TIME-CONDITIONED NEURAL QUANTUM STATES WITH RESIDUAL-BASED FIDELITY BOUNDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning quantum dynamics by minimizing the Schrödinger residual raises a basic question: what does the residual imply about state accuracy? We introduce Abel-tNQS, a time-conditioned autoregressive neural quantum state with an exact initial-state constraint, and relate its Hilbert residual to time-averaged fidelity. The complex conditional-amplitude construction preserves normalization and direct sampling, provided the local normalization vectors remain nonzero. A linear Abel bound supplies the remaining-time weight $(T - t)$ for a separate training surrogate; a cumulative residual form gives a tighter post-training fidelity bound. Full-basis evaluations on small Ising chains yield numerical lower-bound estimates between 0.9972 and 0.9999, without certified numerical error control. In controlled full-gradient $N = 10$ ablations with ten seeds per setting, linear weighting reduces final mean magnetization and nearest-neighbor correlation errors by approximately 85% and 94% relative to uniform weighting. After 3000 optimizer updates, the hard-linear setting also has lower mean errors in both observables than all three tested soft-penalty settings. Observable benchmarks cover chains of up to 20 spins, an open 4×4 lattice, segmented continuation, and time-dependent transverse fields.

1 INTRODUCTION

Real-time quantum many-body dynamics is difficult to simulate: a generic state of N spins has 2^N amplitudes, while entanglement growth limits tensor-network approaches during quenches (Vidal, 2004; Schollwöck, 2011; Paeckel et al., 2019). Neural quantum states (NQS) offer a variational representation of wave functions (Carleo & Troyer, 2017; Lange et al., 2024). A compact representation alone does not establish accuracy: agreement in selected observables can coexist with errors in the underlying quantum state. This raises a central question: how can the accuracy of a learned trajectory be assessed without access to the exact state?

Time-dependent variational Monte Carlo (tVMC) evolves NQS step by step (Carleo et al., 2012; Schmitt & Heyl, 2020), whereas global-in-time methods fit an entire trajectory by minimizing a dynamical residual (Van de Walle et al., 2025; Sinibaldi et al., 2026). Several such methods already incorporate structural initial conditions and residual-based diagnostics or error control.

We introduce **Abel-tNQS**, a normalized autoregressive trajectory with an exact initial-state constraint. A linear bound on time-averaged infidelity exposes the remaining-time weight $(T - t)$ used in our training surrogate. For assessing a trained trajectory, we also use a tighter cumulative bound obtained from the same Hilbert residual. The initial condition is imposed on complex conditional amplitudes, preserving normalization and direct sampling under the stated non-vanishing condition. Figure 1 illustrates how one parameter set represents the full trajectory.

We first evaluate the residual-based fidelity bounds on small chains by full-basis calculation. We then examine observable accuracy on chains of 10 and 20 spins and an open 4×4 lattice. Finally, controlled ablations compare temporal weighting and initial-condition protocols.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

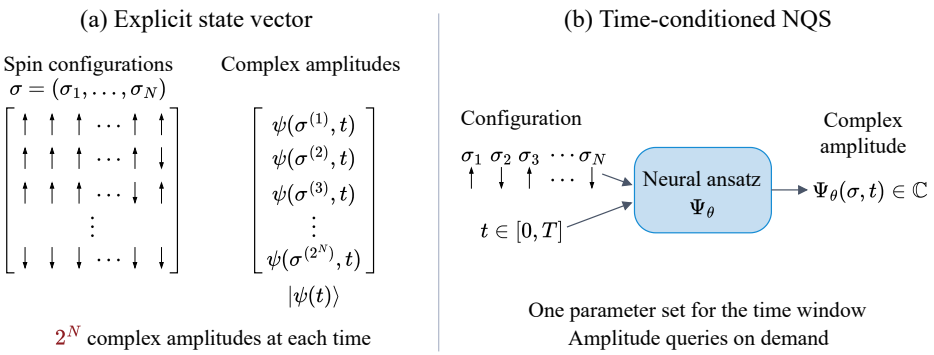

Figure 1: Explicit and time-conditioned representations of a quantum state. (a) For N spin-$1/2$ sites, the state $|\psi(t)\rangle$ is represented in the computational basis by 2^N complex amplitudes $\psi(\sigma, t) = \langle\sigma|\psi(t)\rangle$. Configurations $\sigma^{(k)}$ enumerate the basis; selected rows are shown. (b) The neural ansatz maps a configuration σ and time $t \in [0, T]$ to a complex amplitude $\Psi_\theta(\sigma, t)$, where T is the window duration and θ denotes the shared trainable parameters.

2 Related Work

Variational time evolution. The time-dependent variational principle projects dynamics onto a variational manifold (McLachlan, 1964; Haegeman et al., 2011; Hackl et al., 2020). Standard tVMC implementations repeatedly solve linear systems involving the quantum geometric tensor, with regularization and sampling affecting stability (Schmitt & Heyl, 2020; Hofmann et al., 2022). Midpoint, autoregressive, projected, and unbiased variants address aspects of these difficulties (Gutiérrez & Mendl, 2022; Donatella et al., 2023; Gravina et al., 2025; Krinitsin & Schmitt, 2026). Abel-tNQS instead fits a finite window jointly without forming this tensor.

Autoregressive states. Autoregressive models (Larochelle & Murray, 2011; Germain et al., 2015) and NQS architectures (Sharir et al., 2020; Hibat-Allah et al., 2020; Viteritti et al., 2023) provide normalization and direct ancestral sampling through normalized conditionals, with possible expressivity trade-offs (Bortone et al., 2024). Our initial-state constraint operates on complex conditional amplitudes; Section 4.1 gives its normalization condition.

Global-in-time learning. Residual learning (Raissi et al., 2019) and history-state constructions (Vargas-Calderón et al., 2023) offer global-in-time alternatives to time stepping. The closest approaches are those of Van de Walle et al. (2025), who use a time-conditioned autoregressive transformer, a squared Hilbert residual evaluated through Born-sampled log-local terms, and an initialization layer, and Sinibaldi et al. (2026), who combine static NQS with time-dependent coefficients and derive state and observable error bounds from a projective residual. Our construction pairs a complex conditional-amplitude constraint with a remaining-time-weighted training surrogate. Its Hilbert residual supports both the linear Abel relation used to select the weight and a cumulative bound for post-training fidelity assessment. We evaluate residual components without dividing by wave-function amplitudes. Table 2 compares the representations, residual forms, and error diagnostics of these global-in-time approaches. Unlike approaches learning across families of initial states or Hamiltonians (Rende et al., 2025; Qi et al., 2026), we fit one trajectory for a specified initial state and protocol.

3 From Schrödinger Residuals to Fidelity Bounds

3.1 Problem Setting and Hilbert Residual

Dynamical setting. For N spin-$1/2$ sites, we seek a neural approximation to the exact state $|\psi(t)\rangle$ on $t \in [0, T]$, where t is physical time and $T > 0$ is the window duration. In units where $\hbar = 1$, the

exact state evolves according to

$$i\partial_t|\psi(t)\rangle = \hat{H}(t)|\psi(t)\rangle, \qquad |\psi(0)\rangle = |\psi_0\rangle, \tag{1}$$

where i is the imaginary unit, $\hat{H}(t)$ is the Hermitian Hamiltonian, and $|\psi_0\rangle$ is a known normalized initial state. The computational basis has dimension $D = 2^N$ and configurations $\sigma = (\sigma_1, \ldots, \sigma_N)$ with local spin labels $\sigma_j \in \{\uparrow, \downarrow\}$, $j = 1, \ldots, N$. We write $\Psi_\theta(\sigma, t) = \langle\sigma|\Psi_\theta(t)\rangle$ for the amplitude of configuration σ in the learned state $|\Psi_\theta(t)\rangle$. Here θ denotes the trainable parameters, and $\psi_0(\sigma) = \langle\sigma|\psi_0\rangle$ is the initial-state amplitude.

We use Dirac notation for quantum states, hats for operators, and $\|\cdot\|$ for the Hilbert-space norm. Bold upright letters denote coordinate vectors and matrices, while $\mathcal{X}$ denotes higher-order feature tensors; σ denotes a configuration tuple, and θ denotes the trainable parameters.

Hilbert residual. Training uses no target states for $t > 0$. Let $|\psi_t\rangle = |\psi(t)\rangle$ and $|\Psi_t\rangle = |\Psi_\theta(t)\rangle$ be normalized, absolutely continuous trajectories on the finite-dimensional Hilbert space, with $|\psi_t\rangle$ satisfying Eq. 1 almost everywhere. Assume that $\hat{H}(t)$ is measurable and Hermitian with integrable operator norm $\|\hat{H}(t)\|_{\mathrm{op}}$ on $[0, T]$. Define the Hilbert residual by

$$|d_\Psi(t)\rangle = \partial_t|\Psi_t\rangle + i\hat{H}(t)|\Psi_t\rangle. \tag{2}$$

The scalar components are $d_\Psi(\sigma, t) = \langle\sigma|d_\Psi(t)\rangle$ and require no division by $\Psi(\sigma, t)$; below, $\Psi(\sigma, t) = \Psi_\theta(\sigma, t)$ suppresses the parameter subscript. The bound applies to time-dependent Hamiltonians in any lattice dimension under these assumptions.

3.2 Time-Averaged Fidelity Bound

Abel residual functional. Let $r(t) = \big\||d_\Psi(t)\rangle\big\|$. We define

$$\mathcal{C}_{\mathrm{Abel}}(\Psi) = \int_0^T (T - t)\, r(t)\, dt. \tag{3}$$

We write $B_{\mathrm{Abel}} = \mathcal{C}_{\mathrm{Abel}}/T$ for the corresponding upper bound on average infidelity when the initial state is exact.

Proposition 1 (time-averaged fidelity bound). Under the assumptions in Section 3.1, let $F(t) = |\langle\psi_t|\Psi_t\rangle|^2$ be the instantaneous fidelity and $\bar{F} = T^{-1}\int_0^T F(t)dt$ its time average. Then

$$1 - \bar{F} \le 1 - F(0) + \frac{\mathcal{C}_{\mathrm{Abel}}(\Psi)}{T}. \tag{4}$$

To see this, normalization cancels the real contribution of the residual parallel to $|\Psi_t\rangle$, giving, almost everywhere,

$$|\dot{F}(t)| \le 2\sqrt{F(t)(1 - F(t))}\,\big\||d_\Psi(t)\rangle\big\| \le \big\||d_\Psi(t)\rangle\big\|, \tag{5}$$

where $\dot{F} = dF/dt$. Abel integration by parts applied to $1 - F(t)$ proves Eq. 4; Appendix A.1 gives the full proof. Equation 4 controls the continuous-time average of the fidelity and does not by itself provide an endpoint or pointwise-in-time guarantee.

Bounds with exact initialization. The conditional constraint in Section 4.1 enforces $|\Psi_0\rangle = |\psi_0\rangle$, hence $F(0) = 1$ and

$$\bar{F} \ge \max\left\{0, 1 - \frac{\mathcal{C}_{\mathrm{Abel}}(\Psi)}{T}\right\}. \tag{6}$$

The exact initial state also permits cumulative residual control. With $R(t) = \int_0^t r(s)\, ds$, unitary propagation gives $\big\||\Psi_t\rangle - |\psi_t\rangle\big\| \le R(t)$. Since $1 - F(t)$ is at most both this squared distance and 1, we obtain

$$1 - \bar{F} \le B_{\mathrm{clip}} := \frac{1}{T}\int_0^T \min\{1, R(t)^2\}\, dt \le \frac{1}{T}\int_0^T R(t)\, dt = \frac{\mathcal{C}_{\mathrm{Abel}}}{T}. \tag{7}$$

We use the tighter cumulative bound for post-training assessment. The linear Abel form makes the remaining-time weight $T - t$ explicit; we use this weight in the separate training surrogate

of Section 4.2 and test its effect in Section 7. Appendix A.1 proves Eq. 7 and compares related squared-residual bounds; Appendix A.2 derives an observable-error bound.

Numerical interpretation. We denote numerical estimates by $\widehat{B}_{\mathrm{clip}}$ and $\widehat{B}_{\mathrm{Abel}} = \widehat{\mathcal{C}}_{\mathrm{Abel}}/T$. Full-basis or full-support sampling estimators of the residual norm are described in Appendix A.3. Substituting point estimates into Eqs. 6 and 7 does not give numerically certified guarantees unless normalization, differentiation, quadrature, and any sampling errors are controlled. The small-chain experiments use full-basis summation (Section 5).

4 Abel-tNQS with Exact Initial Conditions

4.1 Autoregressive Trajectory and Exact Initial Condition

We remove the initial-infidelity term $1 - F(0)$ in Eq. 4 by enforcing $\Psi(\sigma, 0) = \psi_0(\sigma)$ in the architecture, so $F(0) = 1$ without a separate initial-condition penalty. The construction also preserves normalization, as required by the fidelity-rate bound, and direct autoregressive sampling under the condition below. It requires tractable initial-state conditionals.

Autoregressive representation. We represent the wave function as an autoregressive product:

$$\Psi_\theta(\sigma, t) = \prod_{i=1}^{N} \psi_i(\sigma_i \mid \sigma_{<i}, t), \tag{8}$$

where $\sigma_{<i} = (\sigma_1, \ldots, \sigma_{i-1})$ is the preceding spin configuration (empty for $i = 1$), and each ψ_i assigns a complex conditional amplitude to each of the $d = 2$ local spin values. In the ideal construction, local normalization,

$$\sum_{v \in \{\uparrow,\downarrow\}} |\psi_i(v \mid \sigma_{<i}, t)|^2 = 1 \quad \forall i, \sigma_{<i}, t, \tag{9}$$

gives $\sum_\sigma |\Psi_\theta(\sigma, t)|^2 = 1$. Ancestral sampling with $p_i(v \mid \sigma_{<i}, t) = |\psi_i(v \mid \sigma_{<i}, t)|^2$ therefore samples $|\Psi_\theta(\sigma, t)|^2$ without a global partition function.

