \tag{S19}$$

with the normalization constant $\mathcal{N}$ that satisfies the constraint $\sum_{m=-\infty}^{\infty} \Omega_m^\dagger \Omega_m = I$. Suppose no information is obtained before the measurement and the initial state is completely mixed as $\rho \propto I$, then the observation is a random variable with Gaussian distribution. After the measurement, the state becomes

$$\rho_f = \frac{\Omega_m \rho \Omega_m^\dagger}{\mathrm{Tr}[\Omega_m \rho \Omega_m^\dagger]} = \frac{1}{\mathcal{N}} \sum_n e^{-k(n-m)^2/2} |n\rangle\langle n|. \tag{S20}$$

This indicates that, when the initial state ρ is an equal probability distribution over all eigenstates, the state after the weak measurement has a Gaussian distribution over all the eigenstates, where the mean value of the Gaussian weights corresponds to an eigenstate and the distribution spreads with a finite uncertainty. Consequently, only partial information can be extracted from this kind of measurement, because it only partially projects onto an eigenstate of the observable with uncertainty. The standard deviation of the final state is $1/\sqrt{k}$. The larger the measurement strength k, the more complete information can be extracted with reduced uncertainty about the quantum state, leading to strong measurement. On the contrary, a small measurement strength generates weak measurement.

We can now describe WCM. In general, continuous measurement means that information is continually extracted from a system over time. To realize WCM, time is divided into a series of intervals of size Δt, and a weak measurement is carried out in each interval. The Hermitian observable is denoted as $\hat{O}$, and the measurement operator with the index α is given by

$$\hat{A}(\alpha) = \left(\frac{4k\Delta t}{\pi}\right)^{1/4} \int_{-\infty}^{\infty} e^{-2k\Delta t(O-\alpha)^2} |O\rangle\langle O| dO, \tag{S21}$$

where the measurement strength is determined by k and Δt. If we set $\Delta t = dt$, then it is a WCM. The mean of the continuous index α is

$$\langle\alpha\rangle = \int_{-\infty}^{\infty} \alpha Tr[\hat{A}^\dagger(\alpha)\hat{A}(\alpha)|\psi\rangle\langle\psi|] d\alpha = \langle\hat{O}\rangle. \tag{S22}$$

The probability distribution of α is

$$\begin{aligned} P(\alpha) &= Tr[\hat{A}^\dagger(\alpha)\hat{A}(\alpha)|\psi\rangle\langle\psi|] \\ &= \sqrt{\frac{4k\Delta t}{\pi}} \int_{-\infty}^{\infty} |\psi(O)|^2 e^{-4k\Delta t(O-\alpha)^2} dO. \end{aligned} \tag{S23}$$

The value of Δt is infinitesimal due to the inherent property of the WCM. As a result, the exponential term in Eq. (S23) is a slow oscillation compared with the wave function under the variable O. Based on this, the wave function can be approximated as $|\psi(O)|^2 \approx \delta(O - \langle O\rangle)$ and we have

$$P(\alpha) \approx \sqrt{\frac{4k\Delta t}{\pi}} e^{-4k\Delta t(\alpha - \langle O\rangle)^2}. \tag{S24}$$

Effectively, α is a stochastic quantity:

$$\alpha_s = \langle \hat{O} \rangle + \frac{\Delta W}{\sqrt{8k}\Delta t}, \tag{S25}$$

where ΔW is a zero-mean, Gaussian random variable with variance Δt. The time evolution of the quantum state under WCM is given by

$$|\psi(t+\Delta t)\rangle \propto \hat{A}(\alpha)|\psi(t)\rangle \propto e^{-2k\Delta t(\alpha-\hat{O})^2}|\psi(t)\rangle. \tag{S26}$$

Substituting Eq. (S25) into this equation, applying Taylor's expansion into the exponential term to first order in Δt and defining $|\psi(t+dt)\rangle \equiv |\psi(t)\rangle + d|\psi\rangle$, we obtain the following stochastic differential equation:

$$d|\psi\rangle = \{-k(\hat{O} - \langle \hat{O} \rangle)^2 dt + \sqrt{2k}(\hat{O} - \langle \hat{O} \rangle) dW\}|\psi(t)\rangle. \tag{S27}$$

Defining $\rho(t+dt) \equiv \rho(t) + d\rho$, we have

$$\begin{aligned} d\rho &= (d|\psi\rangle)\langle\psi| + |\psi\rangle(d\langle\psi|) + (d|\psi\rangle)(d\langle\psi|) \\ &= -k[\hat{O}, [\hat{O}, \rho]]dt + \sqrt{2k}(\hat{O}\rho + \rho\hat{O} - 2\langle \hat{O} \rangle \rho) dW. \end{aligned} \tag{S28}$$

If we redefine the observable as

$$\hat{c} \equiv \sqrt{\eta}\hat{O} \equiv \sqrt{2k}\hat{O},$$

the first term can be rewritten as

$$[\hat{c}\rho\hat{c} - \frac{1}{2}(\hat{c}^2\rho + \rho\hat{c}^2)]dt \tag{S29}$$

and the second term is

$$(\hat{c}\rho + \rho\hat{c} - 2\langle \hat{c} \rangle \rho) dW, \tag{S30}$$

which are consistent with the Lindblad operator $\mathcal{D}$ and the measurement superoperator $\mathcal{H}$ in the SME from the Method part in the main text, respectively. Here, the measurement rate η is proportional to the measurement strength k.

