npj | Quantum Information

www.nature.com/npjqi

ARTICLE OPEN

# Parameter estimation in quantum sensing based on deep reinforcement learning

Tailong Xiao[1], Jianping Fan[2] and Guihua Zeng[1]✉

Parameter estimation is a pivotal task, where quantum technologies can enhance precision greatly. We investigate the time-dependent parameter estimation based on deep reinforcement learning, where the noise-free and noisy bounds of parameter estimation are derived from a geometrical perspective. We propose a physical-inspired linear time-correlated control ansatz and a general well-defined reward function integrated with the derived bounds to accelerate the network training for fast generating quantum control signals. In the light of the proposed scheme, we validate the performance of time-dependent and time-independent parameter estimation under noise-free and noisy dynamics. In particular, we evaluate the transferability of the scheme when the parameter has a shift from the true parameter. The simulation showcases the robustness and sample efficiency of the scheme and achieves the state-of-the-art performance. Our work highlights the universality and global optimality of deep reinforcement learning over conventional methods in practical parameter estimation of quantum sensing.

*npj Quantum Information* (2022)8:2 ; https://doi.org/10.1038/s41534-021-00513-z

## INTRODUCTION

Precise measurement plays a key role in physics and other sciences which have been widely studied. Parameter estimation is directly better benefited from more precise measurement[1]. Quantum mechanics, fortunately, offers a huge potential advantage toward enhancing the precision of measurement, which naturally spawns a new subject called quantum sensing[2,3]. The basic task of quantum sensing is parameter estimation which has a wide application for imaging[4] and spectroscopy[5].

One of the main goals in quantum sensing is to identify the highest precision of quantum parameter estimation with given resources. Generally, quantum parameter estimation consists of three steps : (1) preparing optimal probe states, (2) experiencing an unknown Hamiltonian evolution, and (3) execute optimal measurement. Numerous seminal works have been concentrated on finding optimal probe state and measurement[6–9]. Recently, increasing researches[10–13] propose to search optimal quantum control signals for step (2). For time-independent Hamiltonian evolution, control-enhanced proposals are proved to be useful to obtain optimal quantum Fisher information matrix (QFIM), especially in multiparameter noncommutative Hamiltonian dynamics[14]. However, parameter estimation in time-dependent quantum evolution is investigated much less. In ref. [15,16], the ideal bound of parameter estimation in time-dependent quantum evolution is derived scaling as $T^4$ ($T$ is the duration of evolution) which is larger than the time-independent case of $T^2$ scaling. This promising result relies heavily on the quantum coherent control, which is not readily implemented in practice. In ref. [17], an experiment of the time-dependent parameter estimation in a simplified physical model is demonstrated.

Optimal control signals are highly crucial to complex quantum sensing situations. Conventional methods for calculating optimal quantum control sequences such as gradient ascent pulse engineering (GRAPE)[18] and chopped random basis (CRAB)[19] performs well in some simple quantum evolutions. However, these methods are sensitive to noise and the calculated pulse shape is hard to engineer. Particularly for some complex evolutions such as time-dependent or multiparameter qubit cases, these methods demand a huge computation cost to converge or sometimes not converging[20]. Machine learning, however, is promising to overcome these shortcomings. In ref. [21–23], traditional machine learning methods are proposed to obtain the feedback control signals. In ref. [24–26], deep reinforcement learning-based methods such as Q learning and policy-gradient network are used to learn the optimal control for gate design or quantum memory. These works demonstrate the potential merits of machine learning in finding optimal control sequences. In previous works[27,28], reinforcement learning approaches are successfully applied to time-independent parameter estimation and achieves impressive results. However, their approaches are not promising in time-dependent quantum parameter estimation and even fail in relatively longer evolutions. Moreover, these approaches require a large number of episodes to converge. These facts limit the application of the reinforcement learning approach in quantum sensing.

In this work, we mainly focus on parameter estimation in the time-dependent Hamiltonian evolution of quantum sensing. We utilize the state-of-the-art deep reinforcement learning (DRL) framework to train the artificial agent to find optimal quantum controls for quantum sensing. We call our proposed protocol DRLQS. Firstly, we present the theoretical bound of QFI of parameter estimation from a geometrical perspective in noise-free and noisy situations. The derived QFI bounds are treated as a variable of the reward function. Then, we provide a unique control ansatz for the DRL agent involving the weak physical prior information about the time-dependent Hamiltonian. Moreover, we also design a general reward function for quantum sensing problems, which is the most important component in our DRL. By conducting rich simulations, the results show that our DRL agent can effectively produce optimal control signals and the precision of time-dependent and time-independent parameter estimation can fast reach the theoretical limit. More importantly, we also

[1]State Key Laboratory of Advanced Optical Communication Systems and Networks, and Center of Quantum Sensing and Information Processing, Shanghai Jiao Tong University, Shanghai 200240, China. [2]Department of Computer Science, University of North Carolina-Charlotte, Charlotte, NC 28223, USA. ✉email: ghzeng@sjtu.edu.cn

Published in partnership with The University of New South Wales

validate the robust performance of DRL-based control under dephasing (DP) noise and spontaneous emission (SE) noise dynamics. Finally, we evaluate the transferability of the DRL method. The simulation results show that our DRLQS protocol performs notably efficiently and has a great potential capability in practical situations.

## RESULTS

### Physical model

Consider a generic time-dependent Hamiltonian interacted with a single spin under quantum control given by

$$\hat{H}_{\text{sen}}(t) = -\hat{H}_g(t) + \hat{H}_c'(t), \tag{1}$$

where $g$ represents unknown parameter, $\hat{H}_c'(t)$ denotes the control Hamiltonian functioned on the unknown Hamiltonian of the targeting system. It is worth pointing out that the optimal Hamiltonian form of quantum coherent control relies on $\hat{H}_g(t)$[10,15]. Besides, the control Hamiltonian must be independent of $g$ since $g$ is not known a-priori and thus no explicit value can be chosen to design the control pulse. The physical model of quantum sensing can be regarded as a quantum sensor as Fig. 1 shows. Particularly, in case that the control Hamiltonian is nonlinear, this system is referred to as quantum chaotic sensor[28,29]. The unitary evolution of the quantum sensor can be simply characterized by the Schrödinger equation when the evolution is noise-free. However, complete isolation of any realistic quantum systems from their environment is not typically feasible. Open quantum systems evolve in a non-unitary fashion, inevitably leading to processes of losses, relaxation, and phase decoherence. For simplicity, we only consider the Markovian dissipative process caused by the qubit spontaneous emission or dephasing in our work. These dissipative processes have no memory effect. Therefore, we can make use of the Lindblad master equation to characterize the evolution of the quantum sensor.

### Time-dependent Hamiltonian parameter estimation

In most quantum sensing problems, the first issue we need to consider is to obtain the QFI of the Hamiltonian parameter estimation. The QFI quantifies the ultimate precision of estimating a parameter from a quantum state over all possible quantum measurements. Then, the next step involves finding the optimal probe states and the optimal measurements. However, it is hard to prepare these optimal probe states and implement these quantum measurements practically. Additionally, when the Hamiltonian of the quantum sensor becomes more complex, the calculation of exact QFI also becomes harder. In our quantum sensor, the Hamiltonian is time-dependent whose calculation of QFI should be distinguished from the time-independent case[30].

Firstly, we consider the noiseless case, i.e., our quantum sensor has no energy dissipation to the environment, no decoherence and relaxations such that we are able to make use of unitary matrices to characterize the system evolution. Suppose the parameter to be estimated is denoted by $g$, the precision of estimating $g$ from a set of parameter-encoded quantum state $\hat{\rho}_g$ is determined by the Bruce distance between $\hat{\rho}_g$ and its neighboring states $\hat{\rho}_{g+dg}$. The relation of QFI and Bruce distance obeys the equation[10]

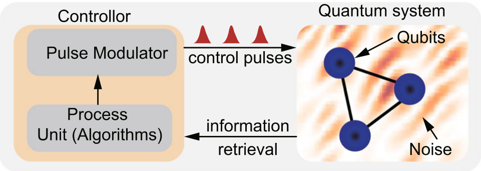

**Fig. 1 Schematic diagram of quantum sensing.** Hamiltonian can be time-dependent or time-independent. Unknown Parameters can be single and multiple. The background noise will cause the the qubit relaxation and dephasing.

$$d_{\text{B}}^2(\hat{\rho}_g, \hat{\rho}_{g+dg}) = \frac{1}{4}\mathcal{F}_g^{(Q)}\,\mathrm{d}g^2, \tag{2}$$

where $d_B(\cdot)$ is the Bruce distance, d$g$ is a small shift from $g$. Through Eq. (2), we can derive that the maximum of QFI for estimating $g$ for time-dependent parameter estimation is given by

$$\max_{\hat{\rho}_0} \mathcal{F}_g^{(Q)} \le \left[\int_0^t \lambda_{\max}(s) - \lambda_{\min}(s)\mathrm{d}s\right]^2, \tag{3}$$

where $\lambda_{\max}(s)(\lambda_{\min}(s))$ is the largest (smallest) eigenvalue of $\partial_g H_g(s)$, $t$ is the total evolution time and $\rho_0$ denotes the probe state. In open noisy quantum evolution, the geometrical framework still works. The maximum QFI for estimating $g$ in noisy situation is given by

$$\max_{\rho_0} \mathcal{F}_g^{(Q)} = \lim_{dg\to 0} \frac{8\left(1 - \int_0^t \max_{\|W\| \le 1} \frac{1}{2}\lambda_{\min}\left[K_W(s) + K_W(s)^\dagger\right] \mathrm{d}s\right)}{dg^2}, \tag{4}$$

where $K_W(s) = \sum_{ij} w_{ij} F_{1i}(s)^\dagger F_{2j}(s)$, $F_{1i}$ and $F_{2j}$ denote the Kraus operators of Kraus evolution $K_g$ and $K_{g+dg}$ respectively. $w_{ij}$ represents the $ij$th entry of $d \times d$ matrix $W$ with $||W|| \le 1$ where $|| \cdot ||$ denotes the operator norm indicating that its largest singular value dose not beyond 1. More derivation details can be seen in Supplementary Note 1.

