

Quantum Wavefunction Dynamics via Hard-Constrained Probability-Preserving Transformers

Mushtaq Ali

`mushtaq.ali@ue.edu.pk`

University of Education Lahore

Muzamil Tariq

University of Education Lahore

Niaz Ali Khan

Xiamen University

Article

Keywords:

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11153117/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Quantum Wavefunction Dynamics via Hard-Constrained Probability-Preserving Transformers

Mushtaq Ali,^{1,*} Muzamil Tariq,¹ and Niaz Ali Khan²

¹*Department of Physics, Division of Science and Technology, University of Education Lahore, Pakistan*

²*Department of Physics, Xiamen University, Xiamen 361005, China*

Solving the time-dependent Schrödinger equation (TDSE) accurately over horizons remains computationally demanding. We propose a generalizable probability-preserving transformer (PPT) that enforces unit probability through a hard architectural constraint rather than a soft penalty, and we train it to predict future complex-valued wavefunctions in a one-dimensional infinite square well from short temporal histories. Using Crank–Nicolson solutions as reference data, we assess generalization across ground, excited, Gaussian wavepacket, and superposition initial states. The PPT generalizes across these families while conserving probability to numerical precision, establishing a physically constrained transformer framework for TDSE evolution.

I. INTRODUCTION

The Schrödinger equation is the canonical equation governing non-relativistic quantum dynamics, parameterizing the evolution of quantum states and providing the basis for predicting physical observables such as position, momentum, and energy [1–3]. For a quantum system characterized by the complex wavefunction $\psi(x, t)$, the squared magnitude $|\psi(x, t)|^2$ signifies the probability density, while its temporal evolution is governed by the time-dependent Schrödinger equation (TDSE),

$$i\hbar \frac{\partial \psi(x, t)}{\partial t} = \hat{H}\psi(x, t) \quad (1)$$

where $\hbar$ represents the reduced Planck constant and $\hat{H}$ denotes the system Hamiltonian operator [1, 2]. The TDSE is paramount to the study of quantum dynamics, encompassing non-equilibrium processes, many-body systems, light–matter interactions, and ultrafast phenomena [4–8]. The governing constraint for physically valid quantum evolution is the conservation of the total probability, $\langle \psi | \psi \rangle = 1$.

Numerical methods remain indispensable for accurately propagating the TDSE and generating reference solutions. The Crank–Nicolson (CN) method provides a well-tested and ubiquitous implicit time-integration scheme for discretized quantum systems [9–11], while finite-difference and spectral approaches provide standard frameworks for representing spatial operators [5, 10, 12, 13]. Although these methods provide reliable solutions, repeated sequential propagation can become computationally demanding when simulations are required for many initial states or trajectories.

Machine learning methods offer a compelling alternative by learning mappings between quantum states and their temporal evolution, potentially reducing the computational cost of repeated predictions [14–16]. Neural

approaches have been applied to quantum states and dynamics [17–21], while transformers have emerged as an active area of study for physical systems, partial differential equations, and quantum dynamics because of their ability to model temporal and long-range dependencies [22–29]. However, unconstrained neural models may produce predictions that violate fundamental physical properties such as wavefunction normalization. Physics-informed methods incorporate physical knowledge into learning [30, 31], and hard architectural constraints provide a parallel scheme for enforcing physical properties directly [32].

In the proposed approach, we employ a probability-preserving transformer (PPT) to learn the temporal evolution of the TDSE. The architecture imposes probability normalization as a hard output constraint by normalizing the predicted complex wavefunction on each forward pass, while its multi-token formulation uses temporal histories to learn state evolution [22–24, 27, 29]. Rather than restricting the investigation to a single initial state, we assess the generalization performance of the PPT across multiple initial-condition families: ground states, first-excited states, Gaussian wavepackets, and superposition states. This allows the model to be tested on diverse quantum trajectories.

High-precision Crank–Nicolson solutions serve as reference trajectories, and the learned predictions are assessed through wavefunction reconstruction accuracy together with physically meaningful constraints, including probability conservation, energy stability, and quantum-state fidelity [1, 2, 4, 33]. The study therefore probes both predictive accuracy and the ability of the probability-preserving architecture to maintain physical consistency across different initial-condition families [20, 27, 28, 34]. This establishes a quantifiable audit of a physically constrained transformer for TDSE propagation.

The remainder of the paper is organized as follows. Section II presents the TDSE formulation and the CN method used to generate reference trajectories for the one-dimensional infinite square well. Section III describes the proposed probability-preserving transformer architecture, dataset construction, and training method-

* mushtaq.ali@ue.edu.pk

ology. Section IV evaluates probability preservation, prediction accuracy, and generalization across unseen trajectories and initial-condition families. Section V compares the computational performance of the standard PPT with the CN method. The concluding section highlights the findings and offers scope for future work.

