An energy gap develops near quantum critical point (QCP) of quantum phase transition (QPT) in a finite many-body (MB) system, facilitating adiabatic ground state transformation by parameter change. In real application scenarios, however, the efficacy for such adiabatic protocol is compromised by the need to balance finite system life time with adiabaticity, as exemplified in a recent experiment that prepares three-mode balanced Dicke state near deterministically [PNAS 115, 6381 (2018)]. Instead of following the instantaneous ground state as unanimously required for most adiabatic crossing, this work reports a faster sweeping policy taking advantage of fast dynamics in the excited levels. It is obtained from deep reinforcement learning (DRL) based on a multi-step training scheme we develop. In the absence of loss, a fidelity ≥ 99% between the prepared and target Dicke state is achieved over a small fraction of the adiabatically required time. When loss is included, training is carried out according to an operational benchmark, the interferometric sensitivity of the prepared state, leading to better sensitivity while using about half of the time previously reported. Implemented in a Bose-Einstein condensate (BEC) of ∼ 10^4 ^{87}Rb atoms, the balanced three-mode Dicke state exhibiting an improved number squeezing of 13.02 ± 0.20 dB is observed within 766 ms, highlighting the potential of DRL for quantum dynamic control and quantum state preparation in interacting MB systems.

One celebrated hallmark for a quantum system lies at its discrete level (eigen-energy) and associated state (orthogonal eigen-state). According to quantum adiabatic theorem, under slow and continuous change of a parameter, the state of a (Hamiltonian) system stays at the level it starts with, e.g., remaining in the ground state. The rate of parameter change has to be much less than the corresponding level spacing in order to avoid excitation. However, gap size or level spacing at quantum critical point (QCP) becomes diminishingly small for increasingly larger system approaching thermodynamic limit. Limited by finite life time, sweeping cannot proceed as slowly as one wishes in an actual experiment. Various shortcuts to adiabaticity (STA) [1] are developed and demonstrated for quantum state preparation [2–4] in diverse systems ranging from thermal [5] to Bose-Einstein condensate (BEC) gases [6] and trapped ions [3, 8] etc, albeit often restricted to a few levels and augmented by counter adiabatic terms in the Hamiltonian [1]. Otherwise no generally applicable strategy is known for crossing quantum phase transition (QPT) to arrive at a transformed ground state, except for adiabatically sweeping over the finite sized gap.

This work reports a sweeping protocol for faster crossing of QCP benefited from excited states by deep reinforcement learning (DRL) [9, 10]. It beats adiabatic crossing through a more sophisticated parameter tuning profile while keeping the form of the Hamiltonian intact without requiring counter adiabatic driving terms. By directly learning from simulated spin mixing dynamics, perspectives and insights are learnt for controlling dynamics of many coupled spins, significantly extending the long list of previous DRL applications in quantum technology of small systems [9–17] to an interacting MB spin systems. More specifically, from the point of view on balanced Dicke state preparation, our work accomplishes a MB optimization (of its dynamics) with DRL, which constitutes a key research direction in noisy intermediate-scale quantum technology of the near future [5, 18–20, 22].

The system we study is a ^{87}Rb atomic BEC in the ground hyperfine F = 1 manifold with the dominant interaction between atoms being symmetric among spin components (m_F = 0, ±1). This facilitates an approximate treatment with a common spatial mode [23–25] and gives rise to the model Hamiltonian (ℏ = 1 hereafter),

\[ H = \frac{c_2}{2N} \mathbf{L}^2 - q(t)N_0, \]

with the first term containing spin exchange interaction while the second term proportional to an effective quadratic Zeeman shift (QZS) q, tunable by magnetic field and/or dressing microwave field. \( N = \sum_{m_F} N_{m_F} \) (\( N_{m_F} = a^\dagger_{m_F} a_{m_F} \)) denotes the total atom number (of all \( m_F \) components) with \( a_{m_F} (a^\dagger_{m_F}) \) the atomic annihilation (creation) operator, and \( \mathbf{L} = \sum_{\mu,\nu} a^\dagger_{\mu} \mathbf{F}_{\mu\nu} a_{\nu} \) refers to the collective spin with \( \mathbf{F}_{\mu\nu} \) the spin-1 matrix element. \( L_z = N_{+1} - N_{-1} \) measures the system magnetization, which is conserved as is \( N \) in the absence of loss. The spin exchange interaction creates (annihilates) paired atoms in \( m_F = ±1 \) at the expense (gain) of \( m_F = 0 \) atoms. In a single spatial mode condensate, it represents an all-to-all interaction because \( \mathbf{L}^2 = 2(a^\dagger_{1} a^\dagger_{-1} a_{0} a_{0} + \text{h.c.}) + (2N_0 - 1)(N - N_0) + 2N + L_z^2 \). Its strength is \( c_2 < 0 \).
or ferromagnetic [26–28] for $^{87}$Rb atoms. Hence, in the absence of QZS or $q = 0$, the MB ground state takes the maximum $L^2 = N(N+1)$ or largest $L = N$, and is $(2N + 1)$-fold degenerate spanning the Dicke state subspace $\{|L,L_z\} : L_z = -N, -N + 1, \ldots, N\}$. [23–25].

The balanced or the $L_z = 0$ Dicke state is one of the most entangled states and enables measurement precision approaching Heisenberg Limit [29]. Zhang and Duan [30] studied its preparation through QPT by adiabatically sweeping from $q(t) > 0$ to $q(t) = 0$ at a constant rate within a sweeping time $\tau$. At $q/c_2 \gg 2$, the BEC ground state is polar with all atoms in $m_F = 0$, and is denoted by $|0,N,0\rangle$ in the Fock state notation $|N-1,N_0,N_1\rangle$. Upon adiabatically sweeping to $q = 0$, the ground state smoothly evolves into the balanced Dicke state $|\psi(0)_{\text{Dicke}}\rangle \equiv |N,L_z = 0\rangle$, after crossing QPT at $q/c_2 \approx 2$ into broken-axisymmetry (BA) phase. The level spacing near QCP scales as $\propto N^{-1/3}$ [10, 30, 32, 33], which erects a speed limit for crossing QCP.

In the following, we shall first discuss how DRL agent is applied to a moderate sized system of up to $N \approx 10^3$ atoms in the absence of loss. An optimized policy is found capable of preparing balanced Dicke state with a theoretical fidelity ($\geq 0.99$) using a much shorter $\tau$ than adiabatic sweep required for by following full quantum dynamics of Hamiltonian (1). Next, loss is modeled as a single atom effect and the DRL agent is retrained with an experimental sized system ($N \approx 10^4$) according to open system dynamics, by starting with the high-fidelity policy from without loss as prior knowledge. The resulting profile is subsequently affirmed experimentally, leading to the improved performance we report here.

Reinforcement Learning.—The key task for the DRL agent is to find a time-dependent $q(t)$ profile that prepares a state $|\psi(\tau)\rangle$ as close as possible to a target state $|\psi(0)_{\text{Dicke}}\rangle$ within a given $\tau$, starting from polar state. The agent learns from its interaction with environment: the numerically simulated spin mixing dynamics following Hamiltonian (1). With $\tau$ discretized into subintervals, the training process is as illustrated in Fig. 1(a), where at each $t$, the agent observes the environment through state $s_t \in S$ and takes action $q_t = q(t) \in A$ according to the current policy $\pi(q_t|s_t)$. The environment then evolves to $s_{t+1}$ and returns a scalar reward $r_t \in R$ back to the agent. The proximal policy optimization (PPO) algorithm [34] is then employed to update $\pi(q_t|s_t)$ due to its ability in sophisticated control of continuous action space over training processes. An optimal stochastic policy arises eventually which returns a normalized distribution in action space $\pi(q_t|s_t)$. More details can be found in the Supplemental Material (SM) [35].

