Entanglement engineering of optomechanical systems by reinforcement learning

Li-Li Ye,[1] Christian Arenz,[1] Joseph M. Lukens,[2,3] and Ying-Cheng Lai[1,4,a]

[1] *School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, Arizona 85287, USA*

[2] *Research Technology Office and Quantum Collaborative, Arizona State University, Tempe, Arizona 85287, USA*

[3] *Quantum Information Science Section, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA*

[4] *Department of Physics, Arizona State University, Tempe, Arizona 85287, USA*

(Dated: 14 August 2024)

Entanglement is fundamental to quantum information science and technology, yet controlling and manipulating entanglement — so-called entanglement engineering — for arbitrary quantum systems remains a formidable challenge. There are two difficulties: the fragility of quantum entanglement and its experimental characterization. We develop a model-free deep reinforcement-learning (RL) approach to entanglement engineering, in which feedback control together with weak continuous measurement and partial state observation is exploited to generate and maintain desired entanglement. We employ quantum optomechanical systems with linear or nonlinear photon-phonon interactions to demonstrate the workings of our machine-learning-based entanglement engineering protocol. In particular, the RL agent sequentially interacts with one or multiple parallel quantum optomechanical environments, collects trajectories, and updates the policy to maximize the accumulated reward to create and stabilize quantum entanglement over an arbitrary amount of time. The machine-learning-based model-free control principle is applicable to the entanglement engineering of experimental quantum systems in general.

I. INTRODUCTION

Entanglement[1–5] is fundamental to all fields in quantum information science such as quantum sensing[6], quantum computation[7], and quantum networks[8–13]. However, the inherent fragility of quantum entanglement and coherence[14] poses significant challenges for experimental applications. For example, in quantum computing, the application of quantum gates to quantum states needs to last for a finite amount of time[15–20], making it critical to maintain the entanglement after its creation. Moreover, the transition from noisy intermediate-scale systems[21] to large-scale, fault-tolerant systems[16] requires sophisticated entanglement engineering strategies to establish and maintain entanglement through optimal control protocols in the presence of noise and decoherence.

At the present, a major limitation/challenge in entanglement engineering is the experimental observation design. Existing machine-learning based works use the full fidelity, i.e., the overlap between the current and target quantum states, as the observation metric. Applications range from the generation of two[22] and multi-qubit entangled states[23,24] to specific many-body states[25–27] and single-particle quantum state engineering via deep reinforcement learning (RL)[28,29]. However, full fidelity observation is not universally applicable in experiments. Moreover, obtaining the relationship between the entanglement and experimental observables is difficult. So far there have been no systematical methods to extract quantitative entanglement from experimental observation for arbitrary quantum systems[30–32], in spite of some initial exploration for specific systems. For example, an entanglement criterion for non-Gaussian states in coupled harmonic oscillators was developed[30]. Under the strong laser approximation, a Bell inequality was tested with photon counting[31], and stationary entanglement for Gaussian states was inferred from the continuous measurement of light only[32].

In this paper, using quantum optomechanical systems with linear or nonlinear photon-phonon interactions as a paradigm, we develop a deep RL approach to entanglement engineering. For quantum control of optomechanical systems, most existing theoretical studies focused on Gaussian states or the linear interaction regime[33–45], with the primary goal of generating entanglement as quickly as possible (entanglement enhancement)[38,43–45]. Previous control methods are mostly model-based: prior information about the system model is needed, such as the pulse method[33–37], time-continuous laser-driven approaches[38,39], periodic modulations[40–42], optimal pulse protocols[43], linear quadratic-Gaussian (LQG) methods[38], and coherent feedback methods using auxiliary optical components[44,45]. We note that there were two previous works[46,47] on model-free RL for controlling and stabilizing a quantum system with an inverted harmonic potential and a double-well nonlinear potential, respectively, to a target state using weak-current measurements (WCMs) and partial state observation. However, these two works did not address entanglement control, while our work is developing a model-free deep-RL method

[a] Electronic mail: Ying-Cheng.Lai@asu.edu

to realize non-Gaussian entanglement engineering using only photon number counting from WCMs. (Backgrounds about WCM, deep RL, and quantum control are presented in Appendix A.) To our knowledge, prior to our work, model-free deep RL feedback control to create and stabilize the entanglement with WCM observations had not been available.

The particular aspects of our work that go beyond the existing works are briefly described, as follows. In our work, in the linear (nonlinear) interaction regime, the observation is the WCM photocurrent (the expectation value of the photon number). We note a previous work[29] that employed a proximal policy optimization (PPO)[48] RL agent, to generate different Fock states and the superposition of a single cavity mode based on observing the density matrix and a fidelity-based reward function. In contrast, the observable in our work is the photocurrent that is more experimentally accessible[49]. For quantum measurement, we use WCM in real-time feedback control, taking into consideration the resulting quantum stochastic process[47,50,51], and identify a numerical relationship between the entanglement and photocurrent. In both the linear and nonlinear regimes, we focus on non-Gaussian state control because, according to the nonlinear quantum master equation resulting from WCM, the time evolving quantum states are intrinsically non-Gaussian. Our deep RL control scheme is model-free[52], where policies or value functions are directly learned from the interactions with the quantum environment without any explicit model of this environment. This should be contrasted to the model-based deep RL methods[53], where a pre-built model of the environment for policy decision-making is needed. We demonstrate that, under the actions of the well-trained PPO or recurrent PPO RL agent, entanglement between the quantum optical and mechanical modes can be created and maintained about the target entanglement.

Our main results are as follows. First, under the strong laser approximation, the interaction resulting from the radiation pressure between the cavity and the mechanical oscillator modes can be linearized and described by the beam-splitter Hamiltonian. During the training phase, the PPO agent interacts with parallel quantum environments and collects the subsequent data by episodic learning, with the observation being the WCM photocurrent. The deep-RL method can extract useful information from the measurement photocurrent, which is encoded in the Wiener process, and achieve the target entanglement engineering in a model-free manner for the quantum system that is dissipative due to coupling to the vacuum bath and is driven by a laser. In the testing phase, with the agent interacting and observing a single quantum environment, we demonstrate that the entanglement-engineering performance of our deep-RL method with WCM observation greatly exceeds that of both state-based Bayesian methods[47,54] and random control. Second, when the driving laser field is not strong, the quantum optomechanical interaction is nonlinear[55,56]. In this case, we articulate two training phases for nonlinear entanglement engineering. The first phase is utilized to infer the entanglement by the model-free deep RL, dubbed as the target-generating phase, where the observation of the PPO agent [with multilayer perceptions (MLPs)] is the logarithmic negativity and the reward function is constructed to limit the high-level excitation and facilitate entanglement learning. (Direct experimental measurement of the logarithmic negativity is currently not available.) The time series of the expected photon number in the regime from converged training episodes is selected as the target for the next phase. The second phase is then the target-utilization phase, where the recurrent PPO (with long short-term memory (LSTM)[57] added after MLPs) observes the expected photon number and obtains the reward only based on the target expected photon number obtained from the last phase. In this framework, the recurrent PPO controls the quantum state in the low-energy regime with the desired entanglement created and stabilized.

II. RESULTS

A. Experimental proposal for entanglement engineering

Our goal is achieving entanglement engineering between the optical cavity and mechanical oscillator modes using deep RL. Based on the current experimental progress, we articulate an experimental proposal to achieve this goal, as shown in Fig. 1. Consider a Fabry-Perot cavity that consists of a single-mode cavity and a movable end mirror. The optical cavity has the frequency ω_c and the optical field exerts a radiation pressure on the mirror. The cavity mode is driven externally by a coherent laser field with frequency ω_L. The mirror's quantized center-of-mass motion is described by a harmonic oscillator of frequency ω_m. In the rotating frame of the laser, the Hamiltonian describing the coupling between the optical cavity and mechanical oscillator modes is given by[55,56]

$$\tilde{H}_{nl} = -\hbar\Delta\hat{a}^\dagger\hat{a} + \hbar\omega_m\hat{b}^\dagger\hat{b} + \hbar g_0(\hat{b}^\dagger + \hat{b})\hat{a}^\dagger\hat{a} + \hbar\alpha_L(\hat{a}^\dagger + \hat{a}),$$

where $\hat{a}$ and $\hat{b}$ are the annihilation operators of the cavity and mechanical mode, respectively, $\hat{a}^\dagger$ and $\hat{b}^\dagger$ are the corresponding creation operators. The frequency detuning of the cavity is $\Delta \equiv \omega_L - \omega_c$. The nonlinear coupling g_0 arises from the radiation pressure force between the light and the movable mirror (details given in Appendix B), and α_L is the real amplitude of the driven electromagnetic field. We set $g_0 > \kappa$ so that the single-photon optomechanical coupling rate g_0 exceeds the coupling strength κ between the cavity and the vacuum bath. This condition guarantees observable nonlinear quantum effects[61]. Under the strong laser approximation: $|\bar{\alpha}_c| \gg 1$, where $|\bar{\alpha}_c|$ is the amplitude of the light field inside the cavity induced by the strong laser, we have $\hat{a} \approx \bar{\alpha}_c + \delta\hat{a}$

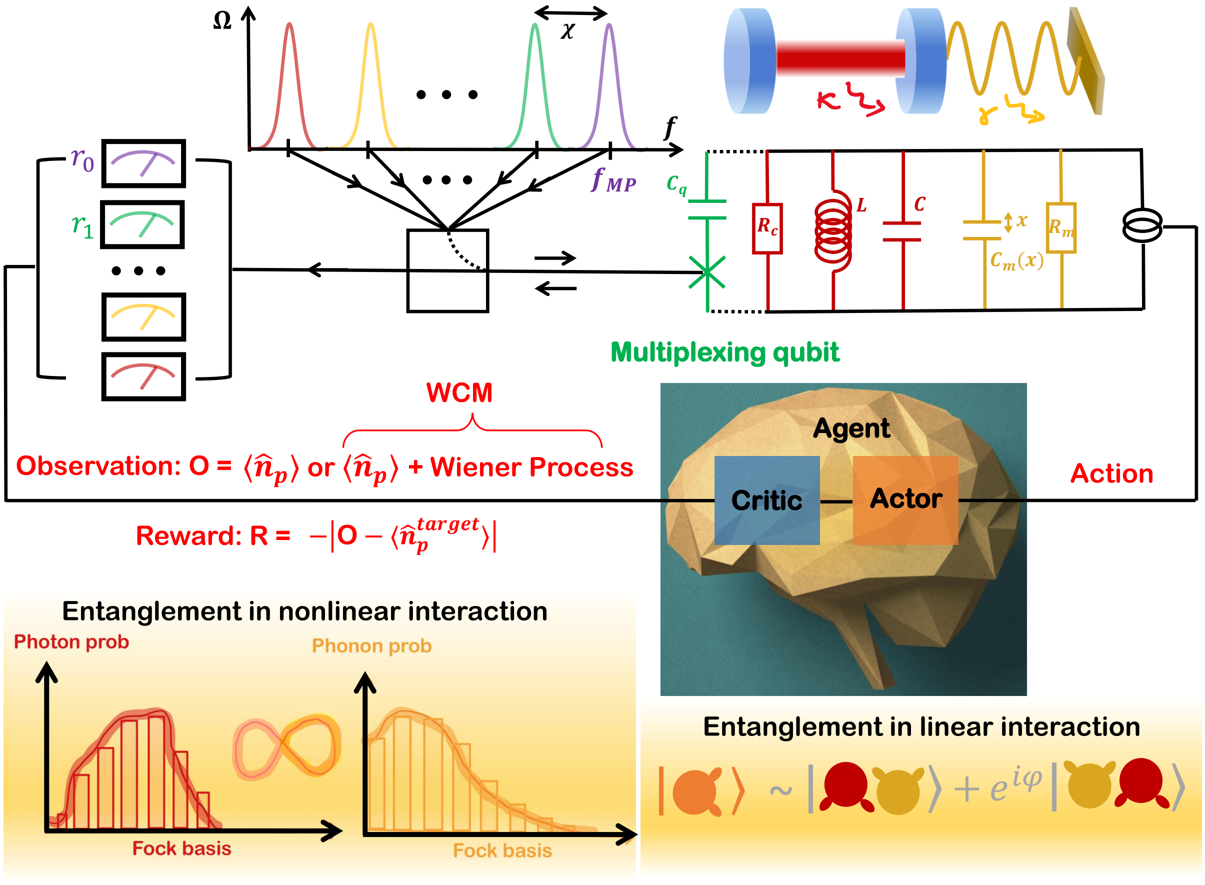
Ω
χ
f
f_{MP}
r_0
r_1
C_q
R_c
L
C
x
$C_m(x)$
R_m
κ
γ
Multiplexing qubit
WCM
Observation: O = $\langle \hat{n}_p \rangle$ or $\langle \hat{n}_p \rangle$ + Wiener Process
Reward: R = $-|O - \langle \hat{n}_p^{target} \rangle|$
Agent
Critic
Actor
Action
Entanglement in nonlinear interaction
Photon prob
Phonon prob
Fock basis
Fock basis
Entanglement in linear interaction
$| \; \rangle \sim | \; \rangle + e^{i\varphi} | \; \rangle$

FIG. 1. Experimental proposal of measurement-based feedback control of deep RL to create and stabilize entanglement in an open quantum optomechanical system dissipatively coupled to the vacuum bath. Quantum optomechanics was experimentally realized in a microwave electromechanical system[58–60], where the multiplexing qubit was used to weakly couple to the microwave resonator for extracting the photon number statistics through weak measurements[49]. The RL agent acts in one or multiple parallel quantum optomechanical environments according to the parameterized policy and collects data in one episode consisting of T time steps: observations O_t, reward R_t, and actions, after which the quantum optomechanical environment is reset. After one or several episodes, the policy is updated using minibatch data to maximize the accumulated reward. The aim is to achieve the desired entanglement $E_N \sim \log 2 \sim 0.7$ (in the natural logarithmic base) between the cavity-optical and mechanical modes. Entanglement engineering of this type can be achieved in both the linear and nonlinear interaction regimes. In the linear case, the task is similar to that of achieving an entangled Bell state of the beam-splitter Hamiltonian or "swap" Hamiltonian. In the nonlinear regime, the entangled states from entanglement engineering can be complicated. Illustrated are the resulting photon and phonon number distributions of the entangled states.

with $\delta\hat{a}$ denoting the excitation or the shifted oscillator on top of the large coherent state with the amplitude $\bar{\alpha}_c$. The resulting linearized beam-splitter or "swap" Hamiltonian [55,62] is

$$\tilde{H}_{bs} \approx \hbar\omega_m \delta\hat{a}^\dagger\delta\hat{a} + \hbar\omega_m \hat{b}^\dagger\hat{b} + \hbar G(\delta\hat{a}^\dagger\hat{b} + \hat{b}^\dagger\delta\hat{a}),$$

which is obtained in the red-detuned regime $\Delta = -\omega_m$, where the coefficient $G \equiv g_0\bar{\alpha}_c$ can be tuned by the amplitude of the incoming laser (a time-dependent modulation)[63]. The interaction term describes the state transfer between photons and phonons in the strong coupling regime for $G > \kappa$, with κ (γ) being the decay rate of the cavity (mechanical) mode to the vacuum bath at zero temperature.

Our control strategy was developed based on considering the current experimental capability. Previous works on the microwave regime of the optomechanical systems[58–60] suggested the feasibility of the experimental implementation of our RL control scheme. In particular, a one-to-one correspondence between the Fabry-Perot cavity and the microwave electromechanical system was demonstrated[58,59,64]. As shown in Fig. 1, the microwave resonator of an LC circuit is equivalent to the Fabry-Perot optical cavity mode with the movable capacity[64] $C_m(x)$ corresponding to the flexible mirror in the optical cavity. The resistors R_c and R_m can be related to the decay rate κ, γ to the vacuum bath[64]. Based on the experimental results, we can compare the typical parameter configurations between the optomechanical and electromechanical systems. The decay rate of the optical cavity mode is $\kappa = 0.01\,\omega_m$ in the linear regime and $\kappa = 0.1\,\omega_m$ in the nonlinear regime, with the better quality of the mechanical oscillator mode $\gamma = 0.01\,\kappa$. Consequently, we have $\gamma \approx 10^{-3}\,\omega_m \sim 10^{-4}\,\omega_m$. The typical experimental decay rate of the microwave resonator is[58,59,61,65–67] $\kappa \approx 0.01\,\omega_m \sim 0.1\,\omega_m$ with $\gamma \approx 10^{-3}\,\omega_m \sim 10^{-9}\,\omega_m$. In our work, the nonlinear coupling is set to be $g_0 = 0.2\,\omega_m$, whereas the typical coupling in the strong coupling regime in a previous work[58] was about $g_0 = 0.1\,\omega_m$. The strength of the laser in our work is $G \in [-5, 5]\,\omega_m$ for the linear system in the red-detuned regime $\Delta = -\omega_m$ and $\Delta, \alpha_L \in [-5, 5]\,\omega_m$ for the nonlinear system. In the microwave version, this range can be adjusted by the pump's strength[58,59,61,65–67].