II. METHODOLOGY

A. Physical System and TDSE Formulation

We consider a one-dimensional infinite square well as a validation protocol for evaluating the Probability-Preserving transformer (PPT) [1–3]. The time-dependent Schrödinger equation (TDSE) is

$$i\hbar \frac{\partial \psi(x,t)}{\partial t} = -\frac{\hbar^2}{2m} \frac{\partial^2 \psi(x,t)}{\partial x^2} + V(x)\psi(x,t) \quad (2)$$

where the infinite-well potential is defined as

$$V(x) = \begin{cases} 0, & 0 < x < L, \\ \infty, & \text{otherwise} \end{cases} \quad (3)$$

Within the well, $V(x) = 0$, yielding

$$i\hbar \frac{\partial \psi(x,t)}{\partial t} = -\frac{\hbar^2}{2m} \frac{\partial^2 \psi(x,t)}{\partial x^2} \quad (4)$$

Natural units $\hbar = m = L = 1$ are used throughout. The wavefunction adheres to Dirichlet boundary conditions,

$$\psi(0,t) = \psi(1,t) = 0 \quad (5)$$

and satisfies the probability normalization condition [1–3],

$$P(t) = \int_0^L |\psi(x,t)|^2 dx = 1 \quad (6)$$

B. Reference Data Generation

High-fidelity reference trajectories are generated using the Crank–Nicolson method [9–12]. The spatial domain is sampled using $N_x = 100$ uniformly spaced points,

$$\Delta x = \frac{L}{N_x - 1} = \frac{1}{99} \approx 0.01010101 \quad (7)$$

while each trajectory contains $N_t = 200$ stored temporal states with $\Delta t = 0.001$, giving a final stored simulation time of $t = 0.199$.

The Crank–Nicolson discretization gives

$$A\psi^{n+1} = B\psi^n \quad (8)$$

where

$$A = \text{tridiag}(-r, 1 + 2r, -r), \quad (9)$$

$$B = \text{tridiag}(r, 1 - 2r, r) \quad (10)$$

with

$$r = \frac{i\hbar\Delta t}{4m\Delta x^2} \quad (11)$$

The derived tridiagonal systems are solved at each time step to obtain the reference wavefunction [9, 10].

1. Standard Dataset for PPT Validation

The first dataset provides the baseline configuration for validating the PPT and its probability-preserving behavior. The initial state is the normalized ground state of the infinite square well,

$$\psi_1(x,0) = \sqrt{\frac{2}{L}} \sin\left(\frac{\pi x}{L}\right) \quad (12)$$

which reduces to $\psi_1(x,0) = \sqrt{2} \sin(\pi x)$ for $L = 1$. Its analytical time evolution is

$$\psi_1(x,t) = \psi_1(x,0) \exp\left(-\frac{iE_1 t}{\hbar}\right) \quad (13)$$

where

$$E_n = \frac{n^2 \pi^2 \hbar^2}{2mL^2} \quad (14)$$

A single Crank–Nicolson trajectory containing 200 temporal states is generated and used for the initial PPT validation.

2. Multi-Initial-Condition Dataset for Generalization

The second dataset is constructed to quantify the generalization of the PPT architecture across physically distinct initial conditions. Four initial-condition families are considered: ground states, first-excited states, Gaussian wavepackets, and superposition states. The infinite-well eigenstates are

$$\psi_n(x,0) = \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi x}{L}\right) \quad (15)$$

with the ground state corresponding to $n = 1$ and the first-excited state to $n = 2$. For the first-excited state,

$$\psi_2(x,0) = \sqrt{\frac{2}{L}} \sin\left(\frac{2\pi x}{L}\right) \quad (16)$$

The Gaussian wavepacket is represented by

$$\psi_G(x,0) = A \exp\left[-\frac{(x-x_0)^2}{4\sigma^2}\right] \exp(ik_0 x) \quad (17)$$

where A is selected to satisfy normalization, x_0 is the initial packet center, σ controls the spatial width, and k_0 quantifies the initial momentum [1, 2]. Superposition

states are constructed from multiple infinite-well eigenstates,

$$\psi_S(x, 0) = \sum_{n=1}^N c_n \psi_n(x, 0) \quad (18)$$

where the coefficients satisfy

$$\sum_{n=1}^N |c_n|^2 = 1 \quad (19)$$

A total of 60 independent trajectories are generated: 10 ground-state, 10 first-excited-state, 20 Gaussian-wavepacket, and 20 superposition trajectories. Each trajectory contains 100 spatial points over 200 temporal states, resulting in 1,200,000 space-time samples. Trajectory identifiers and initial-condition classes are preserved for trajectory-level and family-level generalization analysis.

C. Numerical Validation of Reference Trajectories

The CNM trajectories are validated before being used as reference data. The discrete probability is calculated as

$$P(t_n) = \Delta x \sum_{i=1}^{N_x} [\text{Re}(\psi_i, t_n)^2 + \text{Im}(\psi_i, t_n)^2] \quad (20)$$

and the maximum probability deviation is

$$\Delta P_{\max} = \max_t |P(t) - 1| \quad (21)$$

The relative energy drift is evaluated as

$$\Delta E_{\text{rel}} = \frac{|E(t) - E(0)|}{|E(0)|} \quad (22)$$

where the expectation value of the Hamiltonian is

$$E(t) = \frac{\langle \psi(t) | \hat{H} | \psi(t) \rangle}{\langle \psi(t) | \psi(t) \rangle} \quad (23)$$

For comparison between predicted and reference quantum states, the state fidelity is calculated as

$$F(t) = |\langle \psi_{\text{true}}(t) | \psi_{\text{pred}}(t) \rangle|^2 \quad (24)$$

For the complete 60-trajectory reference dataset, the maximum probability deviation is 5.9508×10^{-14} and the maximum relative energy drift is 7.5163×10^{-15} , with zero boundary error. These results establish the CNM trajectories as numerically consistent reference solutions.

