ATTENTION-ENHANCED NEURAL OPERATOR FOR VARIABLE-TIMESTEP PREDICTION OF PDES

Oluwaseun E. Coker
Department of Computer Science
University of Leeds
Leeds, LS2 9JT.
scoc@leeds.ac.uk

He Wang
AI Centre, Department of Computer Science
University College London
London, WC1E 6BT.
he_wang@ucl.ac.uk

Amirul Khan
School of Civil Engineering,
University of Leeds
Leeds, LS2 9JT.
A.Khan@leeds.ac.uk

Peter K. Jimack
Department of Computer Science
University of Leeds
Leeds, LS2 9JT.
P.K.Jimack@leeds.ac.uk

ABSTRACT

Current neural-based solvers for partial differential equations (PDEs), such as autoregressive neural operators, often rely on fixed-time-stepping schemes that can struggle with multiscale dynamics and long-term error accumulation. To address these limitations, this paper introduces an attention-enhanced variable-timestep prediction framework based on the Fourier Neural Operator (FNO). By leveraging a cross-attention mechanism and temporal embeddings, our approach enables variable-timestep predictions by modulating feature maps based on the requested temporal interval. We evaluate various embedding strategies across four benchmark PDEs - Advection, Viscous Burgers', Diffusion-Reaction, and Diffusion-Sorption - demonstrating that variable-timestep sampling can effectively manage the trade-off between computational efficiency and long-term stability. Our results show that this attention-based conditioning consistently outperforms traditional additive or concatenation methods, providing a more robust solution for complex, time-dependent physical systems.

1 INTRODUCTION

The numerical solution of partial differential equations (PDEs) is a cornerstone of scientific and engineering efforts, and obtaining accurate, efficient solutions is crucial for safety-critical applications (Zhang et al., 2023). While current deep learning-based PDE solvers, specifically autoregressive neural operators, have shown great promise, their reliance on fixed-time-stepping schemes presents significant limitations. These models can be inefficient when dealing with multiscale dynamics or time-dependent solutions that tend to a steady-state, where different phases of the physical evolution would benefit from varying temporal resolutions.

An important challenge in autoregressive modelling is the accumulation of error. The primary cause of divergence in these models is the propagation of error from earlier to later timesteps across many rollouts. Fixed or constant-timestep methods are inherently constrained by the temporal resolution of their training data; if a model is trained with a constant timestep Δt, it lacks the flexibility to take larger steps during slow dynamics to reduce the number of rollouts or smaller steps during fast dynamics to maintain stability. Consequently, these models often struggle with long-term temporal evolution and time extrapolation (McCabe et al., 2023).

To address these limitations, this work proposes a variable-timestep prediction framework. We argue that variable-timestep strategies, which dynamically combine smaller and larger steps, can mitigate error accumulation by reducing the total number of rollouts required for a given prediction. While continuous-time models, such as Neural Ordinary Differential Equations (Neural ODEs), offer one

solution by modelling the derivative of the state, they are often computationally expensive due to the requirements of the underlying ODE solvers (Chen et al., 2018).

Instead, we enhance the Fourier Neural operator (Li et al., 2020) by using an **Attention-based Temporal Conditioning** mechanism. Unlike constant-timestep sampling, which often constrains a model to a specific temporal discretisation, our approach leverages temporal embeddings and a cross-attention mechanism to modulate the model's feature maps based on the requested timestep Δt. This allows the model to capture temporal dependencies, enabling accurate predictions across a continuous range of timesteps.

The contributions of this work are as follows:

- We introduce an attention-based conditioning architecture that enables variable-timestep sampling using Fourier Neural Operators.
- We evaluate various temporal embedding strategies, including Sinusoidal Temporal Embeddings (STE), Embedding Layer (EL) and Multi-Layer Perceptron (MLP).
- We demonstrate the effectiveness of this approach across four benchmark PDEs, showing that variable-timestep sampling can effectively manage the trade-off between efficiency and long-term stability.

1.1 Problem Statement

We consider the task of learning an explicit time-stepping operator for a general class of time-dependent PDEs over the domain $(t, x) \in [0, t_s] \times D$. The system, subject to Dirichlet boundary conditions B, is defined as:

$$\begin{aligned} \partial_t u(t,x) &= \mathcal{F}^\dagger\left(u(t,x)\right), \qquad (t,x) \in (0,t_s] \times D, \\ u(0,x) &= u_0(x), \qquad x \in D, \\ u(t,x) &= B, \qquad (t,x) \in (0,t_s] \times \partial D, \end{aligned} \tag{1}$$

where $u : [0, t_s] \times D \to \mathbb{R}^d$ is the solution state, $u_0(x)$ is the initial condition at $t = 0$, and $\mathcal{F}^\dagger$ denotes a spatial differential operator.

Our objective is to learn a nonlinear parametric operator $\mathcal{F}_\theta$ that maps an input function $u(t_1)$ to a predicted solution $\tilde{u}(t_2)$ for any $t_2 > t_1$. We formulate this autoregressive model as:

$$\tilde{u}(t_2) = \mathcal{F}_\theta\left(\mathbf{C}\left(u(t_1), \mathbf{E}_\phi(t_1), \mathbf{E}_\phi(t_2)\right)\right), \qquad \theta \in \mathbb{R}^p, \tag{2}$$

where t_1 and t_2 represent the temporal coordinates of the input and output states, respectively. $\mathbf{E}_\phi$ denotes a temporal embedding, which may be a fixed mapping or a learned representation with parameters ϕ, and $\mathbf{C}$ represents a conditioning mechanism that integrates the state u with the temporal information. The optimisation problem seeks to find the optimal parameters θ^* that minimise the discrepancy between the predicted and ground-truth states.

Training is performed on a finite collection of PDE trajectories originating from diverse initial conditions. While these trajectories are provided as sequences with even spacing(i.e., constant-timesteps), we employ a **variable-timestep sampling** strategy during training to simulate unevenly spaced sequences (Appendix C). This ensures the model learns to generalise across varying timesteps. The loss is minimised using empirical risk minimisation over a training set, while the generalisation error is evaluated on a disjoint test set of unseen initial conditions. At inference, the trained operator $\mathcal{F}_{\theta^*}$ is applied recursively to generate long-term temporal evolutions.

2 Temporal Integration Preliminaries

The efficacy of variable-timestep neural operators depends on how temporal information is integrated into the architecture. This integration occurs in two stages: **Time Embedding**, which transforms temporal scalars into latent vectors, and **Time Conditioning**, which incorporates these vectors into the model's hidden states.

2.1 TEMPORAL EMBEDDING STRATEGIES

Embeddings map raw timestamps into high-dimensional representations E_ϕ to provide models with sequence awareness. We categorise these into three primary types:

- **Sinusoidal Temporal Embedding (STE):** A non-learnt, deterministic approach using predefined waves of varying frequencies to encode periodic temporal structures, aiding generalisation to unseen timesteps.
- **Embedding Layer (EL):** A learnt, discrete strategy that uses one-hot encoding to optimise representations for specific training intervals.
- **Multi-Layer Perceptron (MLP):** A learnt, continuous mapping capable of evaluating any timestamp within a continuous domain, though it may be prone to overfitting the training range.

2.2 CONDITIONING MECHANISMS

Conditioning refers to the architectural fusion of temporal embeddings with spatial feature maps $u(t)$. Standard Fourier Neural Operator (FNO) baselines typically employ two methods:

- **Concatenation (Co):** Appends the embedding as an additional feature dimension. While flexible enough to learn non-linear relationships, it increases input dimensionality and computational cost.
- **Addition (Ad):** Sums the embedding directly with spatial features, acting as a residual correction. This aligns with the iterative nature of traditional PDE solvers, in which the next state is a refinement of the current state.

3 RELATED WORK

3.1 VARIABLE-TIMESTEP MODELLING

In contrast to traditional numerical methods for time-dependent problems, where adaptive time-stepping is a mature and widely used technique (Söderlind, 2002), effective variable-timestep prediction for PDEs remains relatively under-explored in the context of neural operators. Much of the existing literature focuses on constant-timestep prediction, with advancements primarily targeting improved learning strategies (Lippe et al., 2023; Takamoto et al., 2023) and architectural refinements (McCabe et al., 2023; Rahman et al., 2023). For variable-timestep prediction, current approaches typically utilise embedding modules to transform temporal information into a latent representation. These embeddings are then integrated with the input function through either concatenation (Dang et al., 2022) or additive operations (Gupta & Brandstetter, 2022).

3.2 NEURAL OPERATORS

Neural operators, most notably DeepONet (Di Leoni et al., 2021) and the Fourier Neural Operator (FNO) (Li et al., 2020), have emerged as a dominant class of models for learning PDE solution operators. Numerous studies have sought to enhance these foundational architectures; for instance, Diab & Al Kobaisi (2025) proposed a temporal-branch modification to DeepONet to enable variable-timestep prediction. The present work focuses on the FNO due to its efficacy in capturing global spectral information and its iterative architecture, which employs learnt integral kernels to approximate solution operators.

3.3 ATTENTION MECHANISMS

The attention mechanism, famously introduced within the Transformer architecture (Vaswani et al., 2017), has proven highly effective at capturing long-range dependencies. Consequently, several works have adapted attention-based layers to PDEs (Peng et al., 2022; Takamoto et al., 2022; Kissas et al., 2022; Dang et al., 2022; Li et al., 2022). These applications generally categorise attention into spatial or temporal contexts. For example, Li et al. (2022) utilised attention to encode spatial

relationships between input and query points, while Han et al. (2022) employed a graph neural network with an attention mechanism to capture temporal dependencies. In this work, we leverage the attention mechanism specifically to condition the model on temporal dependencies through our framework.