To saturate the maximum QFI, the optimal quantum control signals are required to steer the evolution of the quantum system. For noise-free evolution, Eq. (3) indicates that if we can prepare the probe state in the superposition of the eigenvectors corresponding to $\lambda_{\max}(s)$ and $\lambda_{\min}(s)$ at $s = 0$ and steer the evolution of the quantum state along the fixed track, we can saturate the optimal QFI. The optimal evolution that corresponds to obtain the maximum QFI gain can be mapped to an evolution according to Schrodinger's equation of unitary propagators $U$ as curves on a manifold $U \in G$, as Fig. 2 shows. The red line represents the steered propagator evolution based on DRL control, which aims to approaches the dotted line. It demonstrates that the quantum control is crucial for time-dependent Hamiltonian estimation to stature the optimal QFI or quantum speed limit in terms of evolution. It is worth pointing out that although quantum controls will not increase maximum QFI, it is necessary to manipulate the quantum evolution and guide the probe state to the right flow of obtaining the maximum information gain. As for the open system, the optimal control signals can be reduced to optimizing a semidefinite programming problem in each time slot

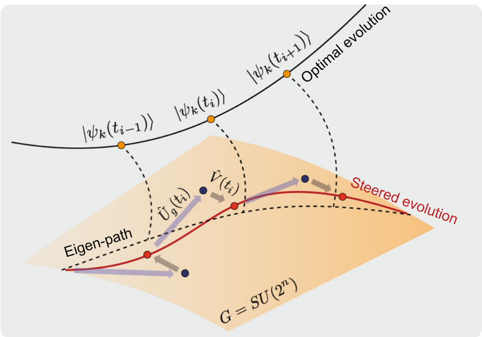

**Fig. 2 Quantum control $\hat{V}(t)$ steers the free evolution $\hat{U}_g(t)$ to evolute along the "eigen-path" approximately, where $|\psi_k\rangle$ denotes the $k$th eigenstate of $\partial_g \hat{H}_{sen}(t)$.** $G$ denotes a Lie group and the manifolds $U, V \in G$. The blue points are the propagator without control and the red points, on the contrary, is the DRL manipulated propagators.

(see Supplementary Note 1). In reality, this optimization for the time-independent case is relatively easier. However, it is impractical for the time-dependent parameter estimation to search for optimal matrix $W$ in each time slot through the convex optimization technique since this optimization process typically becomes a highly non-convex problem in a global sense. Therefore, machine learning becomes the promising method we resort to.

### Deep reinforcement learning for quantum sensing

Before analyzing how DRL has been applied to quantum control, we should present the general ansatz of the control form. Even though we can calculate the optimal coherent form of the quantum control based on the complete knowledge of the Hamiltonian, it is still time-consuming as each time we need to recalculate the control ansatz in different quantum sensing protocols. In particular, we even cannot provide the explicit form when the system Hamiltonian is complicated. One can overcome this issue by an adaptive algorithm but its optimal QFI is reduced compared to the maximum QFI[15]. On the other hand, we also require considering the sufficiency of quantum control ansatz when manipulating the state evolution. Fortunately, we can model a general form in two-dimensional single spin systems as[27,31] $\hat{H}_c^{'}(t)=\sum_j u_j(t)\hat{\sigma}_j$, where $\hat{\sigma}_j, j=1,2,3$ are Pauli $X, Y, Z$ operators, $u_j$ represents the amplitude of the control fields for Pauli operator $j$ which can be assumed to be strong and instantaneous in our model[14]. This control Hamiltonian is sufficient in controlling a Bloch evolution of a single qubit since any $\hat{H}_g(t)$ is composed of the linear combination of Pauli operators.

We note that $\hat{H}_c^{'}(t)$ is designed for time-independent parameter estimation. However, this control form may not be effective in our research. Instead of adopting the optimal coherent control form, we design a general control ansatz for time-dependent Hamiltonian given by

$$\hat{H}_{td}^{'}(t)=\sum_j h_j(t)\hat{\sigma}_j=\sum_j (u_j(t)\cdot f_j(t))\hat{\sigma}_j, \tag{5}$$

where $f_j(t)$ denotes the additional control ansatz which is determined by the physical prior information about the system, i.e., the specific time correlation terms of Pauli operators. The additional control terms should reduce the noncommutativity feature between $\partial_g\hat{H}_g(t)$ and $\hat{H}_g(t)$. More formally, we consider the quantum Hamiltonian of a physical system is denoted as $\hat{H}_g(t)=\sum_{ij} l_j(g_i,z_i(t))\hat{\mathcal{O}}_j$, where $z_i$ denotes the time correlation function. Suppose the unknown parameter and the time function is linearly coupled, i.e., $l_j(g_i, z_i(t)) = l_j(g_i z_i(t))$. The partial derivative over one unknown parameter is

$$\partial_{g_i}\hat{H}_g(t)=\sum_j \partial_{g_i} l_j(g_i z_i(t))z_i(t)\hat{\mathcal{O}}_j, \tag{6}$$

The coefficients $f_j(t)$ should be chosen as $z_i(t)$ for $i$the parameter. The total Hamiltonian can be written as

$$\hat{H}\mathrm{sen}(t)=\sum_{ij} l_j(g_i z_i(t))\hat{\mathcal{O}}_j+\sum_{ij} z_i(t)u_j(t)\hat{\mathcal{O}}_j. \tag{7}$$

In quantum sensing, the partial derivative of the system Hamiltonian over $i$th parameter $g_i$ plays an important role in determining the final precision of the estimation. The chosen coefficient $f_j(t) = z_i(t)$ makes the control form match $\partial_{g_i}\hat{H}_g(t)$ as long as let $u_j(t)=\partial_{g_i} l_j(g_i z_i(t))$. Physically, the chosen explicit function $f_j(t)$ can reduce the noncommutativity feature so that the neural agent only requires learning a relatively simplified function, which is an incidental benefit. Especially in model-free RL, the hardness gap in learning the two functions is amplified since the sample efficiency is notably lower than supervised learning[32]. We note that sample efficiency denotes the number of actions it takes and a number of resulting states and rewards it observes during training in order to reach a certain level of performance. This choice is also potentially beneficial for the neural agent driving the quantum state evolving along the "eigen-path". In case the unknown parameter couples a linear time dependence for which $z_i(t) = ct$ (the linear factor $c$ can be absorbed into $u_j(t)$), then the control ansatz coefficient $f_j(t)$ is written as $f_j(t) = t$. Mathematically, for highly small time slot $\Delta t$, any complex fu nction $h(\Delta t)$ can be first-order Taylor expanded, i.e., $h(\Delta t)=h(t_0)+h'(t_0)(\Delta t-t_0)+o(\Delta t^2)$. Let $t_0 = 0$, we neglect the second and higher-order terms. The constant term $h(0)$ does not affect the actual optimization process in a neural network. Then we have $h(\Delta t)\approx h'(0)\Delta t$, which exactly meets the form of the control ansatz. Compared with $\hat{H}_c^{'}(t)$, we provide an explicit time-dependence for control signals which only requires weak prior information about the quantum sensor. Remarkably, we find that $\hat{H}_c^{'}(t)$ becomes a special case of our proposed ansatz when the unknown parameter is time-independent for which $z_i(t) = 1$. Moreover, our ansatz does not require the exact Hamiltonian expression. On the contrary, the coherent control is constructed based on the complete knowledge of Hamiltonian. Therefore, our DRL control ansatz will be more universal in practical quantum parameter estimation.

DRL has achieved many promising results especially in games such as AlphaGo[33,34], StarCraft II[35] etc. These impressive results boom the development of RL. In RL, states $\mathcal{S}$ are referred to as the position set of the agents at a specific time-step in the environment. Rewards $\mathcal{R}$ are the numerical values that the agent receives on performing some action at some states in the environment. The numerical value can be positive or negative based on the actions of the agent. Thus, whenever an agent performs an action $\mathcal{A}$ the environment provides the agent a reward and a new state where the agent reached by performing the action. The probability that the agent moves from one state to its successor state is called state transition probability obeying the distribution $p(\cdot)$ with which the environment updates the states. $p(\cdot)$ is updated according to the action the agent performed. Notably, a usual MDP is exactly defined that one state moves to another state with transition probability when given an action[36,37]. At the same time, a reward value is also calculated. A MDP can also be described by a tuple $(\mathcal{S},\mathcal{A},p(\cdot),\mathcal{R},\gamma)$, where $\gamma\in(0,1)$ denotes discount rate that balance the importance of current reward and future reward. In RL, the problem to resolve is described as an MDP. Theoretical results in RL rely on the MDP description being a correct match to the problem[38,39]. If the problem is well described as an MDP, then RL may be a good framework to use to find solutions.

A critical task in quantum sensing is to estimate the physical parameter such as frequency, phase of quantum system as precise as possible. In general, a physical system is considered to be a Hamiltonian time evolution, which can be mapped into a MDP. Specifically, a MDP is finite when the sets of $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{R}$ all have a finite number of elements. In this case, the random variables $R_{t_i}, A_{t_i}$ and $S_{t_i}$ have well-defined discrete probability distributions. At time $t_i$, there is a probability of $s'$ and $r$ occures given the preceding state and action:

$$p(s',r|s,a)=\ \Pr\{S_{t_i}=s',R_{t_i}=r|S_{t_{i-1}}=s,A_{t_{i-1}}=a\}. \tag{8}$$

for all $s', s\in\mathcal{S}, r\in\mathcal{R}(s,a)$ and $a\in\mathcal{A}(s)$. In our setup, the totol evolution time $t_i = i\Delta t$, where $i = \{0, 1, \cdots, N\}$ and $\Delta t=\frac{T}{N}$. Formally, the MDP and agent together thereby give rise to a trajectory:

$$\langle S_{t_0},A_{t_0},R_{t_1},S_{t_1},A_{t_2},\cdots,R_{t_N},S_{t_N}\rangle \tag{9}$$

Ultimately, the optimal quantum measurement is executed on the final state to obtain the most precise parameter estimation. The schematic of DRL for quantum parameter estimation is displayed as Fig. 3 shows. DRL aims to maximize the cumulated reward (also called returns in RL) for all time steps. In order to achieve this target, policy $\pi(a|s)$ and state value function $V_\pi(s)$ is introduced. Specifically, $\pi$ is defined as the probability of obtaining one action

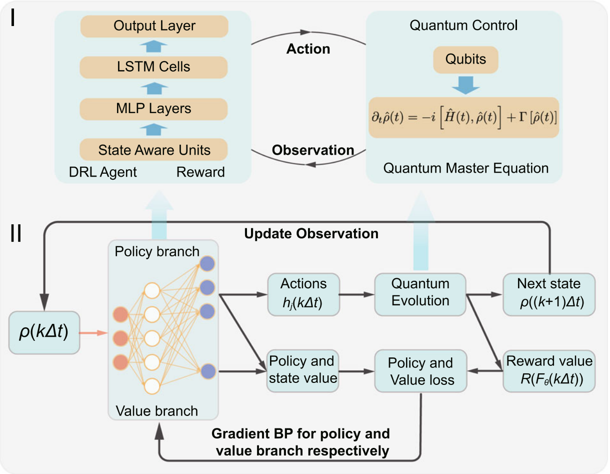

**Fig. 3 Illustration of DRL with (I) agent-environment interaction (II) state-aware policy and value networks with LSTMCells[57] for quantum sensing protocols.** Generally, the quantum evolution can be characterized by quantum master equation both for pure and mixed states. The joint network is divided into the policy/value branch at the final neural layer. The policy and value gradient are updated to the policy and value branch, respectively. The reward $\mathcal{R}$ is a function of the QFI (given by the quantum evolution) which can be calculated by the current control sequence and state. BP refers to backpropagation.