D. Data Representation and Preprocessing

The complex wavefunction is represented by its real and imaginary components,

$$\psi(x, t) = \text{Re}[\psi(x, t)] + i \text{Im}[\psi(x, t)] \quad (25)$$

Each temporal state is therefore represented by a 200-dimensional real vector,

$$\psi(t_n) = [\text{Re}(\psi(x_1, t_n)), \dots, \text{Re}(\psi(x_{100}, t_n)), \text{Im}(\psi(x_1, t_n)), \dots, \text{Im}(\psi(x_{100}, t_n))] \quad (26)$$

For the standard validation, temporal states are first vectorized into a single-token representation and subsequently into five-step temporal windows. The same five-step window representation is used for the multi-initial-condition dataset, guaranteeing that generalization is evaluated using the same input structure and model architecture.

E. Probability-Preserving transformer

The PPT parses short temporal histories of the quantum state and predicts the subsequent wavefunction. State, temporal, and positional embeddings are used to typify the input sequence and model temporal dependencies [22–24, 27, 29].

The forecasted real and imaginary components are combined to reconstruct the complex wavefunction. A hard normalization layer then enforces unit probability:

$$\Psi(x_i, t_n) = \frac{\psi(x_i, t_n)}{\sqrt{P(t_n)}} \quad (27)$$

Here, $P(t_n)$ is calculated from the predicted wavefunction using Eq. (20). Thus, unit probability normalization is explicitly enforced by the architecture rather than through a soft loss penalty [32].

The same PPT architecture and probability-preserving normalization are used in both stages. The first stage establishes the baseline predictive and probability-preserving behavior, while the second stage evaluates whether the learned temporal dynamics generalize across unseen trajectories and multiple initial-condition families.

III. ALGORITHMIC IMPLEMENTATION

The computational framework was architected to progressively evaluate the effects of temporal sequence modeling and the hard probability constraint. Three model configurations were employed: an unconstrained Multi-Layer Perceptron (MLP), a single-token transformer with hard probability normalization, and the final multi-token Probability-Preserving transformer (PPT). The MLP provides an unconstrained neural baseline, while the transformer configurations assess the effects of hard probability preservation and temporal context.

A. Unconstrained MLP Baseline

The first configuration is a fully connected MLP lacking an explicit probability-preserving layer. The input encompasses the temporal coordinate, while the output represents the complete quantum state through the real and imaginary components of the wavefunction. For $N_x = 100$ spatial points, the output dimension is therefore $2N_x = 200$.

The MLP serves as an unconstrained baseline for determining whether the TDSE mapping can be learned without explicitly imposing probability conservation. Since no normalization operation is applied to its output, the predicted wavefunction is not structurally constrained to satisfy $P = 1$.

B. Single-Token Probability-Preserving transformer

The second configuration introduces the hard probability-preserving mechanism within a transformer architecture. Each temporal state is encoded as an individual token, represented by a 200-dimensional real vector corresponding to the full wavefunction. The transformer maps the temporal input to the corresponding quantum state and subsequently deploys the hard normalization layer.

For an unconstrained prediction $\hat{\psi}$, the discrete probability is calculated as

$$P = \Delta x \sum_{i=1}^{N_x} \left[\text{Re}(\hat{\psi}_i)^2 + \text{Im}(\hat{\psi}_i)^2 \right] \quad (28)$$

The normalization factor is then defined as

$$\alpha = \sqrt{P + \varepsilon} \quad (29)$$

where $\varepsilon = 10^{-12}$ provides numerical stability. The normalized real and imaginary components are

$$\text{Re}(\Psi) = \frac{\text{Re}(\hat{\psi})}{\alpha}, \quad \text{Im}(\Psi) = \frac{\text{Im}(\hat{\psi})}{\alpha} \quad (30)$$

The normalized components are subsequently recombined to form the final predicted quantum state. Thus, probability conservation is enforced directly on the model output rather than through an additional penalty term in the training objective [32].

C. Multi-Token Probability-Preserving transformer

The final configuration expands the single-token model by providing explicit temporal context to the transformer. Five consecutive temporal states are grouped into a sliding window,

$$X_i = [\psi(t_i), \psi(t_{i+1}), \psi(t_{i+2}), \psi(t_{i+3}), \psi(t_{i+4})] \quad (31)$$

and the model predicts the subsequent state,

$$\hat{\psi}(t_{i+5}) \quad (32)$$

Each state is represented by 200 real-valued components associated with the real and imaginary parts of the wavefunction. The temporal sequence is therefore processed as a five-token input, allowing self-attention to capture dependencies between preceding quantum states [22–24, 27]. As shown in Fig. 1, the multi-token PPT consists of four main stages: input embedding, transformer-based temporal feature extraction, quantum-state reconstruction, and hard probability normalization.

1. Input Embedding

Each input wavefunction is projected into a $d_{\text{model}} = 64$ -dimensional state representation. Temporal coordinates are separately embedded and concatenated with the state representation. Learnable positional embeddings are added to retain the ordering of the five temporal tokens. The resulting input to the transformer encoder has the form

$$\mathbf{Z} \in \mathbb{R}^{\text{batch} \times 5 \times 64} \quad (33)$$

2. transformer Encoder

The embedded sequence is encoded by two stacked transformer encoder layers with four attention heads and a feed-forward dimension of 128. Multi-head self-attention facilitates interactions between the five preceding temporal states, allowing the model to learn temporal dependencies in the evolving quantum wavefunction [22–24].

The encoded sequence retains the dimensions

$$\mathbf{Z}_{\text{enc}} \in \mathbb{R}^{\text{batch} \times 5 \times 64} \quad (34)$$

The representation corresponding to the final temporal token is used for subsequent quantum-state reconstruction.