For our model system, the following four observables: $\rho_0 = \langle N_0 \rangle / N$, $\langle \delta N^2_0 \rangle / N^2$, $|\langle a_{-1}^\dagger a_{-1} a_{0}^2\rangle|/N^2$, and $\theta_s = \text{arg} \langle a_{-1}^\dagger a_{1}^\dagger a_{0}^2\rangle$ constitute input state $s_t$. Two types of rewards are employed, based either on fidelity between the current and the target state $F = |\langle \psi(t)|\psi(0)_{\text{Dicke}}\rangle|^2$ or entanglement enhanced three-mode SU(2) interferometric sensitivity of the current state [14]. The dimension of concerned Hilbert space is $\propto N$. If DRL is directly implemented, the large distance between initial (polar) state and the target (balanced Dicke) state points to failed training due to sparse reward for $N \gtrsim 10^2$. Hence, a multi-step training approach is developed, with policies from consecutively completed tasks fed forward successively to larger sized systems as illustrated in Fig. 1(b). In explicit detail, a small-sized system of $N = 10$ first trains network until $F \gtrsim 0.99$, the trained network is subsequently used as pre-trained to initialize the next system one size up, e.g. $N = 100$, leading successively to a converged policy in mean-field state space for larger system size $N$ [see Figs. 1(b1)-(b3)].

The multi-step training ends at $N = 2000$ as subsequent training with larger system size is beyond the computational resource we have [37]. The corresponding optimal policy achieves target Dicke state with $F \approx 0.99$ over $|c_2|\tau = 15.5$, which is significantly shorter than lin-
ear adiabatic sweep that requires $|c_2|\tau \gtrsim 600$, or a nonlinear sweep optimized for local adiabaticity that demands $|c_2|\tau \gtrsim 350$ for the same fidelity level [3, 4, 35].

The complete evolution is composed of three main stages, partitioned by the vertical dashed lines as illustrated in Fig. 2. Its understanding heavily relies on the $q$-dependent features in the excitation spectra of Hamiltonian (1), many of which are shared by the broad class of Lipkin-Meshkov-Glick (LMG) model. They exhibit a characteristic critical gap curve (CGC) [35, 39], which connects successive level spacing minima (also scaling as $N^{-1/3}$ [10, 32, 33]) and is approximately described by,

$$E_\text{g}(q) - E_\text{g}(q) = N(q - 2|c_2|)^2/(8|c_2|), \quad (q/|c_2| \leq 2),$$

for our model as shown by gray dashed lines in Figs. 2(d) with $E_\text{g}(q)$ the ground state energy of $H(q)$. Sweeping quickly into either direction, the most likely trajectories for states near CGC will travel diabatically along CGC in the same direction. For instance, starting from the ground state at a high $q$ and rapidly sweeping down, the state will simply ascend along CGC by diabatically crossing successive level spacing minima to reach increasingly higher excited states. In other words, the initial polar ground state at high $q$ can be regarded as a superposition state in the CGC region at any $q$ produced by an abrupt change of $q$. As shown in Fig. 2(a), the starting $q$ of the DRL profile sits on the left-hand-side of QCP to access faster spin mixing dynamics, as otherwise starting from the right-hand-side ($q > q_\text{c}$) would be ineffective due to insufficient excited state components. Their excitations facilitate faster dynamics as illustrated in the following three stages:

(i) For $|c_2|t \in [0, 2.6]$, this stage is denoted by the black solid line segment starting from $q < q_\text{c}$ and marked by a circle in Fig. 2(d1), during which $q(t)$ first ascends quickly to cross $q_\text{c}$ and arrives at $q_{\text{max}}/|c_2| \simeq 2.3$. The current state nearly tracks CGC to successively lower excited states, but does not fall all the way down to the ground state at $q_{\text{max}}$. It is transformed into a superposition of several low lying states due to the $q$-controlled spin-mixing [35]. The multi-level amplitudes during subsequent descending of $q$ to near $q_\text{c}$ can be easily controlled to interfere constructively into the ground state level and destructively to higher levels [5], which suppresses increased spreading of excitations along CGC.

(ii) The middle interval $|c_2|t \in [2.6, 10]$ (blue solid line segment) covers the actual crossing of QCP, where the state (from end of the first stage) dominated by the ground and the first excited states evolve slowly into a slightly excited form dominated by the first few low-lying levels in the instantaneous eigen-basis $|\psi_n(t)\rangle$ (insets in Fig. 2(b)). At the end of this stage $q/|c_2| \simeq 1.86$, the state becomes a Gaussian-like wave packet in the Fock state basis $|k, N - 2k, k\rangle$ (Fig. 2(c2)). It closely matches the ground state of a slightly larger $q/|c_2| \simeq 1.92$, representing a displaced Gaussian packet in a harmonic trap anticipating for rapid translation over the next stage. Strict adiabatic condition is observed for this stage, reflecting the adiabatic speed limit of level spacing ($\sim N^{-1/3}$) near the QCP, in agreement with the paradigm of adiabatic Landau-Zener crossing [35, 40, 41].

(iii) During the final stage of $|c_2|t \in [10, 15.5]$, the red solid line segment, the profile corresponds to a rapid translation of the Gaussian wave packet from the small $k$ region to balanced Dicke state at $k \sim N/4$ as shown in Fig. 2(c3), accompanied by rapidly increasing fidelity.

FIG. 2. Results from DRL policy without atom loss for $N = 2000$. (a) The sweeping profile (in three line segments of black, blue, and red) and corresponding simulated fidelity (dot-dashed line referred to right axis); (b) Density plot of probability distribution $W_n = \langle \psi_n(t) | \psi(t) \rangle^2$ for simulated state in the instantaneous eigen-state basis $|\psi_n(t)\rangle$ (vertical axis); Insets show zoomed-in distributions on the boundaries $|c_2|t = 2.6$ and $10$ between stages; (c1), (c2), and (c3), respectively show evolution of state expanded in the Fock state basis $|k, N - 2k, k\rangle$ over the three stages; (d1) Their corresponding mean energies $\langle \psi(t) | H(q) | \psi(t) \rangle - E_\text{g}(q)$ shown by the same colored lines as in (a) with one standard deviation in shaded regions, and (d2) for an expanded view around QCP. Arrows indicate direction of increasing time and gray dashed lines denote CGC.
shown by black dot-dashed line in Fig. 2(a). Such an overall center of mass translation is well described by a tunable harmonic oscillator model [10, 35],

\[ H/(N/2) = (1 - \lambda^2)(\mu - \mu_0)^2/(2\lambda^2) - \lambda^2\varphi^2/2, \]

with frequency \( \sqrt{1 - \lambda^2} \) and effective mass \( 1/\lambda^2 \) for \( \lambda = q/(2c_2) \), where \( \mu \) and \( \varphi \) are the canonical ‘position’ and ‘momentum’ respectively and \( |\mu_0| = \sqrt{1 - \lambda^2} \) denotes the center of the ground state Gaussian wave packet. Hence, during this stage the DRL profile simply shifts the Gaussian wave packet right after crossing QCP at \( \mu_0 \approx 0 \) (\( q/(c_2) \approx 2 \) or \( \lambda \approx 1 \)) to \( |\mu_0| \approx 1 \) (\( q \approx 0, \lambda \approx 0 \)), following STA for quantum transport with a harmonic trap [42–45]. This is accomplished by first accelerating the trap, followed by decelerating to a standstill, demonstrated by the rise and fall of the average energy (Fig. 2(d1)).

**Experiment.**—The generalization ability of our DRL policy is outstanding. The \( N = 2000 \) policy discussed above achieves a fidelity \( F \gtrsim 0.85 \) for \( 10^4 \) atoms in the absence of atom loss [35]. However, the minimal level spacing near QCP reduces to a few Hertz for a typical \( ^{87}\text{Rb} \) BEC with \( 10^5 \) atoms. QPT from adiabatic level crossing therefore demands a sweeping duration of seconds, during which atom loss, e.g. three-body collisions, can no longer be ignored. In earlier preparation of balanced Dicke state ensembles [14], about 5% of the total atoms were lost over \( \tau = 1.5 \) s.

The DRL profile we find from without loss shows the end result sensitively depends on parameters, especially on the fluctuation of \( N \) [35] (always comes from atom loss). We thus take atom loss into account in the next training process. Operationally, the full quantum simulation based on Hamiltonian (1) is replaced by coupled stochastic differential equations (SDEs) derived from quasi-probability distribution based on truncated Wigner approximation [9, 46–52], with atom loss modeled as one-body decay. Instead of fidelity \( F \), the SU(2) interferometric sensitivity \( (\delta\theta)^2 = (3\langle (zL_z)^2 \rangle_{\theta=0} + 1/2)/\langle L_z^2 \rangle \) [14] for small rotation angle \( \theta \) (from the equatorial plane in the generalized Bloch sphere [54]) is used as a more suitable quantitative figure of merit.