*a* given the current state *s*. The state value $V_\pi(s)$ is defined as the expected cumulated rewards with the discount rate $\gamma$ starting from the current state and going to successor states thereafter, with the policy $\pi$. In DRL, the state value can be calculated by a functional neural network called a value network. The actions can be sampled from the policy $\pi(a|s)$ which is approximated by another functional neural network called policy network. Policy and value networks form a universal architecture for learning interactions with the arbitrary environment. More details can be found in Supplementary Note 4. Moreover, the DRL algorithm has a noble tolerance with noises and action imperfections. Therefore, it will be more suitable for practical and complex quantum sensors.

The DRL state is referred to as the position at a specific time-step in the environment. The quantum state is referred to as the density matrix in open quantum evolution at a time step. In our work, we regard the environment as the quantum evolution (i.e., the environment is quantum), then the DRL state is coincident with the quantum state with the assumption that the DRL agent can be fully aware of the full density matrix of the quantum state, for which we have

$$s_{t_i} = [\Re\{\hat{\rho}_{mn}(t_i)\}, \Im\{\hat{\rho}_{mn}(t_i)\}, m, n \in \{1, 2\}]. \tag{10}$$

The actions of DRL are viewed as the quantum control amplitude array which is denoted by $a_{t_i} = [h_1(t_i), h_2(t_i), h_3(t_i)]$. Each element is used to compose the universal Pauli rotations $\hat{R}_n(\alpha) = \exp\{-i\frac{\alpha}{2}\boldsymbol{n}\cdot\hat{\boldsymbol{\sigma}}\}$, where $\alpha$ being the rotation angle, $n$ denoting a unit vector specifying the rotation axis. Obviously, the general control form is entirely coincided with the general rotation operator per single qubit just by regarding $a_{t_i} = \frac{\alpha}{2}\boldsymbol{n}$. At each time slot, the actions are retrieved from the DRL agent and are used to steer the quantum evolution guiding the quantum state evolving along the "eigen-path" of the system. Therefore, the optimality of the quantum control sequence determines whether parameter estimation is able to reach the maximum QFI. To achieve this goal, it is necessary to offer a well-defined reward function to train the agent to generate optimal actions. The generality and expression as a function of the desired final state are two key features of the reward function. Generality means that the reward function should not implicate the specific information on the characteristics of actions. As for the expression on the desired final state, the goal is to maximize the QFI at the end of quantum evolution. Thus, we define a robust reward function for DRLQS protocols given by

$$r_{t_i} = \begin{cases} \frac{\mathcal{F}^{(Q)}(t_i) - \eta\mathcal{F}^{(Q)}_{\text{nc}}(t_i)}{\mathcal{F}^{(Q)}_{\text{nc}}(t_i)}, & \frac{\mathcal{F}^{(Q)}(t_i)}{\mathcal{F}^{(Q)}_{\text{nc}}(t_i)} < 1 \\ \frac{\mathcal{F}^{(Q)}(t_i) - \zeta\mathcal{F}^{(Q)}_{\max}}{\mathcal{F}^{(Q)}_{\max}}, & \mathcal{F}^{(Q)}_{\text{nc}}(t_i) \leq \mathcal{F}^{(Q)}(t_i) < \mathcal{F}^{(Q)}_{\max}(1-\delta) \\ 1, & \frac{\mathcal{F}^{(Q)}(t_i)}{\mathcal{F}^{(Q)}_{\max}} \geq (1-\delta) \end{cases} \tag{11}$$

for all time step $t_i < N\Delta t$. When the end of the time evolution is reached i.e., $t_i = N\Delta t$, we let $r_{t_i} = r_{t_i} \times C$ to amplify the final reward function which will motivates the agent to emphasize the final state control. In Eq. (11), $\eta, \zeta$ are hyperparameters slightly larger than 1, $\delta$ is a little larger than 0. $\mathcal{F}^{(Q)}_{\text{nc}}$ denotes the QFI without control, $\mathcal{F}^{(Q)}_{\max}$ represents the maximum QFI at the final time. Ideally, the maximum QFI is great larger than the QFI without control. Our design of the reward function firstly aims to ask the agent to provide control signals such that the QFI becomes larger than the QFI without controls. However, QFI that larger than the QFI without control does not indicate the current QFI is optimal. The DRL agent requires approaching the maximum QFI when $\mathcal{F}^{(Q)}(t_i)/\mathcal{F}^{(Q)}_{\text{nc}} \geq 1$ and $\mathcal{F}^{(Q)}(t_i)/\mathcal{F}^{(Q)}_{\max} < 1-\delta$. The reward value is negative both in these two conditions so the agent tries to render the reward value approach to 0. Here we provide a slackness variable $\delta$ aiming to tell the agent that its goal is reached when the QFI is approximately equal to the maximum QFI. The variable $\delta$ can also be used to shrink $\mathcal{F}^{(Q)}_{\max}$ of noise-free case under the noisy conditions. Specifically, in those complex situations that we cannot calculate $\mathcal{F}^{(Q)}_{\max}$ directly, we can adjust $\delta < 1$ since the noisy QFI can be considered as a linear decay of noise-free QFI[12]. Generally, the introduction of the slackness variable will reduce the final target value and make the learning process fast converge. The reward value is set to be 1, indicating that the 'game' is successful during the current episode. It is worth noting that the reward function jumps to a lower value from the first case to the second case. However, the jump does not lead to the sudden drop down of the QFI since parameters of the neural agent do not change suddenly only when a few batches of training data changes which are demonstrated in later simulation results. We set the first stage reward function is mainly to encourage the agent to give controls. The first case large reward will reinforce the control strategy although the agent will be given the second case reward in most episodes. This first case reward is mainly functioned on the time-independent parameter estimation where the gap between $\mathcal{F}^{(Q)}(t_i)$ and $\mathcal{F}^{(Q)}_{\text{nc}}(t_i)$is not large. Additionally, in time-dependent parameter estimation, the QFI is relatively easier larger than the QFI without control because the gap of the scaling with time between them is notably large. The situation should be distinguished from the time-independent case, where both scalings with evolution time have the same order. Thus, the reward function in the time-dependent case requires finer designs to satisfy the effectively training demands. In our implementation, the specific DRL algorithm we used is called A3C LSTM[40,41]. The algorithm details and the reason why we design the reward function of Eq. (11) can be found in Supplementary Note 4. Other interesting RL algorithms that can be applied to quantum controls can refer to[42–44].

## Simulation results

To exemplify the necessity and feasibility of DRL-based quantum control in time-dependent quantum sensor, we consider a single qubit Hamiltonian system of quantum sensor given by[15]:

$$\hat{H}_{\mathrm{sen}}(t) = -A(\cos\omega t\hat{\sigma}_1 + \sin\omega t\hat{\sigma}_3) + \hat{H}'_{td}(t), \tag{12}$$

we first consider estimating the field amplitude $A$. It is easy to verify that the eigenvalues of $\partial_A\hat{H}_{\mathrm{sen}}(t)$ is $\pm 1$ with eigenstates $\left|\psi_{A,1}\right\rangle = \cos\frac{\omega t}{2}|+\rangle + \sin\frac{\omega t}{2}|-\rangle, \left|\psi_{A,-1}\right\rangle = \sin\frac{\omega t}{2}|+\rangle - \cos\frac{\omega t}{2}|-\rangle$, where $|+\rangle = 1/\sqrt{2}(|0\rangle + |1\rangle), |-\rangle = 1/\sqrt{2}(|0\rangle - |1\rangle)$. The optimal probe state, i.e., the superposition of eigenstate corresponding to the largest and smallest eigenvalue, i.e., $|\psi_A(0)\rangle = \frac{1}{\sqrt{2}}(\left|\psi_{A,1}(0)\right\rangle + \left|\psi_{A,-1}(0)\right\rangle)$. The optimal QFI for estimating parameter $A$ within time duration $T$ can be calculated by using Eq. (3) given by

$$\mathcal{F}_A^{(Q)} = 4T^2. \tag{13}$$

When we estimate the field frequency $\omega$, similarly the eigenvalues of the partial derivative of Hamiltonian over $\omega$ is $\pm At$ with eigenstates $\left|\psi_{\omega,+}\right\rangle = \sin\frac{\omega t}{2}|0\rangle + \cos\frac{\omega t}{2}|1\rangle, \left|\psi_{\omega,-}\right\rangle = \cos\frac{\omega t}{2}|0\rangle - \sin\frac{\omega t}{2}|1\rangle$. The optimal probe states can be chosen such as $|\psi_\omega(0)\rangle = \frac{1}{\sqrt{2}}(\left|\psi_{\omega,+}\right\rangle + \left|\psi_{\omega,-}\right\rangle)$. The optimal QFI for estimating $\omega$ can also be calculated by using Eq. (3) given by

$$\mathcal{F}_\omega^{(Q)} = \left[\int_0^T At - (-At)\mathrm{d}t\right]^2 = A^2T^4, \tag{14}$$

where $At$ denotes the largest eigenvalue of $\partial_\omega\hat{H}_{\mathrm{sen}}(t)$ and $-At$ denotes the smallest eigenvalue of $\partial_\omega\hat{H}_{\mathrm{sen}}(t)$. Note that Eqs. (13) and (14) are also known as quantum speed limit (QSL) for time independent and time-dependent parameter estimation, which is an alternative description of Heisenberg uncertainty relation. In case we prepare the optimal probe state (the equal superposition state of largest and smallest eigenstate of $\partial_\theta\hat{H}_{\mathrm{sen}}(t)$), the ultimate state $|\psi(T)\rangle$ will be equal to the probe state. Therefore, the optimal measurement can be chosen as $\hat{\Pi} = \left|\psi_+\right\rangle\left\langle\psi_+\right| - |\psi_-\rangle\langle\psi_-|$ where $|\psi_\pm\rangle = \frac{1}{\sqrt{2}}(|\psi_{\max}(T)\rangle \pm |\psi_{\min}(T)\rangle$ with $\left|\psi_{\max,\min}(T)\right\rangle$ corresponds to the maximum and minimum eigenstates of $\partial_\theta\hat{H}_{\mathrm{sen}}(t)$ at the ultimate time slot. Finally, the best precision of parameter estimation can be obtained. The QFI for estimating $A, \omega$ without quantum control can also be calculated using the rotation frame method. The QFI without control is also important for calculating reward value during training the agent. More details can be seen in Supplementary Note 2.