Conditional constraint. Let $v, w \in \{\uparrow, \downarrow\}$ denote local spin values. The initial and learned conditional amplitudes are $\phi_i(v \mid \sigma_{<i})$ and $f_i(v \mid \sigma_{<i}, t)$, each normalized over the local spin. The initial conditionals satisfy $\prod_i \phi_i(\sigma_i \mid \sigma_{<i}) = \psi_0(\sigma)$. For a fixed prefix and time, suppressing these arguments, define

$$\begin{aligned} \psi_i^{\mathrm{raw}}(v) &= (1 - g(t))\phi_i(v) + g(t) f_i(v), \\ \psi_i(v) &= \frac{\psi_i^{\mathrm{raw}}(v)}{\sqrt{\sum_{w \in \{\uparrow,\downarrow\}} |\psi_i^{\mathrm{raw}}(w)|^2}}. \end{aligned} \tag{10}$$

Here $\psi_i^{\mathrm{raw}}(v)$ is the unnormalized complex amplitude. The denominator sums over both $d = 2$ local values at the same prefix and time, not over all configurations. The normalized scalar $\psi_i(v \mid \sigma_{<i}, t)$ enters Eq. 8. The implementation uses the finite-window power gate

$$g(t) = 1 - \left(1 - \frac{t}{T}\right)^{T/\tau}, \qquad 0 \le t \le T, \quad T > 0, \quad \tau > 0. \tag{11}$$

Here τ is the positive time scale controlling the gate's initial slope. The gate satisfies $g(0) = 0$, $g(T) = 1$, and $g'(0) = 1/\tau$; for the default $T = \tau = 1$, it is linear.

Three properties follow. First, $g(0) = 0$ gives $\Psi(\sigma, 0) = \prod_i \phi_i(\sigma_i \mid \sigma_{<i}) = \psi_0(\sigma)$ exactly. Second, the following non-vanishing condition must hold for every site, prefix, and time in $[0, T]$:

$$\sum_{v \in \{\uparrow,\downarrow\}} |\psi_i^{\mathrm{raw}}(v \mid \sigma_{<i}, t)|^2 > 0 \tag{12}$$

When this condition holds, the ideal conditionals are normalized and $\sum_\sigma |\Psi(\sigma, t)|^2 = 1$. Third, $g(T) = 1$ lets the next window inherit the preceding model's frozen endpoint conditionals, preserving continuity while carrying forward any error relative to exact evolution (Appendix A.9).

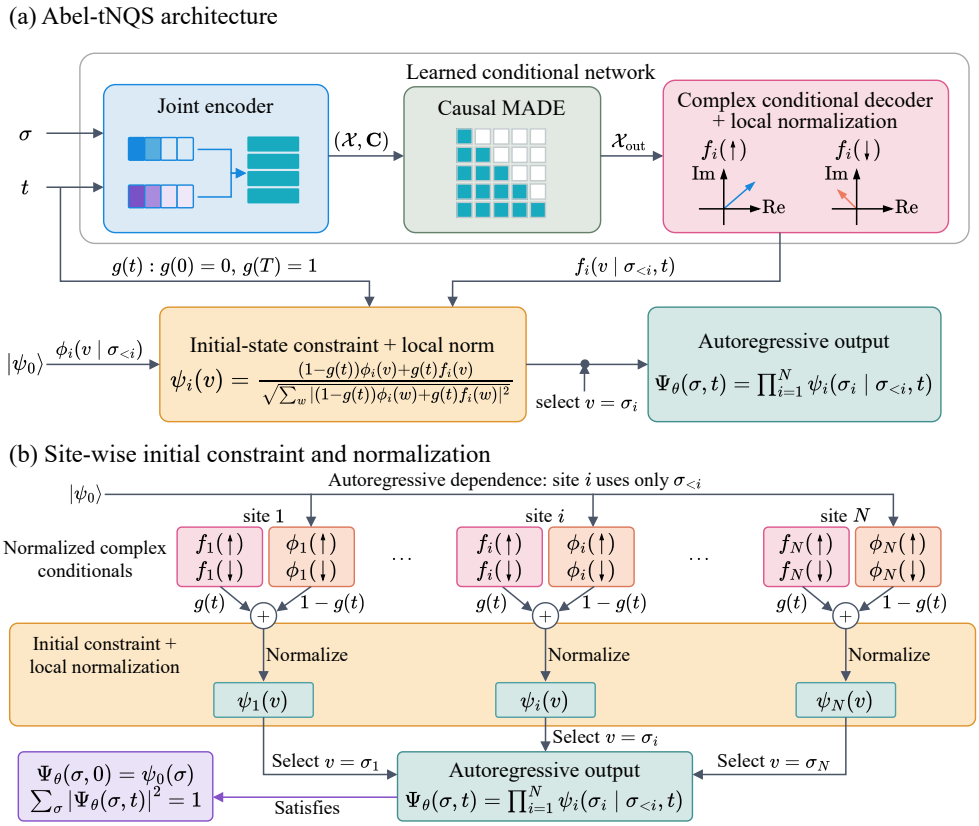

Figure 2: Abel-tNQS architecture and site-wise constraints. (a) MADE maps the encoder's spatial-feature tensor $\mathcal{X}$ and time-conditioning matrix $\mathbf{C}$ to causal features $\mathcal{X}_{\text{out}}$, which are decoded into conditional amplitudes $f_i(v \mid \sigma_{<i}, t)$ normalized over v. Re and Im denote real and imaginary parts; feature shapes appear in Appendix A.7. (b) At each site, complex amplitudes are mixed and normalized jointly over $v \in \{\uparrow, \downarrow\}$ as in Eq. 10. Selecting $v = \sigma_i$ and multiplying the selected scalars yields $\Psi_\theta(\sigma, t)$. Prefix and time arguments are suppressed in the local boxes. The shared initial-state input supplies the time-independent conditionals $\phi_i(v \mid \sigma_{<i})$ to each site. Normalized initial conditionals reconstructing ψ_0, together with $g(0) = 0$, enforce the initial state exactly. Global normalization requires Eq. 12, subject to the numerical caveat discussed in Section 4.1.

In the implementation, local norms are clamped below by ϵ_{norm}. The normalization identities hold up to floating-point error when the raw-vector norms at both normalization stages remain at or above this threshold for every site, prefix, and time in $[0, T]$. Appendix A.5 details the edge cases and admissible initial states.

4.2 Architecture and Training Objective

Network architecture. The network produces the learned conditional amplitudes $f_i(v \mid \sigma_{<i}, t)$ through a joint state–time encoder, a causal masked autoencoder for distribution estimation (MADE) (Germain et al., 2015), and a complex decoder (Figure 2). These amplitudes are mixed with the initial conditionals and locally normalized as in Section 4.1.

The causal mask restricts the ith output to the prefix $\sigma_{<i}$. Tensor shapes and implementation details appear in Appendix A.7.

Training surrogate. Evaluating the Hamiltonian action $\langle\sigma|\hat{H}(t)|\Psi_\theta(t)\rangle$ requires connected configurations or a stochastic estimator. The static-quench benchmarks in Figure 4 use the Abel-weighted

mean-absolute residual surrogate

$$\widehat{\mathcal{J}}_{\mathrm{abs}} = \frac{1}{n_t n_s} \sum_{k=1}^{n_t} \sum_{j=1}^{n_s} (T - t_k) \sqrt{|d_\theta^{\mathrm{FD}}(\sigma_j, t_k)|^2 + \delta_{\mathrm{res}}}, \tag{13}$$

$$d_\theta^{\mathrm{FD}}(\sigma, t) = \frac{\Psi_\theta(\sigma, t_+) - \Psi_\theta(\sigma, t_-)}{t_+ - t_-} + i\langle\sigma|\hat{H}(t)|\Psi_\theta(t)\rangle. \tag{14}$$

Sampled microbatches use n_t uniform times and n_s uniform configurations shared across times; the $N = 10$ continuation instead enumerates the basis and uses stratified uniform times (Appendix A.12). Equation 14 defines the finite-difference residual, with boundary-clipped $t_\pm$ from Eq. 37. The smoothing constant is $\delta_{\mathrm{res}} = 10^{-12}$. The Figure 4 runs evaluate $\widehat{\mathcal{J}}_{\mathrm{abs}}$ without a positive floor on its temporal weight; other loss scales are specified in Appendix A.12. Equation 13 averages residual magnitudes over configurations, whereas both fidelity bounds use the Hilbert-residual norm. We evaluate the bounds separately from the training surrogate.

Optimization protocol. We update the parameters using AdamW (Loshchilov & Hutter, 2019) and approximate time derivatives by finite differences with $\varepsilon_t = 10^{-3}$. Throughout, one training epoch denotes one optimizer update after gradient accumulation. To reduce backpropagation cost, the observable benchmarks in Figure 4 use semi-gradient updates: they evaluate the full residual in Eq. 14 but stop parameter differentiation through its Hamiltonian-action term. The controlled $N = 10$ ablations use full residual gradients under a common optimization protocol. Appendix A.8 specifies the boundary treatment, residual-gradient conventions, and training schedules for each experiment; Table 3 summarizes the optimization protocols.

The analytic bound applies to any learned trajectory satisfying its assumptions, irrespective of the optimization protocol. It motivates the training weight $T - t$ but does not require it; Section 7 compares this weight with uniform weighting.

5 Numerical Evaluation of the Fidelity Bounds

We evaluate Abel-tNQS on the transverse-field Ising model (TFIM) with open boundaries (Pfeuty, 1970; Czischek et al., 2018):

$$\hat{H}(t) = -J \sum_{\langle i,j \rangle} \hat{\sigma}_i^z \hat{\sigma}_j^z - h(t) \sum_{i=1}^{N} \hat{\sigma}_i^x. \tag{15}$$

Here J is the Ising coupling, $h(t)$ is the transverse-field strength, and $\langle i, j\rangle$ ranges over nearest-neighbor pairs. The Pauli operator $\hat{\sigma}_i^\alpha$ acts on site i along axis $\alpha \in \{x, z\}$ and is distinct from the spin label σ_i. We set $J = 1$ and start from $|\psi_0\rangle = |+\rangle_x^{\otimes N}$, where $|+\rangle_x$ is the $+1$ eigenstate of the single-spin Pauli x operator and $\otimes N$ denotes the N-fold tensor product. Static quenches use $h/J = 0.5$ and $JT = 1$ unless specified otherwise; the supplementary driven tests use a linear ramp, a periodic field, and a Gaussian pulse (Appendix A.10).

Evaluation protocol. Exact diagonalization on small chains allows us to compare residual-based estimates with observed fidelity. The reference states are used only for evaluation, not for training or for computing either bound estimate. We train fifteen MADE models, with five seeds (0–4) for each of $N = 4, 5, 6$, for 3000 epochs at $h/J = 0.5$ and $JT = 1$. Full-basis summation gives the residual norm and the overlap with the reference states. From the same residual trace, we compute $B_{\mathrm{Abel}} = \mathcal{C}_{\mathrm{Abel}}/T$ and B_{clip}. Figure 3 compares their numerical estimates with observed average infidelity $E = 1 - \bar{F}$ at the final checkpoints and during training. Evaluation grids and checkpoint schedules are specified in Appendix A.4.

Bound estimates. At the final checkpoints, $\widehat{B}_{\mathrm{clip}}$ ranges from 0.0000976 to 0.000474 for $N = 4$, from 0.000201 to 0.000657 for $N = 5$, and from 0.00162 to 0.00275 for $N = 6$. Its corresponding fidelity lower-bound estimates span 0.99725–0.99990, while observed time-averaged fidelities span 0.998531–0.999980. The largest final cumulative residual $R(T)$ is 0.153, so clipping does not affect these fifteen estimates. Figure 3(a) shows the tighter cumulative and linear Abel estimates alongside the observed error; Table 1 gives the individual values.

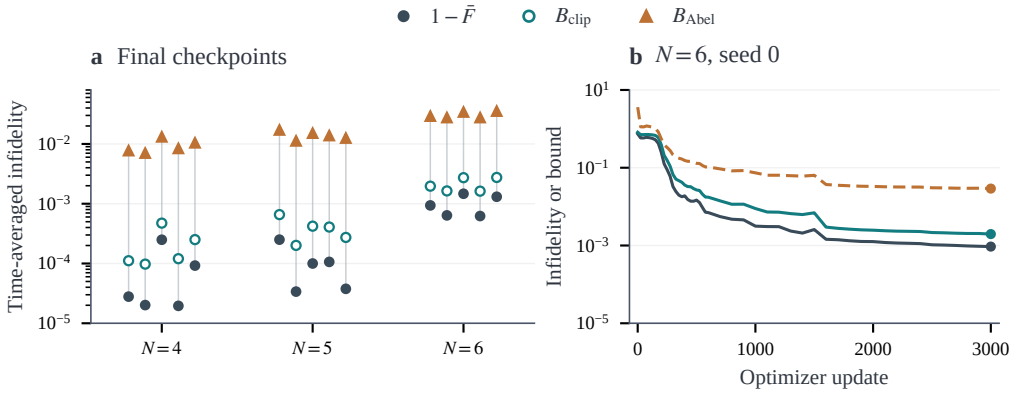

Figure 3: Residual-based estimates of time-averaged infidelity. (a) Observed $E = 1-\bar{F}$, the clipped cumulative estimate $\widehat{B}_{\text{clip}}$, and the Abel estimate $\widehat{B}_{\text{Abel}}$ at the final checkpoint of each of five seeds for $N = 4, 5, 6$; vertical lines connect quantities from the same model and are not error bars. (b) The same three quantities at all 49 evaluated checkpoints for $N = 6$, seed 0. Both panels use logarithmic vertical scales. Estimates use full-basis residual sums and finite-difference time derivatives, without numerical error certification.