In the microwave regime, it was demonstrated that the photon-number statistics of a microwave cavity mode can be detected using multiplexed photon number measurements[49,68,69]. By this method, the multiplexing qubit encodes multiple bits about the photon number distribution of a microwave resonator through dispersive interaction. A frequency comb drive, distributed at $f_{\mathrm{MP}} - k\chi$, reads out all the information about the photon number distribution at once[49], where k denotes the number of photons and χ represents the dispersive qubit-resonator coupling, as shown in Fig. 1. The reduction in the reflection amplitude, $1 - r_k$ with $k = 0, 1, ...$, of the frequency comb, is proportional to the photon-number distribution of the microwave cavity mode over the Fock bases, as detected by the weak measurement[29,49]. In our circuit design of experimental proposal, we add a capacitor C_q to realize the weak coupling to the original electromechanical system. The coupling capacitance is small enough to be neglected in the total Hamiltonian, but it still allows the multiplexing qubit, denoted by the green cross in Fig. 1, to encode the photon number distribution of the microwave resonator through dispersive interaction.

Under weak measurement[29,49], the sequence of the reduced reflection amplitude $1-r_k$ is collected by the PPO agent, which is proportional to the occupied photon number probability. Consequently, the expected photon number is calculated as

$$\langle \hat{n}_p \rangle = \sum_n n \langle \sqrt{\eta} \hat{P}_n \rangle / \sqrt{\eta} = \sum_n n \langle \hat{P}_n \rangle$$

and the WCM photocurrent is

$$\sqrt{\eta}\,\mathcal{I}(t) = \sum_n n \left[\langle \sqrt{\eta} \hat{P}_n \rangle + \frac{dW_n(t)}{\sqrt{4\eta}dt} \right] \qquad (1)$$

with the measurement rate η, where $\hat{P}_n = |n\rangle\langle n|$ is the measurement projector on the Fock state $|n\rangle$, and $dW(t)$ is the Wiener increment with zero mean and variance $dt = 0.01\,\omega_m^{-1}$ (the time step size in our calculations). In the linear quantum optomechanical regime, the Fock space for each mode is limited to $n = 0, 1$. The action is the amplitude modulation of the laser, which is in the range $G \in [-5, 5]\,\omega_m$. In the nonlinear regime, the Fock dimension is $n = 0, 1, \ldots, 9$. The time-dependent control signal consists of the detuning Δ and the amplitude α_L of the driven laser within the fixed range $\Delta, \alpha_L \in [-5, 5]\,\omega_m$.

The open dissipative quantum optomechanics under the WCM obey the stochastic master equation (SME) (see Sec. IV and Appendix C). The number n_{traj} of trajectories simulated from SME can be selected according to the following considerations. If the observable is some expected physical quantity, using one trajectory is sufficient to extract the information about the quantum state: $n_{\mathrm{traj}} = 1$. Experimentally, WCMs are performed, encoding the Wiener process in the observation and resulting in a large variance from the expectation value. To reduce the variance, more quantum trajectories should be used. To make computations feasible, we use five trajectories: $n_{\mathrm{traj}} = 5$.

In the online training phase, for each episode with time steps, e.g., $T = 500$, the PPO agent - the combination of the actor and critic network, collects the sequence of the observations $O(t) = \langle \hat{n}_p \rangle(t)$ or $\mathcal{I}(t)$, the reward value $R(t) = -|O(t) - \langle \hat{n}_p^{\mathrm{target}} \rangle(t)|$, and the resulting actions generated by its policy. After one or several episodes, the policy of the PPO agent is updated using minibatch data to maximize the accumulated reward. The RL agent is designed to interact with a single or multiple parallel quantum environments to make the time evolving observation $O(t)$ align with the target one $\langle \hat{n}_p^{\mathrm{target}} \rangle(t)$. In the online testing phase, the policy of the well-trained agent will not update and only interact with a single quantum environment to give the optimal control protocol to the corresponding observation. To realize entanglement engineering, i.e., achieving the desired entanglement between the cavity-optical and mechanical modes, finding the relation between the experimental observables and entanglement quantities is an unavoidable challenge. In our work, the model-free PPO agent finds the numerical relationship between them and realizes the entanglement engineering in both the linear and nonlinear regimes of quantum optomechanics, as shown in Fig. 1.

A general quantity to measure the entanglement between arbitrary quantum bipartite systems for any mixed states is the logarithmic negativity[70–72], without the influence of the vacuum bath[73]. In contrast, the conventional pure-state entanglement measures, such as the von Neumann and Rényi entropy, capture both quantum and classical correlations. Since the goal of our study is harnessing the entanglement between the cavity and oscillator modes, we focus on the logarithmic negativity: $E_N(\rho) \equiv \log_2 ||\rho^{T_i}||_1$, where $||X||_1 = \mathrm{Tr}\sqrt{X^\dagger X}$ is the trace norm of the partial transpose ρ^{T_i} with respect to the two subsystems $i = 0$ (quantum-optical cavity mode) and 1 (mechanical oscillator mode). The logarithmic negativity measures the degree to which ρ^{T_i} fails to be positive, i.e., the extent of inseparability or entanglement, and it is the upper bound of the distillable entanglement[70,71]. The logarithmic negativity is the full entanglement monotone[71], which satisfies the following criteria[72,74]: (1) E_N is a non-negative functional, (2) E_N

vanishes if the state ρ is separable, and (3) E_N does not increase on average under Gaussian local operations and classical communication[75,76] or positive partial transpose preserving operations[77]. Since E_N quantifies the quantum correlation between the bipartite systems in spite of the coupling to the vacuum bath, the value of E_N calculated from the cavity mode is equal to that of the oscillator mode: $E_N^0 = E_N^1 = E_N$, which can be verified numerically.

To characterize the quantum-entanglement control performance, we use the following three quantities: $\langle E_N \rangle$, $\widetilde{E}_N$, and $\widetilde{R}$ in open quantum optomechanical systems with either linear or nonlinear interaction between the quantum cavity and oscillator modes. In particular, $\langle E_N \rangle$ is the logarithmic negativity averaged over ten successive episodes with a single environment, $\widetilde{E}_N$ is the corresponding average over one episode with T time steps in a single quantum environment, and $\widetilde{R}$ denotes the ensemble-averaged value of the reward R over a small number of multiple parallel quantum environments for each episode. In our computations, all the control actions G, α_L, or detuning Δ, the nonlinear coupling g_0, and the dissipation coefficients (κ, γ) are in units of ω_m. The time unit is ω_m^{-1}.

B. RL in linear quantum optomechanics

A quantum optomechanical system with linear photon-phonon interactions is governed by the beam-splitter Hamiltonian. In an optical experimental platform, a 50:50 beam splitter with the transformation angle $\pi/4$ can create an entangled Bell state between the two input optical modes[78–80]. Similarly, in a quantum optomechanical system, Bell states between photon and phonon modes can be realized by controlling the beam-splitter Hamiltonian. As a result, the maximally attainable value of the logarithmic negativity is $E_N \sim \log 2 \sim 0.7$ (in the natural logarithmic base), corresponding to the maximally entangled Bell state, as shown in Fig. 1. This "best" entangled state can be realized by the model-free PPO agent, regardless of whether the observation is the expectation or WCM photocurrent. To see this, we note that, in the beam-splitter model, the initial quantum state is set as a pure state[81,82]: $|\psi\rangle = |10\rangle$, where the photon is in the first excited mode and the phonon is in the vacuum mode. The partial observable of the quantum state for the PPO agent is set as the expectation of the photon number $\langle \hat{n}_p \rangle(t) = \langle \hat{P}_1 \rangle(t)$ or the WCM photocurrent $\sqrt{\eta}\mathcal{I}(t) = \langle \sqrt{\eta}\hat{P}_1 \rangle(t) + \frac{dW(t)}{\sqrt{4\eta}\, dt}$.

TABLE I. Results of entanglement engineering from deep RL-based, Bayesian, and random control. The observations are the expectation of the photon number $\langle \hat{n}_p \rangle$ and the WCM photocurent $\mathcal{I}(t)$ at the measurement rate $\eta = 1$. The Bayesian hyperparameter is $\lambda_{opt} = 10$ for the $\langle \hat{n}_p \rangle$ task and $\lambda_{opt} = 2$ for the $\mathcal{I}(t)$ task. Displayed are the results of the average logarithmic negativity $\langle E_N \rangle / \log 2$ with the standard deviation. For training and testing phases, $\langle E_N \rangle / \log 2$ is averaged over ten end-training or testing episodes, each having $T = 500$ time steps. Each observation is obtained by averaging over $n_{\rm traj} = 1$ for $\langle \hat{n}_p \rangle$ and $n_{\rm traj} = 5$ for $\mathcal{I}(t)$ through simulating the SME, and $n_{\rm traj}$ denotes the number of independent trajectories from SME simulations.

Controller	Condition	$\langle \hat{n}_p \rangle$ $n_{\rm traj} = 1$	$\mathcal{I}(t)$ $n_{\rm traj} = 5$
Deep RL (%)	Training	83.81 ± 1.85	$\mathbf{64.81 \pm 1.47}$
	Testing	84.95 ± 1.99	$\mathbf{65.01 \pm 1.76}$
Bayesian (%)	$\lambda = 1$	56.89 ± 6.40	35.48 ± 5.34
	λ_{opt}	$\mathbf{93.21 \pm 0.89}$	49.24 ± 0.44
Random (%)		38.15 ± 9.46	33.46 ± 4.27

Experimentally, directly measuring the entanglement, e.g., in terms of logarithmic negativity, for arbitrary entangled states is generally not viable. Identifying an experimentally feasible quantity to characterize the entanglement in arbitrary quantum systems remains challenging. We focus on the relationship between logarithmic negativity and the expected photon number, based on recent experiments on multiplexed photon number measurement[29,49,68,69,83–85]. To proceed, we note that the beam-splitter Hamiltonian is limited to a four-level basis, due to the following reasons: (1) only one energy level in the cavity mode of the initial state has been excited from the vacuum state, i.e., $|\psi\rangle = |10\rangle$, (2) the linear interaction serves only to transfer the quantum states between the cavity and mechanical mode (i.e., no quantum excitation), and (3) the system couples to the vacuum bath only at absolute zero temperature (i.e., without any thermal excitation), thereby blocking any interactions between higher-level quantum states. In this case, the maximum logarithmic negativity $E_N \sim \log 2 \sim 0.7$ implies that the attained quantum state is the following Bell state

$$|\Phi^{\varphi}\rangle = \frac{1}{\sqrt{2}} \left[|10\rangle + e^{i\varphi} |01\rangle \right],$$

with the associated expected photon number $\langle \hat{n}_p^{\rm target} \rangle = \langle \hat{P}_1 \rangle = 0.5$. Consequently, the reward function can be set as $R_t \equiv -|O_t - 0.5|$, regardless of whether the observation O_t is $\langle \hat{n}_p \rangle(t)$ or $\mathcal{I}(t)$. Because of the relatively small target value of the expected photon number: $\langle \hat{P}_1 \rangle = 0.5$, the variance $\mathcal{I}(t)$ in the WCM photocurrent can be reduced by a Gaussian filter[86] with the weak measurement rate $\eta \leq 1$. The Gaussian kernel parameters of the filter such as the filter interval and the variance can be numerically chosen to reduce the standard deviation of the measurement photocurrent into a certain range, e.g., about ten times larger than the mean value (details in Subsec. IV B). The PPO agent applies an updated stochastic policy to the quantum optomechanical environment to maximize the accumulated reward, where the action $G(t)$ is proportional to the amplitude of the cavity mode: $G(t) = g_0 \bar{\alpha}_c(t)$. The action can be controlled by an incident laser[63] and is continuous in a certain range, e.g., $G \in [-5, 5]\omega_m$. The decay rate of cavity and mechanical modes $\kappa = 0.01\,\omega_m$, $\gamma = 0.01\,\kappa$, respectively, because the quality of the mechanical oscillator mode is generally better than that of the optical cavity or microwave resonator mode[58,59,61,65–67].

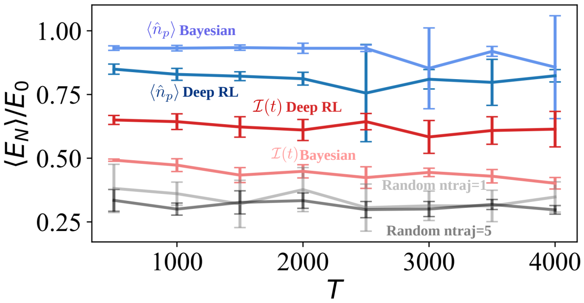

FIG. 2. Performance in terms of $\langle E_N\rangle/\log 2$ over a long time interval, compared for deep RL-based, Bayesian, and random control methods with respect to two observable options: the expected value $\langle\hat{n}_p\rangle$ and the WCM photocurrent $\mathcal{I}(t)$. The deep RL controller is trained with $T = 500$ time steps. For all three control methods, displayed are results from the testing phase for the following set of time steps: $T = [500, 1000, 1500, 2000, 2500, 3000, 3500, 4000]$ at the measurement rate $\eta = 1$. The conventions, which apply to this and all subsequent figures, are as follows. If the vertical axis is labeled as $\langle E_N\rangle/E_0$, it represents the normalized logarithmic negativity, with $E_0 = \log 2 \sim 0.7$ (in the natural logarithmic base) as the target entanglement value. Otherwise, when the vertical axis is labeled as $\langle E_N\rangle$, $\widetilde{E}_N$, or $E_N(t)$, it represents the original value of the logarithmic negativity.

Our deep RL, a model-free learning method, is implemented in the measurement-based feedback control framework for entanglement engineering in open quantum optomechanics. Details about the PPO algorithm applied in the linear quantum optomechanics are presented in Appendix D. To appreciate its performance, we employ two benchmark methods for comparison: Bayesian[47,54] and random control. Bayesian control[47,54] is a state-based feedback control of the stochastic process as governed by the SME. In our case, the control law is given by $G(t) = -\lambda|\langle\hat{n}_p\rangle(t) - 0.5|\omega_m$ with $\langle\hat{n}_p\rangle(t)$ being the observation, where the hyperparameter λ can be numerically optimized based on the performance. If the observation is $\mathcal{I}(t)$, the control flow will be in the form $G(t) = -\lambda|\mathcal{I}(t) - 0.5|\omega_m$, in which the Wiener process blocks the performance to some degree. In Bayesian control, the smaller the variance in the measured photocurrent, the better the performance. For the random control method, the flow is generated by a uniform distribution in the action range $G \in [-5, 5]\,\omega_m$. Note that the actions G of random control and deep RL are in the same range while the one of Bayesian control is determined by the hyperparameter λ and the state-based observation value or the WCM photocurrent. To make a fair comparison, λ_{opt} is optimized within the action range $G \in [-5, 5]\omega_m$. Specifically, the optimized hyperparameter λ_{opt} corresponds to the best performance of Bayesian control in the set $\lambda \in \{1, 2, \ldots, \lambda_{max}\}$, where λ_{max} is the maximum integer of λ to guarantee the action range $G \in [-5, 5]\,\omega_m$.