We first consider the DP noise, therefore the master equation of Eq. (20) preserves the following form

$$\frac{\mathrm{d}\hat{\rho}(t)}{\mathrm{d}t} = -i\left[\hat{H}_{\mathrm{sen}}(t), \hat{\rho}(t)\right] + \frac{\Gamma}{2}[\hat{\sigma}_{\boldsymbol{n}}\hat{\rho}(t)\hat{\sigma}_{\boldsymbol{n}} - \hat{\rho}(t)], \tag{15}$$

where $\Gamma$ is the dephasing rate for qubit $i$. In addition, the dephasing along a general direction is given by $\boldsymbol{n} = (\sin\vartheta\cos\phi, \sin\vartheta\sin\phi, \cos\vartheta)$ and $\hat{\sigma}_{\boldsymbol{n}} = \boldsymbol{n}\cdot\hat{\boldsymbol{\sigma}}^{(i)}$. By choosing specific angles, we are able to obtain the pure parallel and transverse DP noise.

When we consider the SE noise, the evolution can be described by the Lindblad master equation of Eq. (20)

$$\begin{aligned}\frac{\mathrm{d}\hat{\rho}(t)}{\mathrm{d}t} = \; & -i\left[\hat{H}_{\mathrm{sen}}(t), \hat{\rho}(t)\right] + \Gamma^+\left[\hat{\sigma}_+\hat{\rho}(t)\hat{\sigma}_- - \tfrac{1}{2}\{\hat{\sigma}_-\hat{\sigma}_+, \hat{\rho}(t)\}\right] \\ & + \Gamma^-\left[\hat{\sigma}_-\hat{\rho}(t)\hat{\sigma}_+ - \tfrac{1}{2}\{\hat{\sigma}_+\hat{\sigma}_-, \hat{\rho}(t)\}\right],\end{aligned} \tag{16}$$

where $\hat{\sigma}_\pm = (\hat{\sigma}_1 \pm i\hat{\sigma}_2)/2$ are the ladder operators for spins, $\Gamma^\pm$ are the qubit SE relaxation rate.

In noisy cases, the optimal QFI can be calculated by optimizing Eq. (4). This optimization process will provide optimal control signals. Since each time slot optimization of the matrix, $W$ involves semidefinite programming, which in practice the conventional methods are hard to operate. In this work, the QFI with control signals in different time slots will be carried out numerically with Eq. (2) such that the reward function can be calculated. The QFI without control signals is derived by approximately solving the Lindblad master equation. More calculating details can be seen in Supplementary Note 3.

## Noise-free results

In the context of our quantum sensor, there are two parameters to be estimated. However, these two parameters are not able to be estimated simultaneously. In the following, we will estimate $A$ and $\omega$ separately to illustrate the performance of our DRL-based quantum control. We first estimate the time-independent parameter $A$ whose similar investigation can be found in ref. [27]. The optimal probe state is $\hat{\rho}_0 = |1\rangle\langle 1|$. In contrast, our DRL algorithm adds the LSTM cell, and the reward function is more refined in terms of the controller. The simulation results of estimating $A$ are shown in Fig. 4, and the control ansatz adopts a time-independent form. From the Bloch evolution Fig. 4a-c, the final state under explicit control[15] and our DRL control is in a similar position, which demonstrates DRL control is feasible in producing optimal control signals. However, in case there is no control, the final state is far away from the optimal position. More precisely, we have shown the QFI of the final time slot for each episode in Fig. 4d, we can see that the QFI is fast approaching the QSL with nearly 100 episodes. In Fig. 4d, Xu's proposal[27] (here is added with LSTM layer) also showcases a fast convergence. We note that Xu's proposal is equivalent to our reward function in the time-independent case since the maximum QFI is in the same order with the QFI without control. They can be adjusted to be equal by tuning the hyperparameters $\eta, \zeta$ in the reward function. However, the learning curve generated by A3C is not stable compared with A3C LSTM and the learned QFI drops down to a lower value easily. The cross-entropy RL (CERL) method (see Supplementary Note 5, similar to Schuff's proposal[28]) is a baseline method that performs well in Tetris game[45]. Here we make a comparison with the CERL method under the same physical system for estimating $A$. The results indicate that the CERL method cannot converge into the QSL. The generated QFI is even lower than the QFI without control since the initial controls are not small and render the probe state deviate from the optimal evolution path. The results imply that the CERL method cannot show competitive performance with A3C LSTM. Here we do not benchmark the performance with conventional GRAPE algorithm since GRAPE has been demonstrated to work well in time-independent algorithm although its time complexity is relatively higher and the transferability is much lower[27]. We also find that the optimal controls are not unique, i.e., our DRL control is not the same as the explicit control but is still able to obtain the optimal QFI as Fig. 4e shows. Figure 4f demonstrates our DRL algorithm is capable of learning optimal control signals for different time durations by choosing appropriate hyper-parameters.

We then estimate the time-dependent parameter $\omega$. The optimal probe state is $\hat{\rho}_0 = |+\rangle\langle +|$. The control ansatz is adopted as a linear time correlation form given by Eq. (5). The simulation results are displayed in Fig. 5. We also plot the Bloch sphere evolution of the qubit under DRL control, no control, and coherent control respectively, as Fig. 5a-c shows. The visualization of qubit evolution presents a direct sense of how the control signals manipulate the quantum evolution. We can find that the final qubit positions under our DRL control and optimal coherent control are the same. Moreover, the QFI value at each final time slot implies that our DRL control can approach the QSL fast as Fig. 5d shows even in time-dependent cases (200 episodes). However, the QFI curve with the coherent control form learns relatively slow. The learning capability of the agent is restricted by the coherent control form when $T$ is large. One reasonable explanation is that the coherent control form corresponds to a theoretical result that assumes the time slot is infinitely small. However, in our

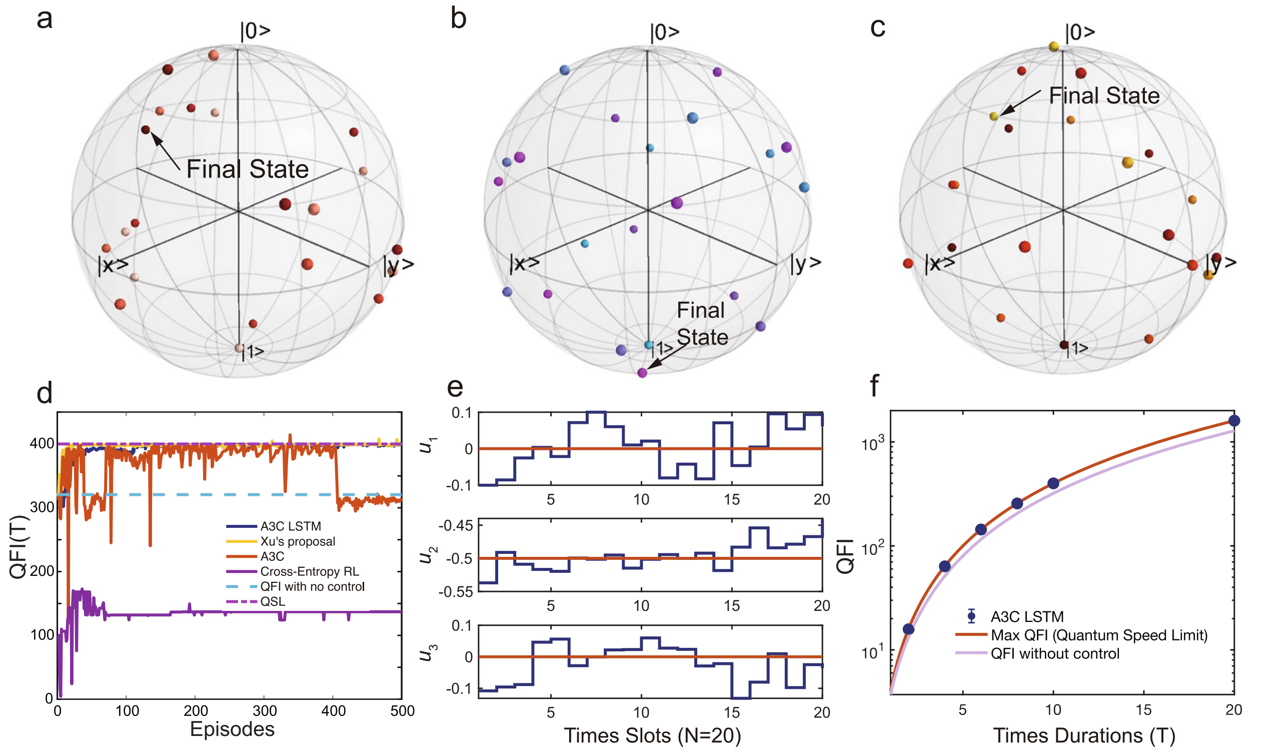

**Fig. 4 Simulation results of field amplitude estimation where $A = 1, \omega = 1$.** (**a**–**c**) are the Bloch sphere evolution of the qubit where (**a**) the quantum state under the DRL-optimized controlled sequences, (**b**) is none control case and (**c**) is the coherent optimal control in ref. [15]. The dark color dot represents the final evolution of the qubit. **d** displays the learning procedure of QFI varied with each episode under different proposals where $T = 10$, $\Delta t = 0.5$ for A3C LSTM, A3C and Xu's proposal, elite ratio is 10% and number of agents is 50 for Cross-Entropy method. **e** shows the optimized control signals over all episodes produced by DRL (dark blue) and explicit coherent control strategy (orange). **f** is result of QFI with different $T$ under DRL control signals where $\Delta t = 1$ for $T = 20$ and $\Delta t = 0.1$ for other $T$. Error bar is the standard error of the mean.

simulation, the time slot cannot be too small to keep an efficient training process. Besides, the coherent control form is likely to limit the imagination of the DRL agent. Also, the optimal control form is not the only alternative as Fig. 5e shows. The agent can select any control sequences as long as the maximum QFI can be acquired. The coherent control form for the DRL agent excludes all possible control sequences. Thus, this coherent control form will cause a serious decline in learning speed, especially for a long evolution time. However, we have shown in Supplementary Note 5 that when $T$ is small, both two control ansatz learn fast and well. In our protocol, the controls are bounded within the range $[-4, 4]$. In fact, the actions are unlikely beyond such constraints. From the coherent Hamiltonian control form, the maximum and minimum control amplitudes are $\pm 1$. However, this optimal control form has two demerits: (1) require knowing the full Hamiltonian knowledge, (2) the practical performance is not good as the purple line shows in Fig. 5(d) of the main text since it limits the "imagination ability" of the neural agent as we have argued. In contrast, the simplified linear-time ansatz is beneficial for practical training performance. The control pulses during the early phase are suppressed by the function $f_j(t) = t$ because $t$ is small during this phase. We consider that the large control in the early phase will render the evolution of the probe state deviate from the optimal evolution path ("eigen-path"). In addition, we also evaluate Xu's proposal, CERL method, and our proposal with no LSTM. From the simulation results shown in Fig. 5d, we find that the CERL method and Xu's proposal cannot find the optimal controls to maximize the final QFI within 600 episodes. Xu's reward design cannot work well in time-dependent parameter estimation since the scaling gap between the maximum QFI and QFI without control is highly large such that the agent will be given relatively a good reward even when the QFI is smaller than the QSL. We also evaluate the performance of simplified reward i.e., the difference of two successive QFIs. The curves are not presented in Fig. 5d and details can be found in Supplementary Note 5 (Reward Comparisons). The CERL method also cannot work efficiently in our time-dependent parameter estimation. The QFI learning curve with no LSTM is highly unstable but it can still approach the QSL, which is similar to the results of estimating $A$. LSTM evaluates the performance over the history observations which can increase the stability of the learning curve. More discussions about the comparisons can be found in Supplementary Note 5. In Fig. 5f, we present the DRL-enhanced QFI with different time durations and the simulation result is well coincident with the theoretical results.