4 ATTENTION-BASED TIME CONDITIONING

We present our proposed attention-based temporal conditioning framework within the standard FNO architecture. This approach enables flexible variable-timestep prediction and sampling, as illustrated in Figure 1. The modified FNO architecture incorporates an attention module alongside the two fundamental components of the standard FNO: spectral convolutions and spatial convolutions. Each layer of the proposed model is defined as:

$$U_{t+1} := \sigma\left(\text{Attention}(U_t, T_t, T_{t+1}) + \text{SpectralConv}(U_t) + \text{SpatialConv}(U_t)\right), \tag{3}$$

where U_t and U_{t+1} denote the hidden function embeddings at the current and subsequent states, respectively. T_t and T_{t+1} represent the corresponding temporal embeddings, and σ denotes a non-linear activation function.

The central philosophy of this design is the parallel modelling of spatial and temporal dependencies. Specifically, the spectral convolution efficiently captures spatial patterns in the frequency domain, while the attention mechanism is leveraged to model temporal dependencies and modulate the state transition based on the requested timestep (Peng et al., 2022; Kissas et al., 2022). The spatial convolution, implemented as a point-wise (1×1) kernel, serves as a learnable bias or skip connection. A key refinement in our architecture is the sharing of temporal embeddings across all iterative layers, a design choice that has been shown to enhance training stability and predictive accuracy.

4.1 INPUT LIFTING AND EMBEDDING

The spatial and spectral modules operate on the lifted state embedding $U_t(x)$, whereas the attention block integrates the lifted state $U_t(x)$ with the temporal embeddings T_t and T_{t+1}. These embeddings are defined as:

$$\begin{aligned} U_t &= L_{\phi,u}(u_t), \\ T_t &= L_{\phi,t}(\hat{t}), \\ T_{t+1} &= L_{\phi,t+1}(\widehat{t+1}), \end{aligned} \tag{4}$$

where $\hat{t}$ and $\widehat{t+1}$ are generated by the initial embedding layers $E_{\phi,t}$ and $E_{\phi,t+1}$. These primary embeddings transform the raw timestamps into continuous or discrete vector spaces using either an MLP (learnt, continuous), an EL (learnt, discrete), or an STE (non-learnt, discrete). The lifting maps $L_{\phi,u}$, $L_{\phi,t}$, and $L_{\phi,t+1}$ are linear transformations that project these representations into a higher-dimensional space d.

4.2 THE ATTENTION MECHANISM

The attention module maps a query (derived from the target temporal state T_{t+1}) and a set of key–value pairs (derived from the current state U_t and T_t) to a modulated output. The output is computed as a weighted sum of the values, with weights determined by a compatibility function between the query and the corresponding keys. Our architecture formulates the Query (Q), Key (K), and Value (V) matrices as:

$$\begin{aligned} Q &= W_q(T_{t+1} + U_t), \\ K &= W_k(T_t + U_t), \\ V &= W_v(U_t), \\ \text{Attention}(Q, K, V) &= \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \end{aligned} \tag{5}$$

where d denotes the model width. Empirically, we observed that incorporating the state embedding U_t into both the query and key vectors significantly improved convergence during training.

4.3 Spectral Convolution and Output Projection

The spectral convolution transforms the input embedding into the frequency domain via the Fast Fourier Transform (FFT) to apply a complex weight tensor W_r:

$$\text{SpectralConv}(U_t) = \mathcal{F}^{-1}\left(W_r \cdot (\mathcal{F}U_t)\right)(x), \tag{6}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the Fourier and Inverse Fourier transforms, respectively. Finally, the updated hidden representation U_{t+1} is projected back to the physical space to yield the prediction u_{t+1}:

$$u_{t+1} = L_{\phi,u_{t+1}}(U_{t+1}). \tag{7}$$

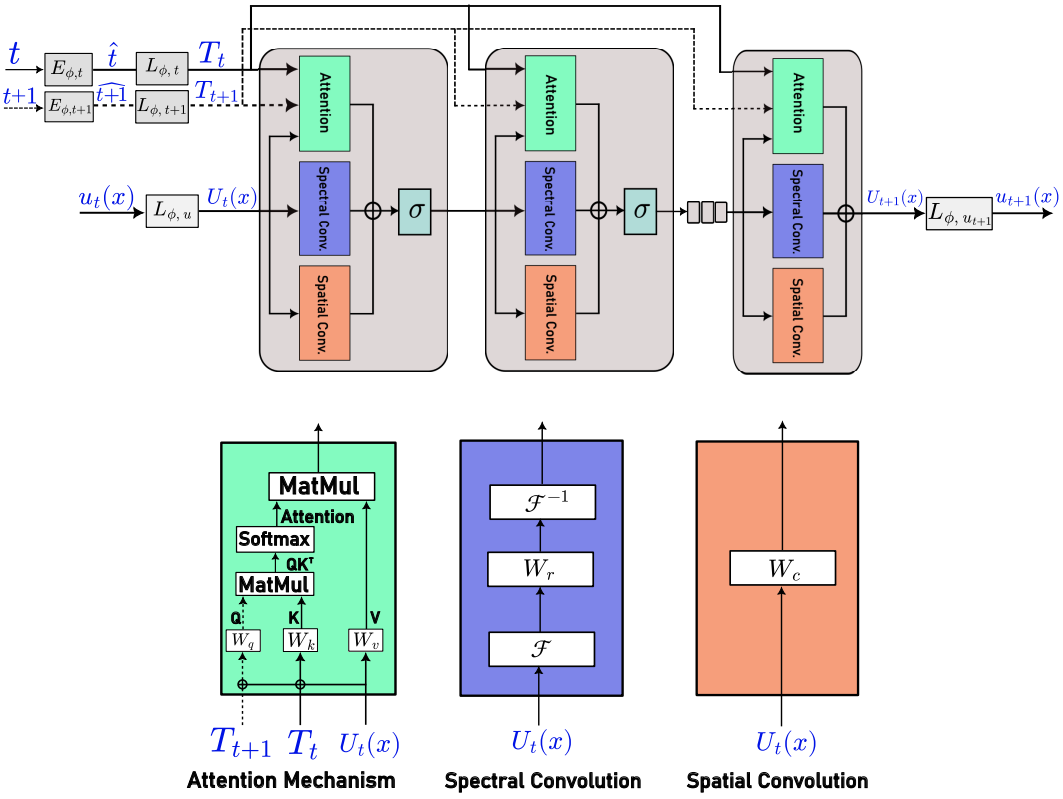

Figure 1: Schematic of the proposed Attention-based Fourier Neural Operator for variable-timestep prediction. The architecture processes inputs through three parallel streams: (i) a spectral convolution for frequency-domain patterns, (ii) a spatial convolution (1×1) acting as a learnable bias, and (iii) an attention mechanism that uses lifted temporal embeddings (T_t, T_{t+1}) to modulate the function embedding U_t. The outputs are summed and passed through a non-linear activation σ to produce the next hidden state U_{t+1}, which is projected back to the physical space u_{t+1}.

5 Experiments

This section details the experimental setup used to evaluate the temporal conditioning and embedding strategies. We categorise these strategies based on their conditioning mechanism, **Addition (Ad)**, **Concatenation (Co)**, and our proposed **Attention (At)**, and their respective embedding types: **Sinusoidal Temporal Embedding (STE)**, **Embedding Layer (EL)**, and **Multi-Layer Perceptron (MLP)**. Table 1 summarises the characteristics of each strategy, encompassing all architectural combinations except for non-learnt continuous embeddings.

The performance of these strategies is benchmarked across four fundamental PDEs:

- **Advection (Av):** A baseline for simple translation.
- **Viscous Burgers' (vB):** A non-linear system where advection and diffusion are balanced, introducing shock formation.
- **Diffusion-Reaction (DR):** Describes substance diffusion combined with chemical interactions, often leading to non-linear error growth.
- **Diffusion-Sorption (DS):** Models molecular permeation and sorption into material structures, characterised by high initial-phase complexity.

Detailed mathematical formulations for each system are provided in Appendix A. These benchmarks were selected to assess the models' capacity to handle linear translation, nonlinear evolution, and complex source-driven dynamics. The training details are provided in Appendix B. Detailed results for each PDE are provided in Appendix G, H, I and J.

Table 1: **Taxonomy of the nine temporal conditioning and embedding strategies.** Checkmarks (✓) indicate the architectural combinations of conditioning mechanisms and embedding properties evaluated across the benchmark PDEs.

Strategy	Conditioning			Embedding			
	Addition	Concatenation	Attention	Non-learnt	Learnt	Discrete	Continuous
Ad	✓			-	-	-	-
Co		✓		-	-	-	-
At (**ours**)			✓	-	-	-	-
STE-Ad	✓			✓		✓	
STE-Co		✓		✓		✓	
STE-At (**ours**)			✓	✓		✓	
EL-Ad	✓				✓	✓	
EL-Co		✓			✓	✓	
EL-At (**ours**)			✓		✓	✓	
MLP-Ad	✓				✓		✓
MLP-Co		✓			✓		✓
MLP-At (**ours**)			✓		✓		✓

6 Results and Discussion

We present a comprehensive evaluation of the proposed variable-timestep framework across four benchmark partial differential equations (PDEs): linear Advection, Viscous Burgers', Diffusion-Reaction, and Diffusion-Sorption. These benchmarks represent a diverse range of physical phenomena, including pure propagation, shock formation, and non-linear source interactions.

Our analysis is structured to validate three core hypotheses: (i) that attention-based conditioning is superior to additive or concatenative methods for temporal modulation; (ii) that the optimal temporal sampling strategy is intrinsically linked to the underlying physics of the system; and (iii) that a single variable-timestep model can maintain high physical fidelity while offering significant gains in parameter efficiency compared to specialised constant-timestep solvers.