Table I displays the values of the averaged logarithmic negativity $\langle E_N\rangle/\log 2$ from the deep RL, Bayesian, and random control methods. From the SME simulations, when the observation is the expectation of the photon number, the Bayesian control with the optimized hyperparameter outperforms the deep RL method. However, when the observation is the WCM photocurrent, the deep RL control outperforms the Bayesian method. This is promising as the WCM photocurrent is directly experimentally accessible while the expected photon number is not. Regardless of the observation, random control is generally ineffective. The results by deep RL control from the observation of WCM photocurrent tend to reduce the performance by about 20% compared to that based on the expected photon number. For Bayesian control, the reduction is about 40%. Moreover, Fig. 2 compares the long-time entanglement engineering for three control methods. Especially, for deep RL control, the PPO agent is trained with $T = 500$ time steps but tested with a longer time horizon, e.g., $T = 4000$ steps, including the unexplored regime by the PPO agent. It is worth noting from Fig. 2 that the performance of deep RL with the observation of WCM photocurrent exhibits a more stable and smaller variance compared to the case where the observation is the expected photon number, especially after $T = 2000$. Overall, with the experimentally feasible observation of WCM, the deep RL controller stands out as the choice of entanglement control for quantum optomechanical systems.

We characterize the performance of our deep-RL-based control method in terms of the dissipation rate, measurement rate, and the randomness effect for the initial state. For the measurement rate $\eta = 1$, the PPO agent is sequence-wise trained with the WCM photocurrent. Figures 3(a) and 3(c) show the average logarithmic negativity $\widetilde{E}_N$ and the mean reward $\widetilde{R}$, respectively, versus the episode during the training phase, in which $\widetilde{E}_N$ and $\widetilde{R}$ are averaged over one and five parallel quantum environments, respectively. Both quantities ultimately converge due to the properly designed reward function $R(t) = -|\mathcal{I}(t) - 0.5|$. Note that the variance of $\widetilde{E}_N$ is suppressed with the episodes, implying the mixture-robust nature of entanglement in the quantum optomechanical system. The testing phase is longer ($T = 4000$ time steps) than the training phase ($T = 500$ time steps) and the corresponding performance measures are shown in Figs. 3(b) and 3(d). In addition to the variance in the learning of the deep RL agent with the stochastic policy, the Gaussian Wiener process in the WCM photocurrent and the stochastic collapse process stipulated by the SME also contribute to the variances of the performance measures. However, the deep RL control still manages to maintain the solid traces of the testing $\mathcal{I}(t)$ around the target value $\langle\hat{n}_p^{\text{target}}\rangle = 0.5$ in Fig. 3(d) and the resulting entanglement quantity $E_N(t)$ is displayed

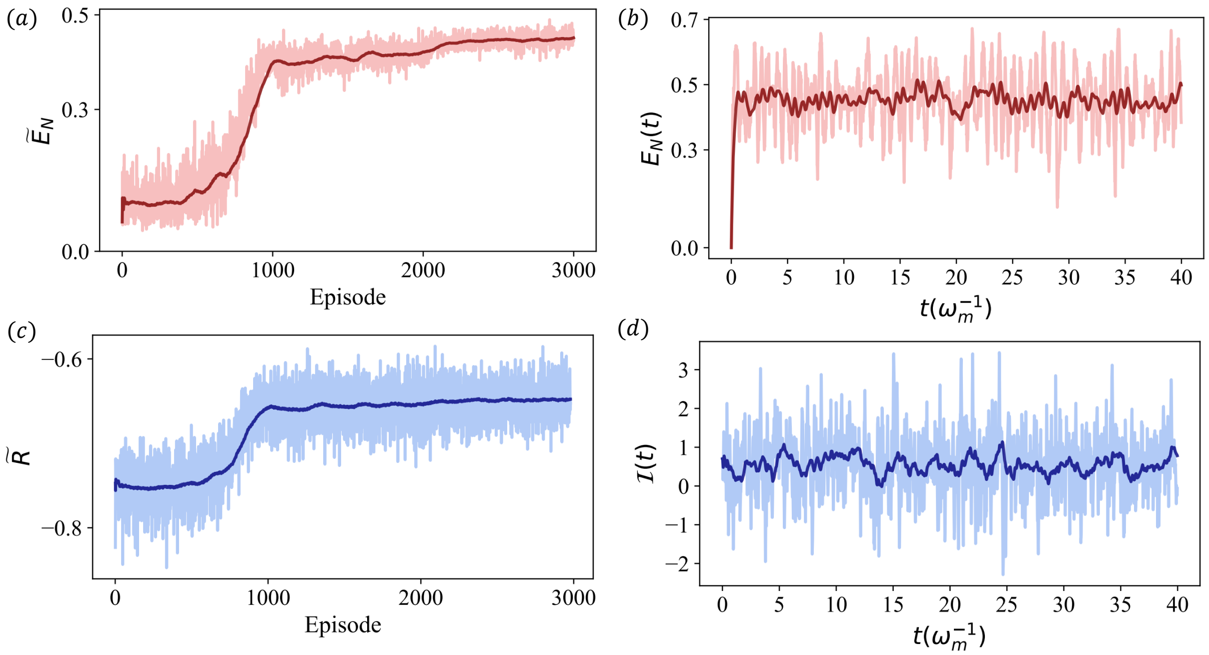

FIG. 3. Performance of deep-RL agent in the online training and testing phase. The characterizing quantities are the logarithmic negativity E_N and the reward function R with the measurement rate $\eta = 1$. (a,c) Performance measures in the online training phase, where the mean $\widetilde{E}_N$ is over one episode with $T = 500$ time steps on the fifth quantum environment (only one environment) and the mean reward $\widetilde{R}$ is obtained from $\mathbb{N} = 5$ parallel quantum environments. (b,d) Performance measures during the testing phase, where the logarithmic negativity $E_N(t)$ and WCM photocurrent $\mathcal{I}(t)$ are obtained with $T = 4000$ time steps. The solid traces represent the moving-window average over 100 episodes for (a,c) and 100 time steps for (b,d).

in Fig. 3(b).

Since the quantum optomechanical system is coupled to the vacuum bath, the coupling strength or disturbance between the classical and quantum environments will affect the control performance, as exemplified in Fig. 4(a). Previous experiments[58,59,61,64] demonstrated that the quality of the mechanical oscillator is generally better than that of the optical cavity or microwave resonator, i.e., $\gamma < \kappa$, so we set the decay rate of the oscillator at two orders of magnitude smaller than that of the cavity[56]: $\gamma = 0.01\,\kappa$. Figure 4(a) shows, for both the expectation and the measurement flow observations, the performances of the training and testing processes, which are consistent with each other in the sense that their mean values decrease and the variances increase with the decay rate. The origin of the performance fluctuations is the classical dissipation to the vacuum bath, rendering the system less controllable by laser.

The uncertainty in the classical information extracted from the quantum system depends on the discrete-time step size dt and the measurement rate η, which directly determines the degree of the quantum-state stochastic collapse and quantum decoherence from the WCM term in the SME. If the expectation of the photon number is the observation, the stronger the measurement rate (proportional to the measurement strength), the poorer the performance of deep-RL control as characterized by a decrease in the mean values and an increase in the uncertainties of E_N, as shown in Fig. 4(b), which originate from the intrinsic random process in the SME induced by the measurement process. However, if the observation is the WCM photocurrent, the weaker measurement rate will introduce larger variances in the observation signal and reduce the stochasticity of the process due to the incomplete/partial extracted information as described by the SME. In our case, the target mean value, $\langle \hat{n}_p^{\text{target}} \rangle = 0.5$, is on the order of 10^{-1}, rendering necessary introducing a Gaussian filter to reduce the uncertainty. The resulting performance of deep-RL control is approximately the same for $\eta \in [0.05, 1]$, as shown in Fig. 4(b).

Experimentally, mixed quantum states are more realizable than pure states due to the quantum decoherence with the classical environment, e.g., the vacuum bath. To address this issue, and referring to the previous work[87], we assume that the initial state is a mixed state in the form of $\rho = (1-p)|10\rangle\langle 10| + p|01\rangle\langle 01|$, where the parameter p is fixed or a random variable $p \in [0, 1]$ because of the coupling to the classical environment. The beam-splitter Hamiltonian stipulates that the photon and phonon modes are symmetric to each other, allowing p to be rescaled to the interval $p \in [0, 0.5]$. Figure 5(a) shows the performance with respect to the initial mixed quantum state with the same parameter p for

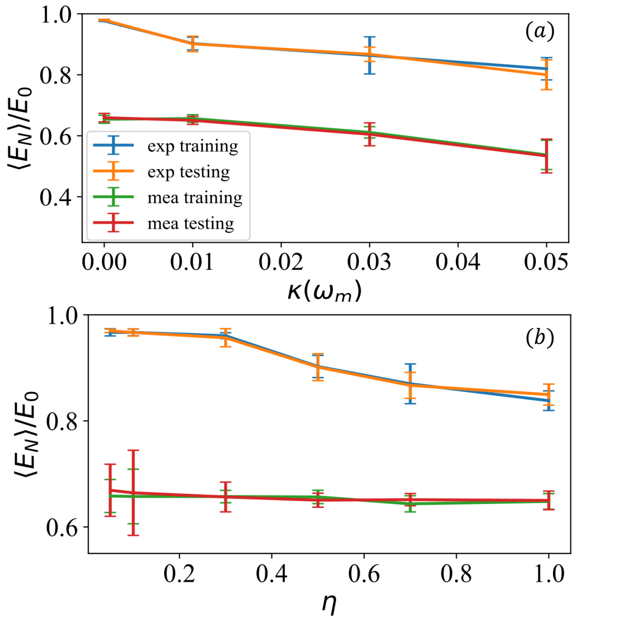

FIG. 4. Effects of decay and measurement rates on the control performance. Shown are the values of the average logarithmic negativity for (a) decay rates $\kappa = [0, 0.01, 0.03, 0.05]\,\omega_m$ with $\eta = 0.5$ and $\gamma = 0.01\,\kappa$, and (b) measurement rates $\eta = [0.05, 0.1, 0.3, 0.5, 0.7, 1]$ with $\kappa = 0.01\,\omega_m$ and $\gamma = 0.01\,\kappa$. The error bars represent the standard deviation of the data points. The average operation is over ten end-training or testing episodes. The training and testing time steps are the same: $T = 500$.

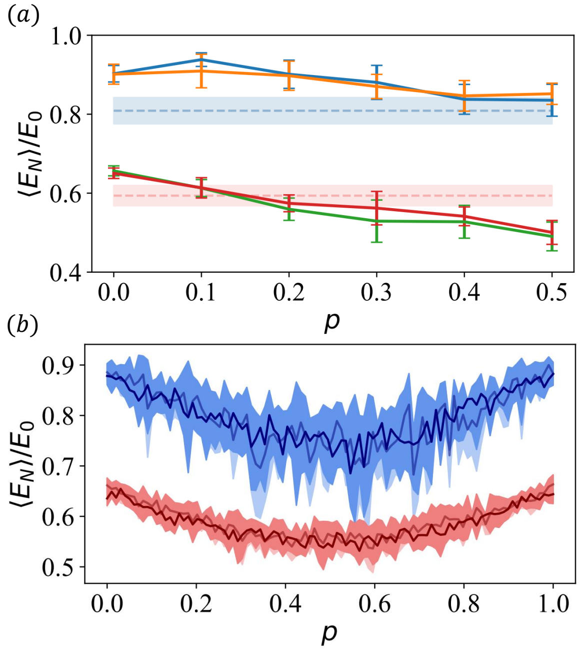

FIG. 5. Robustness of deep-RL method trained with pure or mixed states. (a) In the training and testing phase, performance of $\langle E_N \rangle / E_0$ for different initial mixed states (solid traces): $\rho = (1-p)|10\rangle\langle 10| + p|01\rangle\langle 01|$ with $p = [0, 0.1, 0.2, 0.3, 0.4, 0.5]$. The dashed traces indicate the performances trained with random initial mixed states with the random variable $p \in [0, 0.5]$. (b) Testing performance of two kinds of trained agents with $p \in [0, 1]$: one trained with the pure initial state $|\psi\rangle = |10\rangle$ and another with random initial mixed states, which are distinguished by the color depth of the curve and the error bars. The blue and red curves denote the performances with the observation $\langle \hat{n}_p \rangle$ and $\mathcal{I}(t)$, respectively, with error bars. The measurement rate is $\eta = 0.5$, and the training and testing time steps are $T = 500$.

each training and testing episode (solid traces), where the complete mixed case with $p = 0.5$ leads to the worst performance but still possesses entanglement to a significant extent. The reason lies in the inherent property of the beam-splitter Hamiltonian, which can create the maximum entangled states: $[|10\rangle + e^{i\varphi}|01\rangle]/\sqrt{2}$, with respect to the part of the initial quantum state, such as $|10\rangle$ or $|01\rangle$ through the linear interactions, regardless of whether it acts on a pure or a mixed state. In Fig. 5 (a), the dashed traces display the performance during the training phase with a random initial mixed quantum state, which is generated by the random variable p with the uniform distribution in the range of $p \in [0, 0.5]$. The error bar characterizes the uncertainty over ten end-training episodes.

Figure 5(b) shows the testing performance of two kinds of trained models, one trained by the initial state $|\psi\rangle = |10\rangle$ and another by the random initial mixed-state $\rho = (1-p)|10\rangle\langle 10| + p|01\rangle\langle 01|$ (distinguished by dark and light colors, respectively). Note that the beam-splitter Hamiltonian transforms the initial state $|10\rangle$ or $|01\rangle$ to a Bell state with the corresponding expected photon number: $\langle \hat{n}_p^{\mathrm{target}} \rangle = 0.5$, where the dissipative degree to the vacuum bath is much weaker than the beam-splitter interaction. However, if the initial state is the mixed state, the $|10\rangle\langle 10|$ and $|01\rangle\langle 01|$ components will become independently entangled, resulting in the total quantum state being a mixture of two entangled Bell states. As a result, a nontrivial entanglement value is expected for the initial mixed state governed by the beam-splitter Hamiltonian. With the mixed probability $p = 0.5$, it results in an equal mixture of the Bell states, as shown in Fig. 5. In the testing phase, the two trained models use the same initial state for a fixed value of p. The two models have a comparable performance, suggesting that the deep RL method is robust to the initial randomness in a mixed state. More specifically, during the testing phase, the observation is the expected photon number or WCM photocurrent. The worse performance occurs for $p = 0.5$ and for other values of p, the performance is symmetric about $p = 0.5$ due to symmetric role of the photon and photon modes in the beam splitter Hamiltonian. Note that the model trained with the observation being the measurement photocurrent displays a small difference in the performance measure $[\langle E_N \rangle / E_0$ over the whole probability interval $p \in [0, 1]]$ between the best and worst cases, with less uncertainties than the case where the

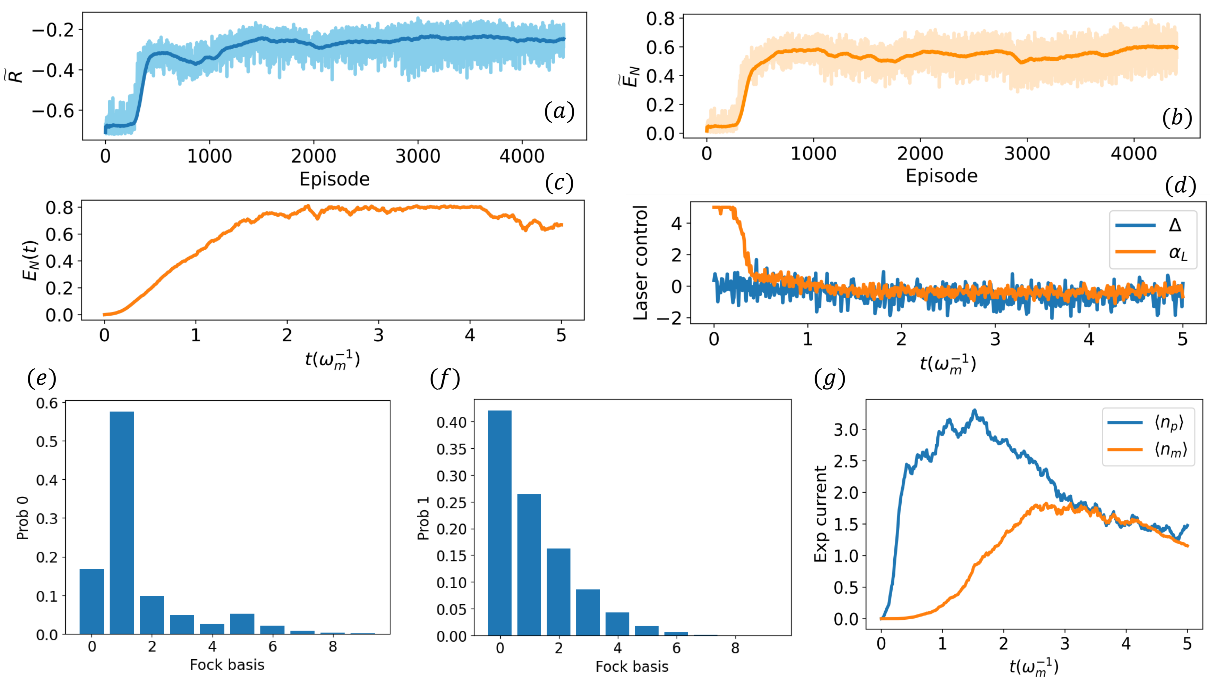

FIG. 6. Generating target for deep-RL based creation and stabilization of entanglement in a nonlinear open quantum optomechanical system. (a,b) Trained quantities $\widetilde{R}$ and $\widetilde{E}_N$ converge to a certain value as the episode number increases, as illustrated by the light-color curves, where the dark blue and orange traces represent the data averaged over 100 previously consecutive episodes. (c,d) Time-dependent series of $E_N(t)$ and the driven laser signals Δ, α_L at a certain episode selected from the training converged regime in (a,b). (e,f) The corresponding photon and phonon statistics on the Fock basis at the end of the time point of the selected training episode in (c,d). (g) The time evolution of the corresponding expected quantities, including the expected numbers $\langle \hat{n}_p \rangle$ and $\langle \hat{n}_m \rangle$ in the Fock basis, where the time series of $\langle \hat{n}_p \rangle$(t) serves as the target to construct the reward function in the next phase, i.e., the experimental version shown in Fig. 7.

observation is the expected photon number. Taken together, our deep-RL model trained by the weak measurement photocurrent holds a lower mean performance but possesses robustness against mixed quantum states compared with the scenario based on observing the expected value of the photon number, due to the strong capability of RL in learning randomness and executing accurate high-dimensional data-fitting.