For $N = 6$, seed 0, the final observed infidelity is 0.000937, compared with $\widehat{B}_{\text{clip}} = 0.00197$ and $\widehat{B}_{\text{Abel}} = 0.0291$ (Figure 3(b)). The estimated clipped bound exceeds the observed infidelity at all 735 archived checkpoint evaluations. Figure 6 shows these records and the seed-0 training histories at each size.

6 Observable Accuracy in Ising Dynamics

Static-quench setup. We assess static-quench observables against reference evolution for open chains with $N \in \{10, 20\}$ spins and an open 4×4 square lattice over $Jt \in [0, 1]$. The chain reference uses Jordan–Wigner/Bogoliubov–de Gennes evolution (Jordan & Wigner, 1928; Lieb et al., 1961); the lattice reference uses Krylov evolution. We evaluate the transverse magnetization

$$\hat{M}_x = \frac{1}{N}\sum_{i=1}^{N}\hat{\sigma}_i^x, \qquad M_x(t) = \langle \hat{M}_x \rangle_t, \tag{16}$$

and the pair-averaged spin correlation at graph distance r,

$$C_{zz}(r,t) = \frac{1}{|\mathcal{P}_r|}\sum_{(i,j)\in\mathcal{P}_r}\langle \hat{\sigma}_i^z\hat{\sigma}_j^z\rangle_t, \tag{17}$$

where $\mathcal{P}_r$ is the set of unordered site pairs at distance r: ordinary separation on a chain and Manhattan distance on the square lattice. Magnetization probes the mean single-spin response; distance-resolved correlations probe spatial structure. Here $\langle\hat{O}\rangle_t$ denotes an observable's expectation in the evaluated state at time t; correlations are raw, without subtracting products of one-site expectations, and $r = 1$ gives the bond-averaged nearest-neighbor correlation. Figure 4(d–f) shows $1 \le r \le 5$ for all three systems. Table 11 reports magnetization, nearest-neighbor, and distance-averaged correlation errors; Appendix A.14 specifies their evaluation grids.

Abel-tNQS reproduces the decay and late-time upturn of the reference magnetization (Figure 4(a–c)). The magnetization mean absolute errors (MAEs) are 0.00113 and 0.00770 for the $N = 10, 20$ chains and 0.0378 for the 4×4 lattice (Table 11); the displayed circles are subsets of the evaluation grids.

The correlation heatmaps reveal errors beyond the aggregate MAEs. For $N = 20$, late-time values are overestimated across separations, with an $r = 1$ endpoint error of 0.0610 (Figure 4(e)).

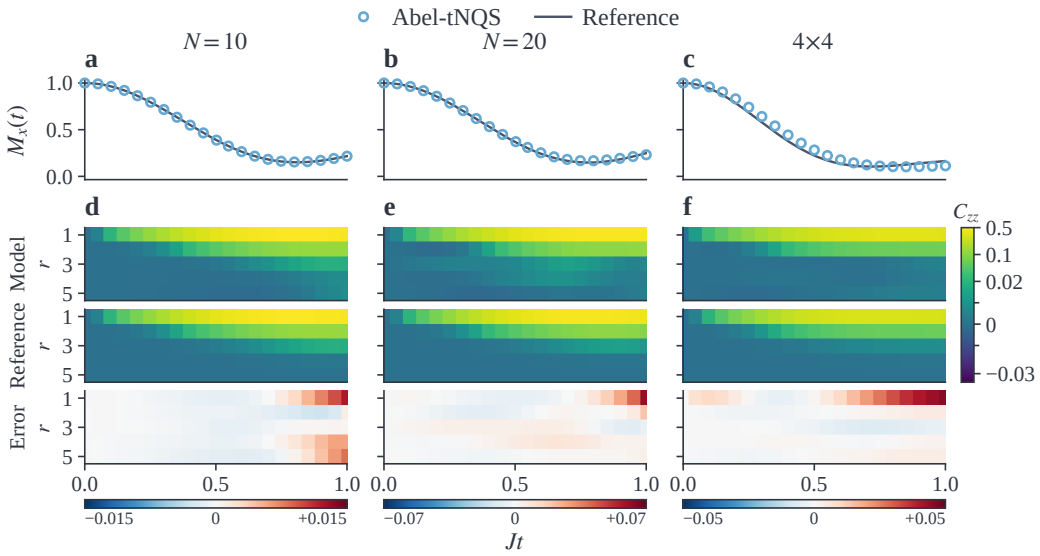

Figure 4: Static-quench dynamics at $h/J = 0.5$ from the product-x state for open chains with $N = 10, 20$ and an open 4×4 lattice. (a–c) Transverse magnetization $M_x(t)$: Abel-tNQS (blue circles) and reference (gray lines). (d–f) Distance-resolved correlations $C_{zz}(r, t)$: model, reference, and signed error (model minus reference), from top to bottom. Model and reference correlations share a symmetric-logarithmic color scale. Signed-error scales are linear and differ by column as labeled; error colors should not be compared across columns. Reference solvers and evaluation details are given in Appendix A.13.

On the 4×4 lattice, nearest-neighbor and distance-averaged correlation MAEs are 0.0151 and 0.00442, respectively. The largest deviations again occur at $r = 1$ and late times.

Extended dynamics. A supplementary $N = 10$ experiment extends the learned trajectory to $JT = 2$ through two consecutive windows. Its observables and a single-window comparison are reported in Appendix A.9. A driven-field test is reported in Appendix A.10.

7 Effects of Temporal Weighting and Initial-State Enforcement

We examine the remaining-time weight and initial-state enforcement on an $N = 10$ open chain at $h/J = 0.5$ and $JT = 1$. The five MADE settings are hard-linear, hard-uniform, and soft-linear with penalty coefficient $\lambda \in \{1, 10, 100\}$. Hard settings use the conditional-amplitude constraint; soft settings use a sampled initial-state penalty. We train ten seeds per setting for 3000 updates under matched architectures and base optimization settings (50 runs). All use full residual gradients; soft settings add penalty evaluations to each update. We evaluate 16 saved checkpoints per run by full-basis summation at the same 101 times. Figure 5 compares final metrics and checkpoint trends; Appendix A.11 gives the protocols and additional fidelity results.

Temporal weighting. With the hard constraint fixed, linear weighting at 3000 epochs lowers mean M_x MAE from 0.05401 to 0.00821 (85%) and C_{zz} MAE from 0.19561 to 0.01159 (94%) relative to uniform weighting (Figure 5(a)). The weights $T - t$ and $T/2$ have the same time average, so only their temporal distribution differs. Grid-mean fidelity rises from 0.4745 to 0.9854, and endpoint fidelity from 0.1762 to 0.9082 (Table 6).

Initial-state enforcement. At 3000 updates, hard-linear has the lowest mean observable errors among the tested settings. Relative to soft $\lambda = 100$, its M_x and C_{zz} MAEs are lower by approximately 87% and 95%. Hard-linear also has higher mean endpoint fidelity (0.9082 versus 0.3216) and grid-mean fidelity (0.9854 versus 0.5877; Table 6), while satisfying the initial condition by construction.

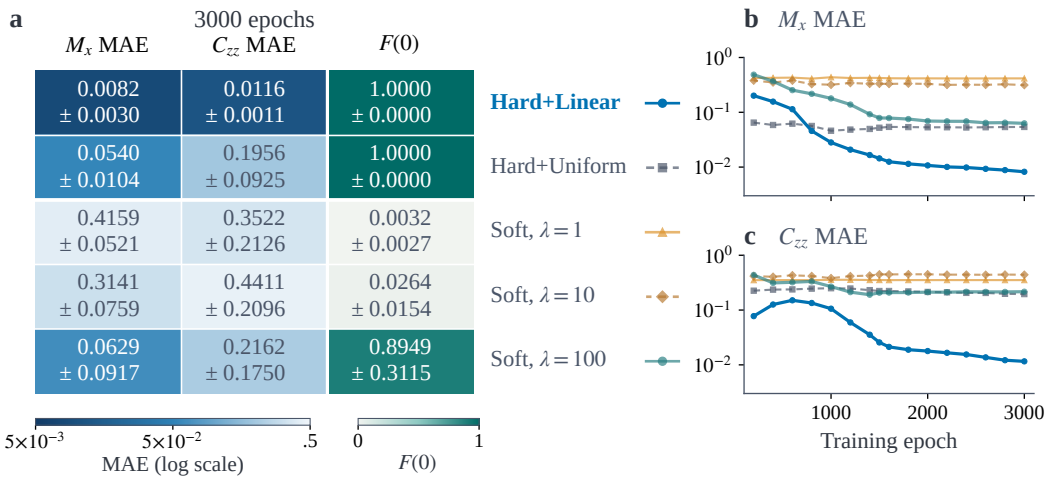

Figure 5: Controlled $N = 10$ MADE ablations with ten seeds per setting and full residual gradients. (a) Results at 3000 epochs: means $\pm$ sample standard deviations. Darker blue denotes lower MAE on a shared logarithmic scale; darker teal denotes higher initial fidelity on a linear scale. Initial fidelities displayed as 1.0000 are rounded numerical values. (b,c) Mean magnetization and nearest-neighbor correlation MAEs versus training epoch. Lines connect all 16 saved checkpoints from 200 to 3000 epochs. The central method labels and symbols identify rows and curves. Observable MAEs use full-basis evaluation on a 101-point time grid; initial fidelity is evaluated at $t = 0$. All soft settings use linear weighting.

For soft $\lambda = 100$, nine seeds reach $F(0)$ between 0.9787 and 0.9977, whereas one seed remains at 0.0086, giving 0.8949 ± 0.3115 overall. The soft baseline can therefore learn the initial state, but does so less consistently across these seeds.

At all 16 saved checkpoints, hard-linear has lower mean observable errors than every tested soft-penalty setting (Figure 5(b,c)). For each of the ten seeds, the final hard-linear model has lower M_x and C_{zz} errors than the corresponding models trained with the tested soft penalties. Compared with hard-uniform, hard-linear has lower mean M_x error at 13 of the 16 checkpoints and lower mean C_{zz} error at all 16.

8 Conclusion and Limitations

Abel-tNQS combines an exact initial-state constraint on complex conditional amplitudes with a remaining-time-weighted residual surrogate. The autoregressive representation also retains direct sampling throughout the time window. The resulting normalized trajectories admit both a linear Abel fidelity bound, which identifies the training weight, and a sharper cumulative bound for post-training assessment. The latter gives numerical time-averaged fidelity lower-bound estimates of 0.99725–0.99990 across the fifteen final small-system models. In controlled full-gradient ablations, linear weighting reduces both observable MAEs and increases grid-mean and endpoint fidelities relative to uniform weighting. Across all 16 evaluated checkpoints, hard-linear has lower mean observable errors than the three tested soft-penalty settings. The main observable benchmarks cover $N = 10, 20$ chains and an open 4×4 lattice.

Turning the numerical lower-bound estimates into certified guarantees would require error control for normalization, residual evaluation, and time integration. The static observable benchmarks in Figure 4 and the relay and driven tests each report one training realization per setting. The construction requires tractable initial-state conditionals and non-vanishing local normalization vectors. Longer windows remain difficult to optimize, and relay continuation can accumulate endpoint error. Appendices A.3, A.5, A.8, and A.9 give further details.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, to draft and revise portions of the manuscript, and to support research ideation and execution. This support included feedback on methodology and experimental design, assistance with code and result analysis, and consistency checks across the text, figures, and reported values. The authors reviewed and verified all AI-assisted text, code, analyses, claims, citations, figures, and numerical results. We take full responsibility for the final content of this work, including all material produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

An anonymized code package accompanies this submission. It contains training and evaluation code, configuration snapshots, selected model weights, and evaluated data behind the quantitative figures. Optimizer-state checkpoints for the long continuation runs are not included; their schedules and optimization conventions are documented in Section 4.2 and Appendix A.12. Appendix A.15 describes the reference solvers used across the experiments: exact diagonalization, Jordan–Wigner/Bogoliubov–de Gennes evolution, Krylov propagation, and Runge–Kutta integration.

REFERENCES

Massimo Bortone, Yannic Rath, and George H. Booth. Impact of conditional modelling for a universal autoregressive quantum state. *Quantum*, 8:1245, 2024. doi: 10.22331/q-2024-02-08-1245.

Giuseppe Carleo and Matthias Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355(6325):602–606, 2017.

Giuseppe Carleo, Federico Becca, Marco Schiró, and Michele Fabrizio. Localization and glassy dynamics of many-body quantum systems. *Scientific Reports*, 2:243, 2012. doi: 10.1038/srep00243.

Stefanie Czischek, Martin Gärttner, and Thomas Gasenzer. Quenches near Ising quantum criticality as a challenge for artificial neural networks. *Physical Review B*, 98:024311, 2018. doi: 10.1103/PhysRevB.98.024311.

Kaelan Donatella, Zakari Denis, Alexandre Le Boité, and Cristiano Ciuti. Dynamics with autoregressive neural quantum states: Application to critical quench dynamics. *Physical Review A*, 108:022210, 2023. doi: 10.1103/PhysRevA.108.022210.

Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. MADE: Masked autoencoder for distribution estimation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 881–889, 2015. URL https://proceedings.mlr.press/v37/germain15.html.

Luca Gravina, Vincenzo Savona, and Filippo Vicentini. Neural projected quantum dynamics: a systematic study. *Quantum*, 9:1803, 2025. doi: 10.22331/q-2025-07-22-1803.

Irene López Gutiérrez and Christian B. Mendl. Real time evolution with neural-network quantum states. *Quantum*, 6:627, 2022. doi: 10.22331/q-2022-01-20-627.