To keep the desirable characteristic features of the high-fidelity profile for \( q(t) \), prior results trained using \( F \) are adopted as pre-trained instead of starting afresh DRL training using \( (\delta\theta)^2 \). The network parameters from a small-sized system \( (N = 100) \) trained with \( F \) in the absence of loss are used to initialize subsequent network. It is followed by solving SDEs using \( (\delta\theta)^2 \) as reward to successively larger systems, analogous to that illustrated in Fig. 1(b). The resulting \( q(t) \) profile for \( N = 11800 \) (Fig. 3(a)) gives a maximal theoretical sensitivity for \( (\delta\theta)^2 \) at \( |c_2|\tau = 13 \). All characteristic features as discussed previously for without loss at \( N = 2000 \) remain.

Experimentally implemented in a \( ^{87}\text{Rb} \) BEC of \( N \approx 11800 \) atoms at \( c_2 = -(2\pi) \times 2.7 \) Hz, the corresponding sweeping time becomes \( \tau \approx 766 \) ms, or about half of the previously reported time [14] based on an empirical piecewise analytic profile after numerical optimization as compared in Fig. 3(a). The corresponding evolutions of the fractional populations \( \rho_{m_F} \) are shown in Fig. 3(b). Including detection noise and other stochastic influences, we measure \( (\delta L_z)_{\theta=0} = 23.88 \pm 0.52 \) dB, which gives a number squeezing of \( \xi_L^{(2)} = -20 \log_{10}(\delta L_z)_{\theta=0}/\sqrt{N} \approx 13.02 \pm 0.20 \) dB (vs. \( 12.56 \pm 0.20 \) dB earlier [14, 55]) below the quantum shot noise of \( \sqrt{N} \approx 106.93 \pm 0.84 \) for the polar state (Fig. 3(c1)). Despite reduced loss from nearly halved sweeping time, the quality for the prepared Dicke state is only marginally improved due to detection noise \( (\delta L_z^{DN}) = 21.5 \) dB, which is dominating but quantitatively well-understood [14]. After taking out detection noise, we infer a number squeezing of \( 20.25 \pm 1.0 \) dB (vs. \( 17.83 \pm 0.63 \) dB earlier [14, 55]). The quality of our prepared Dicke state is further characterized by comparing directly its \( L_z/N \) distribution with the ideal balanced...
Dicke state after a $\theta = \pi/2$ rotation [14]. The measured results are shown respectively by the histogram in Fig. 3(c2), which confirms its extreme close resemblance to the target balanced Dicke state in solid line (see SM [35] for experimental details). Adopting better detection technique, such as fluorescence detection [56], would help harvesting the full advantage of the DRL policy in future.

In summary, we develop a multi-step DRL training scheme, which is pre-trained for small-sized systems without loss to arrive at a high-fidelity policy, and subsequently retrained by including loss for larger systems to quest for enhanced interferometric sensitivity. By adopting the consequent DRL profile to a BEC of $^{87}$Rb atoms, the validity of our policy is affirmed based on observing improved quality Dicke state ensembles. The optimal DRL profile includes three stages for the underline dynamics of spin-mixing. Except in the middle stage, where adiabaticity is maintained to the best possible level constrained by the total sweeping time, faster evolutions are found in the beginning and the ending stages, capitalizing on the features of the excited energy levels. Our work highlights the application of DRL to guide experiments where training with realistic system is difficult or impossible but simulation can be easily performed.

We thank Dr. L.-N. Wu, Dr. P. Xu, and Dr. Y.C. Liu for helpful discussions. This work is supported by the Key-Area Research and Development Program of GuangDong Province (Grant No. 2019B030330001), by the National Key R&D Program of China (Grants No. 2018YFA0306504 and No. 2018YFA0306505), and by the National Natural Science Foundation of China (NSFC) (Grants No. 61606213, No. 11654001, No. 91736311, No. 91836302, and No. U1930201). L.Y. also acknowledges support from BAQIS Research Program (Grant No. Y18G24).

---

* These authors contributed equally to this work.
† mengkhoon.tey@tsinghua.edu.cn
‡ lyou@tsinghua.edu.cn

[1] D. Guéry-Odelin, A. Ruschhaupt, A. Kiedy, E. Torrentegui, S. Martínez-Garaot, and J. G. Muga, Rev. Mod. Phys. 91, 045001 (2019).
[2] K. Takahashi, Phys. Rev. E 87, 062117 (2013).
[3] S. Campbell, G. De Chiara, M. Paternostro, G. M. Palma, and R. Fazio, Phys. Rev. Lett. 114, 177206 (2015).
[4] T. Opatrný, H. Saberi, E. Brion, and K. Molmer, Phys. Rev. A 93, 023815 (2016).
[5] Y.-X. Du, Z.-T. Liang, Y.-C. Li, X.-X. Yue, Q.-X. Lv, W. Huang, X. Chen, H. Yan, and S.-L. Zhu, Nature Communications 7 (2016).
[6] M. G. Bason, M. Viteau, N. Malossi, P. Huillery, E. Arimondo, D. Ciampini, R. Fazio, V. Giovannetti, R. Marnell, and O. Morsch, Nature Physics 8, 147 (2011).
[7] P. Richerme, C. Senko, J. Smith, A. Lee, S. Korenblit, and C. Monroe, Phys. Rev. A 88, 012334 (2013).
[8] C.-K. Hu, J.-M. Cui, A. C. Santos, Y.-F. Huang, M. S. Sarandy, C.-F. Li, and G.-C. Guo, Opt. Lett. 43, 3136 (2018).
[9] M. Bukov, A. G. R. Day, D. Sels, P. Weinberg, A. Polkovnikov, and P. Mehta, Phys. Rev. X 8, 031086 (2018).
[10] P. Mehta, M. Bukov, C.-H. Wang, A. G. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, Physics Reports 810, 1 (2019).
[11] M. Dalgaard, F. Motzoi, J. J. Sørensen, and J. Sherson, npj Quantum Information 6 (2020).
[12] J. Lin, Z. Y. Lai, and X. Li, Phys. Rev. A 101, 052327 (2020).
[13] T. Haug, R. Dumke, L.-C. Kwek, C. Miniatura, and L. Amico, arXiv preprint arXiv:1911.06578 (2019).
[14] M. Bukov, Phys. Rev. B 98, 224305 (2018).
[15] F. Chen, J.-J. Chen, L.-N. Wu, Y.-C. Liu, and L. You, Phys. Rev. A 100, 041801(R) (2019).
[16] M. Y. Niu, S. Boixo, V. N. Smelyanskiy, and H. Neven, npj Quantum Information 5 (2019).
[17] Z. T. Wang, Y. Ashida, and M. Ueda, Phys. Rev. Lett. 125, 100401 (2020).
[18] J. Preskill, Quantum 2, 79 (2018).
[19] P. Doria, T. Calarco, and S. Montangero, Phys. Rev. Lett. 106, 190501 (2011).
[20] J. J. W. H. Sørensen, M. O. Aranburu, T. Heinzel, and J. F. Sherson, Phys. Rev. A 98, 022119 (2018).
[21] P. Xu, S. Yi, and W. Zhang, Phys. Rev. Lett. 123, 073001 (2019).
[22] S. A. Haine and J. J. Hope, Phys. Rev. Lett. 124, 060402 (2020).
[23] C. K. Law, H. Pu, and N. P. Bigelow, Phys. Rev. Lett. 81, 5257 (1998).
[24] T.-L. Ho and S. K. Yip, Phys. Rev. Lett. 84, 4031 (2000).
[25] Y. Kawaguchi and M. Ueda, Physics Reports 520, 253 (2012).
[26] E. G. M. van Kempen, S. J. J. M. F. Kokkelmans, D. J. Heinzen, and B. J. Verhaar, Phys. Rev. Lett. 89, 093201 (2002).
[27] M.-S. Chang, Q. Qin, W. Zhang, L. You, and M. S. Chapman, Nature Physics 1, 111 (2005).
[28] A. Widera, F. Gerbier, S. Fölling, T. Gericke, O. Mandel, and I. Bloch, New Journal of Physics 8, 152 (2006).
[29] M. J. Holland and K. Burnett, Phys. Rev. Lett. 71, 1355 (1993).
[30] Z. Zhang and L.-M. Duan, Phys. Rev. Lett. 111, 180401 (2013).
[31] P. Leyvraz and W. D. Heiss, Phys. Rev. Lett. 95, 050402 (2005).
[32] S. Dusuel and J. Vidal, Phys. Rev. Lett. 93, 237204 (2004).
[33] M. Xue, S. Yin, and L. You, Phys. Rev. A 98, 013619 (2018).
[34] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, arXiv preprint arXiv:1707.06347 (2017).
[35] See supplementary material for a complete description of our DRL task, generalization power of the DRL agent, comparison between DRL protocol and adiabatic sweep, multi-level oscillation in the first stage, derivation of the LMG model and the harmonic model approximation, sensitive dependence on parameters of DRL profile, and a detailed description of our experiment.
[36] Y.-Q. Zou, L.-N. Wu, Q. Liu, X.-Y. Luo, S.-F. Guo, J.-H.
Cao, M. K. Tey, and L. You, Proceedings of the National Academy of Sciences \textbf{115}, 6381 (2018).