C. RL in nonlinear quantum optomechanics

In an open quantum optomechanical system under the strong laser-driven approximation, the radiation pressure on the movable mechanical mirror generates a linear interaction between the optical and mechanical modes. When this approximation does not hold, the interaction between the two modes becomes nonlinear. Entanglement can still be created despite the nonlinear interaction, but control becomes more challenging. In particular, in the standard quantum optomechanical system, the nonlinear coupling term $\hbar g_0 \hat{a}^\dagger \hat{a}(\hat{b}^\dagger + \hat{b})$ can be used to create entanglement, but high-level quantum states can also be excited during the process, making it difficult to stabilize the entanglement within a finite Fock basis. Realistically, the quantum dynamics are governed by the SME due to the WCM, which induces the nonlinear stochastic evolution. The problem then becomes that of creating and stabilizing the entanglement of non-Gaussian states decaying to the vacuum bath. Despite the difficulties, model-free deep-RL can still provide a general approach through some optimal combination of the neural network structure, observable, reward function, and action.

We consider the nonlinear optomechanical system and exploit deep RL to set the control goal of achieving the entanglement near $E_N \sim \log 2$. This nonlinear entangled state shares a similar entanglement value with the maximum entangled Bell state in the corresponding linear system. For entanglement engineering of a nonlinear optomechanical system, a key issue is selecting an effective and experimentally feasible observation quantity. Utilizing a general actor and critic neural network, the deep RL agent can learn the relationship between entanglement and the experimental observables of the optomechanical system in a model-free manner. To achieve control, we articulate a training process consisting of two phases: the target-generating phase and the target-utilization phase, facilitated by deep RL.

The first training step is the target-generating phase, in which numerical SME simulations are used to generate the observation and reward data and the PPO agent

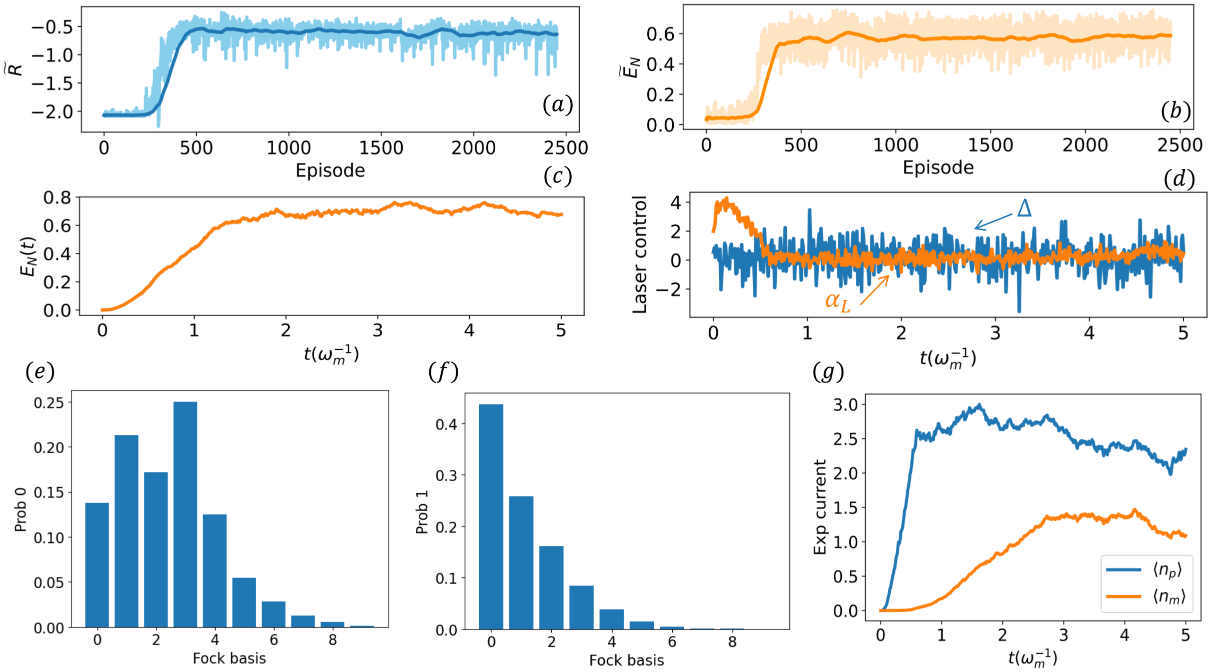

FIG. 7. Entanglement engineering by the recurrent PPO agent. The target generated as described in Fig. 6 is exploited to create entanglement by $E_N \sim \log 2$ from the only partial observation of the expected photon number $\langle \hat{n}_p \rangle(t)$. The reward function is $R(t) = -|\langle \hat{n}_p \rangle(t) - \langle \hat{n}_p^{\rm target} \rangle(t)|$, where the target time series $\langle \hat{n}_p^{\rm target} \rangle(t)$ is from the target-generating process in Fig. 6(g). In this training configuration, while only partial information is extracted from the system, the performance measures in (a-g) display a similar behavior compared with those in Fig. 6. Other aspects of the setting and parameters are the same as in Fig. 6.

interacts with the quantum environment, observes the logarithmic negativity $E_N(t)$ and constructs the reward function combining the expectation number of the photons and phonons: $R(t) = -|E_N(t) - \log 2| - |\langle \hat{n}_p \rangle(t) + \langle \hat{n}_m \rangle(t) - a|/b$ with numerically optimized hyperparameters $a = 1$ and $b = 30$. (Note that direct experimental measurement of the logarithmic negativity is currently not available.) Figure 6 shows the control results, where the excitation of quantum states is limited by the total number $\langle \hat{n}_p \rangle + \langle \hat{n}_m \rangle$. The target time series of the expected photon number is obtained as $\langle \hat{n}_p^{\rm target} \rangle(t)$. The second step is the target-utilization phase, during which the reward function is $R(t) = -|\langle \hat{n}_p \rangle(t) - \langle \hat{n}_p^{\rm target} \rangle(t)|$.

Since it is time-dependent, the recurrent neural network added after the MLPs in the PPO agent displays a strong and stable learning ability, which outperforms the case with only MLPs. The expected photon number $\langle \hat{n}_p \rangle(t)$ is observed by the recurrent PPO agent as $\langle \hat{n}_p \rangle = \sum_n n \langle \hat{P}_n \rangle$, which is experimentally more feasible than the quantity E_N. While the recurrent neural network has some considerable advantages, such as long-term momery[57], it still encounters the challenge of engineering optimization[88] in order to achieve a correct and efficient implementation. In our case, the main challenge is the time cost to optimize the parameters to search for a global minimum or maximum due to the ten stochastic collapse operators, $\hat{P}_n = |n\rangle\langle n|$ with the respective Fock numbers $n = 0, 1, \dots, 9$, in the SME with the measurement rate $\eta = 0.1$, requiring a long simulation time. Our solution is to consider only the $\mathbb{N} = 1$ quantum optomechanical environment, in which the agent collects data and updates the policy every $\mathbb{Z} = 15$ and $\mathbb{Z} = 5$ episodes in two phases (target-generating and target-utilization), respectively, with the time horizon $T = 500$. Note that, using ten stochastic projectors $\hat{P}_n$ can result in a large variance in the WCM photocurrent:

$$\sqrt{\eta}\mathcal{I}(t) = \sum_n n \left[\langle \sqrt{\eta} \hat{P}_n \rangle + \frac{dW_n(t)}{\sqrt{4\eta} dt} \right],$$

where ten independent Wiener processes $dW_n(t)$ are used. In this case, observation of the measured random photocurrent is infeasible. Even if the deep RL agent is trained in two phases with the expected photon number, it can fail during the training process due to the numerical cutoff in the Hilbert space dimension and the strong randomness introduced by the SME. In the nonlinear quantum optomechanical system, the interaction strength is $g_0 = 0.2\,\omega_m$. The PPO agent creates entanglement characterized by $E_N \sim \log 2$ versus time, calculated through the SME with dissipation to the vacuum bath for $\kappa = 0.1\,\omega_m$, and $\gamma = 0.01\,\kappa$. The system is initialized in the vacuum state $|\psi\rangle = |00\rangle$, i.e., the pure state, with 10×10 Fock bases. The time-dependent control signal is the detuning Δ and the amplitude α_L of the driven laser within the fixed range $\Delta, \alpha_L \in [-5, 5]\,\omega_m$.

Representative results are as follows. In the target-generating phase, despite the disturbance of the stochastic process from WCM, the training curves for both the reward $\widetilde{R}$ and the logarithmic negativity $\widetilde{E}_N$ converge with the episode number, as shown in Figs. 6(a,b), indicating that entanglement has been created and stabilized by the well-trained PPO agent, as shown by Fig. 6(c) with the laser control signal displayed in Fig. 6(d). At the end of the time period, the photon and phonon statistics with respect to the Fock basis are shown in Figs. 6(e,f), where the reduced photon state exhibits an oscillating tail that resembles the displaced squeezed state and the reduced phonon state displays the thermal-like state. Figure 6(g) shows the corresponding target pattern $\langle \hat{n}_p^{\mathrm{target}} \rangle(t)$. In the target-utilization phase, the recurrent PPO agent is able to steadily learn to create and stabilize the entanglement, as shown in Fig. 7, where only partial information is extracted from the quantum optomechanical environment. Especially, various entangled states have been created, such as a reduced photon state with the head oscillating on the Fock basis in photon statistics entangled with the thermal-like reduced phonon state, as exemplified in Figs. 7(e,f). Due to the nonlinear and stochastic process in the SME, the entangled states created and controlled are not steady states, rendering infeasible Bayesian control. We thus employ random control as a benchmark, where a uniformly random distribution of actions is taken in a certain range $\Delta, \alpha_L \in [-5, 5]\omega_m$ and the tested values of the measurement rate are $\eta = [0.05, 0.1, 0.3, 0.5, 0.7, 1]$. Figure 8 shows that, as the measurement rate increases, the random control is unable to harness the entanglement while our well-trained recurrent PPO agent can maintain the entanglement percentage at 50% or higher.

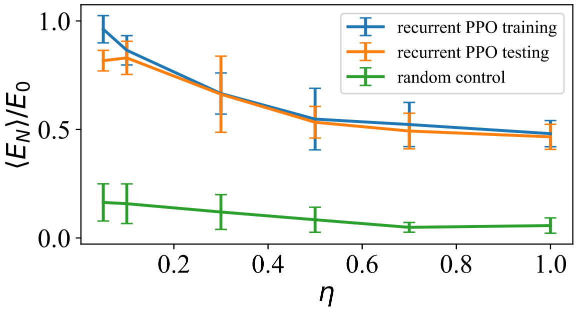

FIG. 8. Target-utilization phase of entanglement engineering of a nonlinear optomechanical systems. Shown are the results of online training and testing of the entanglement measure $\langle E_N \rangle / E_0$ for measurement rates $\eta = [0.05, 0.1, 0.3, 0.5, 0.7, 1]$, in comparison with the benchmark performance of random control. The error bars are the corresponding standard deviation. The results from random control flow are also included for comparison. Other parameters are the same as those in Fig. 6.

D. Physical understanding of entanglement engineering through model-free deep RL

In an experiment, it is usually difficult to directly obtain information about the entanglement. For entanglement engineering of a quantum optomechanical system, one scenario is that the RL agent observes the photon number to steer the laser to create and stabilize entanglement, as illustrated in Fig. 9. Here we provide a physical interpretation of RL control for entanglement engineering in both the linear and the nonlinear interaction regimes. The key physical relationships involved are that between the entanglement and photon number, and that between the photon number and laser driving. We also describe the capability of the RL agent to train the laser driving to modulate the two-mode interaction to reduce quantum decoherence resulting from WCM and the quantum dissipation to the vacuum bath.

1. Linear interaction regime

For the linear quantum optomechanical system, the maximum entanglement corresponds to a Bell state, of which the expected photon number is $\langle \hat{n}_p \rangle = 0.5$. Intrinsically, the beam splitter Hamiltonian is capable of generating Bell states[78–80], a reasonable assumption is that, when the expected photon number reaches the value of 0.5, the maximum entanglement is achieved in a linearly interacting quantum optomechanical system. This assumption provides the base for constructing the reward function $R(t) = -|\langle \hat{n}_p \rangle(t) - 0.5|$, where the deviation in the expected photon number from 0.5 results in a decreasing reward and therefore implies reduced entanglement. As illustrated in Fig. 9, the RL agent is designed to maximize the accumulated reward value, which is equivalent to stabilizing the expected photon number about the value of 0.5 for as long as possible. The testing results shown in Fig. 3 indicate that the maximum entanglement can indeed be created and stabilized by the RL control.

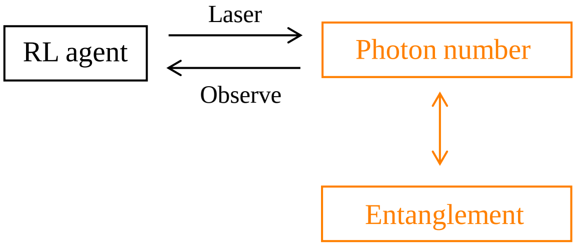

FIG. 9. RL-based entanglement engineering of a quantum optomechanical system.

A central step in RL control is to modulate the laser input based on the measured photon number, which requires the relationship between the laser driving and the photon number. When the frequency of the laser is in the red-detuned regime: $\Delta = \omega_L - \omega_c = -\omega_m$, the quantum state switches between the two modes - the cavity optical and the mechanical oscillator modes, leading to a “swap” Hamiltonian. The coefficient G is proportional to the amplitude of the cavity parameter $\bar{\alpha}_c$ that is determined by the laser. In the linear interaction regime, RL control is achieved via two adjustments of the laser based on the measured photon number: (1) the laser frequency is changed into the red-detuned regime and (2) the laser amplitude is perturbed to modulate the driving strength G to control the two modes of switching, which affects the expected photon number. Note that, during this process, there is no energy gain: there is energy loss due to the dissipation of the cavity and oscillator modes into the vacuum bath with the dissipation rate given by $\gamma = 0.01\,\kappa$. This relation means that the energy loss due to the oscillator mode occurs more slowly than that with the cavity mode. In essence, the working of the laser is to transfer the energy from the oscillator mode to the cavity mode to stabilize the photon number to a desired value. The underlying dissipation process is not beneficial to the entanglement, as it cannot be modulated by the “swap” term in the Hamiltonian, eliminating any possibility of entanglement enhancement in an optomechanical system in the linear interaction regime. It is worth noting that, in the nonlinear interaction regime, entanglement enhancement and dissipation reduction are possible, as will be described below.