3. Quantum-State Reconstruction

The selected encoded representation is projected through a densely connected output layer to emit a 200-dimensional vector,

$$\hat{\mathbf{y}} = [\hat{\psi}_{\text{real}}, \hat{\psi}_{\text{imag}}] \in \mathbb{R}^{200} \quad (35)$$

The first 100 components represent the real part and the remaining 100 components represent the imaginary part of the predicted wavefunction across the spatial grid.

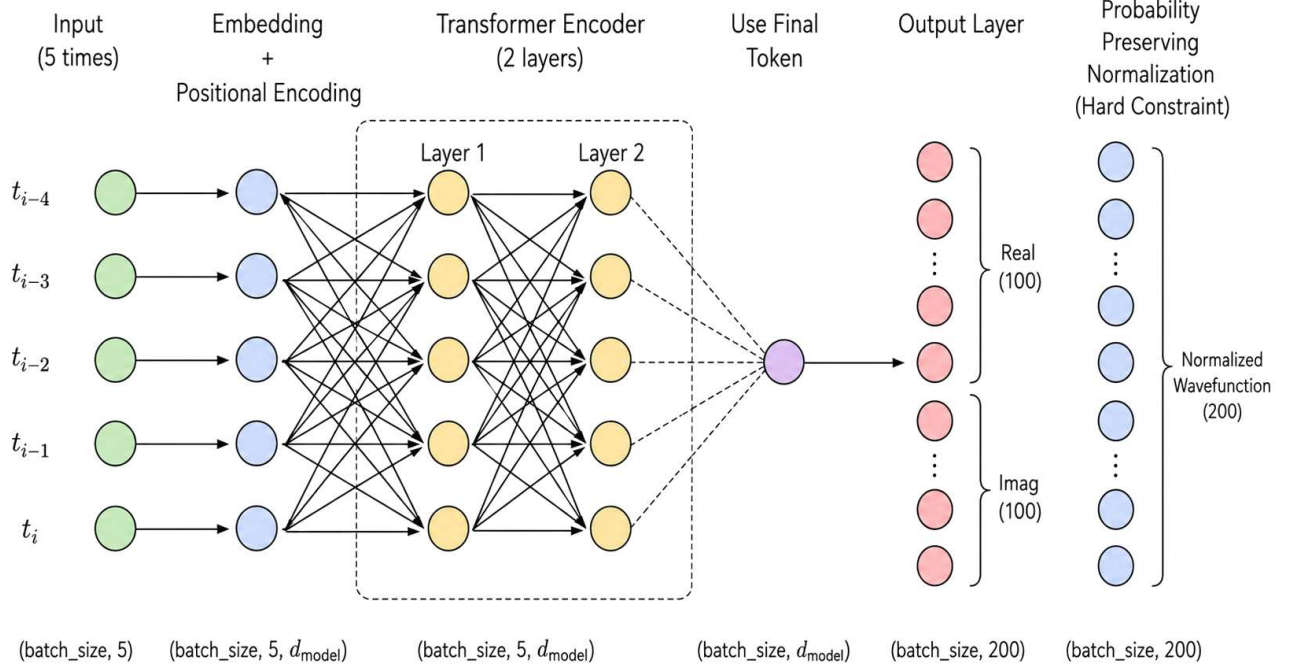


Figure 1. Architecture of the proposed multi-token Probability-Preserving transformer.

4. Hard Probability-Preserving Layer

The unconstrained output is passed through the normalization procedure given by Eqs. (28)–(30). The normalized state is therefore

$$\Psi = [\text{Re}(\Psi), \text{Im}(\Psi)] \quad (36)$$

This operation is differentiable and is utilized during both training and inference. Consequently, the model can be optimized using prediction error while probability conservation is structurally built into the architecture [32].

D. Training Framework

The same multi-token PPT architecture is used for both the standard validation and the multi-initial-condition generalization study. The model is trained using the AdamW optimizer with a optimization step size of 10^{-4} , regularization scale of 10^{-4} , and a batch size of 64.

The training objective combines the wavefunction prediction error with a boundary-aware term,

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \alpha \mathcal{L}_{\text{boundary}} \quad (37)$$

where $\alpha = 0.3$. The loss is assessed on the 200-dimensional real-valued representation containing the real and imaginary components of the wavefunction. The

MSE between the predicted and reference representations is

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{k=1}^N (\hat{\psi}_k - \psi_k)^2 \quad (38)$$

The boundary-aware component assigns greater weight to errors near the well boundaries,

$$\mathcal{L}_{\text{boundary}} = \frac{1}{N} \sum_{k=1}^N w_k (\hat{\psi}_k - \psi_k)^2 \quad (39)$$

where the spatial weighting function increases the contribution of errors near $x = 0$ and $x = 1$.

For the generalization experiment, trajectories are split at the trajectory level using an 80/20 train/test ratio within each initial-condition family using a fixed random seed of 42. To prevent trajectory-level data leakage, all temporal samples from a single trajectory are assigned exclusively to either the training or test set. The same model architecture, normalization procedure, and training framework are retained across all initial-condition families.

IV. NUMERICAL VALIDATION

The numerical validation is divided into two stages. The first appraises probability preservation using the standard dataset. The second examines generalization to

unseen trajectories across multiple initial-condition families.

A. Probability-Preservation Validation

The first experiment reveals whether the proposed architecture preserves quantum probability normalization during prediction. An unconstrained Multi-Layer Perceptron (MLP) is first considered as a baseline. Although the MLP maps the general wavefunction evolution, its predictions do not maintain unit probability, with total probabilities ranging from 1.039418 to 1.045530. This substantiates that minimizing prediction error alone does not guarantee the preservation of the normalization constraint [1, 31, 32].