Supplementary Note 4. REINFORCEMENT LEARNING (RL) IN LINEAR QUANTUM OPTOMECHANICS

Based on the demonstration in the main text about RL in linear quantum optomechanics. This section gives the corresponding details about reinforcement learning for the linear system. During online training, given a fixed training episode length, e.g., $\mathrm{Episode} = 3000$, the RL agent bootstraps itself by executing the procedure described in Supplementary Note 6 A. In the initial preparation process, $\mathbb{N}$ identical and independent quantum optomechanical environments ($\mathbb{N}$ parallel environments) are prepared, where $\mathbb{N} = 5$. In addition, the agent, which has two independent neural networks: actor and critic, is also initialized. The initial quantum state is $|\psi\rangle = |10\rangle$ or

$\rho = (1-p)|10\rangle\langle 10| + p|01\rangle\langle 01|$ with $p \in [0,1]$ and the quantum environments are governed by the SME.

In episodic learning, the quantum environments are reset after each episode. For each set of $\mathbb{Z}$ episodes (e.g., $\mathbb{Z} = 5$), the agent obtains the observation O_t about the photon number and the reward value $R_t = -|O_t - 0.5|$ from $\mathbb{N}$ quantum environments, and independently acts on them by the current stochastic policy $\pi(G_t|O_t;\boldsymbol{\theta})$. Essentially, the policy is the conditional probability distribution on the action space $G_t \in [-5,5]\omega_m$ given the observation O_t and is parameterized through $\boldsymbol{\theta}$. The $\mathbb{N} \times \mathbb{Z}$ independent trajectories, denoted as τ^j with the trajectory index $j = 1, 2, \ldots, \mathbb{N} \times \mathbb{Z}$, are collected with length $T = 500$ (the number of time steps for each episode) and the step size $dt = 0.01\omega_m^{-1}$. Each trajectory τ^j is a sequence of states (observations), actions, rewards, and next states (next observations):

$$\tau^j = (O_0^j, G_0^j, R_0^j, O_1^j, \ldots, O_{T-1}^j, G_{T-1}^j, R_{T-1}^j), \tag{S31}$$

which can be organized as a sub-trajectory tuple

$$\tau_t^j = (O_t^j, G_t^j, R_t^j, O_{t+1}^j) \tag{S32}$$

with the time stage index $t = 0, 1, \ldots, T-2$. At the terminal stage $t = T-1$, we have

$$\tau_{T-1}^j = (O_{T-1}^j, G_{T-1}^j, R_{T-1}^j). \tag{S33}$$

For each sub-trajectory tuple τ_t^j, the generalized advantage estimation (GAE) [45] $\hat{A}_t^j$ uses a value function estimator:

$$\hat{A}_t^j = \delta_t^j + (\gamma\lambda)\delta_{t+1}^j + \ldots + (\gamma\lambda)^{T-t-1}\delta_{T-1}^j \tag{S34}$$

with

$$\delta_t^j = R_t^j + \gamma V(O_{t+1}^j;\boldsymbol{\phi}) - V(O_t^j;\boldsymbol{\phi}), \tag{S35}$$

where the value function $V(O_t^j;\boldsymbol{\phi})$ is utilized to score the quality of O_t^j based on the accumulated reward and parameterized by $\boldsymbol{\phi}$ and δ_t^j is the relative advantage of the current action selected by the stochastic policy $\pi(G_t^j|O_t^j;\boldsymbol{\theta})$ with the discounted factor $\gamma \in (0,1)$ and hyperparameter λ with typical value $\lambda = 0.95$. Intuitively, $\hat{A}_t^j$ is utilized to numerically quantify the relative cumulative advantage of a certain action selected by the current stochastic policy from time t to the terminal stage $T-1$, in which the future impact is included but regarded as less important than the corresponding previous one by the discount factor $\gamma \in (0,1)$. The finite-horizon discounted return $\hat{\mathcal{G}}_t^j$ is defined as

$$\hat{\mathcal{G}}_t^j = \sum_{k=t}^{T-1} \gamma^{k-t} R_k^j, \tag{S36}$$

which can be also obtained from the generalized advantage by

$$\hat{\mathcal{G}}_t^j = \hat{A}_t^j + V(O_t^j;\boldsymbol{\phi}), \tag{S37}$$

where $\hat{\mathcal{G}}_t^j$ denotes the accumulated reward from time t to the terminal stage in the discounted version.

The neural networks constituting the actor and critic are updated from minibatches with size $\mathbb{M}$ from $\mathbb{N} \times \mathbb{Z} \times T$ data points, consisting of the sub-trajectory τ_t^j, the generalized advantage

$\hat{A}_t^j$ and the return $\hat{\mathcal{G}}_t^j$ over $k = 10$ epochs with the Adam algorithm. The typical batch size is $\mathbb{M} = \text{int}(\mathbb{N} \times \mathbb{Z} \times T/10)$. For each epoch, the critic parameters $\boldsymbol{\phi}$ in the loss $L_{critic}(\boldsymbol{\phi})$ and the actor parameters $\boldsymbol{\theta}$ in the loss $L_{actor}(\boldsymbol{\theta})$ need to be updated to minimize the loss function over a random minibatch data. The mean square loss $L_{critic}(\boldsymbol{\phi})$ about the target $\hat{\mathcal{G}}_i$ for the value function $V(O_i;\boldsymbol{\phi})$ is

$$L_{critic}(\boldsymbol{\phi}) = \hat{\mathbb{E}}_i[(V(O_i;\boldsymbol{\phi}) - \hat{\mathcal{G}}_i)^2] \tag{S38}$$

and the clipped loss $L_{actor}(\boldsymbol{\theta})$ is given by

$$L_{actor}(\boldsymbol{\theta}) = \hat{\mathbb{E}}_i\left[-\min(r_i(\boldsymbol{\theta})\hat{A}_i, \text{clip}(r_i(\boldsymbol{\theta}), [1-\epsilon, 1+\epsilon])\hat{A}_i)\right], \tag{S39}$$

where $\hat{\mathbb{E}}_i[\ \ ] = \sum_{i=1}^{\mathbb{M}}[\ \ ]_i/\mathbb{M}$ is the empirical average over a minibatch of the data and $[\ \ ]_i$ denotes the i th element of the minibatch with $i = 0, 1, \ldots, \mathbb{M}-1$, and the clip function clip$(x, [\min, \max])$ returns x clipped to set limits: min $\leq x \leq$ max. The probability ratio $r_i(\boldsymbol{\theta}) > 0$ between the current and old policies is

$$r_i(\boldsymbol{\theta}) = \frac{\pi_{\boldsymbol{\theta}}(G_i|O_i)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(G_i|O_i)}. \tag{S40}$$