In order to benchmark the performance of the GRAPE algorithm in time-dependent parameter estimation, we design two types of gradient-based quantum control optimization algorithms. The core component of the GRAPE algorithm is to calculate the gradient of QFI with respect to control pulses. The two typical GRAPE algorithms can be summarized as follows:

(1) Discretize the whole time evolution into small pieces $\delta t$, and each time slot evolution can be regarded as the approximate time-independent evolution. Then, the GRAPE algorithm is applied in each time slot to optimize the control pulses. The final quantum state in each time slot is viewed as the next probe state of the evolution. Although it can work normally by using the results derived in[14], the ideal QFI realized by this 'sequential' GRAPE can only be $\sim T^3$ scaling (detailed demonstrations can be found in Supplementary Note 5). Here we also assume that the quantum state manipulated by the GRAPE control is the optimal probe state for the next time slot evolution. However, this assumption is highly possible to be not practical.

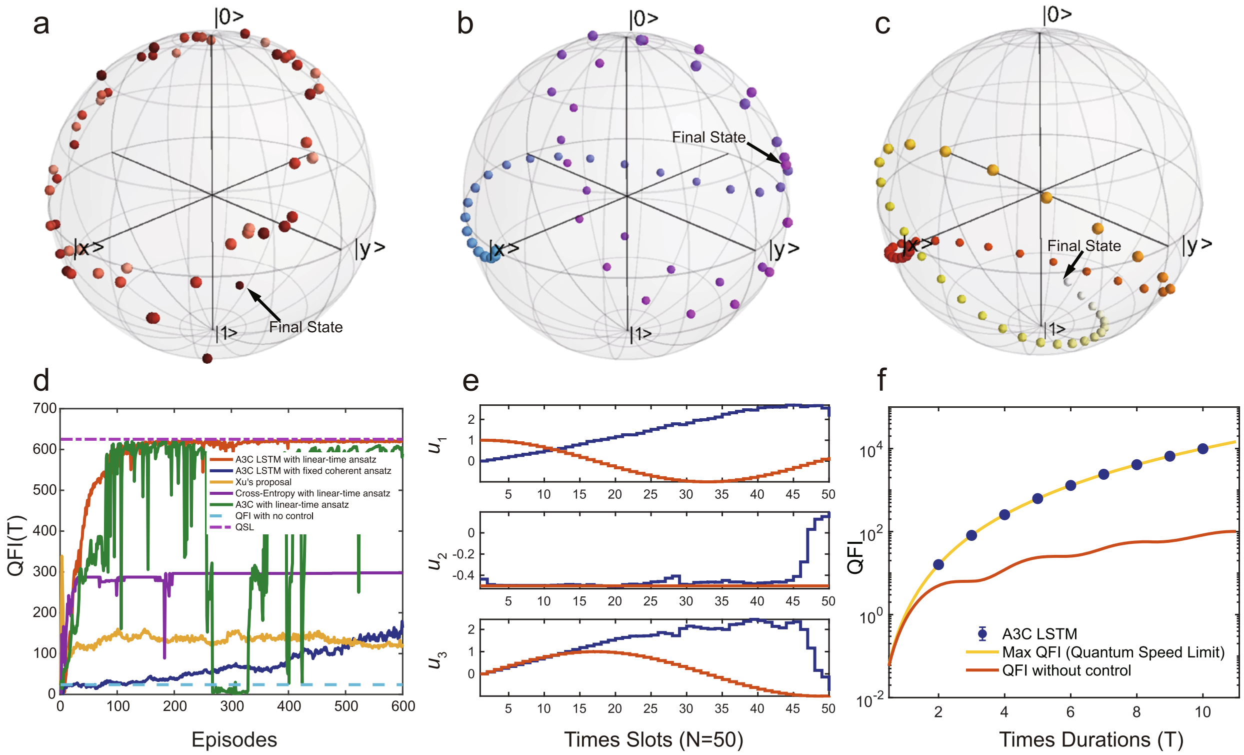

**Fig. 5 Simulation results of quantum field frequency estimation where $A = 1, \omega = 1$. a–c** are Bloch sphere visualization of the qubit evolution where $T = 5, \Delta t = 0.1$. The darkest color dot represent the final evolution of the qubit. **a** denotes evolution with the DRL-based control, **b** denotes no controls are applied, and (**c**) denotes the optimal coherent control (**d**) displays the QFI learning curve at the final time under different schemes varied with episodes where $T = 5, \Delta t = 0.1$ for A3C LSTM, A3C and Xu's proposal, elite ratio is 10% and number of agents is 50 for Cross-Entropy method. **e** plots the optimized control signals produced by DRL and explicit strategy. **f** are the results QFI with different time durations where $\Delta t = 0.5$ for $T = 10$ and $\Delta t = 0.1$ for other $T$. Error bar is the standard error of the mean.

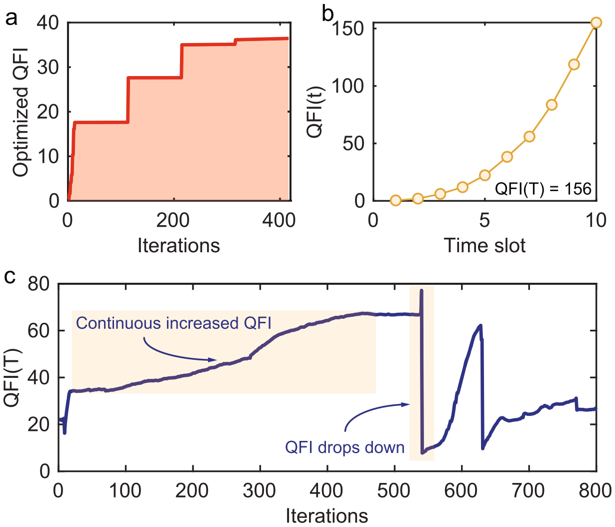

**Fig. 6 The GRAPE algorithm for optimizing quantum control pulses (box shape) in time-dependent parameter estimation. a** denotes the optimized QFI curve in each iteration, (**b**) shows the cumulated QFI varied with the time slot. (**a**) and (**b**) display the simulation results of the first implementation of GRAPE algorithm. **c** represents the final time QFI varied with the optimization iterations. Simulation parameters: $A = 1, \omega = 1, T = 5, \delta t = 0.5, \Delta t = 0.05, \epsilon = 10^{-4}$ for (**a**, **b**), $A = 1, \omega = 1, T = 5, \Delta t = 0.25, \epsilon = 10^{-4}$ for (**c**).

(2) Still adopt the essence of the GRAPE algorithm, i.e. calculating the gradient of the final QFI over the control pulses. The numerical calculation can be implemented based on the first-order numerical differential equation, which is also called the parameter shift rule in numerous variational quantum algorithms[46,47]. The feasibility of this implementation is beneficial from the geometrical perspective of our derivation in calculating the upper bound of the QFI. Then we update the control pulses globally based on the gradients.

Figure 6 a, b display the performance of the first implementation of GRAPE algorithm and the final maximum QFI is equal to 156. The ideal QFI controlled by the 'sequential' GRAPE algorithm is at most $\int_0^T 4t^2\,\mathrm{d}t = \frac{4}{3}T^3$. When $T = 5$, the maximum QFI is approximately equal to 167. Thus, the 'sequential' GRAPE algorithm can reach the suboptimal scaling correctly. We note the overhead of this algorithm is larger compared with the original implementation in[14] since during each time slot, we should execute the GRAPE independently. If the time is divided into smaller pieces, the overhead will be larger but the realized QFI will approach ~ $T^3$ closer. Thus, we can conclude that this 'sequential' GRAPE implementation is a suboptimal algorithm that cannot find the optimal controls to reach the quantum speed limit.

The second implementation of the GRAPE algorithm is more naive, but still obeys the essence of the gradient-based principle. From Fig. 6c, we see that the QFI increases in the early stage of the iteration. However, the QFI enters into a stable area (between two yellow boxes) and does not increase. Then the QFI drops down suddenly and we guess that the gradient escapes from a position similar to the saddle point and enters into a much smaller QFI landscape. We conclude that the scheme of the straightforward gradient updating cannot work well in time-dependent parameter