[37] The training process in this work is carried on two Intel Xeon Gold-6154 CPU (3.0 GHz) processors.

[4] S. Balasubramanian, S. Han, B. T. Yoshimura, and J. K. Freericks, Phys. Rev. A \textbf{97}, 022313 (2018).

[39] P. Solinas, P. Ribeiro, and R. Mosseri, Phys. Rev. A \textbf{78}, 052329 (2008).

[40] L. D. Landau, Phys. Z. Sowjetunion \textbf{2}, 19 (1932).

[41] G. Zener, Proc. R. Soc. London, Ser. A \textbf{137}, 696 (1932).

[42] E. Torrontegui, S. Ibáñez, X. Chen, A. Ruschhaupt, D. Guéry-Odelin, and J. G. Muga, Phys. Rev. A \textbf{83}, 013415 (2011).

[43] T. Schmiedl, E. Dieterich, P.-S. Dieterich, and U. Seifert, Journal of Statistical Mechanics: Theory and Experiment \textbf{2009}, P07013 (2009).

[44] S. An, D. Lv, A. del Campo, and K. Kim, Nature Communications \textbf{7}, 12999 (2016).

[45] M. J. Steel, M. K. Olsen, L. I. Plimak, P. D. Drummond, S. M. Tan, M. J. Collett, D. F. Walls, and R. Graham, Phys. Rev. A \textbf{58}, 4824 (1998).

[46] A. Sinatra, C. Lobo, and Y. Castin, Journal of Physics B: Atomic, Molecular and Optical Physics \textbf{35}, 3599 (2002).

[47] A. A. Norrie, R. J. Ballagh, and C. W. Gardiner, Phys. Rev. A \textbf{73}, 043617 (2006).

[48] B. Opanchuk, M. Egorov, S. Hoffmann, A. I. Sidorov, and P. D. Drummond, EPL (Europhysics Letters) \textbf{97}, 50003 (2012).

[49] P. D. Drummond and B. Opanchuk, Phys. Rev. A \textbf{96}, 043616 (2017).

[50] A. Johnson, S. S. Szigeti, M. Schemmer, and I. Bouchoule, Phys. Rev. A \textbf{96}, 013623 (2017).

[51] C. Gerving, T. Hoang, B. Land, M. Anquez, C. Hamley, and M. Chapman, Nature Communications \textbf{3} (2012).

[52] C. D. Hamley, C. S. Gerving, T. M. Hoang, E. M. Bookjans, and M. S. Chapman, Nature Physics \textbf{8}, 305 (2012).

[53] B. Lücke, M. Scherer, J. Kruse, L. Pezzé, F. Deuretzbacher, P. Hyllus, O. Topic, J. Peise, W. Ertmer, J. Arlt, L. Santos, A. Smerzi, and C. Klempt, Science \textbf{334}, 773 (2011).

[54] A. Qu, B. Evrard, J. Dalibard, and F. Gerbier, Phys. Rev. Lett. \textbf{125}, 033401 (2020).
Supplementary Material

This supplementary provides expanded discussions on several issues left out of the main text due to space limitation. They include: (i) detailed description of deep reinforcement learning (DRL) task in Sec.; (ii) generalization ability of the DRL policy obtained in Sec.; (iii) comparison of adiabatic sweeps with the DRL profile in Sec.; (iv) multilevel oscillation during QCP crossing in Sec.; (v) comparison between full quantum simulations with mean-field simulations using the DRL profile in Sec.; (vi) connection between our model and the Lipkin-Meshkov-Glick (LMG) model in Sec.; (vii) derivation of the approximate simple harmonic oscillator Hamiltonian in the broken-axisymmetry phase in Sec.; (viii) deteriorating quality of the Dicke state with atom loss in Sec.; (ix) analysis of the sensitive dependence on parameters in Sec.; and finally (x) experimental implementation in Sec.

THE DEEP REINFORCEMENT LEARNING TASK

The DRL task is modeled by a Markov decision process (MDP) as illustrated in the main text. At each time instant \( t \), the agent carries out the observation \( s_t \in \mathcal{S} \) of the environment and selects an action \( a_t \in \mathcal{A} \) according to current policy \( \pi(a|s_t) \). The environment then evolves to \( s_{t+1} \) after the action \( a_t \) and a scalar reward \( r_t \in \mathcal{R} \) is computed and returned back to the agent. Policy \( \pi \) is subsequently updated through such experience data to maximize accumulated reward \( R \). The definitions for state \( \mathcal{S} \), action \( \mathcal{A} \), and reward function \( \mathcal{R} \) in our model system of spin-1 Bose-Einstein condensate (BEC) take the following more specific meanings.

- **\( \mathcal{S} \):** In the spin-1 BEC system, the set of four physical observables, including \( \langle N_0 \rangle \), \( \langle \delta N_0^2 \rangle \), \( |\langle a_{-1}^\dagger a_{-1}a_0a_0 \rangle| \), and \( \arg \{\langle a_{-1}^\dagger a_{-1}a_0a_0 \rangle \} \) are chosen to represent the state instead of the wave function \( \psi(t) \), which contains complete information but is too cumbersome to handle at large \( N \). Using these more relevant physical observables as state representation facilitates directly generalizable policy due to their explicit independence of \( N \) after normalization (illustrated in the Sec.), the final policy is also easily interpreted with clear physical insights. For our choice, \( \{\langle N_0 \rangle, \arg \{\langle a_{-1}^\dagger a_{-1}a_0a_0 \rangle \} \} \) corresponds to the mean-field parameters \( \{\rho_0 = \langle N_0 \rangle / \langle N \rangle, \theta_s \} \), while \( \{\langle \delta N_0^2 \rangle, \|\langle a_{-1}^\dagger a_{-1}a_0a_0 \rangle \| \} \) calibrates quantum correlations.

- **\( \mathcal{A} \):** The action space is a continuous range of quadratic Zeeman shift \( q/c_2 \in [-q_{\text{max}}, q_{\text{max}}] \) with \( q_{\text{max}} = 3 \).

- **\( \mathcal{R} \):** Two types of object function \( f(\psi) \) are employed as rewards, overlap (or fidelity) between the current and the target state \( f(\psi) = \mathcal{F} = |\langle \psi(t) | \psi_{\text{Dicke}}^{(0)} \rangle|^2 \) from full quantum simulation in the absence of loss, and \( f(\psi) = \langle (\delta \theta)^2 \rangle_{\text{eff}} / \langle (\delta \theta(t_j))^2 \rangle \) when loss is included with \( \langle (\delta \theta)^2 \rangle_{\text{eff}} = [3 \langle \delta(L_z_{\theta=0})^2 \rangle + 1/2] / \langle L_{\theta=0}^2 \rangle \) the expected interferometric sensitivity of current state, while \( \langle (\delta \theta(t))^2 \rangle_{\text{eff}} = 1/(2 \langle L_{\theta=0}^2 \rangle) \) the optimal (minimum value for an ideal target Dicke state \( |\psi_{\text{Dicke}}^{(0)} \rangle \)). The total reward is calculated as a cumulative sum of all instantaneous rewards, \( R_{\text{tot}} = \sum_{j=1}^{m} r_j \) with \( r_j = f(\psi(t_j)) - f(\psi(t_{j-1})) \). Starting with the initial (polar) state, the object function satisfies \( f(\psi_i) = 0 \), while \( f(\psi_f) \approx 1 \) denotes a complete success with perfect fidelity or sensitivity in the end. Such a dense cumulative partition of reward makes the training process quicker and more stable. In addition, the instantaneous reward can be further revised into

\[
    r_j \rightarrow \log_{10} \left( \frac{1 - f(\psi(t_{j-1}))}{1 - f(\psi(t_j))} \right), \tag{S1}
\]

from which \( f(\psi(t)) \rightarrow 1 \) to prevent deteriorating learning efficiency when the state approaches the target \( f(\psi(t)) \rightarrow 1 \).