If the current policy is the same as the old policy, we have $r_i(\boldsymbol{\theta}_{\text{old}}) = 1$. In general, the ratio $r_i(\boldsymbol{\theta})$ needs to be away from the value one for the policy to be optimized. However, $r_i(\boldsymbol{\theta})$'s deviating too much from the value one will result in many fast policy updates, possibly leading to instabilities and even a collapse of the learning process. To avoid this, the clip function in the actor loss $L_{actor}(\boldsymbol{\theta})$ can be utilized to remove the incentive for $r_i(\boldsymbol{\theta})$ outside of the interval $[1-\epsilon, 1+\epsilon]$ with typical clip range $\epsilon = 0.2$, which decreases the updating speed of policy and improves the learning stability.

Intuitively, the goal of RL is to maximize the cumulative reward. In the linear optomechanical system, the objective is to achieve the entangled Bell state as fast as possible or, as stipulated by the reward function, to achieve the optimal photon number $O_t \to 0.5$ and to maintain this for as long as possible. When the RL agent converges to the optimal policy, the Bellman equation is satisfied [14], so the optimal value function satisfies

$$V^*(O_t^j;\boldsymbol{\phi}) = R_t^j + \gamma V^*(O_{t+1}^j;\boldsymbol{\phi}), \tag{S41}$$

i.e., the optimal value function about O_t^j is equal to the current reward plus the future discounted cumulative reward, in which O_{t+1}^j is determined by the action selected by the optimal policy $\pi^*(G_t^j|O_t^j;\boldsymbol{\theta})$. It guarantees that the agent makes the best possible decisions to maximize the rewards [14]. Moreover, under the optimal policy, it means the zero generalized advantage $\hat{A}_t^j$, so the zero actor loss L_{actor}^* is obtained. It is worth noting that the optimal value function is equal to the discounted accumulated reward from Eq. (S36):

$$V^*(O_i;\boldsymbol{\phi}) = \hat{\mathcal{G}}_i, \tag{S42}$$

which also gives zero critic loss $L_{critic}(\boldsymbol{\phi})$. In the online training process, the RL agent trained as described is called the PPO agent, whose policy is randomly initialized and will gradually converge to the optimal one under the described training scenarios to achieve the maximum accumulated reward. Physically, this enables the entangled Bell state to be created and stabilized. For online testing, the optimized policy is no longer updated and only one quantum environment is involved.