2. Nonlinear interaction regime

When the interactions between the optical and mechanical modes are nonlinear, the relationship between entanglement and photon number can be sophisticated and is currently unknown. However, model-free deep RL can be used to find the relation numerically. To achieve this, we first assume that there is a solution of the one-to-one correspondence between E_N and $\langle \hat{n}_p \rangle$ in the time domain. The reward function is constructed according to the target entanglement $E_N = \log 2$ to train the RL agent to maximize the accumulated reward. In the testing phase, the time-dependent series of the expected photon number controlled by the well-trained PPO in Fig. 6(g) is regarded as the target time series of the expected photon number for the next target-utilization phase. Note that the “best” photon number is no longer simply 0.5: it is now time-dependent. In the next training phase, the reset RL agent will learn to control the system with the observation $\langle \hat{n}_p \rangle(t)$ based on the target's expected photon number $\langle \hat{n}_p^{\text{target}} \rangle(t)$. The performance of the new RL agent in the testing phase, as shown in Fig. 7, validates our initial assumption about the existence of the one-to-one correspondence between E_N and $\langle \hat{n}_p \rangle$, even though it is time-dependent.

In the nonlinear interaction regime, the physical picture of how the laser leverages the radiation-pressure interaction to create and stabilize the photon number and even the entanglement is not straightforward. However, physical insights can be gained by examining the strong laser limit. When the amplitude of the laser is strong: $\bar{\alpha}_c \gg 1$, in the blue-detuned regime with $\Delta = +\omega_m$, the laser can modulate its frequency to create exponential growth of the energies of both the cavity and oscillator modes, accompanied by the generation of strong quantum correlation between the two modes. In the red-detuned regime with $\Delta = -\omega_m$, a switching process between the two modes occurs, which is the same as that in the linear interaction regime.

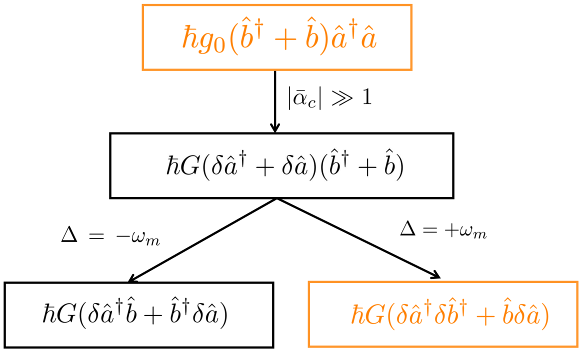
$\hbar g_0(\hat{b}^\dagger + \hat{b})\hat{a}^\dagger \hat{a}$
$|\bar{\alpha}_c| \gg 1$
$\hbar G(\delta\hat{a}^\dagger + \delta\hat{a})(\hat{b}^\dagger + \hat{b})$
$\Delta = -\omega_m$
$\Delta = +\omega_m$
$\hbar G(\delta\hat{a}^\dagger \hat{b} + \hat{b}^\dagger \delta\hat{a})$
$\hbar G(\delta\hat{a}^\dagger \delta\hat{b}^\dagger + \hat{b}\delta\hat{a})$

FIG. 10. Physical insights in the nonlinear regime of cavity-mechanical interaction under the strong laser limit: $|\bar{\alpha}_c| \gg 1$. When the strong laser is in the red-detuned regime with $\Delta = -\omega_m$, the laser controls the two-mode transferring process but, in the blue-detuned regime with $\Delta = +\omega_m$, the laser controls the exponential growth of the two modes in energies and creates the quantum correlation between two modes[61].

The blue- and red-detuned regimes have a competitive relationship with each other in terms of both the photon number and entanglement. In particular, in the blue-detuned regime, photons are excited and the rate of excitation can be larger than that associated with quantum dissipation to the vacuum bath. Furthermore, quantum entanglement is enhanced, overcoming quantum decoherence from the classical environment and even from the SME. However, in the red-detuned regime, no photons are excited and there is only a two-mode energy-transferring process that does not completely suppress the process of quantum dissipation to the vacuum bath, resulting in photon loss and eventually reducing entanglement. Stabilizing the photon number and entanglement requires a balance between the operations in the blue- and red-detuned regimes. In general, the blue-detuned regime is prone to too high photon levels with strong entanglement, which should be balanced by the red-detuned regime operation to reduce the photon number to realize our target entanglement engineering, as shown schematically in Fig. 10. Overall, in the nonlinear interaction regime, laser driving of finite amplitude and frequency modulation can control the photon number and entanglement to a certain extent. An example is shown in Fig. 7(d), where the RL agent finds the optimal action flow with a finite laser amplitude. Note that, the detuning Δ is modulated mainly in the range $\Delta \in [-2\omega_m, 2\omega_m]$, signifying a balance between the blue- and red-detuned operations.

3. Weak continuous measurement

In an open quantum system, under WCM and quantum dissipation into the vacuum bath as well, a Wiener process occurs in the observable. More specifically, the Wiener process arises from the Gaussian-weighted projection over the eigenstates, which weakly extracts the partial information from the quantum system and induces stochastic disturbances in both the dynamical equation and observation. Such disturbances can avoid a complete quantum state collapse and provide the capability to extract the quantum information continuously in the time domain. However, the nonlinear stochastic process occurs in both quantum dynamical trajectories and the measurement photocurrent, making it challenging to control the quantum system through WCM continuously.

For stochastic noise in the WCM photocurrent, the present cutting-edge technology enables the RL agent to extract quantum information through a process resembling noise filtering. Specifically, the observation in the reward function is the WCM photocurrent. We can employ n_{traj} quantum ensembles to reduce the variance and use a Gaussian filter for data pre-processing. The RL agent is trained to maximize the accumulated reward, which serves to average the stochastic term in the measurement photocurrent over time. These noise-filtering processes help extract information about the expected photon number and thus the target quantum entanglement. For the nonlinear quantum stochastic process with quantum dissipation, the RL agent successfully trains the laser to leverage interactions between the optical and mechanical modes, linear or nonlinear, to mitigate quantum decoherence and dissipation to some extent, as exemplified in Figs. 3 and 7.

III. DISCUSSION

Exploiting machine learning for controlling quantum information systems is becoming a promising research realm and is attracting increasing attention. We have developed a model-free deep-RL method for entanglement engineering. We demonstrated its superiority over benchmark quantum control methods in quantum optomechanical systems under WCM. The model-free deep-RL agent sequentially interacts with one or multiple parallel quantum optomechanical environments, collects trajectories, and updates its policy to maximize the accumulated reward to create and stabilize the entanglement. Both linear and nonlinear interacting regimes between the photons in the optical cavity and the phonons associated with the mechanical oscillator in the cavity have been studied. In particular, for linear interactions, the PPO agent directly observes the WCM photocurrent and delivers better performance compared with the benchmark Bayesian and random control methods in the framework of measurement-based feedback control. The performance of deep RL control is tolerant to randomness when initially the system is in some mixed state. For nonlinear interactions, both the model-free PPO and recurrent PPO agents have been tested, where the first was utilized to generate the time series of the target of the expected photon number, and the second one was employed to control entanglement according to an objective. Because of the high degree of randomness in the SME originating from ten stochastic collapse operators, only the observation of the expected photon number is feasible in the nonlinear interaction regime.

More specifically, linear interactions can naturally limit the excitation in the energy levels, providing a mechanism to directly create the entangled Bell states under the premise of strong laser approximation in the red-detuned regime. A disadvantage is that its performance is sensitive to the coupling of the vacuum or thermal bath, even when the decay rate is small (e.g., $\kappa = 0.01\,\omega_m$ with $G \in [-5, 5]\,\omega_m$). This phenomenon is in fact quite common in quantum systems. For instance, in systems with magnon-photon coupling[87], steady Bell states can ideally be generated in the PT-broken phase without dissipation while the entanglement is reduced when the decay rate is not negligible. Another issue with linear interactions is that the time scale associated with generating entangled Bell states[89] tends to be much shorter than the inverse of the coupling strength about the higher-order exceptional points in a system of coupled non-Hermitian qubits with energy loss while the maximum entanglement can only last for a short instant.

In contrast, nonlinear interactions can create and stabilize entanglement and are more robust to the disturbance from the vacuum bath even with a relatively large decay rate, e.g., $\kappa = 0.1\,\omega_m$ with the strong coupling $g_0 = 0.2\,\omega_m$, so $g_0/\kappa = 2 > 1$ to stipulate the nonlinear effect[61]. Potentially, systems with nonlinear coupling thus can outperform those with linear interactions. A caveat is that, in nonlinear optomechanical systems, there is limited experimentally accessible observation. In fact, the relationship between experimental observables and entanglement in nonlinear quantum optomechanical systems has not been well understood, rendering challenging to choose a feasible observable to control entanglement. We have partially relied on the numerical method to create and stabilize entanglement, based on the numerical relation between entanglement and the expected photon number discovered by the deep RL. Another difficulty is that the nonlinear interaction can readily excite the system to high quantum states, which we have overcome by designing a proper reward function.

A previous work[90] studied the acceleration of entanglement generation through feedback weak measurement for two qubits in a four-dimensional Hilbert space, where coupling to a vacuum or thermal bath was not taken into account, nor the interactions between the two qubits, and the control protocol required prior knowledge about the system such as the decoherence-free subspace. In addition, complete observation was needed to design the local Hamiltonian feedback to speed up entanglement. This is in fact a model-based approach. In another study[91], steady-state entanglement between two qubits was achieved using a continuous feedback control method, where the feedback protocol design was informed by a detailed model of the system's dynamics. In contrast, our work creates and stabilizes a two-mode entangled state about a predetermined level of entanglement for both linear and nonlinear interactions via model-free reinforcement learning, with the respective dimensions of the Hilbert space being four and one hundred.

Our work suggests the possibility of exploiting multiagent RL through parallel computation to stabilize entanglement. The agents leverage the decentralized structure of the task and share information via communication. Saliently, if several agents fail in a multiagent system, the remaining agents can take over some of their tasks. In principle, our control framework can be extended to multi-agent RL for multi-mode entanglement engineering of a quantum black box.

IV. METHODS

A. Stochastic Master Equation

An experimental optomechanical system is effectively an open quantum system interacting with the vacuum bath under WCM with the operators[29,49] $\hat{C}_n \equiv \sqrt{\eta}\hat{P}_n$, where $\hat{P}_n = |n\rangle\langle n|$ with $n = 0, 1$ (linear) or $n = 0, 1, \ldots, 9$ (nonlinear) is the measurement operator on the Fock state and η denotes the measurement rate. The quantum dynamics of this system are described by the stochastic master equation (SME)[29,51,92–94]:

$$d\rho = \frac{1}{i\hbar}[\tilde{H}, \rho]dt + \mathcal{L}_{env}\,\rho dt + \sum_n \mathcal{D}(\hat{C}_n)\rho dt + \sum_n \mathcal{H}(\hat{C}_n)\rho dW_n, \qquad (2)$$

where the Hamiltonian is $\tilde{H} = \tilde{H}_{bs}$ or $\tilde{H}_{nl}$ and ρ is a density operator in the Hilbert space. Under the Born-Markov approximation[95,96], which requires the system-bath coupling to be weak and the correlation time of the bath to be much shorter than a characteristic timescale of system-bath interactions, the Markovian master equation, i.e., the first two terms in the right-hand side of Eq. (2), has the Lindblad form[95]. At absolute zero temperature, the following environmental operator $\mathcal{L}_{env}\ \rho$ can be introduced to describe the coupling between the system and vacuum bath: $\mathcal{L}_{env}\ \rho = \kappa\mathcal{D}(\hat{a})\rho + \gamma\mathcal{D}(\hat{b})\rho$, where the cavity and oscillator modes are coupled to the vacuum bath with the strength κ and γ, respectively[56]. The deep RL results in the Lindblad master equation with the nonlinear interaction are presented in Appendix E.

The WCM process described by the last two terms in the right-hand side of Eq. (2) is nonlinear and Markovian in the unconditional master equation[51] in ρ. Under WCM, a Wiener process dW with a Gaussian distribution[51] arises from the Gaussian-weighted projection over the eigenstates that allows the quantum information to be extracted continuously in the time domain, subject to stochastic disturbances in the last term of Eq. (2) and quantum decoherence in the penultimate term of Eq. (2). (Appendix C provides a detailed derivation of the SME.) The Lindblad operator $\mathcal{D}$ and the measurement superoperator $\mathcal{H}$ in Eq. (2) are given by

$$\mathcal{D}(\hat{A})\rho \equiv \hat{A}\rho\hat{A}^\dagger - \frac{1}{2}(\hat{A}^\dagger\hat{A}\rho + \rho\hat{A}^\dagger\hat{A}),$$
$$\mathcal{H}(\hat{A})\rho \equiv \hat{A}\rho + \rho\hat{A}^\dagger - \langle\hat{A} + \hat{A}^\dagger\rangle\rho,$$

with $\langle\hat{A}\rangle \equiv \mathrm{Tr}[\hat{A}\rho]$. The two operators serve to weakly drive the quantum state into the corresponding eigenstates to some degree.

B. Implementation details of deep RL

For simulating the linear or nonlinear quantum optomechanical system described by Eq. (2), we use the "taylor1.5" solver from the SME solver in the QuTip's package[97] with the tolerance tol $= 10^{-6}$ and time step size $dt = 0.01\,\omega_m^{-1}$. The measurement current is simulated with the "homodyne" method, and the custom environment is constructed by the open-source platform OpenAI-Gym[98]. For RL simulations, we construct the PPO agent[48] by "stable-baselines3"[99] in the A2C[100] settings, where stochastic policy (actor) and the value function (critic) are modeled by two independent neural network function approximators, i.e., a set of fully connected feed-forward networks of dimensions 256 × 128 × 64 and the hyperbolic tangent nonlinear activation function for each hidden layer. For the nonlinear quantum optomechanical configuration, in the target-utilization phase, the recurrent PPO agent outperforms the PPO agent, where both independent critic and actor networks are MLPs followed by one independent layer of LSTM with 256 × 128 × 64 fully connected networks and 256 hidden states. More details, especially in the table of Gaussian kernel, are described in Appendix F.

ACKNOWLEDGMENTS

We thank Dr. Kanu Sinha for discussions and comments. This work was supported by the Air Force Office of Scientific Research under Grant No. FA9550-21-1-0438 and by the Office of Naval Research under Grant

No. N00014-24-1-2548. The Quantum Collaborative, led by Arizona State University, also provided valuable expertise and resources for this work through seed funding. The Quantum Collaborative connects top scientific programs, initiatives, and facilities with prominent industry partners to advance the science and engineering of quantum information science.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

AUTHOR CONTRIBUTIONS

L.-L. Y., C. A., J. L., and Y.-C. L. designed the research project, the models, and methods. L.-L. Y. performed the computations. L.-L. Y., C. A., J. L., and Y.-C. L. analyzed the data. L.-L. Y. and Y.-C. L. wrote and edited the manuscript.

DATA AVAILABILITY

The codes used in this paper can be found in the repository: https://github.com/liliyequantum/Entanglement-engineering-by-RL[101]. The data generated in this study about training results can be found in this Zenodo repository: https://doi.org/10.5281/zenodo.12584159[102].

Appendix A: Background related to our work

Quantum control. Quantum control[94] is essential to quantum engineering and technology[103–105], where open-loop control[106] has been successfully demonstrated with methods such as gradient-ascent pulse engineering (GRAPE)[107] in spin systems[108], coupled qubits[109], Jaynes-Cummings systems[110], and qubit-cavity lattices[111]. Recently, the open-loop GRAPE algorithm has been extended to feedback GRAPE[112] 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[113–115].