The proximal policy optimization (PPO) algorithm is employed to find the optimized policy \( \pi^* \) that maximizes cumulative reward \( R \),

\[
    \pi^* = \arg \max_{\pi} R \quad \text{with} \quad R = \sum_{j} \gamma^j r_j, \tag{S2}
\]

and \( \gamma \) a discount factor, which is typically chosen very close to 1 to avoid greedy solutions. The structure of the neural network (NN) for PPO algorithm is shown in Fig. S1(a) and the pseudo-code for PPO algorithm is shown in Table S1 [1]. To facilitate training, we encapsulate the quantum state evolution into a gym environment as suggested by openAI [2]. With PPO algorithm, the policy is stochastic which returns a normalized distribution on action space for a given state \( s \) and satisfies

\[
    \int_{a \in \mathcal{A}} \pi(a|s) da = 1. \tag{S3}
\]
When a policy is reached, one can either choose the action with the maximum probability \( a = \max_{a'} \pi^*(a'|s) \) as a deterministic protocol or select the best one among multiple profiles sampled based on \( \pi^*(a'|s) \). The former approach is taken by us in this work, which leads to the DRL profile \( q(t) \) of the policy.

For a system of \( N \) atoms, a deep NN is adopted to parameterize the actor and the critic networks in PPO, each containing four fully connected hidden layers with \([64, 32, 16, 8]\) neurons respectively. Every learning episode is divided into a few hundred consecutive steps and the total evolution time is limited to \( \tau \) that depends on the total number of atoms \( N \). Other hyper-parameters used in the training are listed in Table S2, some of them are tuned within a range according to system size. Typically, thousands of training epochs are required to reach an optimal policy, while each training epoch contains hundreds of learning episodes. To accelerate the training process and enhance the final performance of optimal policy, the total training epochs are artificially divided into two parts as illustrated in Fig. S1(b). For the first few hundred epochs, a random quantum state is used as an initial state to ignite a learning episode, which helps the agent to learn the basic geography of the state space (exploration for short). Afterwards, the initial state is reset to polar state \(|0,N,0\rangle\) and the agent subsequently finds out a (sub)optimal controlled trajectory in state space \( S \) from the polar state to target (Dicke) state.

Due to the almost independence on \( N \) of the physical observables in state \( s_t \), training tasks in systems of different \( N \) share the same NN structure, i.e., a larger-sized system inherits the trained NN from a smaller-sized system as a pre-trained network, which dramatically reduces the required number of training epochs in the larger-sized system, and the total training process becomes efficient since training in larger-sized systems no longer consumes enormous computational resource. Such a multi-step training process is adopted for training increased system size from \( N = 10 \) to \( N = 2000 \) atoms using full quantum simulation of Schrodinger equation without loss, and from \( N = 1000 \) to \( N = 11800 \) atoms by following stochastic differential equations of the corresponding master equation including atom loss modeled by a one-body decay.

![Figure S1](image)

**FIG. S1.** (a) Schematic of the neural network (NN) structure for PPO algorithm. The hidden layers contain four fully connected ones with \([64, 32, 16, 8]\) neurons respectively. (b) Illustration of the learning process at a specific system size \( N \). Either a randomly initialized network or a trained network can be successively used as input network to the next one.

### GENERALIZATION ABILITY OF THE DRL POLICY

A DRL policy is typically trained at a specific system size of \( N \) atoms. When observables selected for DRL agent are \( N \)-independent or almost \( N \)-independent as in our task, an essentially \( N \)-independent policy results. The performance of such a policy when applied to different conditions of the training process, e.g., different numbers of atoms \( N \), different sweeping time \( \tau \), or range of action space \( q_{\text{max}} \), is measured by the generalization ability of a policy. Here, we focus on the generalization for the number of atoms \( N \) as well as the sweeping time \( \tau \) since we aim to obtain a policy that can handle larger system and prepare target state within shorter time.

Figure S2 illustrates generalization ability for the \( N = 2000 \) policy using \( F \) as reward with \( |c_2|\tau = 15.5 \) (a), while (b) refers to the case of \( N = 11800 \) policy from including atom loss and detection noise using \((\delta \theta)^2\) as reward with \(|c_2|\tau = 13\). Both policies remain effective and perform well for system sizes smaller than the trained \( N \). With increasingly larger system size far above the trained \( N \), however, the performance level gradually tails off and retraining becomes necessary in order to maintain the same calibre of performance. The situation for total sweeping...
TABLE S1. Pseudo-code of PPO algorithm

PPO algorithm
1. Input: initial weights of policy network \( \theta_0 \), initial weights of value function \( \phi_0 \)
2. for \( k = 0, 1, 2, \ldots \) do
3. Collect multiple trajectories of spin state evolution \( D_k = \{ \tau_i \} \) under current policy \( \pi_{\theta_k} \) in self-defined Gym environment.
4. Computes rewards-to-go \( R_t = \sum_{l=0}^{c_2} |c_2|_{t_l} \sum_{t=0}^{c_2} \gamma^l \tau_{t+l} + V_{\phi_k}(s_{|c_2|_{t_l}}) \).
5. Use GAE-\( \lambda \) method and current value function \( V_{\phi_k} \) to estimate advantage function \( A_{\pi_k} \).
6. Maximize PPO-Clip lower bound function and update weights of policy network
\[
\theta_{k+1} = \arg \max_{\theta_k} \frac{1}{|D_k||c_2|_{t_c}} \sum_{\tau \in D_k} \sum_{t=0}^{c_2} L_{\text{clip}}^{\gamma}(s_t, a_t),
\]
usually using gradient descents methods such as Adam and SGD.
7. Minimize mean-squared error and update weights of value function
\[
\phi_{k+1} = \arg \min_{\phi_k} \frac{1}{|D_k||c_2|_{t_c}} \sum_{\tau \in D_k} \sum_{t=0}^{c_2} (V_{\phi_k}(s_t) - R_t)^2,
\]
usually using gradient descents methods such as Adam and SGD.
8. end for

TABLE S2. Training Hyperparameters for PPO

| Hyperparameters          | Value       |
|--------------------------|-------------|
| hidden size              | 64, 32, 16, 8 |
| activation               | tanh        |
| discounted factor \( \gamma \) | 0.99        |
| actor-network learning rate | 3E-4        |
| critic-network learning rate | 5E-4        |
| steps/episode            | 1000 \sim 20000 |
| target KL-divergence     | 0.01 \sim 0.1 |
| clip ratio \( \epsilon \) | 0.2         |
| GAE-\( \lambda \)        | 0.97        |

Due to the large effective range in \( N \), the multi-step training method we develop as described in the main text as well as in the previous section can facilitate enlarged system size efficiently until the limit of computation resource is approached. Similar to enlarged system size, we can also employ multi-step training to shorten sweeping time \( \tau \), according to the previous analysis of generalization ability.

COMPARING ADIABATIC SWEEPS WITH DRL PROTOCOL

In this section, we compare two types of adiabatic sweep with the DRL sweeping protocol for \( N = 2000 \) discussed as an example in the main text. According to quantum adiabatic theorem, a quantum state of a system follows the eigen-state it starts with under slow and continuous parameter change. For the model system we consider, starting from the polar ground state at large \( q > 0 \), sweeping \( q \) down to \( q = 0 \) slowly transforms adiabatically the ground state into balanced target Dicke state. At \( N = 2000 \), we find numerically the maximum achievable fidelity for linear adiabatic sweep approaches the performance of DRL profile (\( F \gtrsim 0.85 \)) at a sweep time \( |c_2|_{\tau} \gtrsim 600 \) [red circles in Fig. S3(a)]. For non-linear sweeping satisfying local adiabatic approximation (LAA) [3, 4], the sweep time reduces to \( |c_2|_{\tau} \gtrsim 350 \) [blue squares in Fig. S3(a)].