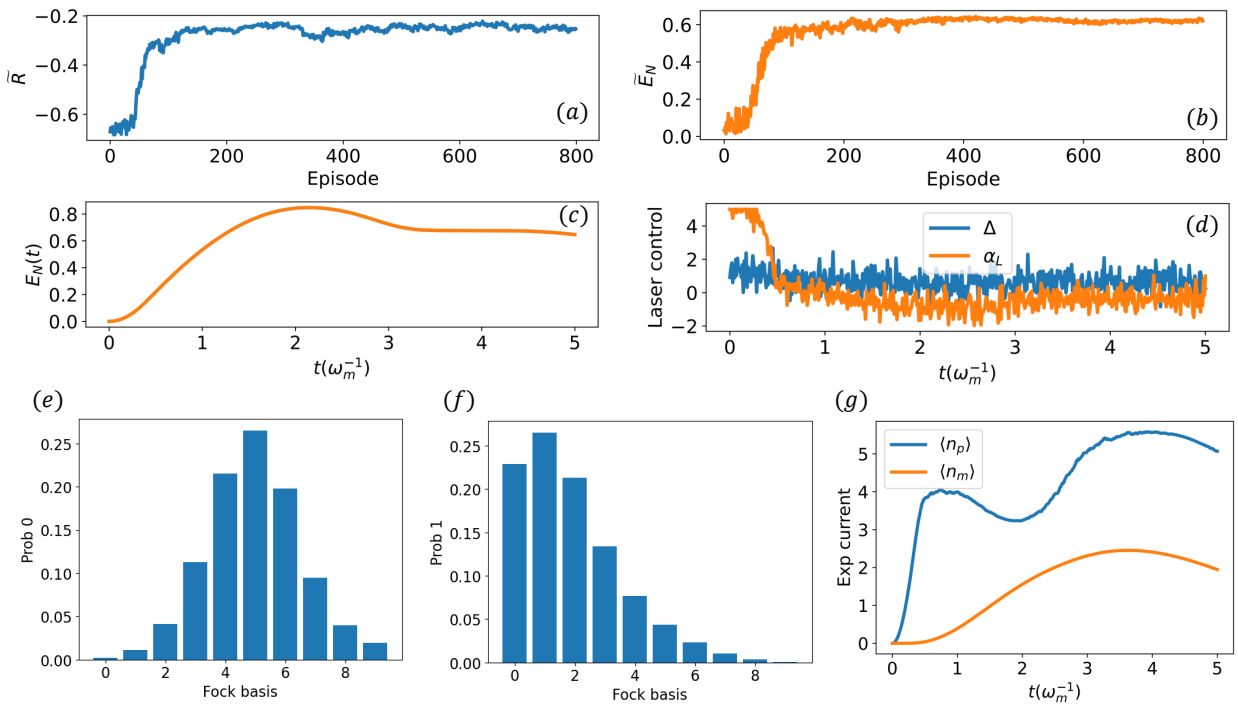

FIG. S1. A detailed account of the target-generating phase in RL control of open optomechanical systems with nonlinear photon-phonon interaction in the framework of Lindblad master equation. Nonlinear interaction of strength $g_0 = 0.2\omega_m$ creates the target entanglement $E_N \sim \log 2$ optimized by the PPO agent from vacuum states with $|\psi\rangle = |00\rangle$ with 10×10 Fock bases. The dissipation rates to the vacuum bath are $\kappa = 0.1\,\omega_m$ and $\gamma = 0.01\kappa$. The time-dependent control signal is the detuning Δ and the amplitude of the driven laser α_L within the range $\Delta, \alpha_L \in [-5, 5]\omega_m$. In the training phase, observation is set as $E_N(t)$. (a,b) Trained $\widetilde{R}$ and $\widetilde{E}_N$ converge to some constant values. (c,d) Time-dependent series $E_N(t)$, where the driven laser signals are shown at the end of the training phase. (e,f) The corresponding coherent- and thermal-shape states expanded in the Fock basis at the end of the time of the selected training episode in (c,d). (g) The time evolution of the corresponding expected measurement current, including the expected number $\langle n_p \rangle$ of photons as well as the expected phonon number $\langle n_m \rangle$ in the Fock basis, where the time series $\langle n_p \rangle$(t) serves as the target to construct reward function in the target-utilization phase shown in Fig. S2.

Supplementary Note 5. RL IN NONLINEAR INTERACTIONS BY THE LINDBLAD MASTER EQUATION

Figs. S1 and S2 display the case where the stochastic process in SME is removed so that the quantum dynamics are reduced to those governed by the Lindblad master equation, in which the decoherence part includes only the dissipation to the vacuum bath. In this setting, the nonlinear coupling represented by $\hbar g_0 \hat{a}^\dagger \hat{a}(\hat{b}^\dagger + \hat{b})$ can still be exploited to create the entanglement. A caveat is that the process can simultaneously generate undesired high-level quantum states. A solution is to apply deep RL to create and stabilize the entanglement $E_N \sim \log 2$, where the problem is how to control the excitation within a limited Fock basis. For this problem, a key is choosing the effective and experimentally feasible observation data.

Here, we describe in detail our two-step training process leading to a solution through the Lind-

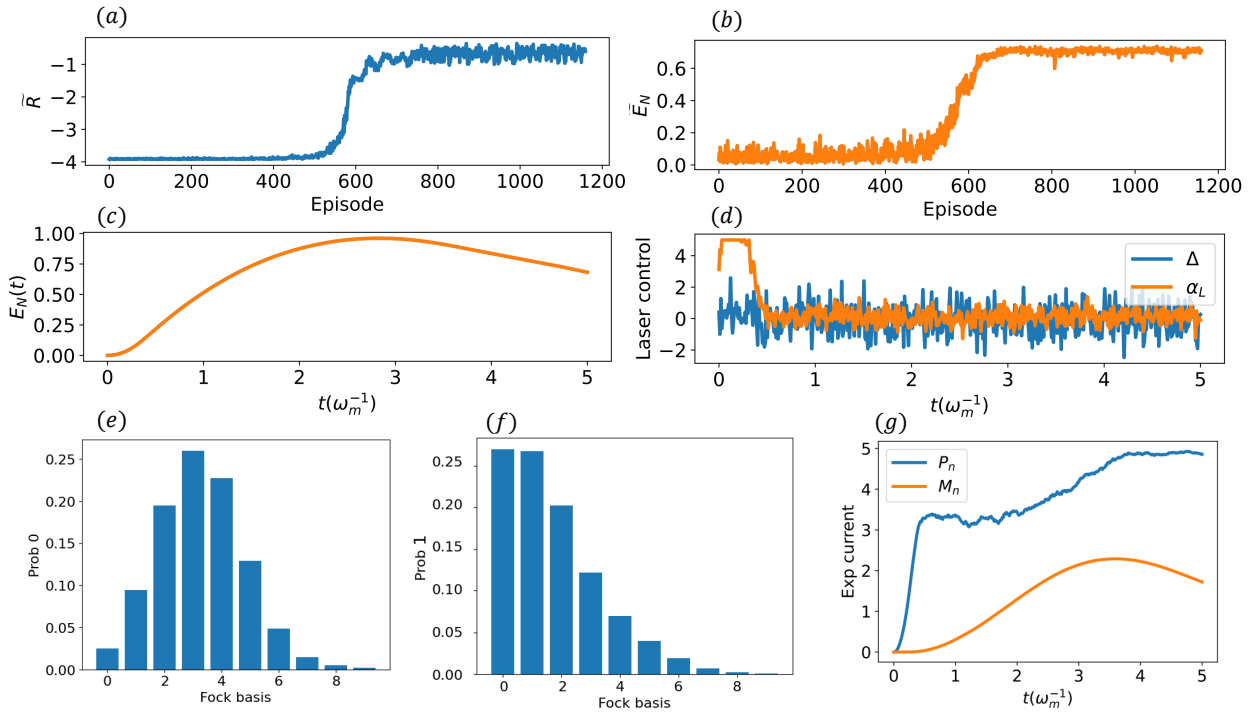

FIG. S2. A detailed account of the target-utilization phase in RL control of open optomechanical systems with nonlinear photon-phonon interaction in the framework of Lindblad master equation. The goal is to create the entanglement about $E_N \sim \log 2$, for which the reward function is $R(t) = -|\langle n_p\rangle(t) - \langle n_p^{\text{target}}\rangle(t)|$, determined by the target time series $\langle n_p^{\text{target}}\rangle(t)$ from Fig. S1(g). In this training configuration, the observation of the recurrent PPO agent is the expected photon number $\langle n_p\rangle(t)$. In spite of the observation being partial and incomplete, all results in (a-g) display similar behavior compared with the ones in Fig. S1, where the entanglement quantity $E_N(t)$ is directly observed. However, such observation is currently not experimentally feasible. The setting and other parameters are the same as those in Fig. S1.