We benchmark these models by comparing the conditioning and embedding options across different sampling strategies (i.e., constant-timestep and variable-timestep sampling) and integrate the normalised root mean squared error (nRMSE) across the full rollout. Constant-timestep sampling is investigated at intervals of $1\Delta t$, $2\Delta t$, and $3\Delta t$, where Δt represents the smallest temporal resolution of the dataset. For variable-timestep sampling, we examine transitions from small-to-large timesteps ($1\Delta t \rightarrow 2\Delta t$ and $1\Delta t \rightarrow 2\Delta t \rightarrow 3\Delta t$) and large-to-small timesteps ($2\Delta t \rightarrow 1\Delta t$ and $3\Delta t \rightarrow 2\Delta t \rightarrow 1\Delta t$). To ensure a balanced comparison, temporal sequences are split evenly between the selected timesteps. For instance, in a simulation of $200\,t_s$ using a $1\Delta t \rightarrow 2\Delta t$ strategy, $1\Delta t$ is employed for the initial 100 steps, followed by $2\Delta t$ for the remainder of the rollout.

6.1 Performance of Conditioning Mechanisms

Our experiments across four benchmark PDEs demonstrate that the proposed **Attention-based (At)** conditioning consistently outperforms standard Concatenation (Co) and Addition (Ad) base-

lines. This superiority is clearly visible in Table 2, where the Attention conditioning (indicated in blue) achieves the lowest overall error in three out of four benchmarks: Viscous Burgers' (1.08%), Diffusion-Reaction (4.4%), and Diffusion-Sorption (0.12%).

The mechanism's strength lies in its ability to dynamically modulate hidden feature maps based on a specific temporal query (Δt). Unlike Addition, which acts as a static residual, or Concatenation, which simply increases dimensionality, the Attention mechanism selectively scales spatial features. This is particularly effective in non-linear systems like **Viscous Burgers'**, where the model must adapt to rapid changes in gradients. Interestingly, for the **Advection** equation, Concatenation remains highly competitive (2.97%), suggesting that for simple linear translations, the added complexity of attention may be less critical than for dissipative or reactive systems.

Table 2: Time Conditioning: Mean normalised root mean squared error (nRMSE) averaged over the full temporal rollout: For each PDE, the smallest value per row and column is highlighted in **bold** and underline, respectively. The overall smallest error is highlighted in blue.

TIME CONDITIONING STRATEGY		VARIABLE-TIMESTEP MODEL						
		CONSTANT SAMPLING			VARIABLE SAMPLING			
		1 (201 t_s)	2 (101 t_s)	3 (67 t_s)	1-2 (150 t_s)	2-1 (150 t_s)	1-2-3 (125 t_s)	3-2-1 (125 t_s)
Av	Ad	147.13	44.62	58.06	83.41	36.93	**24.44**	51.7
	Co	3.78	3.41	4.36	3.30	3.79	**2.97**	4.57
	At (**ours**)	7.14	**5.83**	7.86	6.86	6.17	7.11	6.71
vB	Ad	6.01	**3.15**	4.45	5.85	3.16	5.40	3.52
	Co	3.09	1.30	1.68	3.12	**1.29**	3.16	1.53
	At (**ours**)	2.50	1.21	**1.08**	2.35	1.26	2.25	1.09
DR	Ad	13.2	12.4	12.0	12.2	**11.2**	11.8	13.3
	Co	20.3	14.2	**11.3**	18.8	14.5	18.3	12.3
	At (**ours**)	10.5	5.4	**4.4**	10.1	5.7	9.8	4.7
DS	Ad	1.14	**1.00**	1.20	1.12	1.03	1.10	1.22
	Co	0.17	**0.15**	0.16	0.16	0.16	0.16	0.16
	At (**ours**)	0.13	0.12	0.12	0.13	0.12	0.12	0.12

6.2 Performance of Embedding Mechanisms

The choice of temporal embedding, i.e., how the scalar Δt is represented before conditioning, further improves model performance. Table 3 compares our Attention mechanism across three embedding types: Sinusoidal (STE), Embedding Layer (EL), and Multi-Layer Perceptron (MLP).

The **Embedding Layer (EL)** emerges as the most robust choice for the majority of benchmarks. In the **Advection** and **Diffusion-Reaction** cases, EL-At significantly outperforms the non-learnt STE-At, providing the lowest errors (1.03% and 3.4%, respectively). This indicates that learnt, discrete representations allow the model to better specialise for the specific timesteps encountered during training.

However, for the **Viscous Burgers'** equation, the **MLP** embedding yielded superior results (0.84%). This suggests that for systems with high-frequency shocks and steep gradients, a continuous, non-linear mapping provided by an MLP is better suited to capture rapid physical transitions than discrete embedding layers. For **Diffusion-Sorption**, EL and MLP perform almost identically, both successfully capturing the complex initial sorption dynamics that prove difficult for the fixed sinusoidal (STE) approach

6.3 Strategy Directionality: Temporal Refinement vs. Coarsening

The results in Tables 2 and 3 highlight a vital physical insight: the optimal sampling strategy is intrinsically linked to the system's underlying physics. As summarised in Table 4, we identify two primary patterns:

- **Temporal Refinement (Small-to-Large):** Strategies such as $\Delta t = 1 \rightarrow 2 \rightarrow 3$ are optimal for **Advection** and **Diffusion-Reaction**. In these systems, errors typically accumulate over time due to pure propagation or non-linear source terms. Establishing a stable trajectory with fine resolution early on allows the model to avoid high error in later stages by transitioning to fewer, larger steps, thereby limiting the total number of rollouts.

Table 3: Attention-based Conditioning with different Embeddings: Mean normalised root mean squared error (nRMSE) averaged over the full temporal rollout: For each PDE, the smallest value per row and column is highlighted in **bold** and underline, respectively. The smallest error is highlighted in blue. The EL-At model using a 1-2-3 variable-timestep sampling strategy achieves the lowest overall error.

TIME CONDITIONING AND EMBEDDING STRATEGY		VARIABLE-TIMESTEP MODEL						
		CONSTANT SAMPLING			VARIABLE SAMPLING			
		1 (201 t_s)	2 (101 t_s)	3 (67 t_s)	1-2 (150 t_s)	2-1 (150 t_s)	1-2-3 (125 t_s)	3-2-1 (125 t_s)
Av	STE-At	6.00	9.26	4.65	5.06	8.00	4.62	**4.44**
	EL-At	1.20	1.14	1.93	1.07	1.08	**1.03**	1.29
	MLP-At	2.12	3.87	3.81	1.87	3.90	**1.73**	3.59
vB	STE-At	2.24	1.97	**1.24**	2.27	1.80	2.34	1.33
	EL-At	1.6	1.0	1.0	1.5	1.1	1.5	1.0
	MLP-At	1.8	0.93	**0.84**	1.7	0.96	1.6	**0.84**
DR	STE-At	13.9	9.1	**7.9**	12.0	10.7	10.8	9.3
	EL-At	9.4	4.8	**3.4**	7.8	6.4	6.8	5.1
	MLP-At	11.4	8.0	**7.1**	10.2	9.0	9.4	8.3
DS	STE-At	0.20	0.19	**0.16**	0.22	0.22	0.19	0.18
	EL-At	0.14	0.12	**0.11**	0.13	0.12	0.13	0.12
	MLP-At	0.13	0.12	0.12	0.13	0.12	0.13	0.12

- **Temporal Coarsening (Large-to-Small):** Strategies such as $\Delta t = 3 \rightarrow 2 \rightarrow 1$ are superior for **Viscous Burgers'** and **Diffusion-Sorption**. These systems exhibit high-error initial phases due to shock formation or steep initial gradients. Using larger initial steps allows the model to leap over these unstable regions, effectively mitigating cumulative exposure bias by reducing the total number of autoregressive rollouts (e.g., from 201 steps at $1\Delta t$ to 67 steps at $3\Delta t$).

Across all benchmarks, the variable-timestep framework achieves a significant **Efficiency Advantage**. By adapting the temporal step size to the physical dynamics, these models achieve nRMSE scores comparable to or lower than specialised constant-timestep models while requiring fewer total timesteps (t_s), representing a substantial gain in computational efficiency for long-term PDE evolution.

Table 4: Summary of Optimal Configurations and Temporal Sampling Strategies for Benchmark PDEs based on Total Integrated nRMSE.

Dataset	Best Model	Best Variable Strategy	Physical Justification
Advection	EL-At	$1 \rightarrow 2 \rightarrow 3$	Establishing a stable trajectory allows for larger steps later with minimal error accumulation.
Viscous Burgers'	MLP-At	$3 \rightarrow 2 \rightarrow 1$	Fewer rollouts during high-error shock-formation reduces cumulative exposure bias.
Diff.-Reaction	EL-At	$1 \rightarrow 2 \rightarrow 3$	Transitioning to larger steps bypasses high-error stages as reaction complexities accumulate.
Diff.-Sorption	EL-At	$3 \rightarrow 2 \rightarrow 1$	Larger initial steps reduce the impact of high-error initial sorption dynamics.

6.4 Variable-timestep vs. Constant-timestep Model Performance

We evaluate the generalisation penalty by comparing a single **Variable-timestep (1x)** model against an ensemble of five **Constant-timestep (5x)** models, each specialised for a fixed $\Delta t \in \{1, \ldots, 5\}$. As shown in Table 5, two distinct performance profiles emerge based on the system physics:

- **The Generalisation Penalty:** In **Advection**, **Burgers'**, and **Diffusion-Sorption**, specialised models generally define the accuracy upper bound. Specialisation allows the model to over-fit the spectral transformations required for a specific interval. However, the variable-timestep model maintains exceptionally high Pearson correlation ($R \geq 0.99$),

proving that it captures the correct physical topology and wave phases despite minor numerical offsets.

- **The Multi-task Learning Advantage:** In **Diffusion-Reaction**, the variable-timestep model consistently *outperforms* all specialised models. At $3\Delta t$, its error (0.0192) is nearly 50% lower than the specialised baseline (0.0386). This suggests that training on diverse intervals serves as a form of data augmentation; the model learns a more robust representation of nonlinear reaction terms by observing their evolution across multiple temporal scales.