In the paradigm two-level Landau-Zener transition model, diabatic transition probability is given by,
\[
P = \exp \left( -\frac{2\pi |\Omega|^2}{|\frac{da}{dt} \frac{\partial}{\partial q} (E_2 - E_1)|} \right),
\]
(S4)
FIG. S2. Illustrating generalization ability for (a) the $N = 2000$ fidelity policy (without loss) at $|c_2|\tau = 15.5$ to $N \in [50, 10000]$ and $|c_2|\tilde{\tau} \in [0.4, 30]$, and (b) the $N = 11800$ sensitivity policy (including loss and detection noise) at $|c_2|\tau = 13$ to $N \in [8000, 50000]$ and $|c_2|\tilde{\tau} \in [0.2, 20]$. The sharp boundary (red dashed line) in (a) is fitted to $|c_2|\tau \propto N^{1/\alpha}$ with $\alpha \approx 3.8$, approximately reflecting the speed limit constrained by level spacing ($\sim 1/N^{1/3}$) near QCP in the second stage of our DRL protocol as analyzed in the main text. The deteriorating quality as $\tau$ increases in (b) comes from increased influence of atom loss, which will be discussed in the subsequent Sec.

when parameter $q(t)$ is swept across the avoided level crossing, with $\Omega$ denoting the Rabi frequency for two state coupling, $E_1(q)$ and $E_2(q)$ the parameter $q$-dependent eigen-energies of the two states (as illustrated in Fig. S3(b)). Mapped to our model, the Rabi frequency near quantum critical point (QCP) becomes approximately $\Omega \propto N^{-1/3}$, essentially the gap size. As $\partial (E_2 - E_1)/\partial q$ is determined by the specific $q$-dependent Hamiltonian, the diabatic transition rate is dominantly decided by sweep speed $|\partial q/\partial t|$. For a linear adiabatic sweep, its constant sweeping speed is determined by the minimum energy gap $\Delta E$ between ground and the first excited state, whereas the sweeping speed for non-linear sweep satisfying LAA is proportional to $\Delta E^2$. The DRL sweep profile we obtain is shown in Fig. S3(c1) and (c2). Indeed it exhibits a sufficiently slow speed near QCP region with which both the linear adiabatic sweep and the non-linear LAA sweep [3, 4] can achieve a high-fidelity target state, i.e., satisfying overall adiabaticity. This section of adiabatic sweep in the vicinity of QCP thus constitutes an unavoidable route one must overcome in order to connect the reconfigured state after the first stage of DRL profile to the coherent state like Gaussian wave packet in the Fock basis near the end of the middle stage of DRL profile as analyzed in the main text. The sweep speed slows down at $q \lesssim q_c$ to avoid excitation into higher energy states in the second stage.

MULTI-LEVEL OSCILLATION DURING QCP CROSSING

Typically, with a diabatic $q$-sweep starting at an eigen-state on the right hand side of $q > q_c$, excitation is inevitable as $q$ comes down rapidly close to QCP. The state projections onto the first four eigen-states at $q/|c_2| = 2.2$ illustrates clearly excitation channels in the critical gap curve (CGC) region (as shown in Fig. S4(a)). These excitation channels can cancel out through controlled multi-level destructive interference, e.g. the channels of the first and the second excited states null out such that the probability of staying in the ground state increases. This is akin to Rabi oscillation in a two-level system, where an initial superposition state can be timed to end up in either one of the two levels. Hence, instead of the difficult and time-consuming mission of staying at the ground eigen-state during sweeping down from $q > q_c$, the DRL policy chooses to start at a superposition state of low-lying levels in the first stage before crossing QCP, which is derailed from CGC. Here, we focus on the DRL sweep starting from $q/|c_2| = 2.2$ where the state is dominated by the first five levels. It is known in this case excited level populations can be almost suppressed completely via multi-level oscillation with a quench-$q$ scheme [5]. In our case, DRL finds a similar mechanism in action, significantly reducing excited state populations via a sophisticated control of sweep speed $|\partial q/\partial t|$ (as shown in Fig. S4(b)). The state before entering the QCP region ends up being dominated by the first two energy levels. After crossing QCP, an analogous $q$-controlled dynamical process allows for populating the first few excited levels. Hence, a nearly displaced ground state (a displaced Gaussian wave packet in an approximately simple harmonic trap) is prepared at the end of the second stage, facilitating a rapid translation into high-fidelity target state.
FIG. S3. (a) The dependence of maximum achievable fidelity on sweep time for linear adiabatic sweep (red circles) and non-linear sweep with LAA (blue squares). The initial state for these adiabatic sweeps is the ground state of $q/|c_2| = 3$ very close to the initial polar state (the ground state of $q/|c_2| \gg 2$) used in the DRL sweep profile. The dashed line denotes achievable fidelity level of the $N = 2000$ DRL profile. (b) Illustration of the Landau-Zener transition model in a two-level system of linear level crossing. (c1) Comparing the DRL sweep profile (black dashed line) at $|c_2|\tau = 15.5$ with linear adiabatic ramp at $|c_2|\tau = 600$ (red solid line) and LAA ramp with $|c_2|\tau = 350$ (blue solid line). Their corresponding sweep speeds around QCP are shown in (c2).

FIG. S4. (a) Projections to the ground and the first three excited states at $q/|c_2| = 2.2$ of the eigen-state basis: $P_n = |\langle \psi_n(q)|\psi_k(q/|c_2| = 2.2)\rangle|^2$ ($k = 1, 2, 3, 4$) for small $q$ at $N = 2000$. The red solid line denotes numerically calculated CGC. (b) Evolution of simulated probability distribution $W_n = |\langle \psi_n(q(t))|\psi(q(t))\rangle|^2$ following the $N = 2000$ DRL profile for the first six eigen-states in the instantaneous eigen-state basis $|\psi_n(q(t))\rangle$. The dot-dashed line represents the sweep speed of the DRL $q(t)$ profile. The vertical dashed line separates polar (P) phase ($q > q_c$) and broken-axisymmetry (BA) phase ($q < |q_c|$).

COMPARISON BETWEEN FULL QUANTUM AND MEAN-FIELD SIMULATIONS UNDER THE DRL PROTOCOL

Besides full quantum simulation governed directly by the model Hamiltonian which includes quantum fluctuation, mean-field dynamics neglecting quantum fluctuation is employed for large $N$ by replacing annihilation (creation)
operators with complex numbers: $a_m \rightarrow \sqrt{N} \zeta_m$, $a_m^* \rightarrow \sqrt{N} \zeta_m^*$. Within the $L_z = 0$ subspace, this leads to the following coupled equations

\begin{align}
  i\dot{\zeta}_1 &= c'(|\zeta_0|^2 \zeta_1 + |\zeta_1|^2 \zeta_0 - |\zeta_1|^2 \zeta_1 - \zeta_0^2 \zeta_1), \\
  i\dot{\zeta}_0 &= -q(t)\zeta_0 + [(|\zeta_1|^2 + |\zeta_1|^2)\zeta_0 + 2\zeta_1 \zeta_0^*], \\
  i\dot{\zeta}_{-1} &= c'(|\zeta_{-1}|^2 \zeta_{-1} + |\zeta_0|^2 \zeta_{-1} - |\zeta_1|^2 \zeta_{-1} - \zeta_0^2 \zeta_{-1}),
\end{align}

with $c' = c_2/N$. They are further simplified into two independent equations due to normalization condition $\sum_{m_F = -1,0,1} |\zeta_{m_F}|^2 = 1$ and the conserved magnetization $|\zeta_1|^2 = |\zeta_{-1}|^2$. Setting spinor phase $\theta_s = \chi_1 + \chi_{-1} - 2\chi_0$, the coupled equations reduce to

\begin{align}
  \dot{\rho}_0 &= 2c'\rho_0(1 - \rho_0) \sin \theta_s, \\
  \dot{\theta}_s &= -2q(t) + 2c'(1 - 2\rho_0) (1 + \cos \theta_s),
\end{align}

which behaves as a non-rigid pendulum found earlier [6].