blad master equation. The first step is the target-generating phase, in which numerical simulation is used to generate the observation and reward data, where the PPO agent observes the logarithmic negativity $E_N(t)$ directly and constructs the reward function combining the expected numbers of photons and phonons $R(t) = -|E_N(t) - \log 2| - |\langle n_p\rangle(t) + \langle n_m\rangle(t) - a|/b$. Figure S1 illustrates the target-generating phase, where the range of excited quantum states in the Fock basis is limited by the total number $\langle n_p\rangle + \langle n_m\rangle$ with the optimized hyperparameters $a = 1$ and $b = 40$. The target time series of the expected photon number is $\langle n_p^{\text{target}}\rangle(t)$. The second step is the target-utilization phase, where the reward function is given by $R(t) = -|\langle n_p\rangle(t) - \langle n_p^{\text{target}}\rangle(t)|$. The recurrent PPO will only observe the expected photon number $\langle n_p\rangle(t)$, which is experimentally feasible. During the two-step training, the agent collects data from five parallel quantum optomechanical environments and updates the policy every five episodes.

More specifically, Figs. S1(a) and S1(b) show that both the reward $\widetilde{R}$ and the logarithmic negativity $\widetilde{E}_N$ converge in time during the target-generating phase. The trained agent can create and stabilize the entanglement as shown in Fig. S1(c) controlled by the laser control signal shown in Fig. S1(d). At the end of the time series, entanglement is produced from the coherent-(photon) and thermal-shape (phonon) Fock states as displayed in Figs. S1(e) and S1(f). The corresponding

target pattern $\langle n_p^{\text{target}}\rangle(t)$ is demonstrated in Fig. S1(g). In the target-utilization phase, the target pattern $\langle n_p^{\text{target}}\rangle(t)$ is time-dependent, which is difficult to learn if only MLPs are used. Here a single long short-term memory (LSTM) network is added after the MLPs in both the actor and critic network, so the whole neural-network architecture is able to handle the time-dependent data. Figure S2 illustrates that, with only partial information extracted from the quantum optomechanical environment, the agent can steadily learn to create and stabilize entanglement.

Supplementary Note 6. DEEP RL

There are three main RL approaches [13] based, respectively, on (1) value functions, (2) policy search, and (3) a hybrid actor-critic method that employs both the value functions and policy search. Specifically, the actor-critic method uses the value function as a baseline for policy gradients, based on a trade-off between variance reduction of policy gradients and bias associated with value functions. Incorporating deep neural networks as a powerful function approximator into RL to obtain the optimal value functions and the optimal policy leads to deep RL with the advantage of mitigating the issues associated with high dimensionality (overcoming the curse of dimensionality). A difficulty with deep RL is the local minima in the neural-network dynamics with a large number of parameters when directly searching for the optimal policy [13]. A common solution is to use a trust region that prevents an updated policy from deviating too far from the previous policies, thereby guaranteeing monotonic enhancement in policy search. To implement this, the trust region proximal optimization (TRPO) method [16] can be exploited, which makes the advantage estimate in the surrogate objective function constrained by Kullback–Leibler (KL) divergence. The combination of TRPO and generalized advantage estimation (GAE) is one of the state-of-the-art RL techniques for continuous control.

A. PPO agent

Proximal policy optimization (PPO) [17] agent attains the data efficiency and reliable performance of TRPO with only first-order optimization through a novel objective with clipped probability ratios, which can be readily implemented with reduced complexity. A typical online training process of PPO agent consists of the following steps:

Step 1 - Initialization: initialize the actor $\pi(a|s;\boldsymbol{\theta})$ and the critic $V(s;\boldsymbol{\phi})$ with random parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, respectively. Both the actor and critic are components of the PPO agent. The stochastic policy $\pi(a|s;\boldsymbol{\theta})$ is the conditional probability distribution on action space a given state s. The value function $V(s;\boldsymbol{\phi})$ is utilized to score the quality of state s based on the accumulated reward.

Step 2 - Trajectory collection: The quantum state or quantum environment is initialized for the first episode or reset for the following episodes. The agent interacts independently with $\mathbb{N}$ parallel quantum optomechanical environments (identical and independent) using the current stochastic policy $\pi_{\boldsymbol{\theta}}(a_t|s_t)$ at time t. After $\mathbb{Z}$ episodes, $\mathbb{N}\times\mathbb{Z}$ independent trajectories of length T (the total time steps T for each episode) are collected as sequences of states s_t^j, actions a_t^j, rewards R_t^j, and next states s_{t+1}^j, in which the sub-trajectory tuple τ_t^j is defined as

$$\tau_t^j = (s_t^j, a_t^j, R_t^j, s_{t+1}^j) \tag{S43}$$

with the trajectory index $j = 1, 2, \ldots, \mathbb{N} \times \mathbb{Z}$ and the time index $t = 0, 1, \ldots, T-2$. At the terminal stage $t = T-1$, the following holds:

$$\tau_{T-1}^j = (s_{T-1}^j, a_{T-1}^j, R_{T-1}^j). \tag{S44}$$

The sub-trajectory tuple τ_t^j can be utilized to calculate and evaluate the performance of the agent at each time stage t. The trajectory τ^j of length T is the union of the sub-trajectory tuple τ_t^j in the form of

$$\tau^j = \tau_0^j \cup \tau_1^j \cup ..... \cup \tau_{T-1}^j, \tag{S45}$$

so the trajectory τ^j is given by

$$\tau^j = (s_0^j, a_0^j, R_0^j, s_1^j, ..., s_{T-1}^j, a_{T-1}^j, R_{T-1}^j). \tag{S46}$$