Lucas Hackl, Tommaso Guaita, Tao Shi, Jutho Haegeman, Eugene Demler, and J. Ignacio Cirac. Geometry of variational methods: dynamics of closed quantum systems. *SciPost Physics*, 9:048, 2020. doi: 10.21468/SciPostPhys.9.4.048.

Jutho Haegeman, J. Ignacio Cirac, Tobias J. Osborne, Iztok Pižorn, Henri Verschelde, and Frank Verstraete. Time-dependent variational principle for quantum lattices. *Physical Review Letters*, 107:070601, 2011. doi: 10.1103/PhysRevLett.107.070601.

Tim Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37:185–194, 1995. doi: 10.1080/00401706.1995.10484303.

Mohamed Hibat-Allah, Martin Ganahl, Lauren E. Hayward, Roger G. Melko, and Juan Carrasquilla. Recurrent neural network wave functions. *Physical Review Research*, 2:023358, 2020. doi: 10.1103/PhysRevResearch.2.023358.

Damian Hofmann, Giammarco Fabiani, Johan H. Mentink, Giuseppe Carleo, and Michael A. Sentef. Role of stochastic noise and generalization error in the time propagation of neural-network quantum states. *SciPost Physics*, 12:165, 2022. doi: 10.21468/SciPostPhys.12.5.165.

Pascual Jordan and Eugene Wigner. über das paulische äquivalenzverbot. *Zeitschrift für Physik*, 47: 631–651, 1928. doi: 10.1007/BF01331938.

Wladislaw Krinitsin and Markus Schmitt. Time-dependent variational Monte Carlo without bias. *arXiv preprint arXiv:2605.03930*, 2026.

Hannah Lange, Anka Van de Walle, Atiye Abedinnia, and Annabelle Bohrdt. From architectures to applications: a review of neural quantum states. *Quantum Science and Technology*, 9:040501, 2024. doi: 10.1088/2058-9565/ad7168.

Hugo Larochelle and Iain Murray. The neural autoregressive distribution estimator. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 29–37, 2011. URL https://proceedings.mlr.press/v15/larochelle11a.html.

Elliott Lieb, Theodore Schultz, and Daniel Mattis. Two soluble models of an antiferromagnetic chain. *Annals of Physics*, 16:407–466, 1961. doi: 10.1016/0003-4916(61)90115-4.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://mlanthology.org/iclr/2019/loshchilov2019iclr-decoupled/.

A. D. McLachlan. A variational solution of the time-dependent Schrödinger equation. *Molecular Physics*, 8:39–44, 1964. doi: 10.1080/00268976400100041.

Sebastian Paeckel, Thomas Köhler, Andreas Swoboda, Salvatore R. Manmana, Ulrich Schollwöck, and Claudius Hubig. Time-evolution methods for matrix-product states. *Annals of Physics*, 411: 167998, 2019.

Pierre Pfeuty. The one-dimensional Ising model with a transverse field. *Annals of Physics*, 57: 79–90, 1970. doi: 10.1016/0003-4916(70)90270-8.

Zihao Qi, Christopher Earls, and Yang Peng. Neural operator quantum state: A foundation model for quantum dynamics. *arXiv preprint arXiv:2603.25066*, 2026.

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045.

Riccardo Rende, Luciano Loris Viteritti, Federico Becca, Antonello Scardicchio, Alessandro Laio, and Giuseppe Carleo. Foundation neural-networks quantum states as a unified ansatz for multiple hamiltonians. *Nature Communications*, 16:7213, 2025. doi: 10.1038/s41467-025-62098-x.

Markus Schmitt and Markus Heyl. Quantum many-body dynamics in two dimensions with artificial neural networks. *Physical Review Letters*, 125(10):100503, 2020.

Ulrich Schollwöck. The density-matrix renormalization group in the age of matrix product states. *Annals of Physics*, 326(1):96–192, 2011.

Or Sharir, Yoav Levine, Noam Wies, Giuseppe Carleo, and Amnon Shashua. Deep autoregressive models for the efficient variational simulation of many-body quantum systems. *Physical Review Letters*, 124:020503, 2020.

Alessandro Sinibaldi, Clemens Giuliani, Giuseppe Carleo, and Filippo Vicentini. Unbiasing time-dependent variational Monte Carlo by projected quantum evolution. *Quantum*, 7:1131, 2023. doi: 10.22331/q-2023-10-10-1131.

Alessandro Sinibaldi, Douglas Hendry, Filippo Vicentini, and Giuseppe Carleo. Time-dependent neural galerkin method for quantum dynamics. *Physical Review Letters*, 136:120402, 2026. doi: 10.1103/kqvx-dl54.

Anka Van de Walle, Markus Schmitt, and Annabelle Bohrdt. Many-body dynamics with explicitly time-dependent neural quantum states. *Machine Learning: Science and Technology*, 6:045011, 2025. doi: 10.1088/2632-2153/ae0f39.

Vladimir Vargas-Calderón, Herbert Vinck-Posada, and Fabio A. González. An empirical study of quantum dynamics as a ground state problem with neural quantum states. *Quantum Information Processing*, 22:165, 2023. doi: 10.1007/s11128-023-03902-9.

Guifré Vidal. Efficient simulation of one-dimensional quantum many-body systems. *Physical Review Letters*, 93:040502, 2004. doi: 10.1103/PhysRevLett.93.040502.

Luciano Loris Viteritti, Riccardo Rende, and Federico Becca. Transformer variational wave functions for frustrated quantum spin systems. *Physical Review Letters*, 130:236401, 2023. doi: 10.1103/PhysRevLett.130.236401.

A Appendix

A.1 Fidelity Bounds from the Hilbert Residual

Under the assumptions in Section 3.2, absolute continuity gives an integrable state derivative; normalization and the integrable operator norm give $\|\hat{H}(t)|\Psi_t\rangle\| \le \|\hat{H}(t)\|_{\mathrm{op}}$. Hence the residual is integrable, and the overlap and fidelity are absolutely continuous. All derivative identities below hold almost everywhere. Let $c(t) = \langle\psi_t|\Psi_t\rangle$ and $F(t) = |c(t)|^2$. Using $\partial_t|\psi_t\rangle = -i\hat{H}(t)|\psi_t\rangle$ and $\hat{H}(t)^\dagger = \hat{H}(t)$, where $\dagger$ denotes the Hermitian adjoint, gives

$$\dot{c}(t) = \langle\partial_t\psi_t|\Psi_t\rangle + \langle\psi_t|\partial_t\Psi_t\rangle = \langle\psi_t|d_\Psi(t)\rangle. \tag{18}$$

Writing Re for the real part, normalization of $|\Psi_t\rangle$ and Hermiticity imply

$$\mathrm{Re}\langle\Psi_t|d_\Psi(t)\rangle = \frac{1}{2}\partial_t\big\||\Psi_t\rangle\big\|^2 + \mathrm{Re}\left(i\langle\Psi_t|\hat{H}(t)|\Psi_t\rangle\right) = 0. \tag{19}$$

This sets the real part to zero, not necessarily the complex inner product itself. Writing $|\eta_t\rangle$ for the component orthogonal to $|\Psi_t\rangle$ and $c(t)^*$ for the complex conjugate of $c(t)$, decompose the exact state as

$$|\psi_t\rangle = c(t)^*|\Psi_t\rangle + |\eta_t\rangle, \qquad \langle\Psi_t|\eta_t\rangle = 0, \qquad \big\||\eta_t\rangle\big\|^2 = 1 - F(t). \tag{20}$$

Substituting into the fidelity derivative and using Eq. 19 yields

$$\begin{aligned}\dot{F}(t) &= 2\,\mathrm{Re}\left(c(t)^*\langle\psi_t|d_\Psi(t)\rangle\right)\\ &= 2F(t)\,\mathrm{Re}\langle\Psi_t|d_\Psi(t)\rangle + 2\,\mathrm{Re}\left(c(t)^*\langle\eta_t|d_\Psi(t)\rangle\right)\\ &= 2\,\mathrm{Re}\left(c(t)^*\langle\eta_t|d_\Psi(t)\rangle\right).\end{aligned} \tag{21}$$

Cauchy–Schwarz therefore gives

$$\begin{aligned}|\dot{F}(t)| &\le 2|c(t)|\,\big\||\eta_t\rangle\big\|\,\big\||d_\Psi(t)\rangle\big\|\\ &= 2\sqrt{F(t)(1-F(t))}\,\big\||d_\Psi(t)\rangle\big\| \le \big\||d_\Psi(t)\rangle\big\|,\end{aligned} \tag{22}$$

where $2\sqrt{F(t)(1-F(t))} \le 1$ follows from $(2F(t)-1)^2 \ge 0$. No division by $c(t)$ or $\sqrt{1-F(t)}$ is used, so the argument includes $F = 0$ and $F = 1$. In particular, no nonzero-overlap assumption is needed. Abel integration by parts applied directly to $1 - F(t)$ now gives

$$\begin{aligned}1 - \bar{F} &= 1 - F(0) - \frac{1}{T}\int_0^T (T-t)\dot{F}(t)\,dt\\ &\le 1 - F(0) + \frac{1}{T}\int_0^T (T-t)\big\||d_\Psi(t)\rangle\big\|\,dt.\end{aligned} \tag{23}$$

This proves Eq. 4. Setting $F(0) = 1$ and using $\bar{F} \ge 0$ proves Eq. 6. Equality of the initial states up to a constant phase already suffices for $F(0) = 1$; the architectural constraint imposes the stronger equality of their amplitudes. The argument applies to time-dependent Hamiltonians under the stated

assumptions; no commutation between Hamiltonians at different times is required. This is a bound on the continuous-time average, not on endpoint fidelity or every instantaneous fidelity.

A related state-distance bound follows for the same normalized trajectories. Define the integrated squared state error and its initial value by

$$\mathcal{L} = \int_0^T \big\||\psi_t\rangle - |\Psi_t\rangle\big\|^2 dt = 2T - 2\ \mathrm{Re}\int_0^T c(t)dt, \qquad I_{\mathrm{state}}(0) = \big\||\psi_0\rangle - |\Psi_0\rangle\big\|^2. \tag{24}$$

Integration by parts yields

$$\int_0^T c(t)dt = Tc(0) + \int_0^T (T-t)\dot{c}(t)dt, \tag{25}$$

and therefore

$$\mathcal{L} = T\big\||\psi_0\rangle - |\Psi_0\rangle\big\|^2 - 2\ \mathrm{Re}\int_0^T (T-t)\langle\psi_t|d_\Psi(t)\rangle dt. \tag{26}$$

Because $\big\||\psi_t\rangle\big\| = 1$, Cauchy–Schwarz implies the separate inequality

$$\mathcal{L} \le TI_{\mathrm{state}}(0) + 2\mathcal{C}_{\mathrm{Abel}}. \tag{27}$$

The state distance and infidelity obey

$$\big\||\psi_t\rangle - |\Psi_t\rangle\big\|^2 - (1 - F(t)) = |1 - c(t)|^2 \ge 0. \tag{28}$$

For an exact initial state, Eqs. 27 and 28 give the linear state-distance bound $1 - \bar{F} \le 2\mathcal{C}_{\mathrm{Abel}}/T$; the direct fidelity argument above improves its coefficient. A cumulative state-error argument gives a different comparison. Let $U(t,s)$ be the unitary propagator of $\hat{H}(t)$, set $|\Delta_t\rangle = |\Psi_t\rangle - |\psi_t\rangle$, and write $r(t) = \big\||d_\Psi(t)\rangle\big\|$. When $|\Delta_0\rangle = 0$, Duhamel's formula and unitarity give

$$|\Delta_t\rangle = \int_0^t U(t,s)|d_\Psi(s)\rangle\, ds, \qquad \big\||\Delta_t\rangle\big\| \le \int_0^t r(s)\, ds. \tag{29}$$

Using Eq. 28, $1 - F(t) \le 1$, and averaging over time yields

$$1 - \bar{F} \le B_{\mathrm{clip}} \le B_{\mathrm{int}} := \frac{1}{T}\int_0^T R(t)^2 dt. \tag{30}$$

For every $x \ge 0$, $\min\{1, x^2\} \le x$; exchanging the order of integration in $T^{-1}\int_0^T R(t)dt$ proves $B_{\mathrm{clip}} \le \mathcal{C}_{\mathrm{Abel}}/T$, as stated in Eq. 7. If $r \in L^2([0,T])$, Cauchy–Schwarz inside the square and exchange of integration order further give

$$B_{\mathrm{int}} \le B_{\mathrm{sq}} := \frac{1}{2T}\int_0^T (T^2 - s^2) r(s)^2\, ds. \tag{31}$$

The B_{sq} integrand is also time-separable. For constant $r(s) = r_0$, $B_{\mathrm{Abel}} = r_0 T/2$ and $B_{\mathrm{int}} = B_{\mathrm{sq}} = r_0^2 T^2/3$; clipping further restricts the pointwise infidelity estimate when $R(t) > 1$. The Abel form remains linear in $r(t)$ and displays the $T - t$ weight used in our training surrogate. Related accumulated-residual state-error bounds appear in Sinibaldi et al. (2026). The Hilbert residual remains sensitive to time-dependent global phase, as discussed in Appendix A.5.

A.2 Observable Error Bound

Under the assumptions of Section 3.2, additionally assume $F(0) = 1$. For a bounded Hermitian observable $\hat{O}$, define $\Delta O(t) = \langle\Psi_t|\hat{O}|\Psi_t\rangle - \langle\psi_t|\hat{O}|\psi_t\rangle$ and let $\|\hat{O}\|_{\mathrm{op}}$ denote its operator norm and

$\|\cdot\|_1$ the trace norm. For normalized pure states, the trace-norm identity $\||\Psi_t\rangle\langle\Psi_t| - |\psi_t\rangle\langle\psi_t|\|_1 = 2\sqrt{1-F(t)}$ gives $|\Delta O(t)| \le 2\|\hat{O}\|_{\mathrm{op}}\sqrt{1-F(t)}$. Cauchy–Schwarz in time and Eq. 7 then imply

$$\begin{aligned}\frac{1}{T}\int_0^T |\Delta O(t)|\,dt &\le 2\|\hat{O}\|_{\mathrm{op}}\sqrt{1-\bar{F}} \\ &\le 2\|\hat{O}\|_{\mathrm{op}}\sqrt{B_{\mathrm{clip}}} \le 2\|\hat{O}\|_{\mathrm{op}}\sqrt{B_{\mathrm{Abel}}}.\end{aligned} \tag{32}$$

The observable bound can be loose, and its converse does not hold: agreement for a few observables does not establish high state fidelity. Equation 32 bounds the continuous-time average of the absolute difference between quantum expectation values; MAEs estimated from samples on a discrete time grid also include evaluation error.