Starting in polar state with $N_1/N = N_{-1}/N \approx 0$, the initial evolutions for $m_F = \pm 1$ components are dominated by quantum fluctuations, which are neglected unfortunately in mean-field theory. Hence, in the beginning near $|c_2|t = 0$, mean-field evolution (red dot-dashed line in Fig. S5(a)) is totally different from full quantum simulation (blue solid line). The situation changes from the beginning of the second stage as shown in Fig. S5(b), where evolutions begin to agree with each other although differences remain in their detailed structures, especially concerning $\theta_s(t)$ and $\dot{\theta}_s(t)$. However, beginning with the third stage as shown in Fig. S5(c), evolutions from the above two methods match perfectly with each other as the mean field is well established. Since $\psi(|c_2|t = 10)$ is already nearly an ideal coherent state like Gaussian wave packet in the Fock basis, its dynamics is well described semi-classically by the mean-field approach. Hence, the control problem for the DRL agent at this stage is analogous to the classical control of a non-rigid pendulum.
 CONNECTION BETWEEN OUR MODEL AND THE LIPKIN-MESHKOV-GLICK MODEL

The spin-1 system we consider is fully described by the SU(3) symmetry group and its correspondingly generated Lie algebra. The infinitesimal generators we employ are the Cartesian dipole-quadrupole decomposition of the Lie algebra su(3) \[7\]. This leads to three dipole (or angular momentum) operators that satisfy SU(2) commutation relationships, \[ (1 + a_0 a_0^\dagger ) / 2 \] with \( \lambda \). Hence, if we set \( \gamma / N \approx \gamma _x / \sqrt{2} \) and focus on the \( L_z = 0 \) subspace, the Hamiltonian (S13) can be simplified into, \( \hat{H} = c_2 / 2N \mathbf{L}^2 - q(t)N_0 \), which is exactly the Hamiltonian of our spin-1 system.

AN APPROXIMATE SIMPLE HARMONIC OSCILLATOR HAMILTONIAN IN THE BROKEN-AXISYMMETRY PHASE

As mentioned in the previous section, in the \( L_z = 0 \) subspace and for \( N \simeq N_0 \), our spin-1 Hamiltonian can be described through \( \{ L_x, N_{yz}, (N_{zz} - N_{yy}) \} \) as,

\[
H = - \frac{1}{N} L_x^2 + \lambda (N_{zz} - N_{yy}),
\]

with \( \lambda = q/2|c_2| \). Since the commutator \( [H, L_x^2 + N_{yy}^2 + (N_{zz} - N_{yy})^2] \) vanishes and \( \langle L_x^2 + N_{yy}^2 + (N_{zz} - N_{yy})^2 \rangle \simeq N/2(N/2 + 1) \), we can take a semi-classical approximation by considering the Hamiltonian Eq. (S17) on a sphere with radius \( N/2 \) \[10\],

\[
H = \frac{N}{2} (-\lambda \sin \theta \cos \varphi - \frac{1}{2} \cos^2 \theta),
\]
with polar angle $\theta$ and azimuthal angle $\varphi$, which give $L_x = N \cos \theta / 2$, $N_{yz} = N \sin \theta \sin \varphi / 2$, and $N_{zz} - N_{yy} = -N \sin \theta \cos \varphi$. Setting $\mu = \cos \theta$, this Hamiltonian can be further simplified into,

$$K \equiv \frac{2H}{N} = -\lambda \sqrt{1 - \mu^2} \cos \varphi - \frac{1}{2} \mu^2,$$

(S19)

satisfying the Poisson bracket $\{\mu, \varphi\} = 2 / N$. For low-lying states, we expand $K$ around its minimum which is found from

$$\sin \varphi_0 = 0, \text{ and } \mu_0 = \begin{cases} 0, & (\lambda > 1) \\ \pm \sqrt{1 - \lambda^2}, & (\lambda \leq 1) \end{cases},$$

(S20)

and the corresponding minimum value of $H$ is

$$H_0 = \begin{cases} -\frac{\lambda N}{2}, & (\lambda > 1) \\ -\frac{\lambda}{4} (1 + \lambda^2), & (\lambda \leq 1) \end{cases}.$$

(S21)

Focusing on the broken-axisymmetry (BA) phase region ($\lambda \leq 1$), and expanding around the minimum of $\mu = \mu_0$ and $\varphi = 0$, and keeping the lowest order terms, the Hamiltonian (S19) becomes,

$$K = -\frac{1}{2} - \frac{\lambda^2}{2} + \frac{1}{2\lambda^2} \epsilon^2 - \frac{\lambda^2}{2} \varphi^2$$

(S22)

for low-lying levels, where $\epsilon = \mu - \mu_0$. If $\mu$ and $\varphi$ are taken as canonical conjugate variables for ‘position’ and ‘momentum’, satisfying the commutator $[\mu, \varphi] = i\hbar$ according to Poisson bracket, the harmonic spectra is obtained from Hamiltonian (S22) with frequency $\omega = \sqrt{1 - \lambda^2}$ and mass $m = 1 / \lambda^2$. The comparison between this mean-field spectrum and results from exact diagonalization are shown in Fig. S5.

![Graph](image.png)

**FIG. S6.** Comparing the nearest neighbour level spacing for the $q$-dependent eigen-energies $\Delta E_j = E_{j+1}(q) - E_j(q)$ (gray lines) for the first ten low-lying levels with the approximate harmonic frequency $\sqrt{1 - \lambda^2}$ (red dashed line) of Hamiltonian (S22) at $N = 2000$.

**DETERIORATING DICKE STATE QUALITY WITH ATOM LOSS**

In our experiments, both detection noise $\sigma_{DN}$ and atom loss degrade the observed final Dicke state quality. This section addresses the influence of atom loss by varying the loss rate $\gamma$ in the simulation with $N_f = N_i \exp(-\gamma \tau)$ and using the $N = 2000$ DRL policy with fidelity as reward for $N = 11800$, since detection noise is well understood and can be directly calibrated. The simulated interferometric sensitivity deteriorates is shown in Fig. S6(a) and analogously for entanglement depth of the final state [11–13] shown in Fig. S6(b). The empirical protocol adopted previously [14] demands 1.5 s to complete the sweep, which results in 5% loss of atoms (marked by a cross in Fig. S6). At $N = 11800$ and including atom loss, the sweep time for the DRL protocol we find shortens to 766 ms with the corresponding atom loss reduced to 2.5% (marked by a circle in Fig. S6). As shown in Fig. S6, this reduction in atom loss leads to about 2.5 dB improvement to sensitivity and number of entangled atoms with one standard deviation confidence level increased from less than 1000 to about 3000 atoms.
FIG. S7. Numerically calculated deteriorating Dicke state quality due to atom loss for \( N = 11800 \) using the \( N = 2000 \) DRL policy with fidelity as reward (for without atom loss). (a) The corresponding interferometric sensitivity \( (\delta \theta)^2 \), and (b) entanglement depth. The squares and circles mark the locations for 5% and 2.5% atom lost respectively.

**SENSITIVE DEPENDENCE ON PARAMETERS OF DRL PROFILE**

As mentioned in section , due to favorable generalization ability, the DRL agent trained at \( N = 2000 \) without loss can be directly applied to a system of \( 10^4 \) atoms to achieve a final state fidelity \( \mathcal{F}(\tau) > 0.8 \). For our experiment carried out at \( N = 11800 \), such a DRL policy gives a target state fidelity \( \mathcal{F} \simeq 0.85 \) within \( |c_2|\tau = 24 \) (about 1500 ms for \( c_2 = -2\pi \times 2.7 \) Hz), which can facilitate an interferometric sensitivity of \(-10 \log(4N(\delta \theta)^2) \simeq 37.7 \) dB, approaching to Heisenberg limit. In contrast, the profile learnt from environment including loss discussed in the main text with \( |c_2|\tau = 13 \) only provides \(-10 \log(4N(\delta \theta)^2) \simeq 22.1 \) dB improvement in theory.