Stability and Efficiency: Constant-timestep models typically show improved accuracy at larger Δt due to reduced autoregressive error propagation (fewer rollouts). While the variable-timestep model is more complex when mapping large jumps in a single step, it offers **5x better parameter efficiency**. By replacing multiple specialised solvers with a single, physically consistent architecture, the variable-timestep framework provides a robust and computationally efficient alternative for long-term PDE evolution.

Table 5: Accuracy and Correlation: Comparison between five constant-timestep models and one variable-timestep model. The five constant-timestep models correspond to models trained and tested at $\Delta t = 1$, 2, 3, 4, and 5, respectively. A single variable-timestep model is evaluated at the same timesteps for comparison. Pearson's Correlation for $\Delta t = 1$ and 5 are included in brackets.

DATASET	MODEL	CONSTANT SAMPLING				
		$1\Delta t$	$2\Delta t$	$3\Delta t$	$4\Delta t$	$5\Delta t$
Advection (Av)	Constant-timestep (5x)	**0.0034** (1.0)	**0.0028**	**0.0027**	**0.0027**	**0.0025** (1.0)
	Variable-timestep (1x)	0.0062 (0.99)	0.0061	0.010	0.0337	0.0499 (1.0)
Vis. Burgers' (vB)	Constant-timestep (5x)	**0.0028** (1.0)	**0.0025**	**0.0021**	**0.0013**	**0.0013** (1.0)
	Variable-timestep (1x)	0.0099 (1.0)	0.0051	0.0048	0.0062	0.0085 (1.0)
Diff-React (DR)	Constant-timestep (5x)	0.0515 (0.99)	0.0399	0.0386	0.0349	0.0361 (0.99)
	Variable-timestep (1x)	**0.0489** (0.99)	**0.0258**	**0.0192**	**0.0203**	**0.0316** (0.99)
Diff-Sorpt (DS)	Constant-timestep (5x)	**0.0011** (1.0)	**0.0010**	**0.0010**	**0.0010**	**0.0010** (1.0)
	Variable-timestep (1x)	0.0016 (1.0)	0.0014	0.0014	0.0015	0.0018 (1.0)

7 CONCLUSION

In conclusion, this paper presents a novel attention-enhanced variable-timestep prediction framework for neural operators, specifically designed to overcome the limitations of fixed-timestep schemes in solving partial differential equations (PDEs). The key contributions and findings are as follows:

- **Architectural Innovation:** By integrating a cross-attention mechanism with temporal embeddings into the Fourier Neural Operator (FNO), the model effectively modulates its internal feature maps according to the requested timestep, allowing for accurate predictions across a continuous range of temporal resolutions.
- **Performance and Stability:** The proposed attention-based conditioning (At) consistently outperforms standard concatenation and additive methods across four benchmark PDEs, including Advection and Diffusion-Sorption.
- **Efficiency:** The framework demonstrates a significant efficiency advantage; by utilising variable-timestep sampling strategies, the model reduces the total number of autoregressive rollouts needed, thereby mitigating error accumulation and lowering computational costs for long-term simulations.
- **Robustness:** The results indicate that the Embedding Layer (EL) combined with attention (EL-At) is particularly effective, achieving the lowest overall error by successfully managing the trade-offs between temporal efficiency and long-term stability.

While the attention-enhanced framework improves predictive accuracy, it possesses some limitations. Notably, this work has not explored methods to automatically select optimal timesteps; instead, sampling strategies were manually predefined rather than adaptively determined by the

model based on the evolving physical dynamics. Additionally, the effectiveness of these strategies is problem-dependent, as the cumulative error from increased rollout steps can sometimes offset the benefits of variable-time flexibility.

From a computational standpoint, the cross-attention mechanism introduces a trade-off, requiring approximately 30% more execution time and 8% more memory than simpler concatenation methods. Furthermore, the model's ability to extrapolate over very long rollouts remains constrained, particularly in systems with high-frequency shifts that are difficult for spectral layers to capture over extended gaps. Future work should focus on automating timestep selection and optimising the computational efficiency of the attention layers.

REFERENCES

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Yuchen Dang, Zheyuan Hu, Miles Cranmer, Michael Eickenberg, and Shirley Ho. Tnt: Vision transformer for turbulence simulations. *arXiv preprint arXiv:2207.04616*, 2022.

P Clark Di Leoni, Lu Lu, Charles Meneveau, George Karniadakis, and Tamer A Zaki. Deeponet prediction of linear instability waves in high-speed boundary layers. *arXiv preprint arXiv:2105.08697*, 2021.

Waleed Diab and Mohammed Al Kobaisi. Temporal neural operator for modeling time-dependent physical phenomena. *Scientific Reports*, 15(1):32791, 2025.

Robert Eymard, Thierry Gallouët, and Raphaele Herbin. Finite volume methods: handbook of numerical analysis. *PG Ciarlet and JL Lions (Eds)*, 2000.

Jayesh K Gupta and Johannes Brandstetter. Towards multi-spatiotemporal-scale generalized pde modeling. *arXiv preprint arXiv:2209.15616*, 2022.

Xu Han, Han Gao, Tobias Pfaff, Jian-Xun Wang, and Li-Ping Liu. Predicting physics in mesh-reduced space with temporal attention. *arXiv preprint arXiv:2201.09113*, 2022.

Georgios Kissas, Jacob H Seidman, Leonardo Ferreira Guilhoto, Victor M Preciado, George J Pappas, and Paris Perdikaris. Learning operators with coupled attention. *Journal of Machine Learning Research*, 23(215):1–63, 2022.

Zijie Li, Kazem Meidani, and Amir Barati Farimani. Transformer for partial differential equations' operator learning. *arXiv preprint arXiv:2205.13671*, 2022.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

Phillip Lippe, Bastiaan S Veeling, Paris Perdikaris, Richard E Turner, and Johannes Brandstetter. Pde-refiner: Achieving accurate long rollouts with neural pde solvers. *arXiv preprint arXiv:2308.05732*, 2023.

Michael McCabe, Peter Harrington, Shashank Subramanian, and Jed Brown. Towards stability of autoregressive neural operators. *arXiv preprint arXiv:2306.10619*, 2023.

Wenhui Peng, Zelong Yuan, and Jianchun Wang. Attention-enhanced neural network models for turbulence simulation. *Physics of Fluids*, 34(2), 2022.

Md Ashiqur Rahman, Zachary E Ross, and Kamyar Azizzadenesheli. U-NO: U-shaped neural operators. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=j3oQF9coJd.

Gustaf Söderlind. Automatic control and adaptive time-stepping. *Numerical Algorithms*, 31(1): 281–310, 2002.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. Pdebench: An extensive benchmark for scientific machine learning. *Advances in Neural Information Processing Systems*, 35:1596–1611, 2022.

Makoto Takamoto, Francesco Alesiani, and Mathias Niepert. Learning neural pde solvers with parameter-guided channel attention. In *International Conference on Machine Learning*, pp. 33448–33467. PMLR, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Xuan Zhang, Limei Wang, Jacob Helwig, Youzhi Luo, Cong Fu, Yaochen Xie, Meng Liu, Yuchao Lin, Zhao Xu, Keqiang Yan, et al. Artificial intelligence for science in quantum, atomistic, and continuum systems. *arXiv preprint arXiv:2307.08423*, 2023.

A Partial Differential Equation

We define the partial differential equations used in this work and how the datasets for training, evaluating, and testing our models were generated. We use benchmark datasets by Takamoto et al. (2022) for Advection (Av), viscous Burgers' (vB), Diffusion-Reaction (DR) and Diffusion-Sorption (DS).

A.0.1 Advection Equation (Av)

The advection equation models pure advection behaviour without non-linearity.

$$\begin{aligned} \partial_t u(t,x) + \beta \partial_x u(t,x) = 0, \quad x \in (0,1), t \in (0,2], \\ u(0,x) = u_0(x), \quad x \in (0,1), \end{aligned} \tag{8}$$

where $\beta = 0.5$ is the advection speed. Note that the exact solution of the system is given as: $u(t,x) = u_0(x - \beta t)$. The boundary conditions are periodic, and the initial conditions are generated by a superposition of sinusoidal waves as:

$$u_0(x) = \sum_{k_i = k_1, \cdots, K_N} A_i \sin(k_i x + \phi_i), \tag{9}$$

where $k_i = 2\phi\{ni\}/L_x$ are wave numbers whose $\{ni\}$ are integer numbers selected randomly in $[1, n_{max}]$, N is the integer determining how many waves to be added, L_x is the calculation domain size, A_i is a random float number uniformly chosen in $[0, 1]$, and ϕ_i is the randomly chosen phase in $(0, 2\pi)$. In all examples included in this work, we select $k_{max} = 2$ and $N = 2$. For the numerical setup, the spatial domain [0, 1) is partitioned into $n_x = 1024$ equidistant points, while the temporal interval $[0, 2]$ is discretised using $n_t = 201$ uniform steps. The numerical solution was computed using a 2nd-order upwind finite difference scheme in both time and space. For training and inference, the spatial domain is downsampled to $n_x = 256$.

A.0.2 Viscous Burgers' Equation (vB)

The viscous Burgers' equation is a non-linear parabolic partial differential equation (PDE) with applications in fluid dynamics, nonlinear acoustics, and gas dynamics. It is the simplest PDE that combines nonlinearity and diffusivity. The equation often appears in simplified models of complex systems, thereby providing insight into the general behaviour of complex processes. The one-dimensional version of a viscous fluid is given as

$$\begin{aligned} \partial_t u(t,x) + u(t,x)\partial_x u(t,x) = \nu \partial_{xx} u(t,x) \quad x \in (0,1),\, t \in (0,1], \\ u(0,x) = u_0(x), \qquad x \in (0,1), \end{aligned} \tag{10}$$

where u is a scalar velocity and is a function of space x and time t. The parameter $\nu = 0.01$ is the viscosity term.