Table I. Total probability obtained from the unconstrained MLP predictions.

Test	Total Probability (P)
Sample 0	1.039418
Sample 1	1.041417
Sample 2	1.043455
Sample 3	1.045530

The systematic deviation from unity necessitates the hard probability-preserving layer, which explicitly normalizes each predicted quantum state [32].

For the standard multi-token PPT experiment, the resulting prediction accuracy and probability statistics are summarized in Table II.

Table II. Numerical validation of the multi-token PPT on the standard dataset.

Performance Metric	Multi-token PPT
Mean Squared Error (MSE)	3.46796×10^{-4}
Root Mean Squared Error (RMSE)	1.86225×10^{-2}
Relative L_2 Error	2.64688×10^{-2}
Maximum Absolute Error	6.04706×10^{-2}
Minimum Total Probability	0.99999821186
Maximum Total Probability	1.000000119209
Mean Total Probability	1.000000000000
Probability Standard Deviation	8.37516×10^{-8}

The PPT upholds the total probability at unity with numerical precision, with a maximum deviation of approximately 1.79×10^{-7} . The remaining MSE, RMSE, relative L_2 error, and maximum absolute error quantify the approximation error relative to the reference solution, while the probability statistics exhibit the effect of the hard normalization constraint [1, 32].

Figure 2 and Figure 3 show that both the real and imaginary components closely follow the reference solution across the spatial domain. The corresponding absolute spatial error in Figure 4 remains localized, with the

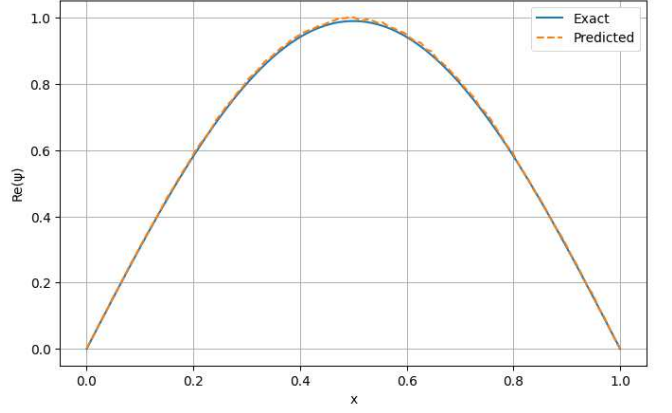


Figure 2. Comparison of the real wavefunction component $\text{Re}(\psi)$ across the spatial domain.

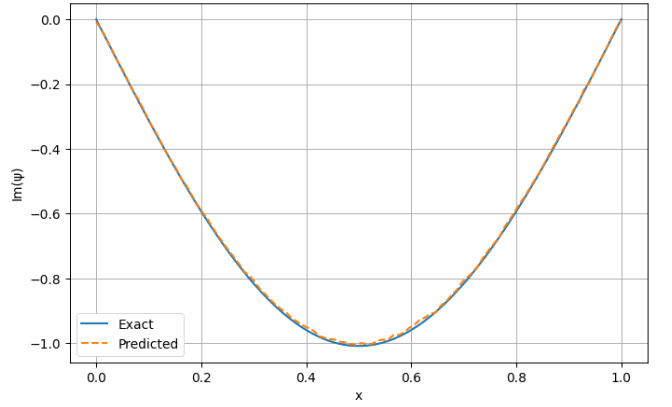


Figure 3. Comparison of the imaginary wavefunction component $\text{Im}(\psi)$ across the spatial domain.

largest deviations occurring near the central region. This distinct concentration of high peaks clearly demonstrates that the largest absolute spatial errors are localized in the central region of the domain.

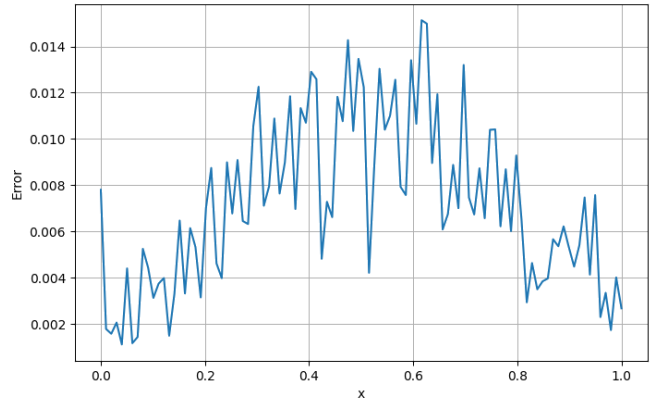


Figure 4. Absolute spatial error of the wavefunction prediction.

B. Generalization Across Initial-Condition Families

The second experiment evaluates the same PPT architecture on previously unseen trajectories from four initial-condition families: ground states, first-excited states, Gaussian wavepackets, and superposition states. The dataset is split at the trajectory level, precluding temporal samples from the same trajectory from appearing in both training and testing sets.

The CNM solutions are retained as the numerical reference for evaluating the learned PPT predictions. The global comparison is summarized in Table III.

Table III. Comparison of the Crank–Nicolson reference solutions and the PPT on held-out trajectories across multiple initial-condition families.