In a real experiment, it is always helpful to take into account the sensitive dependence on parameters before the theoretical DRL sweeping profiles are implemented. According to the analysis in the main text, the third stage of our DRL profile corresponds formally to a STA problem of translating a wave packet of a simple harmonic oscillation in the Fock basis. According to previous studies, such a translational displacement is susceptible to excitation of dipole (sloshing) mode \([15, 16]\) with a STA. For a Gaussian wave packet in a perfect harmonic trap, the residual amplitude of the dipole mode is given by \([15, 16]\),

\[
\mathcal{A} = \left| \int_0^{c_2\tau} \exp(-i\omega t)\dot{q}(t)dt \right|,
\]

which depends on the sweeping time \( |c_2|\tau \) and the profile \( q(t) \). However, the coupling strength \( |c_2| \) fluctuates from experiment to experiment due to initial atom number uncertainty (shot-noise in BEC state preparation) and decays due to atom loss. In the end, for any chosen implementation of sweep profile \( q(t) \), variations of \( |c_2| \) result in dipole mode excitations, which finally lead to (averaged) deteriorating interferometric performance.

In the following, we compare numerically the high-fidelity policy profile from without loss with the one including loss, taking into account for both decaying \( |c_2| \) and shot-noise on initial atom numbers in Fig. S7. Based on the residual sloshing motion in the Fock basis after sweep, we find the profile from without loss is sensitive to variation of \( |c_2| \) (see Figs. S7(a2) and (a3)), while the one including loss is relatively insensitive (see Figs. S7(b2) and (b3)). Thus, the profile trained under more realistic environment including loss is experimentally feasible, and it is chosen for the reported experiments.

**EXPERIMENTAL METHOD**

A BEC with about \( 8.3 \times 10^4 \) atoms is produced in the \( m_F = -1 \) state inside an optical dipole trap formed by two crossed 1064-nm light beams in our system \([14, 17]\). The quantization axis is defined by applying a fixed magnetic field of 815 mG along the direction of gravity, which gives \( q_B \simeq 17|c_2| \). The root-mean-square uncertainty of the magnetic field is suppressed to under 20 \( \mu \)G with the application of magnetic feedback constructed from a temperature-feedback fluxgate magnetometer. A RF \( \pi/2 \)-pulse is then applied to transfer atoms to the \( m_F = 0 \) component and a gradient
FIG. S8. Calculated evolution of excitation spectra $W_k = |\langle \psi(t) | k \rangle|^2$ in the Fock basis $| k, N - 2k, k \rangle$. (a1), (a2), and (a3) denote results from the high-fidelity profile from without loss for 11800 atoms, respectively for a constant $|c_2|$, a decaying $|c_2(t)|$, and a decaying $|c_2(t)|$ with initial atom number deviation ($N \rightarrow N - \sqrt{N}$). (b1), (b2) and (b3) represent the corresponding results from the DRL profile learnt including loss for 11800 atoms. The red dashed-dotted lines denote the mean position in the Fock basis and the black dashed lines represent end of the sweep, after which $q$ is fixed to 0 ($t > \tau$).

magnetic field about 200 G/cm is ramped up to remove remaining atoms in the $m_F = \pm 1$ components. The consequent atom number in the $m_F = 0$ component can be flexibly controlled via changing intensity of the RF pulse. After that, the power of trapping light beams is lowered in 1.8 s to further evaporate and produce a condensate with about 11800 atoms before the gradient magnetic field is switched off. Next, the trap is compressed to its final trapping frequencies of $2\pi \times (237, 112, 183)$ Hz along three orthogonal directions in 300 ms. Meanwhile, $q$ is ramped from $\sim 17|c_2|$ to $3|c_2|$ by linearly tuning up the dressing microwave. This prepares a BEC sample for subsequent experiment to generate balanced Dicke state.

The application of DRL profile starts from abruptly jumping $q$ to the staring value $q(0) < 2|c_2|$ and following the DRL sweep profile for 766 ms. A 8 ms Stern-Gerlach separation is then applied after switching off the trap to obtain normalized atom numbers in each $m_F$ components by taking absorption images. To measure the effective spin length, a RF-resonant $\pi/2$-pulse coupling $|F = 1, m_F = 0\rangle$ and $|F = 1, m_F = \pm 1\rangle$ is applied to rotate the spin-1 Dicke state. The same Stern-Gerlach process is then implemented after rotation to obtain normalized atom numbers.

The calibration of $q$ and $|c_2|$ gives $|c_2| = 2\pi \times 2.7$ Hz [14, 17], at which experimental results match numerical simulations with the same value.

* These authors contributed equally to this work.
† mengkhoon_ney@tsinghua.edu.cn
‡ lyou@tsinghua.edu.cn

[1] OpenAI Spinning Up, URL https://spinningup.openai.com/.
[2] OpenAI Gym, URL https://gym.openai.com/.
[3] P. Richerme, C. Senko, J. Smith, A. Lee, S. Korenblit, and C. Monroe, Phys. Rev. A 88, 012334 (2013), URL https://link.aps.org/doi/10.1103/PhysRevA.88.012334.
[4] S. Balasubramanian, S. Han, B. T. Yoshimura, and J. K. Freericks, Phys. Rev. A 97, 022313 (2018), URL https://link.aps.org/doi/10.1103/PhysRevA.97.022313.
[5] P. Xu, S. Yi, and W. Zhang, Phys. Rev. Lett. 123, 073001 (2019), URL https://link.aps.org/doi/10.1103/PhysRevLett.123.073001.
[6] W. Zhang, D. L. Zhou, M.-S. Chang, M. S. Chapman, and L. You, Phys. Rev. A 72, 013602 (2005), URL https://link.aps.org/doi/10.1103/PhysRevA.72.013602.
[7] Y. Di, Y. Wang, and H. Wei, Journal of Physics A: Mathematical and Theoretical 43, 065303 (2010), ISSN 1751-8121, URL http://dx.doi.org/10.1088/1751-8113/43/6/065303.
[8] I. Carusotto and E. J. Mueller, Journal of Physics B: Atomic, Molecular and Optical Physics 37, S115 (2004), ISSN
[9] C. D. Hamley, C. S. Gerving, T. M. Hoang, E. M. Bookjans, and M. S. Chapman, Nature Physics 8, 305 (2012), ISSN 1745-2481, URL http://dx.doi.org/10.1038/nphys2245.

[10] F. Leyvraz and W. D. Heiss, Phys. Rev. Lett. 95, 050402 (2005), URL https://link.aps.org/doi/10.1103/PhysRevLett.95.050402.

[11] A. S. Sørensen and K. Mølmer, Phys. Rev. Lett. 86, 4431 (2001), URL https://link.aps.org/doi/10.1103/PhysRevLett.86.4431.

[12] B. Lücke, J. Peise, G. Vitagliano, J. Arlt, L. Santos, G. Tóth, and C. Klempt, Phys. Rev. Lett. 112, 155304 (2014), URL https://link.aps.org/doi/10.1103/PhysRevLett.112.155304.

[13] G. Vitagliano, I. Apellaniz, M. Kleinmann, B. Lücke, C. Klempt, and G. Tóth, New Journal of Physics 19, 013027 (2017), URL https://doi.org/10.1088%2F1367-2630%2F19%2F01%2F013027.

[14] Y.-Q. Zou, L.-N. Wu, Q. Liu, X.-Y. Luo, S.-F. Guo, J.-H. Cao, M. K. Tey, and L. You, Proceedings of the National Academy of Sciences 115, 6381 (2018), ISSN 0027-8424, URL https://doi.org/10.1073/pnas.1706504115.

[15] A. Couvert, T. Kawalec, G. Reinaudi, and D. Guéry-Odelin, EPL (Europhysics Letters) 83, 13001 (2008), ISSN 1286-4854, URL http://dx.doi.org/10.1209/0295-5075/83/13001.

[16] G. Ness, C. Shkedrov, Y. Florshaim, and Y. Sagi, New Journal of Physics 20, 095002 (2018), ISSN 1367-2630, URL http://dx.doi.org/10.1088/1367-2630/aadcc1.

[17] X.-Y. Luo, Y.-Q. Zou, L.-N. Wu, Q. Liu, M.-F. Han, M. K. Tey, and L. You, Science 355, 620 (2017), ISSN 0036-8075, https://science.sciencemag.org/content/355/6325/620.full.pdf, URL https://science.sciencemag.org/content/355/6325/620.