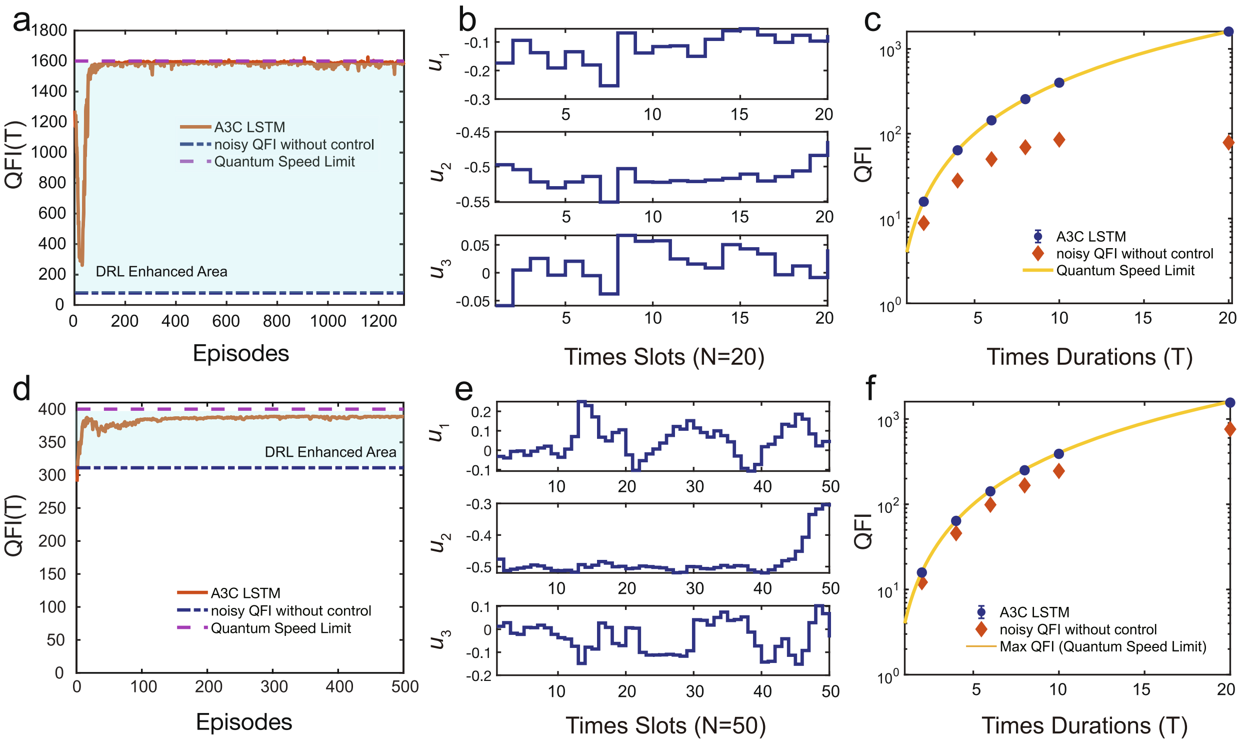

**Fig. 7 Experiment results of estimating $A$ ($A = 1, \omega = 1$). a–c** are for DP noise where $\gamma = 0.1$. **a** displays the QFI at the end evolution ($T = 20$, $\Delta t = 1$) in each episode. (**b**) and (**e**) are the optimized control signals that DRL has learned over all episodes with time-dependent control ansatz. **c** shows the optimized QFIs of different time durations with the DRL control framework. **d–f** similarly are for SE noise. **d** shows the learning procedure of final QFI in each episode where $T = 10$, $\Delta t = 0.2$, $\gamma^{+} = \gamma^{-} = 0.01$. **f** demonstrates the QFI of estimating $A$ with different time durations. Error bar is the standard error of the mean. Error bar is the standard error of the mean.

estimation. But it becomes equivalent to the first case of optimization when the parameter is time-independent except that the gradient is numerical, not analytical.

Based on the simulation results, we find that our proposal showcases the competitive performance over Xu's proposal in estimating time-independent parameter estimation. More significantly, in time-dependent parameter estimation, our proposed protocol is still sample-efficient and optimal in approaching the QSL but previous RL methods and the GRAPE algorithm cannot optimize control pulses to increasingly approach the QSL.

## Noisy dynamics results

In noise-free case, we have evaluated the performance of conventional GRAPE and other previous RL methods to clarify the superiority of our protocol. In noisy dynamics, we only conduct the simulation of our proposal for simplicity. We firstly estimate the field amplitude $A$ under two noisy dynamics respectively. In noisy dynamics, the optimal probe state of estimating $A$ and $\omega$ is the same with the noise-free case. In DP noise, we set $\vartheta = \pi/4, \phi = 0$ which indicates the dephasing are not parallel or vertical. The simulation outcomes are shown in Fig. 7a–c. We can see that the noisy QFI without control shrinks quickly compared to QFI with DRL control. Moreover, the convergence of DRL is highly fast, demonstrating our linear time-correlated control form is efficient in manipulating quantum states. It is important to remark that the noisy QFI also reaches QSL under DRL control rather than a reduced QFI[6]. It can be clarified that quantum control signals compensate for the dissipation of the system and render the qubit remains along the predefined "eigen-path". The control pulses are displayed in Fig. 7b, where we do not present the coherent control pulses as they are not optimal in noisy dynamics although they might be useful in enhancing the precision of parameter estimation. In principle, any quantum control signals might be beneficial since they can to some extent protect the system from dissipation. This clarification can be validated from Fig. 7a where random control signals generated by DRL can obtain a higher QFI than the case with no control. However, random signals cannot saturate the QSL and still require the training procedure. In Fig. 7c, the QFI of different time durations controlled by DRL can perfectly saturate QSL. In contrast, the QFI without control decreases rapidly with the increase of evolution time.

When considering the SE noise in estimating $A$, the results are displayed in Fig. 7d–f. We notice that the final QFI does not entirely saturate the QSL but the gap can be ignored in case more training episodes are given. The effect of SE noise is stronger than DP noise in terms of decelerating DRL training. The set of QFI values with different time duration are also presented in Fig. 7f, where we can find that the DRL-controlled QFI can perfectly saturate the QSL compared with the QFI with no control.

When estimating $\omega$ under DP noise, the results are displayed in Fig. 8a–c. The little gap between QSL and our QFI implies the DP noise effect cannot be eliminated. In contrast, QFI without control decreases dramatically compared to the noise-free case. Similar results can be seen under SE noise. It demonstrates that the time-dependent parameter estimation is more sensitive than the time-independent case. We note that our DRL agent can saturate QSL with small $T$ under noisy dynamics, which can be verified in Fig. 8c–f. Besides, we could find the SE noise is harder to be overcome than DP noise. More significantly, these results demonstrate that the potential capabilities of DRL in quantum control as which can learn the noise feature and generate proper signals to eliminate the noise effect. We also validate the performance of the trivial control ansatz in SE noise (see Supplementary Note 5), the

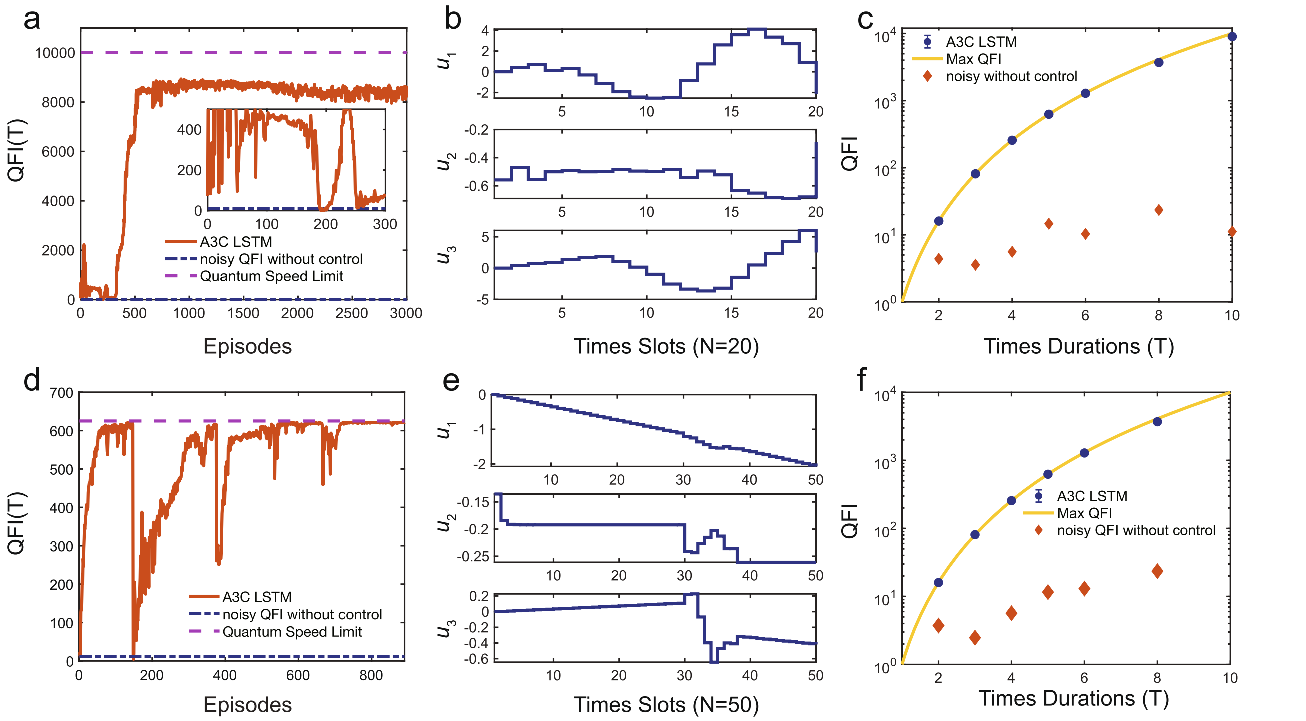

**Fig. 8** **Results of estimating $\omega$ under noisy dynamics with parameters $A = 1, \omega = 1$.** The first row is for DP noise where $T = 10, \Delta t = 0.5, \gamma = 0.1$ and the second row is for SE noise where $T = 5, \Delta t = 0.1, \gamma_+ = 0.1, \gamma_- = 0$. **a** and **d** shows the learning process. **b** and **e** show the optimized control signals over all episodes. **c** and **f** display the QFI value with different time durations. Error bar is the standard error of the mean.

learning procedure is not stable and its QFI does not saturate the QSL with $T = 5$. As a consequence, we can conclude that the linear-time-correlated quantum control can accelerate the convergence speed of DRL in the time-dependent parameter estimation. We speculate that even if LSTM neurons are added to our network[31], the network still seems unable to capture the time dependence of quantum control. We infer that in the time-independent evolution, the time correlation of quantum control can be well solved by simple LSTM neurons. However, in the evolution of a time-dependent quantum system, there is a "double time-dependent relationship" in quantum control, which requires a large number of training samples to capture for the concise LSTM unit, thus reducing the training efficiency of the network. Therefore, we try to add a prior linear-time coupling to the quantum control signals. The behind physical intuition of why this linear-time relation is effective stems from the reduction of the noncommutativity of the Hamiltonian. Similar to human learning, when agents are told the direction and purpose of the learning process, they can give full play to its learning initiative. On the contrary, the overly complex prior information of time relation will limit the active learning process of agents leading to a longer time convergence. The protocol of deep learning with certain physical prior information is well studied in the theory and experiment of quantum control in ref. [48].

## Transferability analysis

The transferability of the parameter estimation algorithm can measure its efficiency and robustness[27]. Here, we analyze the transferability of our DRLQS protocol in estimating $A$ and $\omega$ only with noisy conditions as the noise-free conditions are not such realistic in practical quantum parameter estimation of which exactly conflicts with the intention of transferability analysis. From Fig. 9a, b, the QFI nearly keeps invariant to saturate QSL when $A$ shifts from $[0, 4]$ both for DP and SE noise. These simulation results are in line with our expectations since the amplitude is a linear parameter in our Hamiltonian. The linear relation has no influence on neural networks in generating optimal control signals because the network is also the combination of all linear relationships. More generally speaking, there are a large amount of linear-relation time-independent parameters in quantum sensing that can behave a highly impressive transferability by using our DRLQS protocol.