Metric	CNM	PPT
MSE	2.0693×10^{-6}	2.5284×10^{-5}
RMSE	1.4385×10^{-3}	5.0283×10^{-3}
Maximum Absolute Error	2.0444×10^{-3}	9.6076×10^{-2}
Maximum Probability Deviation	5.9508×10^{-14}	2.9802×10^{-7}
Maximum Relative Energy Drift/Error	7.5163×10^{-15}	2.4875×10^{-2}
Mean Probability	1.0000000000	1.0000000000
Relative L_2 Error	1.4458×10^{-3}	7.1469×10^{-3}
Mean Fidelity	1.000000	0.999949
Minimum Fidelity	1.000000	0.998270

The CNM reference solutions remain close to numerical precision, with a maximum probability deviation of 5.95×10^{-14} and a maximum relative energy drift of 7.52×10^{-15} . The PPT achieves an MSE of 2.5284×10^{-5} , RMSE of 5.0283×10^{-3} , and relative L_2 error of 7.1469×10^{-3} , while maintaining a maximum probability deviation of only 2.98×10^{-7} .

The PPT achieves a mean fidelity of 0.999949 and a minimum fidelity of 0.998270. The mean relative energy error is 1.1675×10^{-3} , with a maximum of 2.4875×10^{-2} .

The largest trajectory-level deviation occurs for gaussian_18, which has a mean RMSE of 0.016202 and a maximum RMSE of 0.029289. This single onerous trajectory contributes disproportionately to the aggregate error and causes the overall mean error to drift upward. The result is retained as part of the generalization assessment rather than treated as an outlier to be removed. Increasing the diversity of Gaussian trajectories will improve performance on such localized wavepacket dynamics.

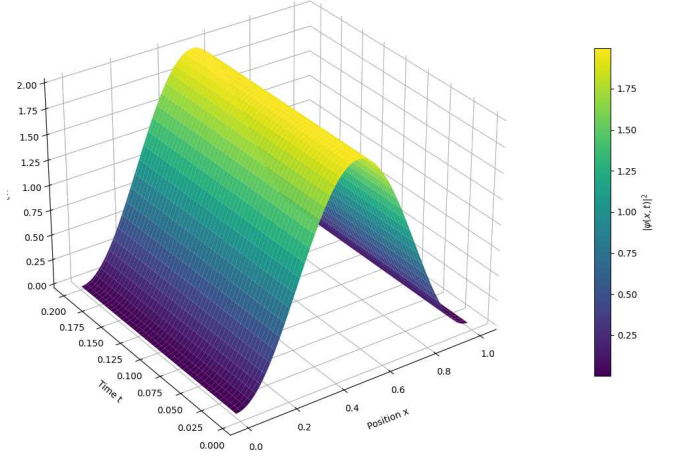


Figure 5. Spatiotemporal evolution of the exact probability density $|\psi(x, t)|^2$.

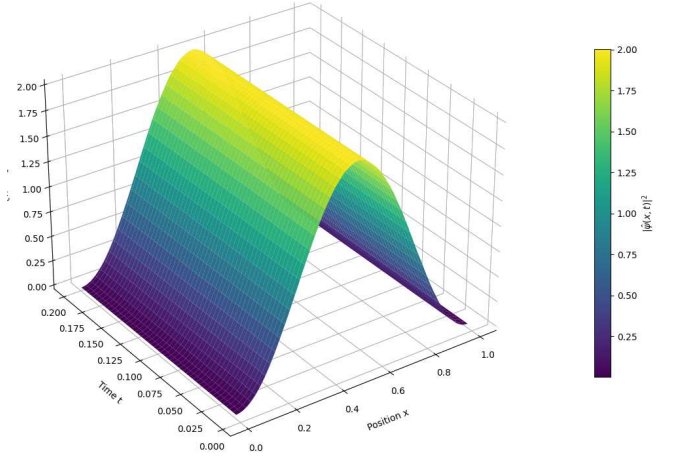


Figure 6. Spatiotemporal evolution of the PPT-predicted probability density $|\psi(x, t)|^2$.

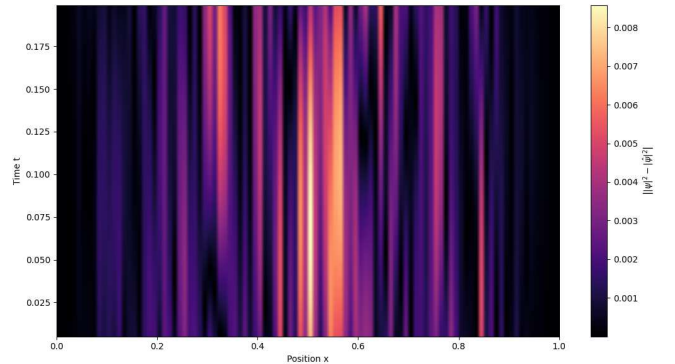


Figure 7. Spatiotemporal error heatmap over time and position.

To elucidate this dynamic behavior, Figure 5 maps the reference exact solution alongside the PPT-predicted so-

lution (Figure 6) over both spatial (x) and temporal (t) dimensions. The results show that the PPT tracks the main structure of the evolving probability density, while the spatiotemporal error heatmap in Figure 7 identifies localized regions of prediction error. The dominant vertical stripe pattern indicates that prediction errors depend primarily on spatial location instead of accumulating or destabilizing over time.

Overall, the standard experiment establishes the probability-preserving behavior of the architecture, while the extended experiment demonstrates that the same constraint remains effective across unseen trajectories and multiple initial-condition families.