A.3 Monte Carlo Estimation and Numerical Certification

At each time, the residual-state norm satisfies $\big\||d_\Psi(t)\rangle\big\| = (\sum_\sigma |d_\Psi(\sigma,t)|^2)^{1/2}$. Enumerating the full basis evaluates this sum without configuration-sampling error. Time quadrature estimates the Abel integral and, after cumulative integration of the residual norm, B_{clip}; finite-difference and quadrature errors remain. Appendix A.4 specifies the full-basis protocol used in the small-chain experiments.

Let $D = 2^N$ be the Hilbert-space dimension, n_t the number of evaluation times, and n_s the number of configuration samples per time. For independent uniform times $t_k \in [0,T]$, $k = 1,\dots,n_t$, let σ_{kj}, $j = 1,\dots,n_s$, be uniformly sampled basis configurations at time t_k. A plug-in estimator of the certificate is

$$\widehat{\mathcal{C}}_{\mathrm{Abel}} = \frac{T}{n_t}\sum_{k=1}^{n_t}(T-t_k)\sqrt{\frac{D}{n_s}\sum_{j=1}^{n_s}|d_\Psi(\sigma_{kj},t_k)|^2}. \tag{33}$$

More generally, let $q(\sigma \mid t)$ be a proposal probability distribution with full support, meaning $q(\sigma \mid t) > 0$ for every configuration, and draw $\sigma_{kj} \sim q(\cdot \mid t_k)$. The residual-norm estimator is

$$\big\|\widehat{|d_\Psi(t_k)\rangle}\big\| = \sqrt{\frac{1}{n_s}\sum_{j=1}^{n_s}\frac{|d_\Psi(\sigma_{kj},t_k)|^2}{q(\sigma_{kj}\mid t_k)}}. \tag{34}$$

For a normalized model, a directly sampled mixture assigns weight $\beta \in [0,1)$ to the Born distribution and weight $1-\beta$ to the uniform distribution:

$$q_\beta(\sigma \mid t) = \beta|\Psi_\theta(\sigma,t)|^2 + \frac{1-\beta}{D}, \qquad 0 \le \beta < 1. \tag{35}$$

Its uniform component guarantees full support, while $\beta = 0$ recovers uniform sampling. This defensive-mixture construction follows the general importance-sampling principle of Hesterberg (1995). Pure Born sampling ($\beta = 1$) need not cover configurations on which Ψ_θ vanishes but d_Ψ does not, a concern also addressed by recent unbiased tVMC estimators (Sinibaldi et al., 2023; Krinitsin & Schmitt, 2026). Proposal probabilities are treated as stop-gradient quantities when optimizing the physical residual.

Because the residual is linear in the state and contains no division by the wave function, uniform sampling remains well defined where $|\Psi|$ is small. Its variance can nevertheless be large in high-dimensional spaces. The Born–uniform mixture concentrates more samples in occupied regions while retaining full support, and neither proposal uniformly dominates the other. For an unbiased estimate of the squared residual norm, Jensen's inequality implies a downward bias after taking the square root. Cumulative integration and clipping do not remove this issue: point estimates of either fidelity lower bound can be too high.

A numerical guarantee from the cumulative bound requires a controlled upper estimate of the residual norm over time, with a stated confidence level if probabilistic, together with control of time integration, differentiation, and normalization errors. An upper estimate $\mathcal{C}_{\mathrm{upper}} \ge \mathcal{C}_{\mathrm{Abel}}$ analogously gives $\bar{F} \ge \max\{0, 1 - \mathcal{C}_{\mathrm{upper}}/T\}$ under the exact-initial-state assumption. Without such error control we report lower-bound estimates, not numerically certified guarantees.

Table 1: Residual-based infidelity estimates at update 3000, with all five seeds for each size. The corresponding numerical fidelity lower-bound estimate is $1 - \widehat{B}_{\mathrm{clip}}$ in every row. Both bounds use the same full-basis residual evaluation; $\bar{F}$ uses the exact-diagonalization reference.

N	Seed	$\widehat{B}_{\mathrm{Abel}}$	$\widehat{B}_{\mathrm{clip}}$	Observed $\bar{F}$
4	0	0.007700	0.000111	0.999972
4	1	0.007002	0.000098	0.999980
4	2	0.013085	0.000474	0.999750
4	3	0.008372	0.000120	0.999980
4	4	0.010404	0.000251	0.999908
5	0	0.017058	0.000657	0.999749
5	1	0.011202	0.000201	0.999966
5	2	0.015087	0.000421	0.999900
5	3	0.013750	0.000407	0.999894
5	4	0.012494	0.000273	0.999962
6	0	0.029091	0.001967	0.999063
6	1	0.027514	0.001638	0.999362
6	2	0.034076	0.002728	0.998531
6	3	0.027510	0.001623	0.999378
6	4	0.035292	0.002751	0.998692

A.4 Small-System Fidelity-Bound Evaluation

The small-system study uses open TFIM chains with $N = 4, 5, 6$, $J = 1$, $h = 0.5$, $T = 1$, and $|\psi_0\rangle = |+\rangle_x^{\otimes N}$. Training uses full residual gradients and a scaled mean-absolute residual surrogate with full configuration enumeration and random training times. Evaluation occurs every 25 epochs through epoch 600 and every 100 epochs thereafter through epoch 3000, including epoch 0. Each of the fifteen training trajectories contributes 49 evaluation records.

At each checkpoint, full-basis summation evaluates the Hilbert-residual norm separately from the training surrogate. The evaluation script computes $\sqrt{\sum_\sigma |d_\theta^{\mathrm{FD}}(\sigma, t)|^2 + 10^{-12}}$ without dividing by $\sqrt{2^N}$. Composite trapezoidal integration on 201 uniformly spaced times gives $\widehat{B}_{\mathrm{Abel}}$ and the cumulative residual $\widehat{R}(t)$; integrating $\min\{1, \widehat{R}(t)^2\}$ on the same grid gives $\widehat{B}_{\mathrm{clip}}$. The time derivative uses $\varepsilon_t = 10^{-3}$ and the boundary stencil in Eq. 37. The overlap with the exact-diagonalization reference is evaluated on 200 uniformly spaced times and trapezoidally integrated to obtain $\bar{F}$. Full-basis summation removes configuration-sampling error, not time-discretization or floating-point error.

Figure 3(a) includes every final model, while panel (b) connects all 49 evaluated checkpoints for $N = 6$, seed 0, chosen by index rather than performance. Figure 6 shows all 735 evaluation records and the seed-0 training curves at each size. The records within each training trajectory are correlated, so Table 1 summarizes final checkpoints separately.

The reported fidelities use raw full-basis overlaps, as in the archived evaluations, and assume normalized model and reference states. The evaluation archives also retain time-resolved state norms, normalized overlaps, and initial-state errors. These diagnostics do not test convergence in the differentiation step or quadrature grid. Five seeds per size provide a limited sample of training variability.

A.5 Properties and Edge Cases of the Conditional Constraint

The power gate remains absolutely continuous when $T/\tau < 1$, but its one-sided derivative diverges at $t = T$. Numerical endpoint differentiation then requires separate treatment; choosing $T/\tau \geq 1$ avoids this particular singularity.

Suppose the learned raw decoder vectors are absolutely continuous in time and never zero for each site and prefix. Their norms then have positive minima on the compact interval $[0, T]$, so normalization preserves absolute continuity. The same argument applies to the pair of raw amplitudes $\psi_i^{\mathrm{raw}}(v \mid \sigma_{<i}, t)$ under Eq. 12: products with the absolutely continuous gate, followed by normalization, give absolutely continuous $\psi_i(v \mid \sigma_{<i}, t)$. The finite autoregressive product is therefore absolutely continuous, as required in Section 3.2. At each fixed time, summing $|\Psi_\theta(\sigma, t)|^2$ first over

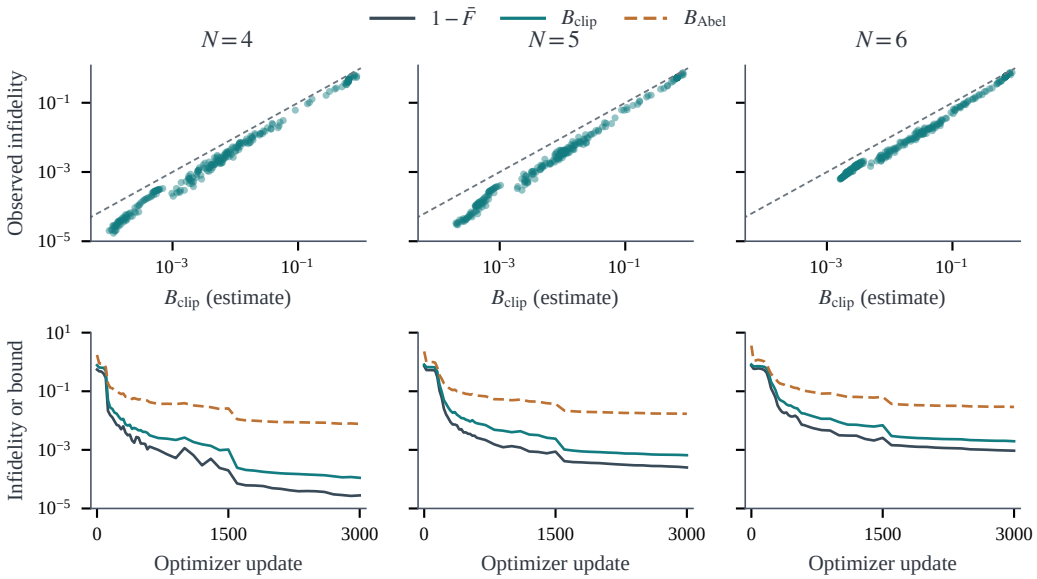

Figure 6: Residual-based infidelity estimates across all checkpoints, grouped by $N = 4, 5, 6$ from left to right. Top: observed $1 - \bar{F}$ versus $\widehat{B}_{\text{clip}}$ for all 245 records per size, pooling five seeds; the dashed line marks equality. Bottom: observed infidelity, $\widehat{B}_{\text{clip}}$, and $\widehat{B}_{\text{Abel}}$ for seed 0 at each size. Lines connect evaluated checkpoints without smoothing or seed averaging. Axes use logarithmic scales for infidelity and bound values.

σ_N, then σ_{N-1}, and so on yields one, since every conditional vector has unit norm. At $t = 0$, $\psi_i(v \mid \sigma_{<i}, 0) = \phi_i(v \mid \sigma_{<i})$ reconstructs ψ_0 exactly. At $t = T$, $\psi_i(v \mid \sigma_{<i}, T) = f_i(v \mid \sigma_{<i}, T)$, so the initial anchor drops out.

At a fixed site and prefix, the gate also sets the initial slope of the normalized conditional. For the unit vectors ϕ_i and $f_i(t)$ in Eq. 10, continuity of f_i at zero gives

$$\begin{aligned} \psi_i'(0) &= \frac{1}{\tau}\left[f_i(0) - \phi_i \operatorname{Re}\langle \phi_i, f_i(0)\rangle\right], \\ \|\psi_i'(0)\| &\le \frac{1}{\tau}. \end{aligned} \tag{36}$$

Thus $f_i'(0)$ does not enter at first order, and choosing $\tau = T$ changes this local slope scale with the window length in the relay comparison.

The non-vanishing condition in Section 4.1 is essential. For example, $f_i(v \mid \sigma_{<i}, t) = -\phi_i(v \mid \sigma_{<i})$ for both local values at $g(t) = 1/2$ makes both raw amplitudes zero, for which normalization is undefined; the map is also ill-conditioned near such cancellation. At both normalization stages, a raw vector $\mathbf{u} \in \mathbb{C}^2$ is mapped to $\mathbf{u}/\max(\|\mathbf{u}\|_2, \epsilon_{\text{norm}})$. In exact arithmetic its output has unit norm when $\|\mathbf{u}\|_2 \ge \epsilon_{\text{norm}}$, but is subnormalized below that threshold. Floating-point evaluation introduces additional rounding error. The analytic certificate assumes a normalized trajectory. When the clamp is active, the output can be subnormalized, so applying the bound requires further analysis. Checking a finite set of sampled prefixes alone does not establish the condition for every configuration and time.

The initial constraint also fixes the global phase at $t = 0$. If $|\Psi_t\rangle = e^{i\varphi(t)}|\psi_t\rangle$ for a differentiable real phase $\varphi(t)$, then $|d_\Psi(t)\rangle = i\dot{\varphi}(t)|\Psi_t\rangle$ and $\big\||d_\Psi(t)\rangle\big\| = |\dot{\varphi}(t)|$. The residual penalizes time-dependent phase drift. At an exact zero-residual solution, $\varphi(0) = 0$ implies a constant zero phase. The phase can still drift when the residual is nonzero.

For the product states used at the beginning of the static-quench and driven experiments, $\phi_i(v \mid \sigma_{<i})$ is analytic. In later relay segments, its conditionals are evaluated from the preceding Abel-tNQS

endpoint. A generic entangled initial state can be used only when its normalized autoregressive conditionals can be evaluated tractably, for example from a compatible NQS or, at small size, by enumeration.