Step 3 - Generalized advantage estimator and return: Estimate the advantages for each sub-trajectory tuple τ_t^j in the collected trajectories. In particular, the generalized advantage estimation (GAE) [45] uses a value function estimator:

$$\hat{A}_t^j = \delta_t^j + (\gamma\lambda)\delta_{t+1}^j + \ldots + (\gamma\lambda)^{T-t-1}\delta_{T-1}^j, \tag{S47}$$

with

$$\delta_t^j = R_t^j + \gamma V(s_{t+1}^j; \boldsymbol{\phi}) - V(s_t^j; \boldsymbol{\phi}), \tag{S48}$$

where δ_t is the relative advantage of the current action selected by the policy $\pi(a_t^j | s_t^j; \boldsymbol{\theta})$ with the discounted factor $\gamma \in (0, 1)$ and hyperparameter λ (typical value $\lambda = 0.95$). The generalized advantage $\hat{A}_t^j$ at time t is the discounted cumulative advantage from time t to the terminal stage $T-1$.

In episodic learning (policy update after each Z number of episodes), the return $\hat{\mathcal{G}}(\tau^j)$ is defined as the cumulative reward over the trajectory τ^j, i.e., $\hat{\mathcal{G}}(\tau^j) = \sum_{t=0}^{T-1} R_t^j$ with the time horizon T. For mathematical convenience, we use the discounted version, i.e., finite-horizon discounted return

$$\hat{\mathcal{G}}(\tau^j) = \sum_{t=0}^{T-1} \gamma^t R_t^j.$$

It implies that future performance is also included but less important than the previous one. The return $\hat{\mathcal{G}}_t^j$ at each time step is the sum of the discounted reward from the current time t,

$$\hat{\mathcal{G}}_t^j = \sum_{k=t}^{T-1} \gamma^{k-t} R_k^j,$$

which can be also obtained from the generalized advantage:

$$\hat{\mathcal{G}}_t^j = \hat{A}_t^j + V(s_t^j; \boldsymbol{\phi}). \tag{S49}$$

Step 4 - Update of the actor and critic from minibatches of training data over k epochs with Adam or stochastic gradient descent. For each epoch, we first sample a random minibatch data set

with size $\mathbb{M}$ from $\mathbb{N} \times \mathbb{Z} \times T$ data points, including the sub-trajectory tuple τ_t^j, the corresponding advantage $\hat{A}_t^j$ and return value $\hat{\mathcal{G}}_t^j$. We then update the critic parameters $\boldsymbol{\phi}$ by minimizing the loss $L_{critic}(\boldsymbol{\phi})$ across all sampled minibatch data, which is given by

$$L_{critic}(\boldsymbol{\phi}) = \hat{\mathbb{E}}_i[(V(s_i;\boldsymbol{\phi}) - \hat{\mathcal{G}}_i)^2], \tag{S50}$$

where $\hat{\mathbb{E}}_i[\ ] = \sum_{i=1}^{\mathbb{M}}[\ ]_i/\mathbb{M}$ is the empirical average over a minibatch of data and $[\ ]_i$ denotes the ith element of the minibatch with $i = 0, 1, ..., \mathbb{M}-1$. After this, we update the actor parameters $\boldsymbol{\theta}$ by minimizing the loss $L_{actor}(\boldsymbol{\theta})$ given by

$$L_{actor}(\boldsymbol{\theta}) = \hat{\mathbb{E}}_i\left[-\min(r_i(\boldsymbol{\theta})\hat{A}_i, \text{clip}(r_i(\boldsymbol{\theta}), [1-\epsilon, 1+\epsilon])\hat{A}_i)\right]. \tag{S51}$$

where the clip function $\text{clip}(x, [\min, \max])$ returns x clipped to set limits, i.e., $\min \leq x \leq \max$. The probability ratio $r_i(\boldsymbol{\theta})$ between the current and old policies is defined as

$$r_i(\boldsymbol{\theta}) = \frac{\pi_{\boldsymbol{\theta}}(a_i|s_i)}{\pi_{\boldsymbol{\theta}_{\text{old}}}(a_i|s_i)}. \tag{S52}$$

If the current policy is the same as the old policy, we have $r_i(\boldsymbol{\theta}_{\text{old}}) = 1$. Otherwise, the ratio $r_i(\boldsymbol{\theta})$ will be away from the value one to get the new optimized policy. The clip function in actor loss $L_{actor}(\boldsymbol{\theta})$ is utilized to remove the incentive for $r_i(\boldsymbol{\theta})$ outside of the interval $[1-\epsilon, 1+\epsilon]$, which decreases the update speed of policy and improves the learning stability.

Step 5 - Repeating Steps (2-4) for a specified number of iterations or until convergence is achieved.

B. Recurrent PPO agent

In general, the dynamical process of RL is Markovian: the future depends only on the present state. While this suitably describes many processes, there are applications where a non-Markovian type of RL is required, e.g., partially observable Markov Decision Processes (POMDPs) or when the physical system to be controlled is in a non-Markovian environment. Leveraging recurrent neural networks (RNNs) for memory-based agent learning provides a solution. In particular, a RNN can store past information as memory by introducing loops in the neural network, in contrast to, e.g., feed-forward neural networks where signals flow only from input to output in a one-way manner. However, conventional RNNs may not be able to efficiently connect the long past information to the present task, a problem known as gap sensitivity or vanishing gradient.