When discussing the transferability of the frequency $\omega$, the results are not notably impressive compared to the time-independent case. Figure 9c, d displays the QFI with different shifts from the true parameter. When $\omega$ has a $\pm 0.1$ deviation from $\omega_0 = 1$, the performance (QFI value) still stays at a high level. The QFI value decreases relatively large when the deviation is greater than 0.3. We note that when $\omega = 0.5$, the QFI value occurs to rebound. The transferability of the time-dependent case is not as impressive as the time-independent case. There are two possibilities: (1) the parameter $\omega$ does not own a linear-relation with the Hamiltonian and (2) a large deviation of frequency will lead to the huge difference of the evolution. While the input of the DRL algorithm varies greatly, the control signals will not be effective in controlling quantum evolution. However, DRL is not entirely useless or causes a much smaller QFI value compared with noisy QFI without control. In reality, we can also use the DRLQS protocol to roughly estimate the shifted time-dependent parameter although sometimes we cannot obtain the optimal QFI in case of large parameter deviation. This result also verifies that the time-dependent parameter is highly sensitive to noise that may lead to a large shift of parameters. Thus, the transferability of time-dependent parameter estimation in quantum sensing should be paid much more attention.

## DISCUSSION

In summary, we have systematically explored the capability of using DRL to generate robust and optimal control signals for

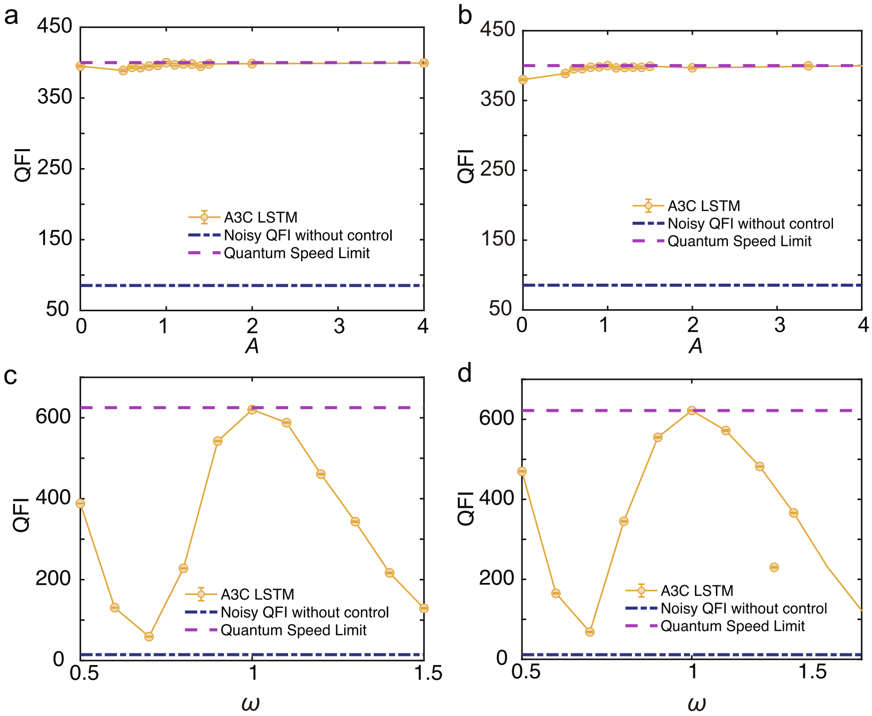

**Fig. 9 Transferability results of DRL in terms of estimating *A*, *ω*.** DRL is trained with $A=1,\omega=1$ and the performance is validated on a shift from a broad range of true value. **a**, **b** are the results of estimating $g$ where $T=10,\Delta t=0.5,\Gamma=0.1$ and **c**, **d** are for estimating $\omega$ where $T=5$, $\Delta t=0.1,\Gamma_{+}=0.1,\Gamma_{-}=0$. Error bar is the standard error of the mean.

quantum sensing, especially in time-dependent parameter estimation. We have presented a new theoretical derivation of QFI under unitary and open evolution from a geometrical perspective. We have also offered a detailed calculation of QFI without control in the open dynamics through approximately solving the Lindblad master equation. The derived bounds and the noisy QFI without control are useful for calculating the reward function which directly determines DRL's performance in generating quantum control signals. The main challenge of DRL for quantum parameter estimation is low efficiency which has been greatly improved by designing a time-correlated control ansatz and a notably instructive reward function. Besides, we add the LSTMCell into the DRL algorithm to learn the relations of each successive quantum state to further enhance the stability of the learning process. By conducting plenty of simulations, our results demonstrate that DRL is capable of controlling quantum sensors perfectly and the QSL of the time-dependent and time-independent cases can be saturated both for noise-free and noisy dynamics. More significantly, our trained DRL algorithm showcases the transferability in controlling quantum sensors when the actual parameter deviates from a broad range of true parameters, especially for time-independent parameter estimation. Compared to previous RL proposals and conventional GRAPE algorithm, our DRLQS protocol exhibits better performance in terms of universality, sample efficiency, and the ability to approach the QSL, particularly in time-dependent parameter estimation.

We remark that the DRL agent is trained based on the full density matrix elements. This assumption requires the state tomography technique in practical quantum sensors. Recently, the classical shadow[49] of quantum state may be beneficial for obviating this issue. The classical shadow does not require the full tomography of quantum state and can also be applied to calculate QFI[50]. Therefore, the DRL agent may be fed with the classical shadow to train the agent in practice. In addition, the generative neural quantum state[51] can also be incorporated into the neural agent to alleviate the issue by only using a polynomial number of measurements. These techniques help deal with the density matrix assumption in practice. When the agent being well-trained, it can be referred to as the "coarse-grained" pulses which are highly instructive in improving the precision of the quantum sensor. Then we can calibrate them in a closed-loop based on the practical measurement through a simple optimization algorithm such as the Nelder-Mead method[52,53]. Also, we can adopt the teacher-student network to make the well-trained agent work in practical sensing. Let the well-trained agent (policy networks) as the teacher network, and the simplified feedforward networks fed with the practical measurements as the student network. Then the student network does not need the full density matrix and generates the control pulses with the help of the teacher networks[25]. The proposed protocol can also be applied in a quantum sensor network combined with the hybrid quantum-classical architecture[54,55].

In addition, the DRLQS protocol can be easily extended to multi-qubit dynamics by designing information complete control Hamiltonian and keeping the full density matrix states and reward function unchanged. Therefore, our investigation suggests that DRL-based quantum control is highly universal and achieves the state-of-the-art performance in practical time-dependent quantum sensing protocols compared to conventional methods and previous RL works. In future work, we will concentrate on multiparameter estimation in time-dependent quantum systems to further exploit the capabilities of the DRLQS protocol.

## METHODS

### Characterizing quantum evolution for quantum sensors

For noise-free case, the time evolution operator can be represented with unitary matrices given interrogation time $T$,

$$\hat{U}(0\to T)=\mathcal{T}\exp\left\{-i\int_0^T \hat{H}_{\rm sen}(t)\mathrm{d}t\right\}, \tag{17}$$

where $\mathcal{T}$ denotes time-order operator. However, this integration is complex and it is extremely hard to calculate the ultimate analytical solution in a time-dependent Hamiltonian evolution. Generally, we discrete the continuous-time evolution into small timepieces $\Delta t = T/N$. When $\Delta t$ is small enough, the evolution can be regarded as time-independent, i.e.,

$$\hat{U}(k\Delta t \rightarrow (k+1)\Delta t) \approx \exp\{-i\hat{H}_{\text{sen}}(k\Delta t)\Delta t\}, \tag{18}$$

where $k \in \{0, 1 \cdots, N-1\}$ denotes the time slot during the evolution. The potential assumption here is that the quantum control is able to operate the quantum sensor instantaneously. For pure state, the unitary time evolution between $k$th and $(k+1)$th time slot is given by

$$\begin{aligned} |\psi((k+1)\Delta t)\rangle &= \hat{U}(k\Delta t \rightarrow (k+1)\Delta t)|\psi(k\Delta t)\rangle \\ &= \exp\{-i\hat{H}_{\text{sen}}(k\Delta t)\Delta t\}|\psi(k\Delta t)\rangle. \end{aligned} \tag{19}$$

For noisy evolution, we use Lindblad master equation to characterize the dynamics given by

$$\frac{\mathrm{d}\,\hat{\rho}(t)}{\mathrm{d}\,t} = \hat{\mathcal{L}}_t[\hat{\rho}(t)], \tag{20}$$

where $\hat{\rho}$ denotes the density matrix of the quantum state in the system, $\hat{\mathcal{L}}_t[\circ]$ is a superoperator called Lindbladian given by[6,56]

$$\hat{\mathcal{L}}_t[\circ] = -i[\hat{H}_{\text{sen}}(t), \circ] + \sum_i \eta_i(t)(\hat{A}_i(t) \circ \hat{A}_i^\dagger(t) - 1/2\{\hat{A}_i^\dagger(t)\hat{A}_i(t), \circ\}),$$

where $i$ denotes the number of noisy quantum channels, $\hat{A}_i(t)$ denotes noise operators. We assume that the quantum Hamiltonian controls are capable of operating the qubit with unknown parameters. The coupling spin qubit and the environment noise operators will not be affected by the control fields[11,14]. As a result, the time evolution of the quantum sensor under noisy quantum environment is given by

$$\hat{\rho}(T) = \mathcal{T}\exp\left\{\int_0^T \hat{\mathcal{L}}_t \,\mathrm{d}\,t\right\}\hat{\rho}(0). \tag{21}$$

We have assumed that $\hbar = 1$. In time-dependent Markovian evolution, it turns out that $\eta_i \geq 0\,\forall\, i, t$. However, if quite a few $\eta_i < 0$ for some time slots, the associated dynamics would be non-Markovian which is beyond the scope of our work. Analogously, to implement the control fields to the coupling system, it is still required to discretize the continuous-time into $N$ small time slots. Therefore, the evolution of the density matrix from $k\Delta t$ to $(k+1)\Delta t$ is given by $\rho((k+1)\Delta t) \approx \exp\{\hat{\mathcal{L}}_k \Delta t\}\hat{\rho}(k\Delta t)$, where $\hat{\mathcal{L}}_k$ denotes the Lindbladian at $k$th time slot. Thus, the quantum controls can be applied to each time slot as constants to steer the evolution to achieve the optimal estimation precision.