A.6 Detailed Comparison with Global-in-Time NQS

Table 2: Comparison with the two closest global-in-time NQS baselines. The comparison concerns residual forms, temporal treatment, and initial-state constructions; state-error bounds can also imply fidelity bounds. In the initial-condition row, c_i are the basis-expansion coefficients and δ_{i0} is the Kronecker delta.

Property	Van de Walle et al.	Sinibaldi et al.	Abel-tNQS (ours)
Trajectory	Time-conditioned transformer NQS	Time-dependent combination of static NQS	Time-conditioned autoregressive NQS
Residual	Squared Hilbert residual with Born-sampled log-local estimator	Projective, gauge-invariant residual	Hilbert-residual norm, evaluated without amplitude division, for the analytic bound; Eq. 13 for main training runs
Error control	Accumulated-residual diagnostic	State and observable error bounds	Linear Abel and cumulative time-averaged fidelity bounds
Initial condition	Architectural initialization layer	Fixed basis state and $c_i(0) = \delta_{i0}$	Conditional-amplitude gate
Sampling	Direct autoregressive Born sampling	Born-distribution MCMC	Uniform or full-basis training; full-basis observable and certificate evaluation
Time treatment	Uniform discrete time grid	Numerical time quadrature	Random training times; $(T - t)$-weighted residual

A.7 Architecture and Implementation Details

The joint encoder shifts the one-hot spin inputs right in the `state_emb` module, inserts a fixed beginning-of-sequence (BOS) token, and applies a shared per-site multilayer perceptron (MLP). This produces the spatial-feature matrix $\mathbf{H}_\sigma \in \mathbb{R}^{N \times d_{\text{hid}}}$, where d_{hid} is the spatial-feature dimension. The `time_encoder` maps scalar t to the time-feature vector $\mathbf{h}_t \in \mathbb{R}^{d_t}$, where d_t is the time-feature dimension. Broadcasting over all $n_t n_s$ pairs of times and configurations gives the third-order spatial-feature tensor $\mathcal{X} \in \mathbb{R}^{(n_t n_s) \times N \times d_{\text{hid}}}$ and the aligned conditioning matrix $\mathbf{C} \in \mathbb{R}^{(n_t n_s) \times d_t}$. These are the features $\mathcal{X}$ and $\mathbf{C}$ in Figure 2.

The causal backbone uses MADE, allowing each site output to depend on the full permitted prefix. An additive projection of $\mathbf{C}$ supplies time conditioning. The `context_mixer` output is the third-order tensor $\mathcal{X}_{\text{out}} \in \mathbb{R}^{(n_t n_s) \times N \times d_{\text{hid}}}$.

An MLP then decodes the causal features into real–imaginary pairs for each local value v and normalizes the two complex amplitudes jointly over v, yielding $f_i(v \mid \sigma_{<i}, t)$. The constraint in Section 4.1 interpolates these with $\phi_i(v \mid \sigma_{<i})$ and again normalizes over both local values to obtain $\psi_i(v \mid \sigma_{<i}, t)$. For a queried configuration, we select the conditional amplitude corresponding to $v = \sigma_i$ at each site and multiply these amplitudes to obtain $\Psi_\theta(\sigma, t) = \prod_i \psi_i(\sigma_i \mid \sigma_{<i}, t)$. These conditionals also support direct ancestral sampling.

A.8 Training Details

The static-quench runs in Figure 4 use the mean-absolute residual surrogate, linear time weighting, and the hard initial-state constraint, with the initial-condition loss disabled. All connected configurations are retained, with neither neighbor subsampling nor entropy regularization. The $N = 10$ MADE run starts with uniform configuration sampling and then uses all 1024 configurations; its final stage uses 32 stratified time points per update. The displayed $N = 20$ MADE checkpoint uses semi-gradients with uniform times and configurations; its final continuation stage uses 128 times and 4096 configurations globally per batch, accumulating eight batches across four GPUs before each update.

Table 3: Training objectives and residual-gradient conventions for the reported experiments. Full gradients include parameter differentiation through $\hat{H}|\Psi\rangle$; semi-gradients stop that contribution. Both conventions use finite-difference time derivatives. Objectives listed here are training quantities; the fidelity bounds are evaluated separately using the Hilbert-residual norm. Initial penalties are added for soft-constraint ablations.

Experiment	Residual training objective	Gradient
Small-system bound evaluation, $N = 4, 5, 6$	Scaled mean-absolute residual	Full
Static dynamics, $N = 10, 20$	Mean-absolute residual	Semi
Static dynamics, 4×4	Mean-absolute residual	Semi
One-dimensional ablations, $N = 10$	Mean-absolute residual	Full
Relay and single-window control, $N = 10$	Mean-absolute residual	Semi
Driven protocols, $N = 10$	Mean-absolute residual	Semi, then full

The 4×4 run in Figure 4 uses the same hard constraint, mean-absolute residual, linear time weight, semi-gradient, finite-difference step, and complete Hamiltonian connectivity. Its MADE has three hidden layers of width 512 and $1{,}541{,}188$ parameters. The initial 3000 epochs use 64 independently sampled uniform times on each of three GPUs and 1536 configurations globally (512 per GPU); the learning rate is 10^{-4} through epoch 1500 and 10^{-5} thereafter. The first continuation, through epoch 5000, uses two GPUs, 64 uniform times per GPU, 2048 configurations globally per microbatch, and two-microbatch gradient accumulation, with learning rates 5×10^{-6} for epochs 3001–4000 and 2×10^{-6} for epochs 4001–5000. The displayed second continuation resumes the model and AdamW state at epoch 5000 with the same sampling and accumulation settings; its learning rate is 2×10^{-6} for epochs 5001–6000 and 10^{-6} for epochs 6001–7000. The final observables are normalized full-basis sums over all 2^{16} configurations at 201 times and are compared with a dimension-120 Krylov evolution.

In the implementation, the residual surrogate and uniform sampling correspond to `res_stat=abs` and `born_beta=0`, respectively. The `system_size=large` backend setting activates `detach_Hpsi=True` by default, including for the $N = 10$ run; this flag selects the backend rather than specifying the physical site count.

The semi-gradient replaces the Hamiltonian-action state $\hat{H}(t)|\Psi_\theta(t)\rangle$ by $\mathrm{sg}(\hat{H}(t)|\Psi_\theta(t)\rangle)$ during backpropagation. Here sg acts componentwise, leaving forward values unchanged and stopping parameter derivatives. The residual-driven gradient vanishes at zero numerical residual, but nonzero-residual stationary points need not match those of the full-gradient objective. This statement concerns the residual update, not the separate AdamW weight-decay step. A full-gradient variant differentiates both residual terms.

For the finite-difference residual, the implemented stencil is

$$\partial_t \Psi(\sigma, t) \approx \frac{\Psi(\sigma, t_+) - \Psi(\sigma, t_-)}{t_+ - t_-}, \qquad t_+ = \min(T, t + \varepsilon_t), \qquad t_- = \max(0, t - \varepsilon_t). \tag{37}$$

For sufficiently smooth trajectories the interior stencil is second order in ε_t, whereas the boundary stencils are generally only first order. This finite-difference error remains even as the number of configuration samples increases and must be included in any numerical certification error budget.

A.9 Segmented Continuation: Construction and Evaluation

At each handoff, the next relay window starts exactly from the preceding learned endpoint, which may differ from the true state. Setting the gate horizon to the segment length ΔT gives $g(\Delta T) = 1$, so the transmitted conditionals contain no earlier anchor. This allows continuation without recursively retaining historical models; segmented continuation itself is shared with prior approaches. Relative to the original exact trajectory, subsequent windows must retain the initial-infidelity term in Eq. 4; a zero initial term applies only relative to exact evolution initialized from that learned endpoint.

We compare relay and single-window training on an $N = 10$ open chain at $h/J = 0.5$, initialized in $|+\rangle_x^{\otimes N}$, over $Jt \in [0, 2]$. Relay uses two unit-length windows with 5000 optimizer updates each. The second network receives the first network's frozen endpoint conditionals as its hard initial anchor. A wider single-window network is evaluated after 10000 updates and after continuation to 20000 updates. These are checkpoints of the same single-window trajectory. An epoch in Figure 7 denotes one optimizer update.

Both architectures have three MADE hidden layers. Each relay network uses MLP hidden widths of 256 and has $755{,}844$ parameters, giving $1{,}511{,}688$ across both windows. The single-window network uses widths of 384, retaining the same layer counts and embedding dimensions, and has $1{,}521{,}668$ parameters (0.66% more than the relay total). These counts include entries disabled by fixed MADE masks; excluding them gives $1{,}247{,}728$ for relay and $1{,}280{,}006$ for the control (2.59% more). The first relay network and the control use initialization seed 1; the second relay network starts afresh with seed 2.

All runs use the mean-absolute residual surrogate, semi-gradients, normalized linear weight $(\Delta T - t)/\Delta T$, and gate parameter $\tau = \Delta T$ in local time. Each update uses 24 stratified uniform times and all 1024 configurations. Full-basis spin-flip indexing retains every Hamiltonian connection; finite differences use $\varepsilon_t = 10^{-3}$ with boundary clipping. AdamW uses $\beta = (0.9, 0.99)$, $\epsilon = 10^{-8}$, weight decay 10^{-4}, and gradient-norm clipping at 1. Each relay window uses learning rate 10^{-4} for 2500 updates and 10^{-5} for 2500 updates, with a fresh optimizer at the handoff. The control uses 10^{-4} through update 5000 and 10^{-5} thereafter. Its continuation from 10000 to 20000 restores both model and optimizer states and retains 10^{-5}.

The relay windows and first 10000 control updates run sequentially on two GPUs, each processing 12 times per update. The control continuation uses six GPUs, each processing four times, with the same global batch. When the GPU count changes, sampling uses new random-number streams, while optimizer moments and step counts are preserved. Checkpoints and evaluations are saved every 200 updates. The first comparison matches total updates and approximately matches aggregate parameter counts; the 20000-update control doubles the update budget. GPU counts, floating-point work, and optimizer histories differ, so these budgets do not establish a wall-clock speedup.

Reference states are obtained by full computational-basis diagonalization and used only for evaluation. Normalized full-basis observables and fidelities are evaluated at 201 common times. Figure 7 displays 21 of these times as markers and heatmap columns, without interpolation. At $Jt = 1$, the plotted relay value comes from the first window; the second window's initial M_x and all five correlations agree with it to within 1.3×10^{-8}.

Table 4: Final-checkpoint errors for Figure 7. MAEs average over all 201 evaluation times; $\mathrm{MAE}_{1:5}$ additionally averages over separations $r = 1, \ldots, 5$. $F(T)$ is endpoint fidelity at $JT = 2$. Each strategy has one training realization; the two single-window rows belong to the same trajectory.

Setting (updates)	M_x MAE	$C_{zz}(1)$ MAE	C_{zz} $\mathrm{MAE}_{1:5}$	$F(T)$
Relay ($5000 + 5000$)	0.00520	0.00307	0.00361	0.9490
Single (10000)	0.10720	0.06307	0.03933	0.4687
Single (20000)	0.10367	0.07442	0.04417	0.4671

The two-window run has lower observable errors than both single-window checkpoints (Table 4). Because each gate uses its window duration as τ, the single-window control has a smaller maximum initial conditional slope ($1/2$, versus 1 in each relay window; Eq. 36). This comparison therefore measures the complete continuation protocols rather than the effect of segmentation in isolation. The relay run retains late-time correlation errors, including a nearest-neighbor endpoint error of 0.01854.

A.10 Dynamics under a Time-Dependent Hamiltonian

We test three time-dependent transverse fields on the open $N = 10$ chain, starting from $|+\rangle_x^{\otimes N}$ over $JT = 1$. With $s = t/T$, the fields are a linear ramp, $h/J = 0.5 + 1.5s$; a periodic drive, $h/J = 1.25 - 0.75\cos(2\pi s)$; and a Gaussian pulse, $h/J = 0.5 + 1.5\exp[-((s - 0.5)/0.15)^2]$. Figure 8 compares magnetization and distance-resolved correlations with independent spin-basis RK4 evolution.

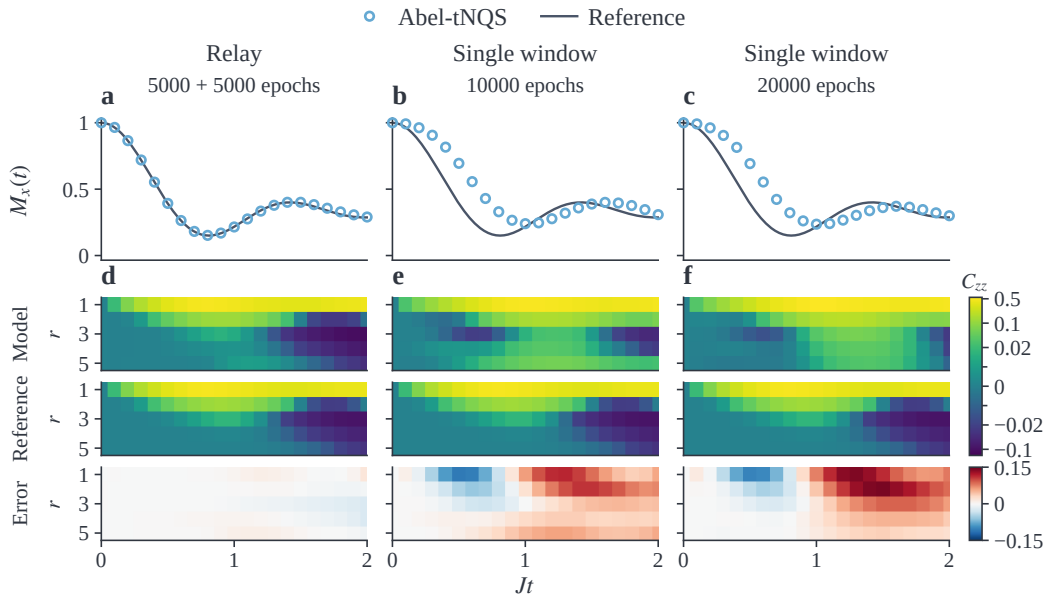

Figure 7: Segmented continuation on the $N = 10$ open chain over $Jt \in [0, 2]$. Columns compare relay ($5000 + 5000$ updates), a wider single window (10000 updates), and the same single window continued to 20000 updates. Aggregate parameter counts are approximately matched. (a–c) M_x: model values at 21 times (blue open circles) and the full 201-time exact reference (gray line). (d–f) Pair-averaged $C_{zz}(r, t)$ for $r = 1, \ldots, 5$: model, reference, and signed error (model minus reference). Heatmaps use 21 times without interpolation. Model and reference share a symmetric-logarithmic scale, linear for $|C_{zz}| \leq 0.02$, with limits $[-0.15, 0.55]$; errors share a linear scale $[-0.15, 0.15]$. All evaluations are full-basis sums, and metrics use all 201 times.