Long short-term memory (LSTM) [18] is capable of learning long-term dependencies, thereby overcoming the vanishing gradient problem. The key component of LSTM is the cell state, which mimics a conveyor belt [46]. Information can be added or removed by the forget input, and output gates. Since the actor and critic networks underlying PPO are multilayer perceptrons (MLPs), e.g. a special class of the feed-forward neural networks with fully connected layers, applying LSTM after MLPs leads to a recurrent PPO agent, where MLPs are responsible for feature learning and LSTM contributes long-term history memorization. For a recurrent PPO, the state s_t is replaced by observation o_t and the hidden states h_t with POMDPs [19].

C. Details in deep RL

Some details about the hyperparameter in the PPO agent are as follows: The discounted factor is $\gamma = 0.99$, the parameter for the generalized advantage estimation(GAE) is $\lambda = 0.95$, the clip range is set $\epsilon = 0.2$, the maximum gradient is set to be 0.5 and the learning rate is 0.5×10^{-3}. Especially, GAE is normalized by subtracting its mean value and dividing by its standard deviation, the stochastic policy is based on the action noise exploration instead of the state-dependent exploration, and the value function is no clipping. Since the observation is the measurement current with large variance, it is necessary to apply a one-dimensional Gaussian filter from the Scipy package, of which the filter interval and standard deviation of the Gaussian kernel are listed in as bellow. In the process of variance reduction for WCM photocurrent, the measurement photocurrent is averaged over five trajectories (an independent ensemble) at each time step, and then averaged over the previously successive five time steps. Finally, the obtained data is filtered by the Gaussian kernel. In the updating phase, the network parameters from actor and critic are updated by Adam with the minibatch size, one-tenth of training data, and epochs $k = 10$.

TABLE S1. Gaussian filter with filter interval and standard deviation of the Gaussian kernel.

measurement rate	filter interval size	standard deviation
1.0	10	3.0
0.7	10	4.5
0.5	10	6.0
0.3	20	6.0
0.1	100	24.0
0.05	150	48.0

Supplementary Note 7. SUPPLEMENTARY REFERENCES

[1] Wiseman, H. M. & Milburn, G. J. *Quantum measurement and control* (Cambridge university press, 2009).

[2] Rosencher, E. *et al.* Quantum engineering of optical nonlinearities. *Science* **271**, 168–173 (1996).

[3] Iannaccone, G., Bonaccorso, F., Colombo, L. & Fiori, G. Quantum engineering of transistors based on 2d materials heterostructures. *Nature nanotechnology* **13**, 183–191 (2018).

[4] Bohn, J. L., Rey, A. M. & Ye, J. Cold molecules: Progress in quantum engineering of chemistry and quantum matter. *Science* **357**, 1002–1010 (2017).

[5] Paz-Silva, G. A. & Viola, L. General transfer-function approach to noise filtering in open-loop quantum control. *Physical review letters* **113**, 250501 (2014).

[6] Machnes, S. *et al.* Comparing, optimizing, and benchmarking quantum-control algorithms in a unifying programming framework. *Phys. Rev. A* **84**, 022305 (2011). URL `https://link.aps.org/doi/10.1103/PhysRevA.84.022305`.

[7] Dolde, F. *et al.* High-fidelity spin entanglement using optimal control. *Nat Commun* **5**, 3371 (2014).

[8] Spörl, A. *et al.* Optimal control of coupled josephson qubits. *Phys. Rev. A* **75**, 012302 (2007). URL `https://link.aps.org/doi/10.1103/PhysRevA.75.012302`.

[9] Heeres, R. W. *et al.* Implementing a universal gate set on a logical qubit encoded in an oscillator. *Nat. Commun.* **8**, 94 (2017).

[10] Fisher, R., Helmer, F., Glaser, S. J., Marquardt, F. & Schulte-Herbrüggen, T. Optimal control of circuit quantum electrodynamics in one and two dimensions. *Phys. Rev. B* **81**, 085328 (2010). URL `https://link.aps.org/doi/10.1103/PhysRevB.81.085328`.

[11] Porotti, R., Peano, V. & Marquardt, F. Gradient-ascent pulse engineering with feedback. *PRX Quantum* **4**, 030305 (2023).

[12] Mnih, V. *et al.* Human-level control through deep reinforcement learning. *nature* **518**, 529–533 (2015).

[13] Arulkumaran, K., Deisenroth, M. P., Brundage, M. & Bharath, A. A. Deep reinforcement learning: A brief survey. *IEEE Signal Process. Mag.* **34**, 26–38 (2017).

[14] Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction* (MIT press, 2018).

[15] Lillicrap, T. P. *et al.* Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* (2015).

[16] Wu, Y., Mansimov, E., Grosse, R. B., Liao, S. & Ba, J. Scalable trust-region method for deep reinforcement learning using Kronecker-factored approximation. *Adv. Neural Inf. Process.* **30** (2017).

[17] Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[18] Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).

[19] Pleines, M., Pallasch, M., Zimmer, F. & Preuss, M. Generalization, mayhems and limits in recurrent proximal policy optimization. *arXiv preprint arXiv:2205.11104* (2022).

[20] Fösel, T., Tighineanu, P., Weiss, T. & Marquardt, F. Reinforcement learning with neural networks for quantum feedback. *Phys. Rev. X* **8**, 031084 (2018). URL `https://link.aps.org/doi/10.1103/PhysRevX.8.031084`.