The initial conditions are generated as in the advection equation. The numerical solution was calculated with the temporally and spatially 2nd-order upwind difference scheme for the advection term, and the central difference scheme for the diffusion term. For the numerical setup, the spatial domain [0, 1) is partitioned into $n_x = 1024$ equidistant points, while the temporal interval $[0, 1]$ is discretised using $n_t = 201$ uniform steps. For training and inference, the spatial domain is downsampled to $n_x = 256$.

A.0.3 DIFFUSION-REACTION EQUATION (DR)

Here, we consider a one-dimensional diffusion-reaction PDE that combines a diffusion process with rapid evolution driven by a source term. The equation is expressed as:

$$\begin{aligned} \partial_t u(t,x) - \nu \partial_{xx} u(t,x) - \rho u(1-u) = 0, \quad x \in (0,1),\, t \in (0,1], \\ u(0,x) = u_0(x), x \in (0,1), \end{aligned} \tag{11}$$

where viscosity $\nu = 0.0005$ and density $\rho = 0.01$. The variable u could grow exponentially due to the force term, which depends on u. As in the 1D advection equation, we use periodic boundary conditions and the same initial condition (Equation 9); however, $k_{max} = 3$ and $N = 3$. The numerical solution was computed using a 2nd-order central-difference scheme in both time and space. For the source term part, we use the piecewise-exact solution (PES) method. For the domain setup, the spatial domain [0, 1] is partitioned into n_x = 1024 equidistant points, while the temporal interval $[0, 1]$ is discretised using n_t = 201 uniform steps. For training and inference, the spatial domain is downsampled to $n_x = 256$.

A.0.4 DIFFUSION-SORPTION (DS)

The diffusion-sorption equation models a diffusion process that is retarded by sorption. The equation is written as;

$$\partial_t u(t.x) = D/R(u) \partial_{xx} u(t,x), \tag{12}$$

where D is the effective diffusion coefficient, R is the retardation factor representing the sorption that hinders the diffusion process.

This nonlinear equation is retarded by the retardation factor R, which is dependent on u based on the Freundlich sorption isotherm:

$$R(u) = 1 + \frac{1-\phi}{\phi} \rho_s k n_f u^{n_f - 1}, \tag{13}$$

where $\phi = 0.29$ is the porosity of the porous medium, $\rho_s = 2880$ is the bulk density, $k = 3.5 \times 10^{-4}$ is the Freundlich's parameter, $n_f = 0.874$ is the Freundlich's exponent, and the effective diffusion coefficient $D = 5 \times 10^{-4}$.

The initial condition is generated with a uniform distribution $u(0,x) \sim \mathcal{U}(0, 0.2)$ for $x \in (0,1)$. The data is discretised into $n_x = 1024$ and $n_t = 501$, and the downsampled version is used for the models' training, with $n_x = 256$, $n_t = 101$. The spatial discretisation is performed using the finite volume method (Eymard et al., 2000), and the time integration is performed using a fourth-order Runge-Kutta method.

B TRAINING

All FNO models use 16 Fourier modes, a hidden dimension $d = 32$, and 4 iterative layers. Training is performed for 200 epochs using the Adam optimiser with a batch size of 128. The initial learning rate of 10^{-3} is modulated by a scheduler with a decay factor of 0.95 per epoch. All models are trained for 200 epochs with 100 iterations per epoch using a normalised Root Mean Square Error (nRMSE) loss function. For each iteration and each sample, we stochastically generate a rollout sequence $\widehat{T}$ using a variable-timestep sampling strategy, thereby exposing the model to a continuous range of temporal scales (See further details in Appendix C). We use a multistep prediction of 5, i.e., the input (history) and output (horizon) are equal.

As shown in Figure 2, each method was trained until convergence was achieved. All nine strategies exhibited stable convergence within 200 epochs, using the viscous Burgers' equation as a representative case. Notably, the **Attention-based (At)** configurations demonstrated a marginally smoother

loss trajectory compared to the Concatenation (Co) baselines. This suggests that the attention mechanism facilitates more stable gradient flow during the early stages of learning, when the model is first exposed to variable timesteps.

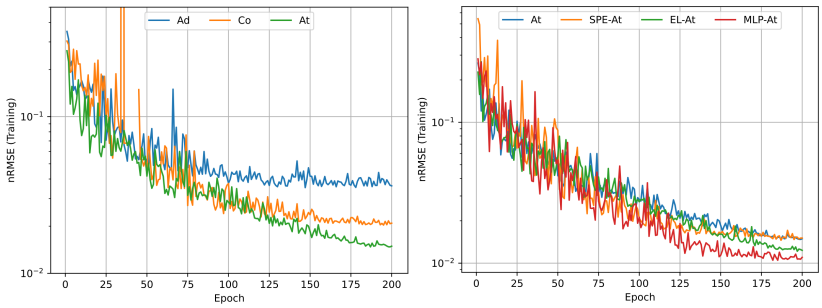

Figure 2: **Training convergence for the Viscous Burgers' equation over 200 epochs**. Each curve represents the mean Normalised Root Mean Squared Error (nRMSE) for one of the nine strategies outlined in Table 1. The consistent downward trend and stabilisation across all strategies indicate robust convergence and provide a reliable baseline for the variable-timestep sampling evaluations.

C Variable-timestep Sampling

A key contribution of this work is the variable-timestep sampling strategy (Table 6) used during training. This procedure ensures the model is exposed to a broad spectrum of temporal scales using a dataset with fixed Δt.

The objective of our variable-timestep sampling strategy is to extract a randomly sampled, ordered subset $\widehat{T}$ of a fixed size $R+1$ from the temporal domain $T = \{0, 1, \ldots, t_s\}$. This subset is constructed such that the intervals between consecutive elements are non-constant. R denotes the number of autoregressive rollouts, i.e., the number of recursive model applications in which the prediction at one step serves as the input to the next.

To implement this, we independently sample a start time t_{start} and an end time t_{end} from T for each iteration, subject to the constraint $t_{\text{end}} \geq t_{\text{start}} + R$. Within the interval $[t_{\text{start}}, t_{\text{end}}]$, we randomly select $R-1$ elements to form the subset $\widehat{T}$. The magnitude of R directly influences the sampling distribution; specifically, as R increases, the probability of observing large temporal gaps between consecutive elements in $\widehat{T}$ decreases. In this work, we keep R constant throughout training. To enhance data diversity, t_{start} is sampled independently for each trajectory within a single batch. This stochastic approach ensures the model is exposed to a broad spectrum of temporal scales.

D Final time prediction

While total integrated error provides a holistic view of model performance, the final-time error (nRMSE at the last timestep of the rollout) evaluates the long-term stability and terminal accuracy of the learned evolution laws. As shown in Tables 7 and 8, the results indicate that multiple temporal paths can converge to an accurate final state.

1. **Conditioning and Architecture Convergence:** The Attention (At) and Concatenation (Co) mechanisms remain the superior conditioning strategies for final time accuracy. In the Diffusion-Reaction and Diffusion-Sorption systems, the Attention mechanism achieves the lowest final-time errors (0.04 and 0.0008, respectively). Interestingly, in the Viscous Burgers' and Advection equations, the performance gap between Concatenation and Attention narrows significantly at the final timestep. This suggests that while Attention is more effective at capturing transient dynamics (such as shock formation), both architectures can converge to the correct steady-state or terminal topology.

Algorithm 1: Training with Variable-timestep Sampling

Input: Epochs E, iterations per epoch I, time domain $T = \{0, 1, \ldots, t_s\}$, training samples $\{u_t\}_{t \in T}$, rollout count R, model $\mathcal{F}_\theta$, loss criterion $\mathcal{L}$.

for $e = 1$ **to** E **do**
  for $i = 1$ **to** I **do**
    $t_{\text{start}} \sim \text{Uniform}(0, t_s - R)$
    $t_{\text{end}} \sim \text{Uniform}(t_{\text{start}} + R, t_s)$
    $\widehat{T} = \text{Sort}(\text{SampleWithReplacement}(\text{range}(t_{\text{start}}, t_{\text{end}}), \text{size} = R + 1))$
    Total Loss $= 0$
    $\tilde{u}_{t_0} = u_{t_{\text{start}}}$
    for $j = 1$ **to** R **do**
      $t_{curr} = \widehat{T}[j-1], \quad t_{next} = \widehat{T}[j]$
      $\tilde{u}_{t_{next}} = \mathcal{F}_\theta(\tilde{u}_{t_{curr}}, t_{curr}, t_{next})$
      Total Loss $=$ Total Loss $+ \mathcal{L}(\tilde{u}_{t_{next}}, u_{t_{next}})$
    end for
    $\theta \leftarrow \text{Optimiser}(\text{Total Loss})$
  end for
end for

Table 6: **Variable-timestep Sampling Strategy for Autoregressive Training:** This procedure generates non-uniform temporal rollout sequences. For each iteration, a random temporal window $[t_{\text{start}}, t_{\text{end}}]$ is selected, from which an ordered subset $\widehat{T}$ of size $R + 1$ is sampled. This ensures the intervals between consecutive elements are non-constant, exposing the model to diverse temporal scales (Δt) to improve long-term stability and capture multiscale dynamics.

2. **Embedding Stability:** Table 8 demonstrates that Embedding Layer (EL) and MLP embeddings provide the most stable long-term rollouts.In Advection, the EL-At configuration using a $1 \rightarrow 2$ variable-timestep strategy achieves a terminal error of 0.008, significantly outperforming the non-learnt Sinusoidal (STE) baseline.In Viscous Burgers', the MLP-At model with a constant $3\Delta t$ sampling yields the best terminal result (0.0024).These findings confirm that learnt embeddings better preserve the solution's physical structure over longer time horizons than deterministic embeddings.

3. **Resilience to Exposure Bias:** A critical observation across both tables is that variable-timestep strategies often match or exceed the terminal accuracy of $1\Delta t$ constant-sampling, despite the $1\Delta t$ model having a higher sampling density. For instance, in Advection, the $1 \rightarrow 2 \rightarrow 3$ strategy achieves a lower final error (0.030) than the constant $1\Delta t$ baseline (0.040).This validates the Exposure Bias Resilience of the variable-timestep framework: by using larger timesteps, the model reaches the final time in fewer discrete steps, thereby reducing the number of opportunities for autoregressive error to accumulate. This demonstrates that for long-term physical modelling, temporal flexibility is not just a computational convenience but a structural advantage for maintaining terminal stability.