V. COMPUTATIONAL PERFORMANCE

The computational efficiency of the devised PPT is benchmarked against the classical CNM in the standard PPT experiment. This calibration is separate from the multi-initial-condition generalization evaluation. The CNM propels the wavefunction through successive temporal steps, whereas the trained PPT predicts a future state from a fixed five-state history in a single forward pass [9, 14, 16, 22]. Both methods use the same spatial resolution, $N_x = 100$ and $\Delta t = 0.001$.

For each prediction horizon, the sequential CNM runtime, batched PPT runtime, and single-state PPT inference time were measured. All measurements were averaged over $R = 1000$ independent runs:

$$T = \frac{1}{R} \sum_{i=1}^R T_i \quad (40)$$

The corresponding relative speedup is defined as

$$S = \frac{T_{\text{CN}}}{T_{\text{PPT}}} \quad (41)$$

where $S > 1$ indicates faster PPT inference than sequential CNM propagation for the specified prediction horizon.

Table IV. Comparative computational performance of the classical CNM and the proposed multi-token PPT.

Prediction Horizon (N)	Crank-Nicolson Sequential (s)	PPT Batched (s)	PPT Constant-Time (s)	Speedup
10	0.000673	0.000950	0.000630	1.07×
20	0.001528	0.001350	0.000609	2.51×
50	0.007220	0.002246	0.000666	10.84×
100	0.010220	0.003773	0.000658	15.54×
150	0.013823	0.005503	0.001038	13.32×
195	0.020106	0.007236	0.000899	22.35×

Both methods were benchmarked on the same CPU under identical computational conditions. The sequential CNM runtime generally increases with the prediction horizon because intermediate states must be computed sequentially, whereas single-state PPT inference remains approximately constant, ranging from 0.000609 to 0.001038 s across all tested horizons. Consequently, the PPT provides an increasingly substantial computa-

tional advantage at longer prediction horizons, reaching a maximum measured speedup of $22.35\times$ at $N = 195$.

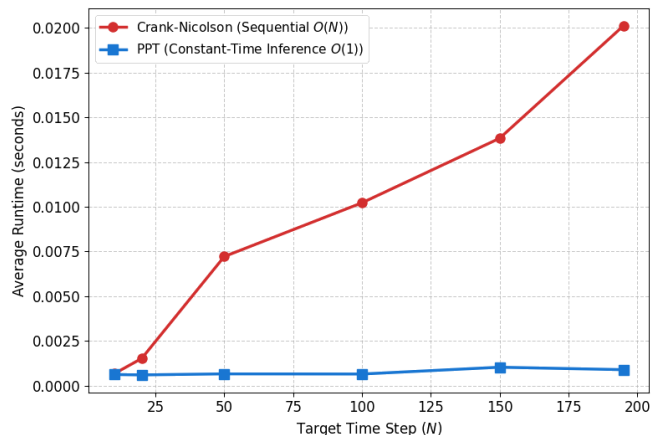


Figure 8. Average runtime comparison between sequential Crank–Nicolson and constant-time PPT inference across target time steps.

Figure 8 illustrates the empirical computational scaling of both approaches as a function of the target time step N . The classical Crank–Nicolson solver exhibits characteristic linear growth, scaling from 0.000673 s at $N = 10$ to 0.020106 s at $N = 195$. In contrast, the PPT model demonstrates approximately constant inference time over the tested horizons, maintaining similar evaluation overhead regardless of how far into the future the state is evaluated. These results highlight the efficacy of the PPT as a high-speed learned surrogate for long-horizon quantum-state evaluations, while CNM serves as the baseline numerical reference solver [9, 14–16].

VI. CONCLUSION

This paper deploys a Probability-Preserving transformer (PPT) for learning the temporal evolution of one-dimensional quantum wavefunctions. By enforcing probability normalization as a hard architectural constraint, the PPT reproduces Crank–Nicolson reference dynamics while strictly preserving total probability [30, 32]. Furthermore, the same architecture generalizes to unseen trajectories across ground states, first-excited states, Gaussian wavepackets, and superposition states. Although prediction accuracy varies across initial-condition families—particularly for certain Gaussian wavepackets—the probability constraint is strictly satisfied in all cases. The computational benchmarks demonstrate the potential of direct PPT inference to provide substantial speed advantages over sequential Crank–Nicolson propagation, particularly for longer prediction horizons. Future work will investigate extended temporal rollouts, derivative-aware Sobolev training, and extensions to higher-dimensional quantum systems [35, 36].

VII. DATA AVAILABILITY

Data will be made available on request. The source code used in this study is publicly available in [Repository](#)

VIII. FUNDING

The authors declare that no funds, grants, or other support were received.