[21] Wang, Z. T., Ashida, Y. & Ueda, M. Deep reinforcement learning control of quantum cartpoles. *Phys. Rev. Lett.* **125**, 100401 (2020). URL https://link.aps.org/doi/10.1103/PhysRevLett.125.100401.

[22] Borah, S., Sarma, B., Kewming, M., Milburn, G. J. & Twamley, J. Measurement-based feedback quantum control with deep reinforcement learning for a double-well nonlinear potential. *Phys. Rev. Lett.* **127**, 190403 (2021). URL https://link.aps.org/doi/10.1103/PhysRevLett.127.190403.

[23] Sivak, V. *et al.* Model-free quantum control with reinforcement learning. *Phys. Rev. X* **12**, 011059 (2022).

[24] Porotti, R., Essig, A., Huard, B. & Marquardt, F. Deep reinforcement learning for quantum state preparation with weak nonlinear measurements. *Quantum* **6**, 747 (2022).

[25] Cao, J.-H. *et al.* Detection of entangled states supported by reinforcement learning. *Phys. Rev. Lett.* **131**, 073201 (2023).

[26] Haldar, S., Barge, P. J., Khatri, S. & Lee, H. Fast and reliable entanglement distribution with quantum repeaters: principles for improving protocols using reinforcement learning. *Phys. Rev. Applied* **21**, 024041 (2024).

[27] Reuer, K. *et al.* Realizing a deep reinforcement learning agent discovering real-time feedback control strategies for a quantum system. *arXiv preprint arXiv:2210.16715* (2022).

[28] Ristè, D., Bultink, C. C., Lehnert, K. W. & DiCarlo, L. Feedback control of a solid-state qubit using high-fidelity projective measurement. *Phys. Rev. Lett.* **109**, 240502 (2012). URL https://link.aps.org/doi/10.1103/PhysRevLett.109.240502.

[29] Lundeen, J. S. & Bamber, C. Procedure for direct measurement of general quantum states using weak measurement. *Phys. Rev. Lett.* **108**, 070402 (2012). URL https://link.aps.org/doi/10.1103/PhysRevLett.108.070402.

[30] Smith, G. A., Chaudhury, S., Silberfarb, A., Deutsch, I. H. & Jessen, P. S. Continuous weak measurement and nonlinear dynamics in a cold spin ensemble. *Phys. Rev. Lett.* **93**, 163602 (2004). URL https://link.aps.org/doi/10.1103/PhysRevLett.93.163602.

[31] Yang, D., Huelga, S. F. & Plenio, M. B. Efficient information retrieval for sensing via continuous measurement. *Phys. Rev. X* **13**, 031012 (2023). URL https://link.aps.org/doi/10.1103/PhysRevX.13.031012.

[32] Giovannetti, V., Lloyd, S. & Maccone, L. Quantum metrology. *Phys. Rev. Lett.* **96**, 010401 (2006). URL https://link.aps.org/doi/10.1103/PhysRevLett.96.010401.

[33] Giovannetti, V., Lloyd, S. & Maccone, L. Advances in quantum metrology. *Nature Photon* **5**, 222–229 (2011).

[34] Puentes, G. *et al.* Bridging particle and wave sensitivity in a configurable detector of positive operator-valued measures. *Phys. Rev. Lett.* **102**, 080404 (2009).

[35] Kanem, J., Maneshi, S., Myrskog, S. & Steinberg, A. Phase space tomography of classical and nonclassical vibrational states of atoms in an optical lattice. *J. Opt. B: Quantum Semiclass. Opt.* **7**, S705 (2005).

[36] Lvovsky, A. I. & Raymer, M. G. Continuous-variable optical quantum-state tomography. *Rev. Mod. Phys.* **81**, 299 (2009).

[37] Lundeen, J. S., Sutherland, B., Patel, A., Stewart, C. & Bamber, C. Direct measurement of the quantum wavefunction. *Nature* **474**, 188–191 (2011).

[38] Thekkadath, G. S. *et al.* Tuning between photon-number and quadrature measurements with weak-field homodyne detection. *Phys. Rev. A* **101**, 031801 (2020). URL https://link.aps.org/doi/10.1103/PhysRevA.101.031801.

[39] Rozema, L. A. *et al.* Violation of Heisenberg's measurement-disturbance relationship by weak measurements. *Phys. Rev. Lett.* **109**, 100404 (2012). URL https://link.aps.org/doi/10.1103/PhysRevLett.109.100404.

[40] Kaneda, F., Baek, S.-Y., Ozawa, M. & Edamatsu, K. Experimental test of error-disturbance uncertainty relations by weak measurement. *Phys. Rev. Lett.* **112**, 020402 (2014). URL https://link.aps.org/doi/10.1103/PhysRevLett.112.020402.

[41] Jacobs, K. & Steck, D. A. A straightforward introduction to continuous quantum measurement. *Contemp. Phys.* **47**, 279–303 (2006).

[42] Liu, Y.-C., Xiao, Y.-F., Luan, X. & Wong, C. W. Dynamic dissipative cooling of a mechanical resonator in strong coupling optomechanics. *Phys. Rev. Lett.* **110**, 153606 (2013).

[43] Qian, J., Clerk, A., Hammerer, K. & Marquardt, F. Quantum signatures of the optomechanical instability. *Phys. Rev. Lett.* **109**, 253601 (2012).

[44] Bowen, W. P. & Milburn, G. J. *Quantum optomechanics* (CRC press, 2015).

[45] Schulman, J., Moritz, P., Levine, S., Jordan, M. & Abbeel, P. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438* (2015).

[46] Olah, C. Understanding LSTM networks (2015). https://colah.github.io/posts/2015-08-Understanding-LSTMs/.