TIME CONDITIONING STRATEGY		VARIABLE-TIMESTEP MODEL						
		CONSTANT-TIMESTEP SAMPLING			VARIABLE-TIMESTEP SAMPLING			
		1 (201 t_s)	2 (101 t_s)	3 (67 t_s)	1-2 (150 t_s)	2-1 (150 t_s)	1-2-3 (125 t_s)	3-2-1 (125 t_s)
Av	Ad	1.57	0.58	0.56	0.55	0.66	**0.18**	0.31
	Co	0.040	0.033	0.045	0.031	0.037	**0.030**	0.041
	At (**ours**)	0.059	0.054	0.058	**0.049**	0.053	0.070	0.038
vB	Ad	0.030	0.017	0.044	0.034	0.015	0.028	**0.013**
	Co	0.0060	0.0027	0.0035	0.0061	**0.0026**	0.0063	0.0029
	At (**ours**)	0.012	0.0047	**0.0031**	0.0096	0.0053	0.0084	0.0042
DR	Ad	0.13	0.15	0.13	0.10	**0.09**	0.11	0.12
	Co	0.21	0.13	**0.10**	0.18	0.14	0.18	0.11
	At (**ours**)	0.10	0.06	**0.04**	0.10	0.06	0.09	0.05
DS	Ad	0.0051	0.0027	**0.0021**	0.0042	0.0048	0.0035	0.0056
	Co	0.0015	0.0010	**0.0009**	0.0012	0.0013	0.0011	0.0011
	At (**ours**)	0.0010	0.0008	0.0008	0.0008	0.0009	0.0008	0.0008

Table 7: Time Conditioning: Mean normalised root mean squared error (nRMSE) at final time. For each PDE, the smallest value per row and column is highlighted in **bold** and underline, respectively. The smallest error is highlighted in blue. The EL-At model using a 1-2-3 variable-timestep sampling strategy achieves the lowest overall error.

TIME CONDITIONING STRATEGY		VARIABLE-TIMESTEP MODEL						
		CONSTANT-TIMESTEP SAMPLING			VARIABLE-TIMESTEP SAMPLING			
		1 (201 t_s)	2 (101 t_s)	3 (67 t_s)	1-2 (150 t_s)	2-1 (150 t_s)	1-2-3 (125 t_s)	3-2-1 (125 t_s)
Av	STE-At	0.059	0.076	0.036	0.031	0.041	**0.027**	0.032
	EL-At	0.011	0.010	0.019	**0.008**	0.009	0.009	0.009
	MLP-At	0.021	0.023	0.024	0.012	0.021	**0.011**	0.018
vB	STE-At	0.0066	0.0081	0.0039	0.0069	0.0051	0.0065	**0.0038**
	EL-At	0.0059	0.0034	**0.0034**	0.0043	0.0048	0.0043	0.0039
	MLP-At	0.01	0.0037	**0.0024**	0.0085	0.0042	0.0071	0.0031
DR	STE-At	0.14	0.087	**0.081**	0.11	0.12	0.098	0.10
	EL-At	0.10	0.052	**0.031**	0.077	0.077	0.066	0.065
	MLP-At	0.12	0.094	**0.066**	0.10	0.10	0.090	0.093
DS	STE-At	0.0013	0.0014	0.0010	0.0020	0.0027	**0.0011**	0.0019
	EL-At	0.0013	0.0011	**0.0008**	0.0011	0.0011	0.0009	0.0010
	MLP-At	0.0009	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008

Table 8: Attention-based Conditioning with different Embeddings: Mean normalised root mean squared error (nRMSE) at final time. For each PDE, the smallest value per row and column is highlighted in **bold** and underline, respectively. The smallest error is highlighted in blue. The EL-At model using a 1-2-3 variable-timestep sampling strategy achieves the lowest overall error.

E EFFECT OF MODEL SIZE

As shown in the Table below (9), we investigate the effect of model size using the viscous Burgers' equation. We double the hidden dimension d of our best-performing embedding method (EL) for the attention (At) and concatenation (Co) conditioning approaches, i.e., EL-At and EL-Co. The results show that Attention scales well and outperforms Concatenation across all time-sampling strategies. However, the Concatenation option is more computationally efficient (30% faster) due to lower memory (8% less) usage than our Attention-based approach.

	Time per Epoch (sec)	Memory (MB)	**Constant-timestep Sampling**			**Variable-timestep Sampling**			
			1 201 t_s	2 101 t_s	3 67 t_s	1-2 150 t_s	2-1 150 t_s	1-2-3 125 t_s	3-2-1 125 t_s
EL-At	-	-	**1.8**	**0.93**	**0.84**	**1.7**	**0.96**	**1.6**	**0.83**
EL-Co	-	-	2.2	1.2	1.6	2.2	1.3	2.2	1.5
EL-At 2x param	71.8	2.49	**0.77**	**0.48**	**0.49**	**0.76**	**0.48**	**0.76**	**0.49**
EL-Co 2x param	**54.7**	**2.30**	0.98	0.64	0.69	0.95	0.65	0.96	0.68

Table 9: Effect of model size of the top two conditioning and embedding approaches.

F VARIABLE-TIMESTEP AND CONSTANT-TIMESTEP MODEL

Figure 3 establishes the scaling properties of our variable-timestep framework, demonstrating that while a generalisation penalty exists at low data volumes, it can be mitigated by increasing the training scale. Most importantly, the scaling law confirms that the flexibility of a single model does not come at the cost of structural stability; as long as sufficient data is provided, the variable-timestep model can recover the performance of a constant-timestep model while offering 5x better parameter efficiency.

We visualise the solution of our four benchmark PDEs from Figure 4 - 7, and show the solution of the constant-timestep and variable-timestep models at $1\Delta t$. The solutions qualitatively look similar and thus are consistent with the correlation table. Consequently, the variable-timestep framework offers a robust and computationally efficient alternative to specialised solvers, particularly in non-linear systems such as Diffusion-Reaction, where multi-scale temporal training enhances predictive performance.

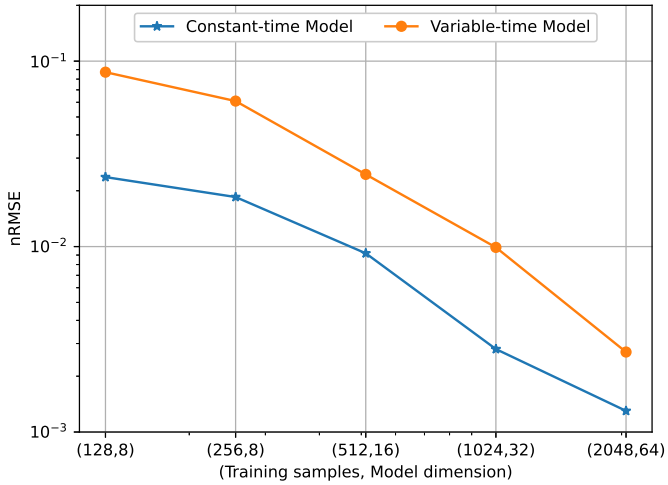

Figure 3: Scaling: The rate of error reduction with an increase in the number of training samples and model dimension.

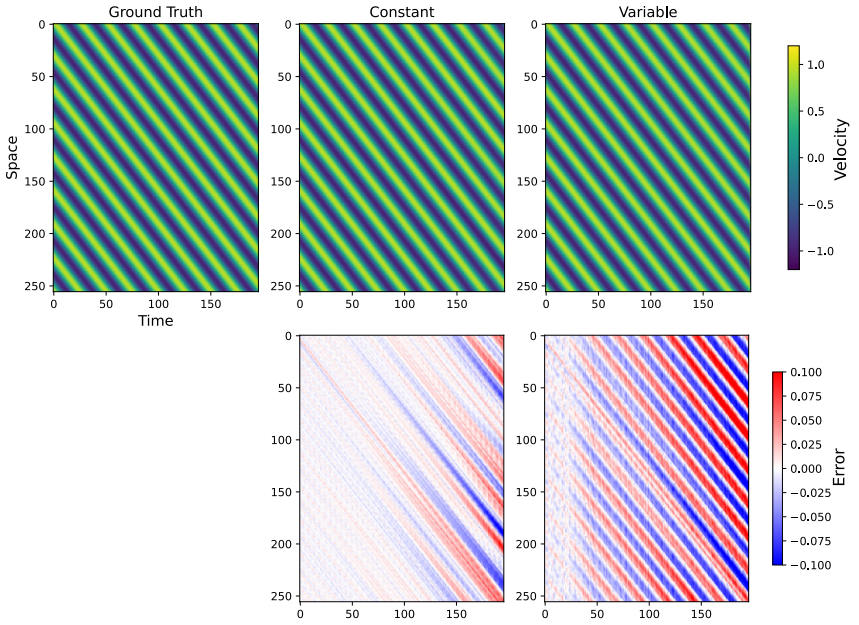

Figure 4: Advection: Visualisation of prediction (top) and error (bottom) for $1\Delta t$ constant-timestep model (middle) and variable-timestep model evaluated at $1\Delta t$ (left). The data sample with the highest error is shown.