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)[113], deep deterministic policy gradient (DDPG)[116], and trust region proximal optimization (TRPO)[117]. A state-of-the-art deep RL algorithm for continuous control is proximal policy optimization (PPO)[48], whose performance can exceed that of TRPO. Incorporating recurrent neural networks[57] into the PPO algorithm leads to improved performance[88]. 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[118], state preparation and stabilization for a single particle[28,29,46,47] with an unstable potential[46] or a double-well potential[47], discrimination between entangled states[119] for quantum meteorology, and long-distance entanglement distribution on quantum networks[120]. 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[121].

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[122]. To avoid a complete collapse, one can exploit weak measurements[50,123], 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[124] of driven dissipative quantum-optical systems - a basic component of quantum meteorology[125,126]. 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[84], Wigner[84] and Husimi Q functions[127] with a homodyne apparatus[85], rendering observing both pure[128] and mixed[123] quantum states experimentally feasible. WCM has been experimentally implemented by a weak-field homodyne detector[83–85] to measure the photon-number statistical distribution over the Fock basis. In another example, WCM has been realized in an atomic spin ensemble[50] 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[129] and error-disturbance uncertainty relation[130]. Because of the typical time scales of the quantum dynamics, WCM cannot be regarded as occurring instantaneously[51]. Theoretically, the impact of WCM on the underlying quantum system can be described by the stochastic master equation[51].

Appendix B: Quantum optomechanical system

The standard Hamiltonian of a quantum optomechanical system in the rotating frame of the laser is given by[55,56]

$$\begin{aligned}\tilde{H} = -\hbar\Delta\hat{a}^\dagger\hat{a} + \hbar\omega_m\hat{b}^\dagger\hat{b} + \hbar g_0(\hat{b}^\dagger + \hat{b})\hat{a}^\dagger\hat{a}\\ + \hbar(\alpha_L\hat{a}^\dagger + \alpha_L^*\hat{a}),\end{aligned} \tag{B1}$$

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:

$$\begin{aligned}\hat{H}_0 &= \hbar\omega_c(\hat{q})\hat{a}^\dagger\hat{a} + \hbar\omega_m\hat{b}^\dagger\hat{b}\\ &= \hbar(\omega_c + (\partial\omega_c(q)/\partial q)\hat{q})\hat{a}^\dagger\hat{a} + \hbar\omega_m\hat{b}^\dagger\hat{b}\\ &= \hbar\omega_c\hat{a}^\dagger\hat{a} + \hbar\omega_m\hat{b}^\dagger\hat{b} + \hbar g_0\hat{a}^\dagger\hat{a}(\hat{b}^\dagger + \hat{b}),\end{aligned} \tag{B2}$$

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[93]:

$$\tilde{H} = \hat{U}^\dagger\hat{H}\hat{U} - \hat{A} \tag{B3}$$

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

$$\begin{aligned}\exp\left(i\omega_L\hat{a}^\dagger\hat{a}\, t\right)\hat{a}\exp\left(-i\omega_L\hat{a}^\dagger\hat{a}\, t\right) = \hat{a}\exp\left(-i\omega_L t\right),\\ \exp\left(i\omega_L\hat{a}^\dagger\hat{a}\, t\right)\hat{a}^\dagger\exp\left(-i\omega_L\hat{a}^\dagger\hat{a}\, t\right) = \hat{a}^\dagger\exp\left(i\omega_L t\right),\end{aligned} \tag{B4}$$

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

$$\tilde{H}_0 = -\hbar\Delta\hat{a}^\dagger\hat{a} + \hbar\omega_m\hat{b}^\dagger\hat{b} + \hbar g_0\hat{a}^\dagger\hat{a}(\hat{b}^\dagger + \hat{b}). \tag{B5}$$

The quantized electromagnetic field can be written as

$$\hat{H}_{driven} = \hbar\left[\alpha_L\exp(-i\omega_L t)\hat{a}^\dagger + \alpha_L^*\exp(i\omega_L t)\hat{a}\right]. \tag{B6}$$

Through the unitary transformation, we obtain

$$\hat{U}^\dagger\hat{H}_{driven}\hat{U} = \hbar\alpha_L\hat{a}^\dagger + \hbar\alpha_L^*\hat{a}. \tag{B7}$$

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

$$\begin{aligned}\tilde{H} &= \hat{U}^\dagger(\hat{H}_0 + \hat{H}_{driven})\hat{U} - \hat{A}\\ &= \tilde{H}_0 + \hat{U}^\dagger\hat{H}_{driven}\hat{U}\\ &= -\hbar\Delta\hat{a}^\dagger\hat{a} + \hbar\omega_m\hat{b}^\dagger\hat{b} + \hbar g_0\hat{a}^\dagger\hat{a}(\hat{b}^\dagger + \hat{b})\\ &\quad + \hbar(\alpha_L\hat{a}^\dagger + \alpha_L^*\hat{a}).\end{aligned} \tag{B8}$$

Appendix C: Quantum stochastic master equation

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

$$\dot{\rho} = \frac{1}{i\hbar}[\hat{H}, \rho] \equiv \mathcal{L}\rho, \tag{C1}$$

where $\mathcal{L}$ is the Liouvillian superoperator. Equation (C1) can be derived from the Schrödinger equation and its conjugate:

$$\begin{aligned}i\hbar\frac{\partial}{\partial t}|\psi\rangle = \hat{H}|\psi\rangle,\\ -i\hbar\frac{\partial}{\partial t}\langle\psi| = \langle\psi|\hat{H},\end{aligned} \tag{C2}$$

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

$$\begin{aligned}i\hbar\dot{\rho} &= \sum_i P_i(i\hbar|\dot{\psi_i}\rangle)\langle\psi_i| - \sum_i P_i|\psi_i\rangle(-i\hbar\langle\dot{\psi_i}|)\\ &= \sum_i P_i\hat{H}|\psi_i\rangle\langle\psi_i| - \sum_i P_i|\psi_i\rangle\langle\psi_i|\hat{H}\\ &= \hat{H}\rho - \rho\hat{H} = [\hat{H}, \rho],\end{aligned} \tag{C3}$$

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)[51,93,94]:

$$d\rho = \frac{1}{i\hbar}[\hat{H}, \rho]dt + \mathcal{L}_{env}\,\rho dt + \mathcal{D}(\hat{c})\rho dt + \mathcal{H}(\hat{c})\rho dW, \tag{C4}$$

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

$$\mathcal{L}_{env}\,\rho = \kappa\mathcal{D}(\hat{a})\rho + \gamma\mathcal{D}(\hat{b})\rho, \tag{C5}$$

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. (C5), the symbols $\mathcal{D}$ and $\mathcal{H}$ denote the Lindblad and measurement superoperators, respectively, which are given by

$$\mathcal{D}(\hat{c})\rho \equiv \hat{c}\rho\hat{c}^{\dagger} - \frac{1}{2}(\hat{c}^{\dagger}\hat{c}\rho + \rho\hat{c}^{\dagger}\hat{c}), \tag{C6}$$

$$\mathcal{H}(\hat{c})\rho \equiv \hat{c}\rho + \rho\hat{c}^{\dagger} - \langle\hat{c} + \hat{c}^{\dagger}\rangle\rho. \tag{C7}$$

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[51]. 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: $\{|n\rangle : n = 1, ..., n_{\max}\}$. 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

$$\rho_f = \frac{P_n\rho P_n}{\text{Tr}[P_n\rho P_n]} = |n\rangle\langle n|. \tag{C8}$$

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

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

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

$$\rho_f = \frac{\Omega_m\rho\Omega_m^{\dagger}}{\text{Tr}[\Omega_m\rho\Omega_m^{\dagger}]}, \tag{C9}$$

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

$$P(m \in [a, b]) = \sum_{m=a}^{b} \text{Tr}[\Omega_m\rho\Omega_m^{\dagger}] = \text{Tr}[\sum_{m=a}^{b} \Omega_m\rho\Omega_m^{\dagger}]. \tag{C10}$$

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:

$$\Omega_m = \frac{1}{\mathcal{N}} \sum_n e^{-k(n-m)^2/4}|n\rangle\langle n|, \tag{C11}$$

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

$$\rho_f = \frac{\Omega_m\rho\Omega_m^{\dagger}}{\text{Tr}[\Omega_m\rho\Omega_m^{\dagger}]} = \frac{1}{\mathcal{N}} \sum_n e^{-k(n-m)^2/2}|n\rangle\langle n|. \tag{C12}$$

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

$$\hat{A}(\alpha) = \left(\frac{4k\Delta t}{\pi}\right)^{1/4} \int_{-\infty}^{\infty} e^{-2k\Delta t(O-\alpha)^2}|O\rangle\langle O|dO, \tag{C13}$$

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

$$\langle\alpha\rangle = \int_{-\infty}^{\infty} \alpha Tr[\hat{A}^{\dagger}(\alpha)\hat{A}(\alpha)|\psi\rangle\langle\psi|]d\alpha = \langle\hat{O}\rangle. \tag{C14}$$

The probability distribution of α is

$$P(\alpha) = Tr[\hat{A}^\dagger(\alpha)\hat{A}(\alpha)|\psi\rangle\langle\psi|] = \sqrt{\frac{4k\Delta t}{\pi}}\int_{-\infty}^{\infty}|\psi(O)|^2 e^{-4k\Delta t(O-\alpha)^2}dO. \quad \text{(C15)}$$

The value of Δt is infinitesimal due to the inherent property of the WCM. As a result, the exponential term in Eq. (C15) 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

$$P(\alpha) \approx \sqrt{\frac{4k\Delta t}{\pi}} e^{-4k\Delta t(\alpha-\langle O\rangle)^2}. \quad \text{(C16)}$$

Effectively, α is a stochastic quantity:

$$\alpha_s = \langle\hat{O}\rangle + \frac{\Delta W}{\sqrt{8k\Delta t}}, \quad \text{(C17)}$$

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

$$|\psi(t+\Delta t)\rangle \propto \hat{A}(\alpha)|\psi(t)\rangle \propto e^{-2k\Delta t(\alpha-\hat{O})^2}|\psi(t)\rangle. \quad \text{(C18)}$$

Substituting Eq. (C17) 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:

$$d|\psi\rangle = \{-k(\hat{O} - \langle\hat{O}\rangle)^2 dt + \sqrt{2k}(\hat{O} - \langle\hat{O}\rangle)dW\}|\psi(t)\rangle. \quad \text{(C19)}$$

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

$$d\rho = (d|\psi\rangle)\langle\psi| + |\psi\rangle(d\langle\psi|) + (d|\psi\rangle)(d\langle\psi|) = -k[\hat{O},[\hat{O},\rho]]dt + \sqrt{2k}(\hat{O}\rho + \rho\hat{O} - 2\langle\hat{O}\rangle\rho)dW. \quad \text{(C20)}$$

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

$$[\hat{c}\rho\hat{c} - \frac{1}{2}(\hat{c}^2\rho + \rho\hat{c}^2)]dt \quad \text{(C21)}$$

and the second term is

$$(\hat{c}\rho + \rho\hat{c} - 2\langle\hat{c}\rangle\rho)dW, \quad \text{(C22)}$$

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.

Appendix D: 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., Episode = 3000, the RL agent bootstraps itself by executing the procedure described in Appendix F 1. 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):

$$\tau^j = (O_0^j, G_0^j, R_0^j, O_1^j, \ldots, O_{T-1}^j, G_{T-1}^j, R_{T-1}^j), \quad \text{(D1)}$$

which can be organized as a sub-trajectory tuple

$$\tau_t^j = (O_t^j, G_t^j, R_t^j, O_{t+1}^j) \quad \text{(D2)}$$

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

$$\tau_{T-1}^j = (O_{T-1}^j, G_{T-1}^j, R_{T-1}^j). \quad \text{(D3)}$$

For each sub-trajectory tuple τ_t^j, the generalized advantage estimation (GAE)[131] $\hat{A}_t^j$ uses a value function estimator:

$$\hat{A}_t^j = \delta_t^j + (\gamma\lambda)\delta_{t+1}^j + \ldots + (\gamma\lambda)^{T-t-1}\delta_{T-1}^j \quad \text{(D4)}$$

with

$$\delta_t^j = R_t^j + \gamma V(O_{t+1}^j;\boldsymbol{\phi}) - V(O_t^j;\boldsymbol{\phi}), \quad \text{(D5)}$$

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

$$\hat{\mathcal{G}}_t^j = \sum_{k=t}^{T-1} \gamma^{k-t} R_k^j, \tag{D6}$$

which can be also obtained from the generalized advantage by

$$\hat{\mathcal{G}}_t^j = \hat{A}_t^j + V(O_t^j;\boldsymbol{\phi}), \tag{D7}$$

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 $\mathbb{M} = \mathrm{int}(\mathbb{N}\times\mathbb{Z}\times T/10)$. 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

$$L_{critic}(\boldsymbol{\phi}) = \hat{\mathbb{E}}_i[(V(O_i;\boldsymbol{\phi}) - \hat{\mathcal{G}}_i)^2] \tag{D8}$$

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

$$L_{actor}(\boldsymbol{\theta}) = \hat{\mathbb{E}}_i\left[-\min(r_i(\boldsymbol{\theta})\hat{A}_i, \mathrm{clip}(r_i(\boldsymbol{\theta}), [1-\epsilon, 1+\epsilon])\hat{A}_i)\right], \tag{D9}$$

where $\hat{\mathbb{E}}_i[\quad] = \sum_{i=1}^{\mathbb{M}}[\quad]_i/\mathbb{M}$ is the empirical average over a minibatch of the data and $[\quad]_i$ denotes the i th element of the minibatch with $i = 0, 1, \ldots, \mathbb{M}-1$, and the clip function $\mathrm{clip}(x, [\min, \max])$ 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

$$r_i(\boldsymbol{\theta}) = \frac{\pi_{\boldsymbol{\theta}}(G_i|O_i)}{\pi_{\boldsymbol{\theta}_{\mathrm{old}}}(G_i|O_i)}. \tag{D10}$$

If the current policy is the same as the old policy, we have $r_i(\boldsymbol{\theta}_{\mathrm{old}}) = 1$. 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[115], so the optimal value function satisfies

$$V^*(O_t^j;\boldsymbol{\phi}) = R_t^j + \gamma V^*(O_{t+1}^j;\boldsymbol{\phi}), \tag{D11}$$

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[115]. 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. (D6):

$$V^*(O_i;\boldsymbol{\phi}) = \hat{\mathcal{G}}_i, \tag{D12}$$

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.