-
- [1] D. J. Griffiths and D. F. Schroeter, *Introduction to Quantum Mechanics*, 3rd ed. (Cambridge University Press, 2018).
 - [2] J. J. Sakurai and J. Napolitano, *Modern Quantum Mechanics*, 3rd ed. (Cambridge University Press, 2020).
 - [3] R. Shankar, *Principles of Quantum Mechanics*, 2nd ed. (Springer Science & Business Media, 1994).
 - [4] N. A. Khan, S. Ye, Z. Zhou, S. Cheng, and G. Xianlong, Chebyshev polynomial approach to loschmidt echo: Application to quench dynamics in two-dimensional quasicrystals, *Physical Review E* **109**, 065311 (2024).
 - [5] N. A. Khan, W. Chen, M. Jan, and G. Xianlong, Linear-scale simulations of quench dynamics, *Computer Physics Communications* **299**, 109132 (2024).
 - [6] B. Fauseweh, Quantum many-body simulations on digital quantum computers: State-of-the-art and future challenges, *Nature Communications* **15**, 2123 (2024).
 - [7] A. Frisk Kockum, A. Miranowicz, S. De Liberato, S. Nori, and F. Nori, Ultrastrong coupling between light and matter, *Nature Reviews Physics* **1**, 19 (2019).
 - [8] C. Heide, P. D. Keathley, and M. F. Kling, Petahertz electronics, *Nature Reviews Physics* **6**, 648 (2024).
 - [9] J. Crank and P. Nicolson, A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type, *Mathematical Proceedings of the Cambridge Philosophical Society* **43**, 50 (1947).
 - [10] R. J. LeVeque, *Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems* (Society for Industrial and Applied Mathematics, 2007).
 - [11] S. A. Chin, Symplectic and unitary integrators for the time-dependent schrödinger equation, *Physical Review E* **109**, 035302 (2024).
 - [12] M. D. Feit, J. A. Fleck Jr, and A. Steiger, Solution of the schrödinger equation by a spectral method, *Journal of Computational Physics* **47**, 412 (1982).
 - [13] N. A. Khan and S. T. Amin, Probing band-center anomaly with the kernel polynomial method, *Physica Scripta* **96**, 045812 (2021).
 - [14] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar, Fourier neural operator for parametric partial differential equations, arXiv preprint arXiv:2010.08895 10.48550/arXiv.2010.08895 (2020).
 - [15] L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis, Learning nonlinear operators via deepnet based on the universal approximation theorem of operators, *Nature Machine Intelligence* **3**, 218 (2021).
 - [16] N. Kovachki, Z. Li, B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, and A. Anandkumar, Neural operator: Learning maps between function spaces with applications to pdes, *Journal of Machine Learning Research* **24**, 1 (2023).
 - [17] G. Carleo and M. Troyer, Solving the quantum many-body problem with artificial neural networks, *Science* **355**, 602 (2017).
 - [18] D. Pfau, J. S. Spencer, A. G. D. G. Matthews, and W. M. C. Foulkes, Ab initio solution of the many-electron schrödinger equation with deep neural networks, *Physical Review Research* **2**, 033429 (2020).
 - [19] J. Hermann, Z. Schätzle, and F. Noé, Deep-neural-network solution of the electronic schrödinger equation, *Nature Chemistry* **12**, 891 (2020).
 - [20] Y. Yao, C. Cao, S. Haas, M. Agarwal, D. Khanna, and M. Abram, Emulating quantum dynamics with neural networks via knowledge distillation, *Frontiers in Materials* **9**, 1060744 (2023).
 - [21] A. Jasinski, J. Montaner, R. C. Forrey, B. H. Yang, P. C. Stancil, N. Balakrishnan, J. Dai, R. A. Vargas-Hernández, and R. V. Krems, Machine learning corrected quantum dynamics calculations, *Physical Review Research* **2**, 032051 (2020).
 - [22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, *Advances in Neural Information Processing Systems* **30**, 5998 (2017).
 - [23] N. Geneva and N. Zabaras, Transformers for modeling physical systems and solving pdes, *Journal of Computational Physics* **460**, 111135 (2022).
 - [24] K. Sprague, S. Viterbi, and J. Carrasquilla, Transformer-based neural network quantum states for time evolution of quantum spin systems, *Physical Review B* **109**, 125105 (2024).
 - [25] P. Cha, C. Giles, and E.-A. Kim, Attention-based neural quantum state tomography and unitary dynamics reconstruction, *Physical Review Letters* **132**, 030401 (2024).
 - [26] H. Ma, H. Shang, and J. Yang, Quantum embedding method with transformer neural network quantum states for strongly correlated materials, *npj Computational Materials* **10**, 220 (2024).
 - [27] L. E. Herrera Rodríguez and A. A. Kananenka, A short trajectory is all you need: A transformer-based model for long-time dissipative quantum dynamics, *The Journal of Chemical Physics* **161**, 171101 (2024).
 - [28] A. Van de Walle, M. Schmitt, and A. Bohrdt, Many-body dynamics with explicitly time-dependent neural quantum states, *Machine Learning: Science and Technology* **6**, 045011 (2025).
 - [29] Z.-W. Wang, L.-A. Wu, and Z.-M. Wang, Time series prediction of open quantum system dynamics by transformer neural networks, *Entropy* **28**, 133 (2026).
 - [30] M. Raissi, P. Perdikaris, and G. E. 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 (2019).
 - [31] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, Physics-informed machine learning, *Nature Reviews Physics* **3**, 422 (2021).

- [32] S. Dong and N. Ni, Exact enforcement of boundary conditions and conservation laws in physics-informed neural networks, *Journal of Computational Physics* **434**, 110234 (2021).
- [33] N. A. Khan, Fidelity susceptibility probes of dynamical quantum criticality, *Chaos, Solitons & Fractals* **183**, 114975 (2024).
- [34] K. Shah, P. Stiller, N. Hoffmann, and A. Cangi, Physics-informed neural networks as solvers for the time-dependent schrödinger equation, *arXiv preprint arXiv:2210.12522* (2022).
- [35] W. M. Czarnecki, S. Osindero, M. Jaderberg, G. Swirszcz, and R. Pascanu, Sobolev training for neural networks, *Advances in Neural Information Processing Systems* **30**, 4278 (2017).
- [36] H. Son, S.-W. Cho, and H. J. Hwang, Sobolev training for physics-informed neural networks with spatial and temporal gradient constraints, *Computer Methods in Applied Mechanics and Engineering* **418**, 116521 (2024).