### Neural network and software specifications

Our DRL is composed of five neural layers and each layer is followed by a Leaky ReLU unit to render activation. The dimension of the input units is eight, which is determined by the full tomography of a single qubit. The dimension of output units is three, which acts as three control signals for the quantum sensor. The software is coded by python language and the deep learning package Pytorch is used to construct and train our neural networks. The quantum evolution is simulated in Qutip environment of which is a prominent integrated package used for quantum mechanics. Also, we validate the performance of traditional methods such as GRAPE and CRAB algorithms in Qutip. We find that these conventional algorithms are easily stuck into the local minima especially in time-dependent quantum evolution even in naive implementations. Since the benchmarking comparison requires the gradient optimization of QFI for control signals which is highly complex and beyond the scope of our work. We mainly aim to demonstrate that conventional algorithms are not stable and universal which are dependent on the specific mathematical derivations. These shortcomings limit their availability. However, DRL can exactly overcome these shortcomings. Thus, the superiority of DRLQS is demonstrated. More quantitive studies of conventional algorithms on generating quantum controls can be found in Supplementary Note 5. The specific model parameters for neural network and qubit simulations can be found in Supplementary Note 6.

## DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Received: 14 August 2020; Accepted: 3 December 2021;
Published online: 10 January 2022

## REFERENCES

1. Helstrom, C. W. *Quantum detection and estimation theory* (Academic press, 1976).
2. Giovannetti, V., Lloyd, S. & Maccone, L. Quantum metrology. *Phys. Rev. Lett.* **96**, 010401 (2006).
3. Giovannetti, V., Lloyd, S. & Maccone, L. Advances in quantum metrology. *Nat. Photon.* **5**, 222–229 (2011).
4. Brida, G., Genovese, M. & Berchera, I. R. Experimental realization of sub-shot-noise quantum imaging. *Nat. Photon.* **4**, 227–230 (2010).
5. Kira, M., Koch, S. W., Smith, R. P., Hunter, A. E. & Cundiff, S. T. Quantum spectroscopy with Schrödinger-cat states. *Nat. Phys.* **7**, 799–804 (2011).
6. Tsang, M. Quantum metrology with open dynamical systems. *N. J. Phys.* **15**, 073005 (2013).
7. Pinel, O., Jian, P., Treps, N., Fabre, C. & Braun, D. Quantum parameter estimation using general single-mode Gaussian states. *Phys. Rev. A* **88**, 040102 (2013).
8. Alipour, S., Mehboudi, M. & Rezakhani, A. Quantum metrology in open systems: dissipative Cramér-Rao bound. *Phy. Rev. Lett.* **112**, 120405 (2014).
9. Brask, J. B., Chaves, R. & Kołodyński, J. Improved Quantum Magnetometry beyond the Standard Quantum Limit. *Phys. Rev. X* **5**, 031010 (2015).
10. Yuan, H. & Fung, C.-H. F. Optimal Feedback Scheme and Universal Time Scaling for Hamiltonian Parameter Estimation. *Phys. Rev. Lett.* **115**, 110401 (2015).
11. Liu, J. & Yuan, H. Control-enhanced multiparameter quantum estimation. *Phys. Rev. A* **96**, 042114 (2017).
12. Yuan, H. & Fung, C.-H. F. Fidelity and Fisher information on quantum channels. *N. J. Phys.* **19**, 113039 (2017).
13. Yuan, H. & Fung, C.-H. F. Quantum parameter estimation with general dynamics. *npj Quantum Inf.* **3**, 1–6 (2017).
14. Liu, J. & Yuan, H. Quantum parameter estimation with optimal control. *Phys. Rev. A* **96**, 012117 (2017).
15. Pang, S. & Jordan, A. N. Optimal adaptive control for quantum metrology with time- dependent Hamiltonians. *Nat. Commun.* **8**, 1–9 (2017).
16. Fiderer, L. J., Fraïsse, J. M. & Braun, D. Maximal quantum Fisher information for mixed states. *Phys. Rev. Lett.* **123**, 250502 (2019).
17. Naghiloo, M., Jordan, A. & Murch, K. Achieving optimal quantum acceleration of frequency estimation using adaptive coherent control. *Phys. Rev. Lett.* **119**, 180801 (2017).
18. Khaneja, N., Reiss, T., Kehlet, C., Schulte-Herbrüggen, T. & Glaser, S. J. Optimal control of coupled spin dynamics: design of NMR pulse sequences by gradient as- cent algorithms. *J. Magn. Reson.* **172**, 296–305 (2005).
19. Caneva, T., Calarco, T. & Montangero, S. Chopped random-basis quantum optimization. *Phys. Rev. A* **84**, 022326 (2011).
20. Glaser, S. J. et al. Training Schrödinger's cat: quantum optimal control. *Eur. Phys. J. D.* **69**, 1–24 (2015).
21. Xiao, T., Huang, J., Fan, J. & Zeng, G. Continuous-variable Quantum phase estimation based on Machine Learning. *Sci. Rep.* **9**, 1–13 (2019).
22. Palittapongarnpim, P., Wittek, P., Zahedinejad, E., Vedaie, S. & Sanders, B. C. Learning in quantum control: High-dimensional global optimization for noisy quantum dynamics. *Neurocomputing* **268**, 116–126 (2017).
23. Lumino, A. et al. Experimental phase estimation enhanced by machine learning. *Phys. Rev. Appl.* **10**, 044033 (2018).
24. Bukov, M. et al. Reinforcement learning in different phases of quantum control. *Phys. Rev. X* **8**, 031086 (2018).
25. Fösel, T., Tighineanu, P., Weiss, T. & Marquardt, F. Reinforcement Learning with Neural Networks for Quantum Feedback. *Phys. Rev. X* **8**, 031084 (2018).
26. Niu, M. Y., Boixo, S., Smelyanskiy, V. N. & Neven, H. Universal quantum control through deep reinforcement learning. *npj Quantum Inf.* **5**, 1–8 (2019).
27. Xu, H. et al. Generalizable control for quantum parameter estimation through reinforcement learning. *npj Quantum Inf.* **5**, 1–8 (2019).
28. Schuff, J., Fiderer, L. J. & Braun, D. Improving the dynamics of quantum sensors with reinforcement learning. *N. J. Phys.* **22**, 035001 (2020).
29. Fiderer, L. J. & Braun, D. Quantum metrology with quantum-chaotic sensors. *Nat. Commun.* **9**, 1–9 (2018).
30. Xie, D. & Xu, C. Optimal control for multi-parameter quantum estimation with time- dependent Hamiltonians. *Results Phys.* **15**, 102620 (2019).
31. August, M. & Hernández-Lobato, J. M. Taking gradients through experiments: LSTMs and memory proximal policy optimization for black-box quantum control. *In International Conference on High Performance Computing* 591–613 (Springer, 2018).
32. Yarats, D. et al. Improving sample efficiency in model-free reinforcement learning from images. *arXiv:1910.01741* (2019).

33. Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
34. Silver, D. et al. Mastering the game of go without human knowledge. *Nature* **550**, 354–359 (2017).
35. Vinyals, O. et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* **575**, 350–354 (2019).
36. Szepesvári, C. *Reinforcement learning algorithms for MDPs* (Morgan & Claypool Publisher, 2009).
37. Kaelbling, L. P., Littman, M. L. & Moore, A. W. Reinforcement learning: A survey. *J. Artif. Intell. Res.* **4**, 237–285 (1996).
38. Arulkumaran, K., Deisenroth, M. P., Brundage, M. & Bharath, A. A. Deep reinforcement learning: A brief survey. *IEEE Signal Process. Mag.* **34**, 26–38 (2017).
39. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction* (MIT press, 2018).
40. Schulman, J., Moritz, P., Levine, S., Jordan, M. & Abbeel, P. High-dimensional continuous control using generalized advantage estimation. *arXiv:1506.02438* (2015).
41. Mnih, V. et al. Asynchronous methods for deep reinforcement learning. In International Conference on Machine Learning 1928–1937 (PMLR, 2016).
42. Lillicrap, T. P. et al. Continuous control with deep reinforcement learning. *arXiv:1509.02971* (2015).
43. Schulman, J., Levine, S., Abbeel, P., Jordan, M. & Moritz, P. *Trust region policy optimization*. In International Conference on Machine Learning 1889–1897, (PMLR, 2015).
44. Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. Proximal policy optimization algorithms. *arXiv:1707.06347* (2017).
45. Szita, I. & Lőrincz, A. Learning Tetris using the noisy cross-entropy method. *Neural Comput.* **18**, 2936–2941 (2006).
46. Beckey, J. L., Cerezo, M., Sone, A. & Coles, P. J. Variational quantum algorithm for estimating the quantum fisher information. *arXiv:2010.10488* (2020).
47. Meyer, J. J. Fisher information in noisy intermediate-scale quantum applications. *arXiv:2103.15191* (2021).
48. Perrier, E., Ferrie, C. & Tao, D. Quantum Geometric Machine Learning for Quantum Circuits and Control. *N. J. Phys.* **22**, 103056 (2020).
49. Huang, H.-Y., Kueng, R. & Preskill, J. Predicting many properties of a quantum system from very few measurements. *Nat. Phys.* **16**, 1050–1057 (2020).
50. Rath, A., Branciard, C., Minguzzi, A. & Vermersch, B. Quantum Fisher information from randomized measurements. *arXiv:2105.13164* (2021).
51. Carrasquilla, J., Torlai, G., Melko, R. G. & Aolita, L. Reconstructing quantum states with generative models. *Nat. Mach. Intell.* **1**, 155–161 (2019).
52. Rabitz, H., de Vivie-Riedle, R., Motzkus, M. & Kompa, K. Whither the future of controlling quantum phenomena? *Science* **288**, 824–828 (2000).
53. Egger, D. J. & Wilhelm, F. K. Adaptive hybrid optimal quantum control for imprecisely characterized systems. *Phys. Rev. Lett.* **112**, 240503 (2014).
54. Xia, Y., Li, W., Zhuang, Q. & Zhang, Z. Quantum-enhanced data classification with a variational entangled sensor network. *Phys. Rev. X* **11**, 021047 (2021).
55. Zhuang, Q. & Zhang, Z. Physical-layer supervised learning assisted by an entangled sensor network. *Phys. Rev. X* **9**, 041023 (2019).
56. Dive, B., Mintert, F. & Burgarth, D. Quantum simulations of dissipative dynamics: Time dependence instead of size. *Phys. Rev. A* **92**, 032111 (2015).
57. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).

## ACKNOWLEDGEMENTS

We thank the fruitful discussions with Jinzheng Huang and Nana Liu. This work is supported by National Natural Science Foundation of China (Grant No. 61701302 and 61631014).

## AUTHOR CONTRIBUTIONS

T.L.X. developed the simulation and implemented the algorithms, T.L.X. and G.H.Z. elaborated on the Hamiltonian framework and the master equation formalism, J.P.F. operated the reinforcement learning analysis, and G.H.Z. and J.P.F. conceived and coordinated the research. All the authors discussed and contributed to the writing of the manuscript.

## COMPETING INTERESTS

The authors declare no competing interests,.

## ADDITIONAL INFORMATION

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41534-021-00513-z.

**Correspondence** and requests for materials should be addressed to Guihua Zeng.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.