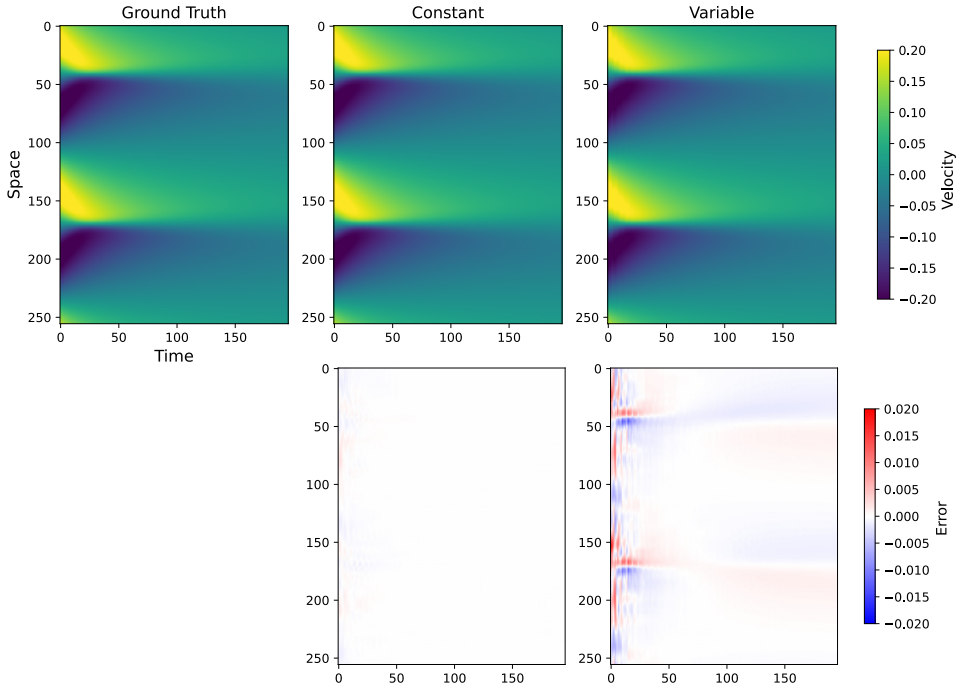

Figure 5: Viscous Burgers: Visualisation of prediction (top) and error (bottom) for $1\Delta t$ constant-timestep model (middle) and variable model evaluated at $1\Delta t$ (left). Data sample with the smallest error.

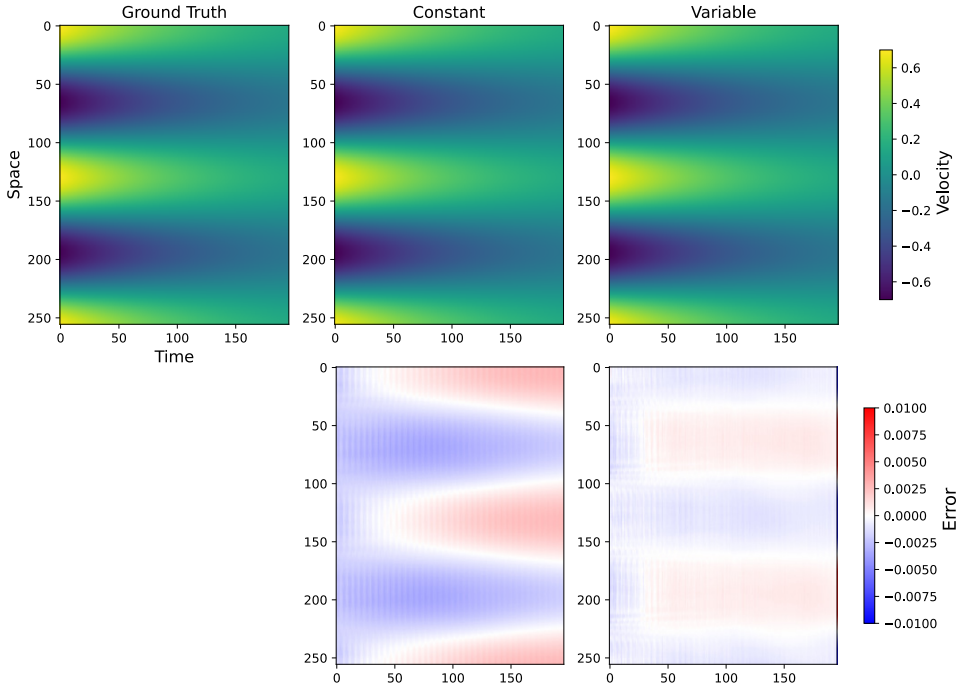

Figure 6: Diffusion-Reaction: Visualisation of prediction (top) and error (bottom) for $1\Delta t$ constant-timestep model (middle) and variable model evaluated at $1\Delta t$ (left)

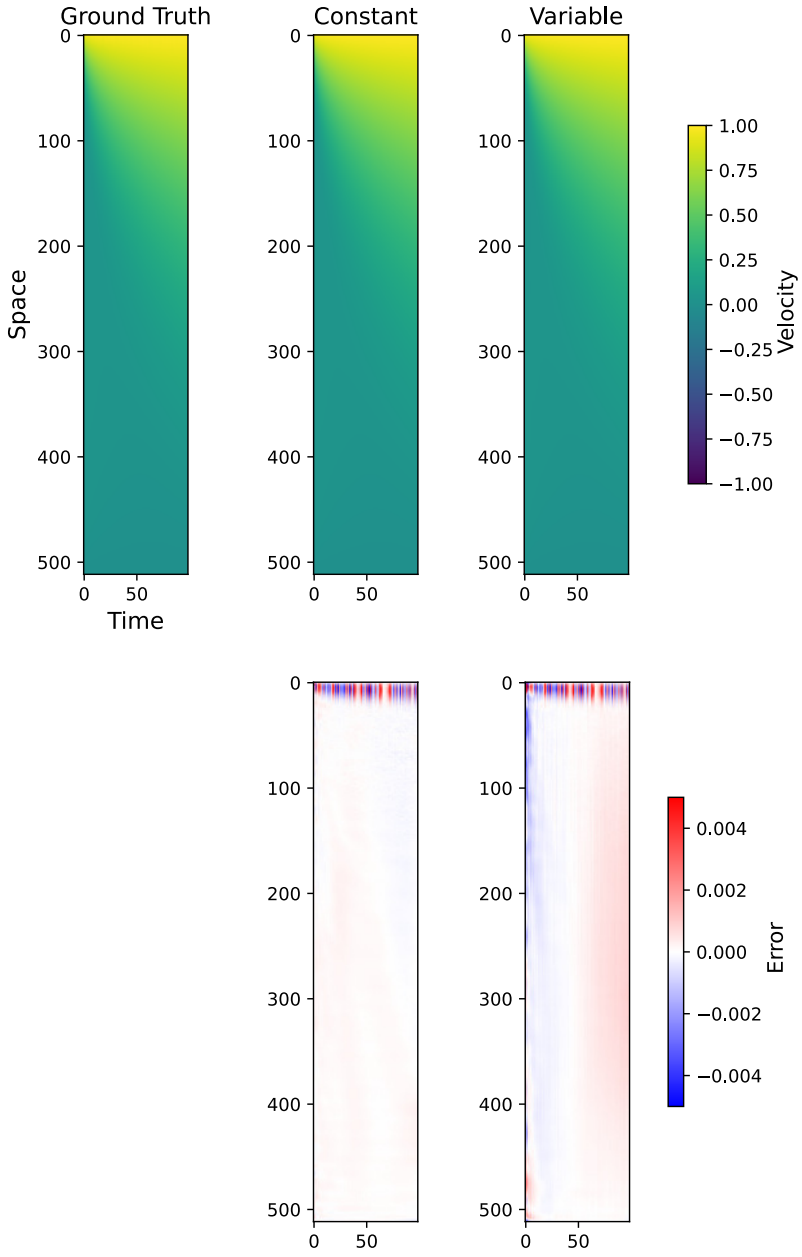

Figure 7: Diffusion-Sorption: Visualisation of prediction (top) and error (bottom) for $1\Delta t$ constant-timestep model (middle) and variable model evaluated at $1\Delta t$ (left)

G ADVECTION (AV)

The 1D advection equation is the simplest test case, in which the solution is purely translated at a constant speed parameter β. Because the system lacks dissipative mechanisms like diffusion, it can be sensitive to the accumulation of numerical or model-induced errors during long-term autoregressive rollouts.

Overall, the **Attention-based (At)** conditioning mechanism combined with the **Embedding Layer (EL)** architecture achieved the most robust performance. As shown in Table 10, the **EL-At** configuration using a **1-2-3** variable-timestep sampling strategy yielded the lowest total error. However, when evaluating the accuracy of the final state (Table 11), the **1-2** variable strategy with EL-At was superior. Across all tests, Attention (At) consistently outperformed Concatenation (Co) and Addition (Ad).

G.1 ANALYSIS OF CONSTANT-TIME SAMPLING

Figure 8 illustrates the error propagation for models evaluated at fixed temporal intervals ($\Delta t = 1$ to 5).

- **Stability at Small Timesteps:** The error propagation for $1\Delta t$ and $2\Delta t$ remains the lowest and follows nearly identical linear trajectories. This suggests that the model accurately captures wave translation at these time resolutions.
- **The $2\Delta t$ Threshold:** A significant performance degradation is observed once the timestep exceeds $2\Delta t$. The error curves for $3\Delta t$, $4\Delta t$, and $5\Delta t$ are significantly worse than those for $1\Delta t$ and $2\Delta t$, with the most pronounced increase occurring between $3\Delta t$ and $4\Delta t$. In the context of advection, this indicates that the model's ability to map the function over longer temporal gaps diminishes, likely due to the spectral layers' limitations in capturing high-frequency shifts at large Δt.

G.2 ANALYSIS OF VARIABLE-TIME SAMPLING

The efficacy of variable-timestep sampling is highly dependent on whether the sampling includes the high-error timesteps identified in the constant-timestep sampling error curves (i.e., $\Delta t > 2$).

- **Increasing timestep sampling (Figure 9):** Transitioning from $1\Delta t \rightarrow 2\Delta t$ and $1\Delta t \rightarrow 2\Delta t \rightarrow 3\Delta t$ (i.e., Small-to-Large) shows an improvement over constant $1\Delta t$. This demonstrates that once the model has established a stable trajectory on small time steps, it can transition to larger steps to complete the rollout with less cumulative error. Also, we observe that $1\Delta t \rightarrow 2\Delta t$ performs better than $1\Delta t \rightarrow 2\Delta t \rightarrow 3\Delta t$, most likely due to the use of error-prone $3\Delta t$ in the latter sampling.