Appendix E: RL in nonlinear interactions by the Lindblad master equation

Figs. 11 and 12 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 Lindblad 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 $R(t) = -|E_N(t) - \log 2| -$

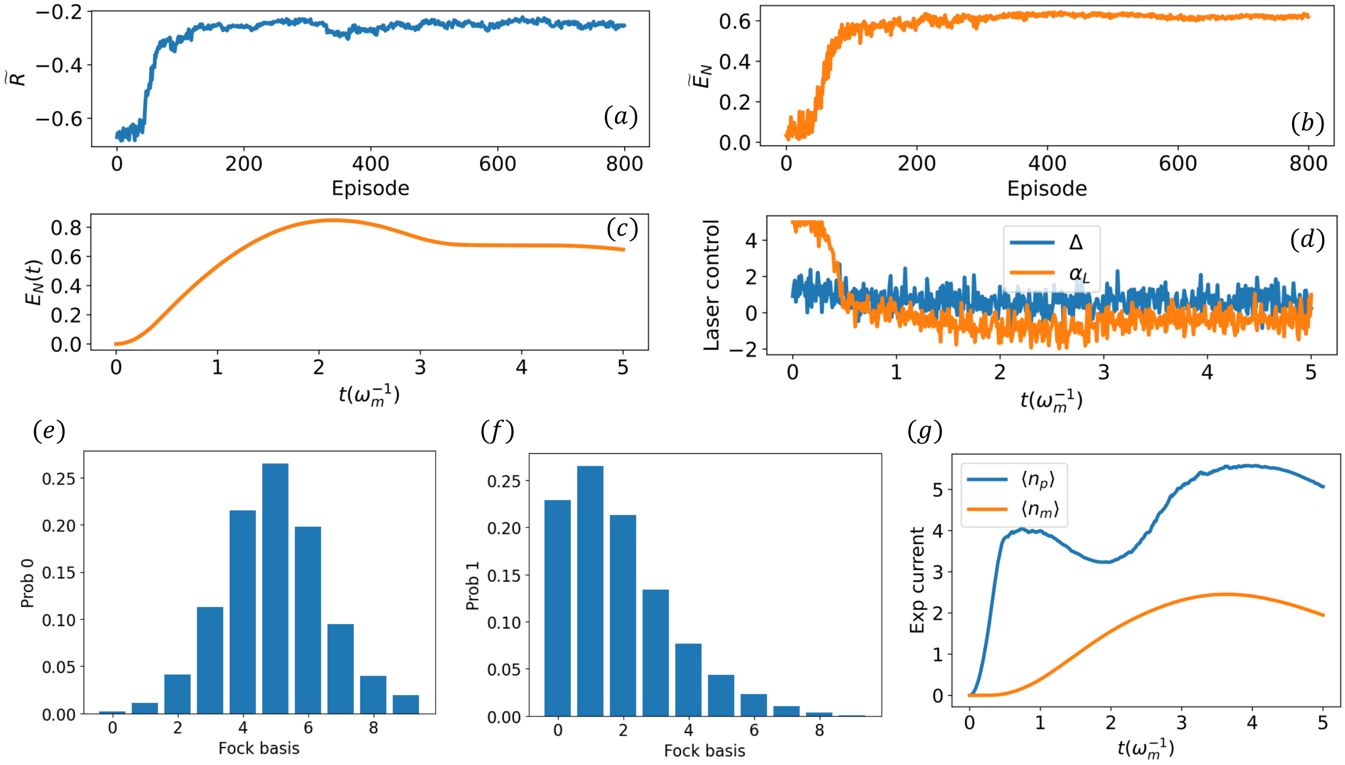

FIG. 11. A detailed account of the target-generating phase in RL control of open optomechanical systems with nonlinear photon-phonon interaction in the framework of Lindblad master equation. Nonlinear interaction of strength $g_0 = 0.2\omega_m$ creates the target entanglement $E_N \sim \log 2$ optimized 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. 12.

$|\langle n_p\rangle(t) + \langle n_m\rangle(t) - a|/b$. Figure 11 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. 11(a) and 11(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. 11(c) controlled by the laser control signal shown in Fig. 11(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. 11(e) and 11(f). The corresponding target pattern $\langle n_p^{\text{target}}\rangle(t)$ is demonstrated in Fig. 11(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 12 illustrates that, with only partial information extracted from the quantum optomechanical environment, the agent can steadily learn to create and stabilize entanglement.

Appendix F: Deep RL

There are three main RL approaches[114] 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

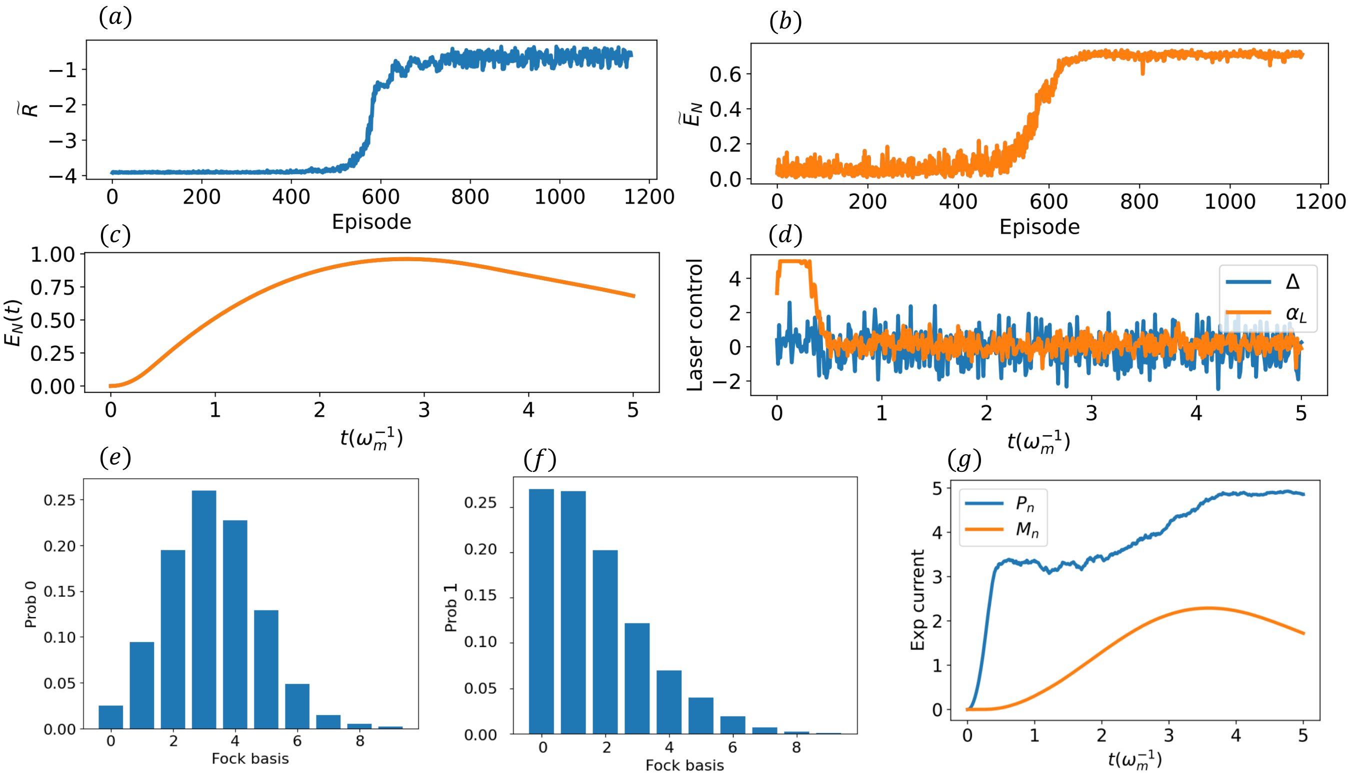

FIG. 12. A detailed account of the target-utilization phase in RL control of open optomechanical systems with nonlinear photon-phonon interaction in the framework of Lindblad master equation. The goal is to create the entanglement about $E_N \sim \log 2$, for which the reward function is $R(t) = -|\langle n_p\rangle(t) - \langle n_p^{\text{target}}\rangle(t)|$, determined by the target time series $\langle n_p^{\text{target}}\rangle(t)$ from Fig. 11(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. 11, 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. 11.

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[114]. 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[117] 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.

1. PPO agent

Proximal policy optimization (PPO)[48] 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

$$\tau_t^j = (s_t^j, a_t^j, R_t^j, s_{t+1}^j) \tag{F1}$$

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:

$$\tau_{T-1}^j = (s_{T-1}^j, a_{T-1}^j, R_{T-1}^j). \tag{F2}$$

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

$$\tau^j = \tau_0^j \cup \tau_1^j \cup ...... \cup \tau_{T-1}^j, \tag{F3}$$

so the trajectory τ^j is given by

$$\tau^j = (s_0^j, a_0^j, R_0^j, s_1^j, ..., s_{T-1}^j, a_{T-1}^j, R_{T-1}^j). \tag{F4}$$

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)[131] uses a value function estimator:

$$\hat{A}_t^j = \delta_t^j + (\gamma\lambda)\delta_{t+1}^j + \ldots + (\gamma\lambda)^{T-t-1}\delta_{T-1}^j, \tag{F5}$$

with

$$\delta_t^j = R_t^j + \gamma V(s_{t+1}^j; \boldsymbol{\phi}) - V(s_t^j; \boldsymbol{\phi}), \tag{F6}$$

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:

$$\hat{\mathcal{G}}_t^j = \hat{A}_t^j + V(s_t^j; \boldsymbol{\phi}). \tag{F7}$$

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

$$L_{critic}(\boldsymbol{\phi}) = \hat{\mathbb{E}}_i[(V(s_i; \boldsymbol{\phi}) - \hat{\mathcal{G}}_i)^2], \tag{F8}$$

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

$$L_{actor}(\boldsymbol{\theta}) = \hat{\mathbb{E}}_i\left[-\min(r_i(\boldsymbol{\theta})\hat{A}_i, \mathrm{clip}(r_i(\boldsymbol{\theta}), [1-\epsilon, 1+\epsilon])\hat{A}_i)\right]. \tag{F9}$$

where the clip function $\mathrm{clip}(x, [\min, \max])$ 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

$$r_i(\boldsymbol{\theta}) = \frac{\pi_{\boldsymbol{\theta}}(a_i|s_i)}{\pi_{\boldsymbol{\theta}_{\mathrm{old}}}(a_i|s_i)}. \tag{F10}$$

If the current policy is the same as the old policy, we have $r_i(\boldsymbol{\theta}_{\mathrm{old}}) = 1$. 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.

2. 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)[57] 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[132]. 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[88].

3. 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 II. 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

[1] C. H. Bennett and D. P. DiVincenzo, "Quantum information and computation," nature **404**, 247–255 (2000).

[2] S. L. Braunstein and P. van Loock, "Quantum information with continuous variables," Rev. Mod. Phys. **77**, 513–577 (2005).

[3] H.-Y. Huang, R. Kueng, and J. Preskill, "Information-theoretic bounds on quantum advantage in machine learning," Phys. Rev. Lett. **126**, 190505 (2021).

[4] H.-Y. Huang, M. Broughton, J. Cotler, S. Chen, J. Li, M. Mohseni, H. Neven, R. Babbush, R. Kueng, J. Preskill, *et al.*, "Quantum advantage in learning from experiments," Science **376**, 1182–1186 (2022).

[5] S. Lee, J. Lee, H. Zhai, Y. Tong, A. M. Dalzell, A. Kumar, P. Helms, J. Gray, Z.-H. Cui, W. Liu, *et al.*, "Evaluating the evidence for exponential quantum advantage in ground-state quantum chemistry," Nat. Commun. **14**, 1952 (2023).

[6] C. L. Degen, F. Reinhard, and P. Cappellaro, "Quantum sensing," Rev. Mod. Phys. **89**, 035002 (2017).

[7] R. Horodecki, P. Horodecki, M. Horodecki, and K. Horodecki, "Quantum entanglement," Rev. Mod. Phys. **81**, 865 (2009).

[8] H. J. Kimble, "The quantum internet," Nature **453**, 1023–1030 (2008).

[9] S. Wehner, D. Elkouss, and R. Hanson, "Quantum internet: A vision for the road ahead," Science **362**, eaam9288 (2018).

[10] D. Bouwmeester, J.-W. Pan, K. Mattle, M. Eibl, H. Weinfurter, and A. Zeilinger, "Experimental quantum teleportation," Nature **390**, 575–579 (1997).

[11] J.-G. Ren, P. Xu, H.-L. Yong, L. Zhang, S.-K. Liao, J. Yin, W.-Y. Liu, W.-Q. Cai, M. Yang, L. Li, *et al.*, "Ground-to-satellite quantum teleportation," Nature **549**, 70–73 (2017).

[12] Y.-A. Chen, Q. Zhang, T.-Y. Chen, W.-Q. Cai, S.-K. Liao, J. Zhang, K. Chen, J. Yin, J.-G. Ren, Z. Chen, *et al.*, "An integrated space-to-ground quantum communication network over 4,600 kilometres," Nature **589**, 214–219 (2021).

[13] S. Pirandola, R. Laurenza, C. Ottaviani, and L. Banchi, "Fundamental limits of repeaterless quantum communications," Nat. Commun. **8**, 15043 (2017).

[14] N. P. De Leon, K. M. Itoh, D. Kim, K. K. Mehta, T. E. Northup, H. Paik, B. Palmer, N. Samarth, S. Sangtawesin, and D. W. Steuerman, "Materials challenges and opportunities for quantum computing hardware," Science **372**, eabb2823 (2021).

[15] A. Noiri, K. Takeda, T. Nakajima, T. Kobayashi, A. Sammak, G. Scappucci, and S. Tarucha, "Fast universal quantum gate above the fault-tolerance threshold in silicon," Nature **601**, 338–342 (2022).

[16] A. B. Magann, C. Arenz, M. D. Grace, T.-S. Ho, R. L. Kosut, J. R. McClean, H. A. Rabitz, and M. Sarovar, "From pulses to circuits and back again: A quantum optimal control perspective on variational quantum algorithms," PRX Quantum **2**, 010101 (2021).

[17] L. Banchi, N. Pancotti, and S. Bose, "Quantum gate learning in qubit networks: Toffoli gate without time-dependent control," Npj Quantum Inf. **2**, 1–6 (2016).

[18] G. Romero, D. Ballester, Y. Wang, V. Scarani, and E. Solano, "Ultrafast quantum gates in circuit QED," Phys. Rev. Lett. **108**, 120501 (2012).

[19] T. Van der Sar, Z. Wang, M. Blok, H. Bernien, T. Taminiau, D. Toyli, D. Lidar, D. Awschalom, R. Hanson, and V. Dobrovitski, "Decoherence-protected quantum gates for a hybrid solid-state spin register," Nature **484**, 82–86 (2012).

[20] J. M. Chow, J. M. Gambetta, A. D. Corcoles, S. T. Merkel, J. A. Smolin, C. Rigetti, S. Poletto, G. A. Keefe, M. B. Rothwell, J. R. Rozen, *et al.*, "Universal quantum gate set approaching fault-tolerant thresholds with superconducting qubits," Phys. Rev. Lett. **109**, 060501 (2012).

[21] K. Bharti, A. Cervera-Lierta, T. H. Kyaw, T. Haug, S. Alperin-Lea, A. Anand, M. Degroote, H. Heimonen, J. S. Kottmann, T. Menke, *et al.*, "Noisy intermediate-scale quantum algorithms," Rev. Mod. Phys. **94**, 015004 (2022).

[22] J. Mackeprang, D. B. R. Dasari, and J. Wrachtrup, "A reinforcement learning approach for quantum state engineering,"

Quantum Mach. Intell. **2**, 1–14 (2020).

[23] S. Giordano and M. A. Martin-Delgado, "Reinforcement-learning generation of four-qubit entangled states," Phys. Rev. Research **4**, 043056 (2022).

[24] Z.-Y. Zhang, Z. Sun, T. Duan, Y.-K. Ding, X. Huang, and J.-M. Liu, "Entanglement generation of polar molecules via deep reinforcement learning," J. Chem. Theory Comput. **20**, 1811–1820 (2024).

[25] F. Metz and M. Bukov, "Self-correcting quantum many-body control using reinforcement learning with tensor networks," Nat. Mach. Intell. **5**, 780–791 (2023).

[26] M. Bukov, A. G. Day, D. Sels, P. Weinberg, A. Polkovnikov, and P. Mehta, "Reinforcement learning in different phases of quantum control," Phys. Rev. X **8**, 031086 (2018).

[27] S.-F. Guo, F. Chen, Q. Liu, M. Xue, J.-J. Chen, J.-H. Cao, T.-W. Mao, M. K. Tey, and L. You, "Faster state preparation across quantum phase transition assisted by reinforcement learning," Phys. Rev. Lett. **126**, 060401 (2021).

[28] V. Sivak, A. Eickbusch, H. Liu, B. Royer, I. Tsioutsios, and M. Devoret, "Model-free quantum control with reinforcement learning," Phys. Rev. X **12**, 011059 (2022).

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

[30] P. Jayachandran, L. H. Zaw, and V. Scarani, "Dynamics-based entanglement witnesses for non-Gaussian states of harmonic oscillators," Phys. Rev. Lett. **130**, 160201 (2023).

[31] M. Ho, E. Oudot, J.-D. Bancal, and N. Sangouard, "Witnessing optomechanical entanglement with photon counting," Phys. Rev. Lett. **121**, 023602 (2018).

[32] C. Gut, K. Winkler, J. Hoelscher-Obermaier, S. Hofer, R. M. Nia, N. Walk, A. Steffens, J. Eisert, W. Wieczorek, J. Slater, *et al.*, "Stationary optomechanical entanglement between a mechanical oscillator and its measurement apparatus," Phys. Rev. Res. **2**, 033244 (2020).

[33] S. G. Hofer, W. Wieczorek, M. Aspelmeyer, and K. Hammerer, "Quantum entanglement and teleportation in pulsed cavity optomechanics," Phys. Rev. A **84**, 052327 (2011).