All three models use the MADE backbone with three width-256 hidden layers, 755844 parameters, and the hard initial-state constraint. Each has one training realization (seed 1). Training uses the linearly time-weighted mean-absolute residual and finite-difference time derivatives with step 10^{-3}. After 28000 semi-gradient updates, the ramp, periodic, and pulse models receive 17000, 23000, and 20040 full-gradient updates, respectively.

Each 6000-update full-gradient block consists of 4000 updates at a learning rate of 10^{-6} with 64 sampled time points, followed by 2000 updates at 5×10^{-7} with 128 sampled time points. Both stages enumerate all 1024 basis configurations. In the final refinement stage, the ramp and periodic models each receive 5000 updates, while the pulse receives 2040 updates. This stage uses a learning rate of 10^{-6} and accumulates gradients over 16 independent microbatches of 128 sampled times and all 1024 basis configurations per optimizer update. Full gradients differentiate through both residual terms; reference observables are not training targets. AdamW moments are reset when switching from semi-gradients to full gradients and restored between subsequent full-gradient blocks.

The figure shows the final saved checkpoints at 45000, 51000, and 48040 total updates for the ramp, periodic drive, and pulse, respectively. The pulse run stopped at update 48040 to release compute resources; the other two completed their schedules. Checkpoints were not selected by reference error. These exploratory schedules are not a matched comparison of gradient conventions.

All reported errors use full-basis evaluation at 201 uniformly spaced times, without sampling or smoothing of observables. We summarize correlation accuracy by the MAE averaged over times and separations $r = 1, \ldots, 5$, and report nearest-neighbor and magnetization errors alongside it (Table 5). The correlation MAEs are 0.01187, 0.01168, and 0.01259 for the ramp, periodic drive, and pulse, respectively. Magnetization retains visible late-time deviations, with maximum absolute errors of 0.11566, 0.09904, and 0.05857. The corresponding maximum absolute correlation errors over all times and separations are 0.13180, 0.08374, and 0.07890.

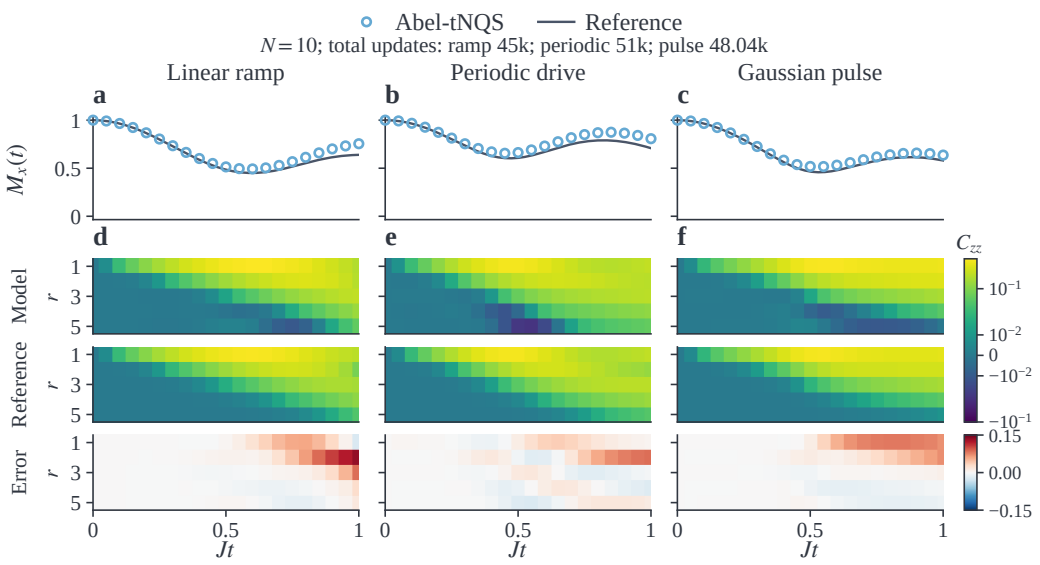

Figure 8: Driven dynamics on the $N = 10$ open chain with MADE. Columns show the linear ramp, periodic drive, and Gaussian pulse. (a–c) M_x: model values at 21 times (blue open circles) and the 201-time RK4 reference (gray line). (d–f) Pair-averaged $C_{zz}(r, t)$ for $r = 1, \ldots, 5$: model, reference, and signed error (model minus reference), displayed at 21 times without interpolation. Model and reference share a symmetric-logarithmic scale, linear for $|C_{zz}| \leq 0.02$; errors use a shared linear scale $[-0.15, 0.15]$, covering all evaluated errors without clipping. Metrics use full-basis sums at all 201 times. Each column represents one training realization.

Table 5: Driven errors at the displayed checkpoints (ramp 45000, periodic 51000, pulse 48040 updates). MAEs average over 201 times; $\mathrm{MAE}_{1:5}$ also averages over the five separations.

Protocol	C_{zz} $\mathrm{MAE}_{1:5}$	$C_{zz}(1)$ MAE	M_x MAE
Linear ramp	0.01187	0.01647	0.03438
Periodic drive	0.01168	0.01237	0.04767
Gaussian pulse	0.01259	0.03332	0.02883

A.11 Controlled Ablation Protocols and Supplementary Results

All runs in Figure 5 use the MADE backbone, AdamW, a finite-difference step of 10^{-3}, full parameter differentiation through $\hat{H}|\Psi\rangle$, and no endpoint-weight floor. Within each study, the hard-linear and hard-uniform configurations differ only in the time weight. The soft variants retain the same network widths and depths, remove the architectural initial constraint, and add a sampled phase-aligned initial penalty. The initial penalty requires additional loss evaluations. All reported summaries include every seed in the specified protocol.

The $N = 10$ runs use the mean-absolute residual surrogate, a loss scale of one in the large-system backend, and uniform configuration sampling. Each microbatch contains 128 sampled times and 64 configurations, with gradients accumulated over two microbatches. The learning rate is 10^{-4} for epochs 1–1500 and 10^{-5} for epochs 1501–3000. The time weights are $T - t$ (linear) and $T/2$ (uniform), both with continuous-time mean $T/2$. Soft variants use 64 uniform initial configurations per microbatch and add $\lambda \widehat{L}_{\mathrm{IC}}/D$, where $D = 2^{10}$ and $\widehat{L}_{\mathrm{IC}}$ is defined below. The coefficients $1, 10, 100$ therefore multiply a per-configuration mean.

Observables and normalized overlaps are evaluated on 101 uniformly spaced times by full-basis summation against exact diagonalization; C_{zz} averages raw nearest-neighbor correlations over open-chain bonds. Figure 5 and Table 6 use all ten seeds, 4–13, for each setting.

Table 6: One-dimensional controlled MADE ablations, $N = 10$, ten seeds (4–13) and 3000 epochs, supplementing Figure 5(a). Initial fidelities for the hard-constraint models are rounded to one; all other entries report means $\pm$ sample standard deviations. All soft settings use linear weighting. F_{grid} is the arithmetic mean over 101 evaluation times, not the trapezoidal time average used in Figure 3.

Setting	M_x MAE	$F(0)$	$F(T)$	F_{grid}
Hard, linear	0.0082 ± 0.0030	1.0000	0.9082 ± 0.0126	0.9854 ± 0.0042
Hard, uniform	0.0540 ± 0.0104	1.0000	0.1762 ± 0.2247	0.4745 ± 0.1926
Soft, $\lambda = 1$	0.4159 ± 0.0521	0.0032 ± 0.0027	0.0041 ± 0.0034	0.0033 ± 0.0028
Soft, $\lambda = 10$	0.3141 ± 0.0759	0.0264 ± 0.0154	0.0164 ± 0.0095	0.0179 ± 0.0099
Soft, $\lambda = 100$	0.0629 ± 0.0917	0.8949 ± 0.3115	0.3216 ± 0.2491	0.5877 ± 0.3203

Figure 5(b,c) uses checkpoints every 200 epochs from 200 to 3000, together with the checkpoint at epoch 1500, giving 16 checkpoints per run and 800 evaluated checkpoints in total. Each is evaluated by summing over all $2^{10} = 1024$ configurations at the same 101 uniformly spaced times. Panel (a) and Table 6 retain the 3000-epoch results. The curves show means across the ten seeds and connect the saved checkpoints without smoothing; panel (a) and Table 6 report the corresponding final sample standard deviations.

For the soft variants, the sampled phase-aligned initial penalty is

$$\widehat{L}_{\mathrm{IC}} = \frac{D}{B_0} \sum_{b=1}^{B_0} \left| \Psi_\theta(\sigma_b, 0) - \widehat{u}\, \psi_0(\sigma_b) \right|^2 ,$$
$$\widehat{u} = \frac{\widehat{z}}{\max(|\widehat{z}|, 10^{-8})}, \qquad \widehat{z} = \frac{1}{B_0} \sum_{b=1}^{B_0} \psi_0(\sigma_b)^* \Psi_\theta(\sigma_b, 0). \tag{38}$$

Here $B_0 = 64$ and the configurations σ_b are sampled uniformly. The factor D converts the sampled mean squared mismatch to an estimate of the full Hilbert-space mismatch before the training objective divides it by D. The safeguarded alignment factor $\widehat{u}$ has unit modulus only when the sampled overlap exceeds the numerical floor; below it, the target amplitude is also attenuated. This penalty is not an exact globally phase-minimized mismatch.

Post-training checks at the 30 final one-dimensional soft checkpoints found a minimum raw local norm of 0.6323 and no normalization-clamp activations in a full-basis evaluation at $t = 0$. Across 30,000 fresh batches of 64 initial configurations (1000 per checkpoint), the phase floor was never reached. These checks do not cover earlier training updates.

The reported one-dimensional summaries use unrounded saved evaluations rather than rounded training logs and exclude the earlier three-seed pilot runs. In the soft training logs, the field labeled L2 contains the residual plus the weighted initial penalty and is not a common accuracy metric across settings.

A.12 Hyperparameters

Tables 7–10 give the settings and complete training schedules for the reported MADE checkpoints, including continuation stages.

Architecture and optimizer. All networks have spatial and time feature dimensions $d_{\mathrm{hid}} = d_t = 32$. The spatial encoder, time encoder, and complex decoder have two, four, and three hidden layers, respectively, followed by their output projections. Their hidden widths are listed as w_{MLP} in Table 7. The MADE backbone has three masked hidden layers of width w_{MADE}, GELU activations, additive time projections, and zero dropout. The other MLPs use LeakyReLU with slope 0.01 and LayerNorm in hidden blocks after the first; no normalization across MADE hidden units is used. Both conditional-amplitude normalization stages use a norm floor of 10^{-12}. The hard-constraint gate uses $\tau = T$ for each window; the soft ablations disable this gate.

Network parameters are stored in float32. AdamW uses $\beta_1 = 0.9$, $\beta_2 = 0.99$, $\epsilon = 10^{-8}$, and weight decay 10^{-4}. Gradients are averaged over the complete effective batch and clipped to Euclidean norm 1 before each update. All experiments use finite differences with $\varepsilon_t = 10^{-3}$, residual

smoothing $\delta_{\mathrm{res}} = 10^{-12}$, and complete Hamiltonian connectivity without neighbor subsampling. The objectives and gradient conventions are listed in Table 3. The small-system bound experiments multiply the configuration-mean absolute residual by 2^N; all other reported mean-absolute residual objectives use multiplier 1. These training scales are separate from the Hilbert-residual norm used for bound evaluation. Linear time weights have zero floor; the ablation's uniform weight and soft penalties are specified in Appendix A.11, and relay weights are normalized by the window duration as in Appendix A.9.

Table 7: Network widths and seeds contributing to the reported results. w_{MLP} is shared by the spatial encoder, time encoder, and decoder. For static dynamics, the seed column identifies the run shown in Figure 4. Each relay window has its own network.

Experiment	w_{MADE}	w_{MLP}	Seed(s)
Bound evaluation, $N = 4, 5, 6$	128	256	0–4 per size
Static chain, $N = 10$	256	256	1
Static chain, $N = 20$	256	256	2
Static 4×4 lattice	512	256	1
Controlled ablations, $N = 10$	128	256	4–13 per setting
Relay, $N = 10$	256	256	1, then 2
Single-window relay control	384	384	1
Driven protocols, $N = 10$	256	256	1 per protocol

Batch conventions. Unless stated otherwise, n_t and n_s denote the global numbers of times and configurations per microbatch, before accumulation and after combining all GPUs. They correspond to `n_sample_t` and `n_L2_states` in sampled configurations; custom continuation stages override these base settings. A is the number of microbatches averaged before an optimizer update, giving An_tn_s residual evaluations. The symbol B marks full-basis enumeration and U marks uniform configuration sampling. Full-basis stages of static $N = 10$, relay, and driven runs use stratified uniform times; other stages use independent uniform times. Accumulation passes resample times and, in sampled stages, configurations. For the lattice only, $n_t = 64$ is per GPU: each rank samples its own time grid and its share of the global n_s configurations, so An_tn_s still counts the total residual evaluations. The chain continuations instead share each sampled time grid across GPUs. Memory chunks used to evaluate one batch do not increase A or change its statistical weight.

Static chains. Table 8 gives the $N = 10$ schedule through update 20000, and Table 9 gives the $N = 20$ schedule through update 29040. Each column of Figure 4 shows one checkpoint from the seed listed in Table 7. AdamW moments are restored between continuation stages. Static $N = 10$ uses one GPU; static $N = 20$ uses one GPU through update 28226 and four GPUs for its final continuation.

Table 8: Static $N = 10$ training schedule. All stages use semi-gradients and $A = 1$. This is also the initial 20000-update schedule for each driven protocol, applied to its own Hamiltonian.

Updates	Learning rate	n_t	n_s	Proposal
1–3000	5×10^{-5}	64	64	U
3001–6000	2×10^{-5}	8	1024	B
6001–8000	5×10^{-6}	16	1024	B
8001–10000	10^{-5}	16	1024	B
10001–11000	5×10^{-6}	16	1024	B
11001–12000	10^{-6}	32	1024	B
12001–17000	10^{-5}	16	1024	B
17001–19000	3×10^{-6}	32	1024	B
19001–20000	10^{-6}	32	1024	B

Bound evaluation, ablations, lattice, and relay. For each $N = 4, 5, 6$ bound-evaluation run, every microbatch uses 128 uniform times and all $2^N = 16, 32, 64$ configurations, with $A = 2$ and full residual gradients. The learning rate is 10^{-4} for updates 1–1500 and 10^{-5} for 1501–3000. The

Table 9: Training schedule for the displayed $N = 20$ checkpoint, including continuation stages. All configurations are sampled uniformly, all rows use semi-gradients, and sample counts are global per microbatch, not per GPU.

System	Updates	Learning rate	n_t	n_s	A
$N = 20$	1–3000	5×10^{-5}	64	256	1
	3001–10000	2×10^{-5}	64	512	1
	10001–17000	10^{-5}	64	1024	1
	17001–19000	3×10^{-6}	64	2048	1
	19001–20000	10^{-6}	64	2048	1
	20001–23000	10^{-5}	64	2048	1
	23001–25000	3×10^{-6}	64	2048	1
	25001–28226	10^{-6}	64	2048	1
	28227–29040	10^{-6}	128	4096	8

$N = 10$ controlled ablations use the same time count, accumulation, learning-rate schedule, and gradient convention, but sample 64 uniform configurations per microbatch. The five settings and initial-condition penalty are given in Appendix A.11.

The static 4×4 lattice always uses $n_t = 64$ per GPU and uniform configurations; n_s below is summed over GPUs. Updates 1–1500 and 1501–3000 use $n_s = 1536$, $A = 1$, and learning rates 10^{-4} and 10^{-5}, respectively. Updates 3001–4000, 4001–6000, and 6001–7000 use $n_s = 2048$, $A = 2$, and rates 5×10^{-6}, 2×10^{-6}, and 10^{-6}, respectively. Three GPUs are used initially and two after update 3000; AdamW moments are restored at both continuation boundaries (3000 and 5000).

Relay and its single-window control use $n_t = 24$, $n_s = 1024$ (full basis), $A = 1$, and semi-gradients. In each 5000-update relay window, the first 2500 updates use 10^{-4} and the rest use 10^{-5}. The single-window control uses 10^{-4} for updates 1–5000 and 10^{-5} for 5001–20000, with results reported at 10000 and 20000. The gate horizons are 1 for each relay segment and 2 for the single window. Optimizer resets and the GPU partition are specified in Appendix A.9.

Driven protocols. Each driven run first follows Table 8, then the continuation in Table 10. All continuation configurations are fully enumerated. The ramp has two 6000-update full-gradient blocks, while periodic and pulse have three; each block uses 4000 updates at 10^{-6} with $n_t = 64$, then 2000 at 5×10^{-7} with $n_t = 128$. These stages have $A = 1$. The final $A = 16$ stage uses $n_t = 128$ and 10^{-6}. AdamW moments are reset once, at the switch from semi-gradients to full gradients at update 28000, and restored between later full-gradient stages. The pulse schedule stops at the displayed update 48040.

Table 10: Driven continuation after update 20000. Here $n_s = 1024$ (full basis) throughout. The block index is $b = 0, 1$ for the ramp and $b = 0, 1, 2$ for periodic and pulse. Times are resampled in every accumulated microbatch.

Protocol	Updates	Learning rate	n_t	A	Gradient
All	20001–25000	10^{-5}	16	1	Semi
All	25001–27000	3×10^{-6}	32	1	Semi
All	27001–28000	10^{-6}	32	1	Semi
All	$28001 + 6000b$–$32000 + 6000b$	10^{-6}	64	1	Full
All	$32001 + 6000b$–$34000 + 6000b$	5×10^{-7}	128	1	Full
Ramp	40001–45000	10^{-6}	128	16	Full
Periodic	46001–51000	10^{-6}	128	16	Full
Pulse	46001–48040	10^{-6}	128	16	Full

A.13 Evaluation and Display of Static Dynamics

Figure 4 uses Jordan–Wigner/BdG references for the open chains and dimension-120 Krylov evolution for the open 4×4 lattice. The $N = 10, 20$ chains and the lattice use full-basis observable sums.

Magnetization is evaluated at 200 times for $N = 10, 20$ and 201 times for the lattice. The figure displays 21 approximately uniformly spaced model markers and the reference on each complete saved time grid. Correlation maps show pair-averaged raw $C_{zz}(r,t)$ for $r = 1, \ldots, 5$ at 21 times without interpolation. Error maps show $C_{zz}^{\text{Abel-tNQS}} - C_{zz}^{\text{ref}}$. Model and reference maps share a symmetric-logarithmic color scale with linear region $|C_{zz}| \leq 0.02$ and limits $[-0.05, 0.5]$; signed-error maps use zero-centered linear limits ± 0.015, ± 0.07, and ± 0.05 from left to right, without clipping.

Each column uses one MADE checkpoint: seed 1, update 20000 for $N = 10$; seed 2, update 29040 for $N = 20$; and seed 1, epoch 7000 for the lattice.

A.14 Supplementary Observable Error Metrics

Table 11 records three error summaries for the checkpoints in Figure 4. For $N = 10, 20$, magnetization MAEs use all 200 evaluation times and correlation errors use the 21 displayed times. The 4×4 entries use all 201 evaluation times for every metric, although its heatmap displays a 21-time subset. The distance-averaged metric is

$$\text{MAE}_{r\leq 5}(C_{zz}) = \frac{1}{5n_t} \sum_{r=1}^{5} \sum_{k=1}^{n_t} \left| C_{zz}^{\text{Abel-tNQS}}(r, t_k) - C_{zz}^{\text{ref}}(r, t_k) \right|. \tag{39}$$

This is an arithmetic mean with equal weight for each distance and evaluation time. Each correlation estimate first averages over all site pairs at the specified separation. Small averages can therefore obscure errors at particular distances and times; the signed-error heatmaps retain these details.

Table 11: Static-quench errors at $h/J = 0.5$ and $Jt \in [0, 1]$ for the systems in Figure 4. References are JW/BdG for the chains and Krylov evolution for the 4×4 lattice. All entries use full-basis sums. Values are rounded to three significant figures and are point estimates from one trained checkpoint per system; no confidence intervals are reported.

System	$\text{MAE}(M_x)$	$\text{MAE}(C_{zz}(1,t))$	$\text{MAE}_{r\leq 5}(C_{zz})$
$N = 10$	0.00113	0.00229	0.00139
$N = 20$	0.00770	0.00940	0.00437
4×4	0.0378	0.0151	0.00442

For the square lattice, $\mathcal{P}_r$ in Eq. 17 contains 24, 34, 32, 20, and 8 pairs for Manhattan distances $r = 1, \ldots, 5$. On the full 201-time grid, the magnetization maximum absolute error is 0.0662. At $Jt = 1$, $\Delta M_x = -0.0511$ and $\Delta C_{zz}(1,t) = +0.0454$. The per-distance correlation MAEs for $r = 2, 3, 4, 5$ are 0.00180, 0.00292, 0.000631, and 0.00170, respectively.

A.15 Reference Solvers

The reference solvers use the same Hamiltonians, boundary conditions, and initial states as the corresponding neural-model experiments. They provide evaluation trajectories independently of the neural ansatz and are not training targets. Full-basis diagonalization and Jordan–Wigner/BdG solve the finite-system equations up to floating-point error; Krylov and Runge–Kutta propagation are numerical approximations, with checks reported below.

Small static chains. For the $N = 4, 5, 6$ certificate study, the $N = 10$ controlled ablations, and the $N = 10$ relay comparison, we construct the Hamiltonian in the full computational basis. Its diagonal entries are $-J \sum_{j=1}^{N-1} \sigma_j^z \sigma_{j+1}^z$, and each single-spin flip has matrix element $-h$. A Hermitian eigendecomposition $H = V \operatorname{diag}(E_\nu) V^\dagger$ gives

$$\psi_{\text{ref}}(t) = V \operatorname{diag}(e^{-iE_\nu t}) V^\dagger \psi_0, \qquad \psi_0(\sigma) = 2^{-N/2}. \tag{40}$$

Full-basis sums then give the reference observables and overlaps on the evaluation grids specified for each experiment.

Static-chain observables. The chain columns in Figure 4 use Jordan–Wigner/BdG evolution. With $\sigma_j^x = 1-2c_j^\dagger c_j$ and the Nambu vector $\boldsymbol{\xi} = (c_1, \dots, c_N, c_1^\dagger, \dots, c_N^\dagger)^\mathsf{T}$, the implementation constructs

$$\begin{aligned} \mathcal{H}_{\mathrm{BdG}} &= \begin{pmatrix} A & B \\ -B & -A \end{pmatrix}, \\ A_{jj} &= 2h, \qquad A_{j,j+1} = A_{j+1,j} = -J, \\ B_{j,j+1} &= -B_{j+1,j} = -J, \end{aligned} \tag{41}$$

with all other entries zero. The product-x initial state is the fermion vacuum, so $G(0) = \langle \boldsymbol{\xi}\boldsymbol{\xi}^\dagger \rangle = \mathrm{diag}(I_N, 0_N)$. Diagonalizing the $2N \times 2N$ matrix yields $U(t) = e^{-i\mathcal{H}_{\mathrm{BdG}}t}$ and $G(t) = U(t)G(0)U(t)^\dagger$. Magnetization follows from $M_x(t) = N^{-1}\sum_j (2\,\mathrm{Re}\, G_{jj}(t) - 1)$. The eigendecomposition uses double precision and the covariance arithmetic uses complex double precision; the archived magnetization output is cast to single precision.

For correlations, define Majoranas $a_j = c_j + c_j^\dagger$ and $b_j = i(c_j^\dagger - c_j)$. Wick's theorem expresses $\langle \sigma_j^z \sigma_{j+r}^z \rangle$ as $(-i)^r \,\mathrm{Pf}(K_{j,r})$, where the upper-triangular entries of the antisymmetric matrix $K_{j,r}$ are the two-point contractions of the ordered string $(b_j, a_{j+1}, b_{j+1}, \dots, a_{j+r})$. We average over the $N - r$ open-chain pairs. The figure evaluator uses pivoted antisymmetric elimination: at each step it selects the largest-magnitude entry in the first active row, swaps the corresponding rows and columns together, and updates the remaining antisymmetric block. An independent six-spin computational-basis diagonalization at $Jt = 0, 0.2, 0.6, 1$ and $r = 1, 2, 3$ gives a maximum correlation discrepancy of 1.86×10^{-15}.

Open 4×4 lattice. The reference spans all $D = 2^{16}$ configurations and all 24 nearest-neighbor bonds. We apply H without storing a dense $D \times D$ matrix: the interaction multiplies each amplitude by its diagonal Ising energy, and the field term sums amplitudes at the 16 single-spin-flipped configurations. Starting from the uniform product-x vector, Hermitian Lanczos iteration constructs a 120-dimensional Krylov basis Q_m and a tridiagonal matrix T_m. The propagated state is approximated by $Q_m e^{-itT_m} e_1$ and evaluated at 201 uniform times over $Jt \in [0, 1]$. The implementation repeats orthogonalization against the current and previous Lanczos vectors and stops early if the next-vector norm is below 10^{-13}; it does not use full reorthogonalization. Magnetization is computed by single-spin-flip overlaps, and $C_{zz}(r, t)$ by probability-weighted Manhattan-shell averages. The archived reference has maximum norm error 5.00×10^{-14}; the separately computed spatial $r = 1$ curve agrees with the bond-average reference to 1.10×10^{-14}. These checks assess normalization and observable consistency, rather than establishing a rigorous Krylov truncation bound.

Driven ten-spin chain. For each of the three fields in Appendix A.10, the reference independently integrates $\dot{\psi} = -iH(t)\psi$ in the computational basis using classical fourth-order Runge–Kutta in complex double precision. The solver evaluates the field at each Runge–Kutta stage, uses 4000 uniform steps over $JT = 1$ ($J\Delta t = 2.5 \times 10^{-4}$), and records states every 20 steps, giving 201 times. Magnetization and pair-averaged correlations at $r = 1, \dots, 5$ are full-basis sums over 1024 configurations. Repeating with 2000 steps changes M_x and C_{zz} by at most 4.11×10^{-11} and 1.09×10^{-10}, respectively, across all three protocols; the maximum norm error of the finer trajectories is 2.15×10^{-13}. This is a step-refinement check, not a rigorous integration-error bound.