[34] Q. Cai, B. Fan, Y. Fan, G. Deng, Y. Wang, H. Song, G. Guo, and Q. Zhou, "Entangling optical and mechanical cavity modes in an optomechanical crystal nanobeam," Phys. Rev. A **108**, 022419 (2023).

[35] J. Clarke, P. Sahium, K. Khosla, I. Pikovski, M. Kim, and M. Vanner, "Generating mechanical and optomechanical entanglement via pulsed interaction and measurement," New J. Phys. **22**, 063001 (2020).

[36] S. Kiesewetter, Q. He, P. Drummond, and M. Reid, "Scalable quantum simulation of pulsed entanglement and Einstein-Podolsky-Rosen steering in optomechanics," Phys. Rev. A **90**, 043805 (2014).

[37] G. Wang, L. Huang, Y.-C. Lai, and C. Grebogi, "Nonlinear dynamics and quantum entanglement in optomechanical systems," Phys. Rev. Lett. **112**, 110406 (2014).

[38] S. G. Hofer and K. Hammerer, "Entanglement-enhanced time-continuous quantum control in optomechanics," Phys. Rev. A **91**, 033822 (2015).

[39] Q. Lin, B. He, R. Ghobadi, and C. Simon, "Fully quantum approach to optomechanical entanglement," Phys. Rev. A **90**, 022309 (2014).

[40] A. Farace and V. Giovannetti, "Enhancing quantum effects via periodic modulations in optomechanical systems," Phys. Rev. A **86**, 013820 (2012).

[41] A. Mari and J. Eisert, "Opto-and electro-mechanical entanglement improved by modulation," New J. Phys. **14**, 075014 (2012).

[42] A. Mari and J. Eisert, "Gently modulating optomechanical systems," Phys. Rev. Lett. **103**, 213603 (2009).

[43] D. Stefanatos, "Maximising optomechanical entanglement with optimal control," Quantum Sci. Technol. **2**, 014003 (2017).

[44] J. Guo and S. Gröblacher, "Coherent feedback in optomechanical systems in the sideband-unresolved regime," Quantum **6**, 848 (2022).

[45] A. Harwood, M. Brunelli, and A. Serafini, "Cavity optomechanics assisted by optical coherent feedback," Phys. Rev. A **103**, 023509 (2021).

[46] Z. T. Wang, Y. Ashida, and M. Ueda, "Deep reinforcement learning control of quantum cartpoles," Phys. Rev. Lett. **125**, 100401 (2020).

[47] S. Borah, B. Sarma, M. Kewming, G. J. Milburn, and J. Twamley, "Measurement-based feedback quantum control with deep reinforcement learning for a double-well nonlinear potential," Phys. Rev. Lett. **127**, 190403 (2021).

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

[49] A. Essig, Q. Ficheux, T. Peronnin, N. Cottet, R. Lescanne, A. Sarlette, P. Rouchon, Z. Leghtas, and B. Huard, "Multiplexed photon number measurement," Phys. Rev. X **11**, 031045 (2021).

[50] G. A. Smith, S. Chaudhury, A. Silberfarb, I. H. Deutsch, and P. S. Jessen, "Continuous weak measurement and nonlinear dynamics in a cold spin ensemble," Phys. Rev. Lett. **93**, 163602 (2004).

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

[52] J. Ramírez, W. Yu, and A. Perrusquía, "Model-free reinforcement learning from expert demonstrations: A survey," Artif. Intell. Rev. **55**, 3213–3241 (2022).

[53] L. Kaiser, M. Babaeizadeh, P. Milos, B. Osinski, R. H. Campbell, K. Czechowski, D. Erhan, C. Finn, P. Kozakowski, S. Levine, *et al.*, "Model-based reinforcement learning for atari," arXiv preprint arXiv:1903.00374 (2019).

[54] J. K. Stockton, R. van Handel, and H. Mabuchi, "Deterministic Dicke-state preparation with continuous measurement and control," Phys. Rev. A **70**, 022106 (2004).

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

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

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

[58] J. M. Dobrindt, I. Wilson-Rae, and T. J. Kippenberg, "Parametric normal-mode splitting in cavity optomechanics," Phys. Rev. Lett. **101**, 263602 (2008).

[59] T. Palomaki, J. Teufel, R. Simmonds, and K. W. Lehnert, "Entangling mechanical motion with microwave fields," Science **342**, 710–713 (2013).

[60] S. Barzanjeh, A. Xuereb, S. Gröblacher, M. Paternostro, C. A. Regal, and E. M. Weig, "Optomechanics for quantum technologies," Nat. Phys. **18**, 15–24 (2022).

[61] M. Aspelmeyer, T. J. Kippenberg, and F. Marquardt, "Cavity optomechanics," Rev. Mod. Phys. **86**, 1391–1452 (2014).

[62] M.-A. Lemonde, N. Didier, and A. A. Clerk, "Nonlinear interaction effects in a strongly driven optomechanical cavity," Phys. Rev. Lett. **111**, 053602 (2013).

[63] F. Marquardt, "Quantum optomechanics," (2014), `https://mpl.mpg.de/fileadmin/user_upload/Marquardt_Division/Teaching/2014_ChapterDraftLesHouches.pdf`.

[64] X. Song, M. Oksanen, J. Li, P. J. Hakonen, and M. A. Sillanpää, "Graphene optomechanics realized at microwave frequencies," Phys. Rev. Lett. **113**, 027404 (2014).

[65] J. Shin, Y. Ryu, M.-A. Miri, S.-B. Shim, H. Choi, A. Alù, J. Suh, and J. Cha, "On-chip microwave frequency combs in a superconducting nanoelectromechanical device," Nano Lett. **22**, 5459–5465 (2022).

[66] Y. Seis, T. Capelle, E. Langman, S. Saarinen, E. Planz, and A. Schliesser, "Ground state cooling of an ultracoherent electromechanical system," Nat. Commun. **13**, 1507 (2022).

[67] Y. Liu, H. Sun, Q. Liu, H. Wu, M. A. Sillanpää, and T. Li, "Long-lived microwave electromechanical systems enabled by cubic silicon-carbide membrane crystals," arXiv preprint arXiv:2401.01020 (2024).

[68] B. Johnson, M. Reed, A. A. Houck, D. Schuster, L. S. Bishop, E. Ginossar, J. Gambetta, L. DiCarlo, L. Frunzio, S. Girvin, *et al.*, "Quantum non-demolition detection of single microwave photons in a circuit," Nat. Phys. **6**, 663–667 (2010).

[69] S. Gleyzes, S. Kuhr, C. Guerlin, J. Bernu, S. Deleglise, U. Busk Hoff, M. Brune, J.-M. Raimond, and S. Haroche, "Quantum jumps of light recording the birth and death of a photon in a cavity," Nature **446**, 297–300 (2007).

[70] G. Vidal and R. F. Werner, "Computable measure of entanglement," Phys. Rev. A **65**, 032314 (2002).

[71] M. B. Plenio, "Logarithmic negativity: A full entanglement monotone that is not convex," Phys. Rev. Lett. **95**, 090503 (2005).

[72] A. Kitagawa, M. Takeoka, M. Sasaki, and A. Chefles, "Entanglement evaluation of non-Gaussian states generated by photon subtraction from squeezed states," Phys. Rev. A **73**, 042310 (2006).

[73] H. Shapourian, S. Liu, J. Kudler-Flam, and A. Vishwanath, "Entanglement negativity spectrum of random mixed states: A diagrammatic approach," PRX Quantum **2**, 030347 (2021).

[74] G. Vidal, "Entanglement monotones," J. Mod. Opt. **47**, 355–376 (2000).

[75] J. Eisert, S. Scheel, and M. B. Plenio, "Distilling Gaussian states with Gaussian operations is impossible," Phys. Rev. Lett. **89**, 137903 (2002).

[76] J. Fiurášek, "Gaussian transformations and distillation of entangled Gaussian states," Phys. Rev. Lett. **89**, 137904 (2002).

[77] E. M. Rains, "A semidefinite program for distillable entanglement," IEEE Trans. Inf. Theory **47**, 2921–2933 (2001).

[78] R. Pakniat, M. H. Zandi, and M. K. Tavassoly, "On the entanglement swapping by using the beam splitter," Eur. Phys. J. Plus **132**, 1–10 (2017).

[79] F. Bouchard, A. Sit, Y. Zhang, R. Fickler, F. M. Miatto, Y. Yao, F. Sciarrino, and E. Karimi, "Two-photon interference: The [Hong–Ou–Mandel] effect," Rep. Prog. Phys. **84**, 012402 (2020).

[80] Y.-H. Kim and W. P. Grice, "Reliability of the beam-splitter–based Bell-state measurement," Phys. Rev. A **68**, 062305 (2003).

[81] R. O. Behunin and P. T. Rakich, "Quantum optomechanics in tripartite systems," arXiv preprint arXiv:2210.14967 (2022).

[82] Y.-D. Wang, S. Chesi, and A. A. Clerk, "Bipartite and tripartite output entanglement in three-mode optomechanical systems," Phys. Rev. A **91**, 013807 (2015).

[83] G. S. Thekkadath, D. S. Phillips, J. F. F. Bulmer, W. R. Clements, A. Eckstein, B. A. Bell, J. Lugani, T. A. W. Wolterink, A. Lita, S. W. Nam, T. Gerrits, C. G. Wade, and I. A. Walmsley, "Tuning between photon-number and quadrature measurements with weak-field homodyne detection," Phys. Rev. A **101**, 031801 (2020).

[84] G. Puentes, J. S. Lundeen, M. P. Branderhorst, H. B. Coldenstrodt-Ronge, B. J. Smith, and I. A. Walmsley, "Bridging particle and wave sensitivity in a configurable detector of positive operator-valued measures," Phys. Rev. Lett. **102**, 080404 (2009).

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

[86] M. Nixon and A. Aguado, *Feature extraction and image processing for computer vision* (Academic press, 2019).

[87] H. Y. Yuan, P. Yan, S. Zheng, Q. Y. He, K. Xia, and M.-H. Yung, "Steady Bell state generation via magnon-photon coupling," Phys. Rev. Lett. **124**, 053602 (2020).

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

[89] Z.-Z. Li, W. Chen, M. Abbasi, K. W. Murch, and K. B. Whaley, "Speeding up entanglement generation by proximity to higher-order exceptional points," Phys. Rev. Lett. **131**, 100202 (2023).

[90] C. Hill and J. Ralph, "Weak measurement and control of entanglement generation," Phys. Rev. A **77**, 014305 (2008).

[91] G. F. Diotallevi, B. Annby-Andersson, P. Samuelsson, A. Tavakoli, and P. Bakhshinezhad, "Steady-state entanglement production in a quantum thermal machine with continuous feedback control," New J. Phys. **26**, 053005 (2024).

[92] A. Blaquiere, S. Diner, and G. Lochak, "Information complexity and control in quantum physics," in *Proceedings of the 4th International Seminar on Mathematical Theory of Dynamical Systems and Microphysics Udine. Vienna: Springer. Springer* (Springer, 1987).

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

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

[95] F. Nathan and M. S. Rudner, "Universal Lindblad equation for open quantum systems," Phys. Rev. B **102**, 115109 (2020).

[96] S.-Y. Bai, C. Chen, H. Wu, and J.-H. An, "Quantum control in open and periodically driven systems," Adv. Phys. X **6**, 1870559 (2021).

[97] J. R. Johansson, P. D. Nation, and F. Nori, "QuTiP: An open-source Python framework for the dynamics of open quantum systems," Comput. Phys. Commun. **183**, 1760–1772 (2012).

[98] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," arXiv preprint arXiv:1606.01540 (2016).

[99] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-Baselines3: Reliable reinforcement learning implementations," J. Mach. Learn. Res. **22**, 1–8 (2021).

[100] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *ICML* (PMLR, 2016) pp. 1928–1937.

[101] L.-L. Ye, "Entanglement engineering by RL," `https://github.com/liliyequantum/Entanglement-engineering-by-RL` (2024).

[102] L.-L. Ye, "Entanglement engineering of optomechanical systems by reinforcement learning," `https://doi.org/10.5281/zenodo.12584159` (2024).

[103] E. Rosencher, A. Fiore, B. Vinter, V. Berger, P. Bois, and J. Nagle, "Quantum engineering of optical nonlinearities," Science **271**, 168–173 (1996).

[104] G. Iannaccone, F. Bonaccorso, L. Colombo, and G. Fiori, "Quantum engineering of transistors based on 2d materials heterostructures," Nat. Nanotechnol. **13**, 183–191 (2018).

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

[106] G. A. Paz-Silva and L. Viola, "General transfer-function approach to noise filtering in open-loop quantum control," Phys. Rev. Lett. **113**, 250501 (2014).

[107] S. Machnes, U. Sander, S. J. Glaser, P. de Fouquières, A. Gruslys, S. Schirmer, and T. Schulte-Herbrüggen, "Comparing, optimizing, and benchmarking quantum-control algorithms in a unifying programming framework," Phys. Rev. A **84**, 022305 (2011).

[108] F. Dolde, V. Bergholm, Y. Wang, I. Jakobi, B. Naydenov, S. Pezzagna, J. Meijer, F. Jelezko, P. Neumann, T. Schulte-Herbrüggen, *et al.*, "High-fidelity spin entanglement using optimal control," Nat. Commun. **5**, 3371 (2014).

[109] A. Spörl, T. Schulte-Herbrüggen, S. J. Glaser, V. Bergholm, M. J. Storcz, J. Ferber, and F. K. Wilhelm, "Optimal control of coupled josephson qubits," Phys. Rev. A **75**, 012302 (2007).

[110] R. W. Heeres, P. Reinhold, N. Ofek, L. Frunzio, L. Jiang, M. H. Devoret, and R. J. Schoelkopf, "Implementing a universal gate set on a logical qubit encoded in an oscillator," Nat. Commun. **8**, 94 (2017).

[111] R. Fisher, F. Helmer, S. J. Glaser, F. Marquardt, and T. Schulte-Herbrüggen, "Optimal control of circuit quantum electrodynamics in one and two dimensions," Phys. Rev. B **81**, 085328 (2010).

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

[113] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," nature **518**, 529–533 (2015).

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

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

[116] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," arXiv preprint arXiv:1509.02971 (2015).

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

[118] T. Fösel, P. Tighineanu, T. Weiss, and F. Marquardt, "Reinforcement learning with neural networks for quantum feedback," Phys. Rev. X **8**, 031084 (2018).

[119] J.-H. Cao, F. Chen, Q. Liu, T.-W. Mao, W.-X. Xu, L.-N. Wu, and L. You, "Detection of entangled states supported by reinforcement learning," Phys. Rev. Lett. **131**, 073201 (2023).

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

[121] K. Reuer, J. Landgraf, T. Fösel, J. O'Sullivan, L. Beltrán, A. Akin, G. J. Norris, A. Remm, M. Kerschbaum, J.-C. Besse, *et al.*, "Realizing a deep reinforcement learning agent discovering real-time feedback control strategies for a quantum system," arXiv preprint arXiv:2210.16715 (2022).

[122] D. Ristè, C. C. Bultink, K. W. Lehnert, and L. DiCarlo, "Feedback control of a solid-state qubit using high-fidelity projective measurement," Phys. Rev. Lett. **109**, 240502 (2012).

[123] J. S. Lundeen and C. Bamber, "Procedure for direct measurement of general quantum states using weak measurement," Phys. Rev. Lett. **108**, 070402 (2012).

[124] D. Yang, S. F. Huelga, and M. B. Plenio, "Efficient information retrieval for sensing via continuous measurement," Phys. Rev. X **13**, 031012 (2023).

[125] V. Giovannetti, S. Lloyd, and L. Maccone, "Quantum metrology," Phys. Rev. Lett. **96**, 010401 (2006).

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

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

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

[129] L. A. Rozema, A. Darabi, D. H. Mahler, A. Hayat, Y. Soudagar, and A. M. Steinberg, "Violation of Heisenberg's measurement-disturbance relationship by weak measurements," Phys. Rev. Lett. **109**, 100404 (2012).

[130] F. Kaneda, S.-Y. Baek, M. Ozawa, and K. Edamatsu, "Experimental test of error-disturbance uncertainty relations by weak measurement," Phys. Rev. Lett. **112**, 020402 (2014).

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

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