  However, strategies starting with a larger steps (e.g., 2-3 or 2-3-4) do not show improvement over constant $2\Delta t$, as the sampling strategy contains timesteps (i.e., $3\Delta t$ and $4\Delta t$) that have high-error (as seen in Figure 8), which compromises the entire rollout's integrity.
- Decreasing timestep sampling (Figure 10): Conversely, decreasing the timestep during the rollout can also be beneficial (i.e., Large-to-Small). The $2\Delta t \rightarrow 1\Delta t$ strategy improves upon a constant $2\Delta t$ rollout. Notably, the $3\Delta t \rightarrow 2\Delta t \rightarrow 1\Delta t$ strategy significantly outperforms the $3\Delta t \rightarrow 2\Delta t$ and $3\Delta t$ sampling. This suggests that correcting the trajectory with smaller, more accurate steps toward the end of a rollout can mitigate some of the error propagation introduced by the early, larger timesteps.

Cumulative Error and Accuracy Trade-offs: Figure 11 further confirms existing trends by showing the cumulative sum of errors. For a fixed sampling budget (i.e., different sampling with the same number of total timesteps 125 t_s), more samples from the lower-error (earlier) region ultimately reduce the total error, as seen in the case where the cumulative error for the $1\Delta t \rightarrow 2\Delta t$ and strategy is lower than that for the $2\Delta t \rightarrow 1\Delta t$ strategy. The difference is more pronounced in the comparison between $1\Delta t \rightarrow 2\Delta t \rightarrow 3\Delta t$ and $3\Delta t \rightarrow 2\Delta t \rightarrow 1\Delta t$ sampling. Starting with a larger timestep ($3\Delta t$) results in much higher total error than with a smaller timestep. We observe a significant increase in the slope of $3\Delta t \rightarrow 2\Delta t \rightarrow 1\Delta t$ at approximately $t \approx 137$, where the timestep transitions

from $2\Delta t$ to $1\Delta t$, while there is no visible change in the slope when transitioning from $2\Delta t$ to $3\Delta t$ in the $1\Delta t \rightarrow 2\Delta t \rightarrow 3\Delta t$ sampling.

	TIME CONDITION AND EMBED STRATEGY	VARIABLE-TIMESTEP MODEL						
		CONSTANT-TIMESTEP SAMPLING			VARIABLE-TIMESTEP SAMPLING			
		1 (201 t_s)	2 (101 t_s)	3 (67 t_s)	1-2 (150 t_s)	2-1 (150 t_s)	1-2-3 (125 t_s)	3-2-1 (125 t_s)
Advection (Av)	Ad	147.13	44.62	58.06	83.41	36.93	**24.44**	51.7
	Co	3.78	3.41	4.36	3.30	3.79	**2.97**	4.57
	At (**ours**)	7.14	**5.83**	7.86	6.86	6.17	7.11	6.71
	STE-Ad	61.70	**11.82**	15.76	57.83	14.42	52.65	17.23
	STE-Co	187.00	**49.39**	83.58	156.66	61.69	130.87	6.73
	STE-At (**ours**)	6.00	9.26	4.65	5.06	8.00	4.62	**4.44**
	EL-Ad	27.68	25.34	**24.50**	27.01	26.08	26.43	24.93
	EL-Co	2.31	1.98	2.76	1.89	**1.79**	1.85	2.46
	EL-At (**ours**)	1.20	1.14	1.93	1.07	1.08	**1.03**	1.29
	MLP-Ad	3.17	3.62	5.07	3.35	**3.02**	3.73	3.63
	MLP-Co	2.42	2.44	3.42	2.35	**2.33**	2.45	3.15
	MLP-At (**ours**)	2.12	3.87	3.81	1.87	3.90	**1.73**	3.59

Table 10: Mean normalised root mean squared error (nRMSE) averaged over the full temporal rollout for the Advection equation: The values in **bold** and underline are the smallest error per row and column, respectively. The smallest error is highlighted in blue. The EL-At model using a 1-2-3 variable-timestep sampling strategy achieves the lowest overall error.

	TIME CONDITION AND EMBED STRATEGY	VARIABLE-TIMESTEP MODEL						
		CONSTANT-TIMESTEP SAMPLING			VARIABLE-TIMESTEP SAMPLING			
		1 (201 t_s)	2 (101 t_s)	3 (67 t_s)	1-2 (150 t_s)	2-1 (150 t_s)	1-2-3 (125 t_s)	3-2-1 (125 t_s)
Advection (Av)	Ad	1.57	0.58	0.56	0.55	0.66	**0.18**	0.31
	Co	0.040	0.033	0.045	0.031	0.037	**0.030**	0.041
	At (**ours**)	0.059	0.054	0.058	**0.049**	0.053	0.070	0.038
	STE-Ad	0.43	0.09	**0.07**	0.41	0.15	0.38	0.16
	STE-Co	1.13	0.61	1.19	**0.33**	0.45	0.37	0.057
	STE-At (**ours**)	0.059	0.076	0.036	0.031	0.041	**0.027**	0.032
	EL-Ad	0.16	0.14	0.14	0.14	0.15	0.14	0.14
	EL-Co	0.018	0.018	0.022	**0.011**	0.011	0.013	0.013
	EL-At (**ours**)	0.011	0.010	0.019	**0.008**	0.009	0.009	0.009
	MLP-Ad	0.023	0.022	0.034	0.017	**0.015**	0.020	0.015
	MLP-Co	0.021	0.016	0.032	0.017	0.016	0.016	0.018
	MLP-At (**ours**)	0.021	0.023	0.024	0.012	0.021	**0.011**	0.018

Table 11: Mean normalised root mean squared error (nRMSE) at **final** time: The value in **bold** and underline are the smallest error per row and column, respectively. The smallest error is highlighted in blue.

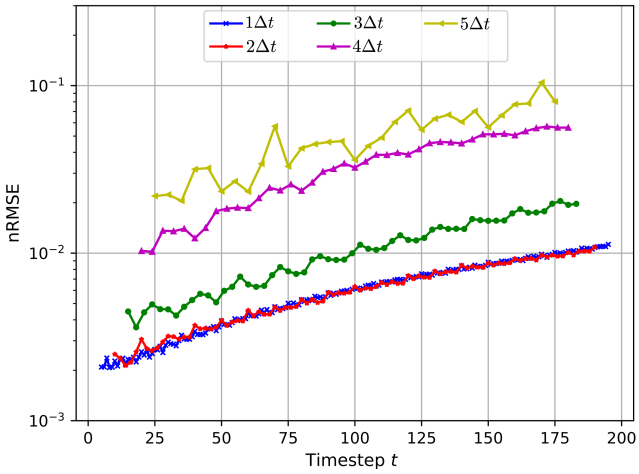

Figure 8: Temporal error propagation for constant-timestep sampling ($\Delta t \in \{1, \ldots, 5\}$). A performance elbow is observed at $\Delta t > 2$, where the error increases significantly.

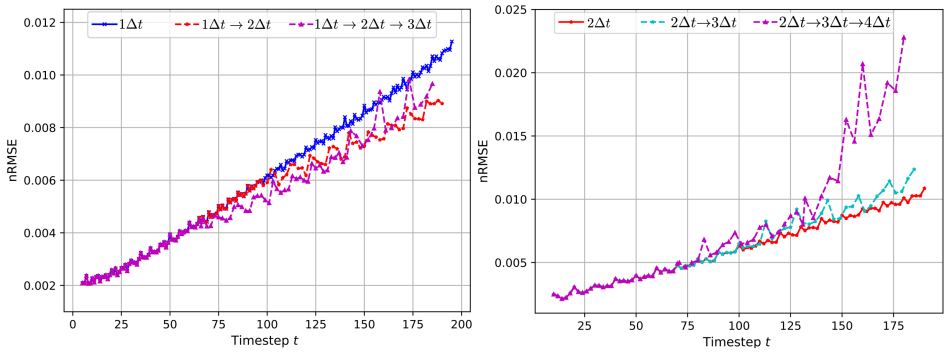

Figure 9: Impact of **increasing** variable-timestep sampling strategies on prediction accuracy. Strategies starting with small timesteps (e.g., 1-2 and 1-2-3) show improved stability over the constant $\Delta t = 1$ baseline, whereas strategies starting at larger intervals (2-3-4) suffer from high initial error injection.

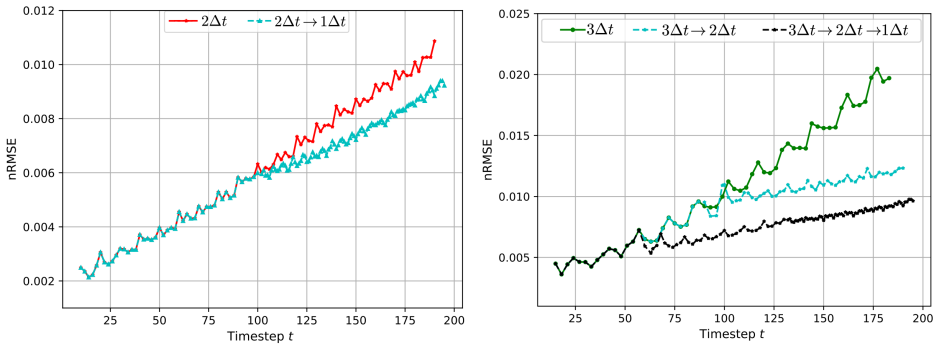

Figure 10: Impact of **decreasing** variable-timestep sampling strategies. The results demonstrate a "correction" effect, in which transitioning from a large to a smaller timestep (e.g., 3-2-1) mitigates cumulative error more effectively than maintaining a constant large timestep.

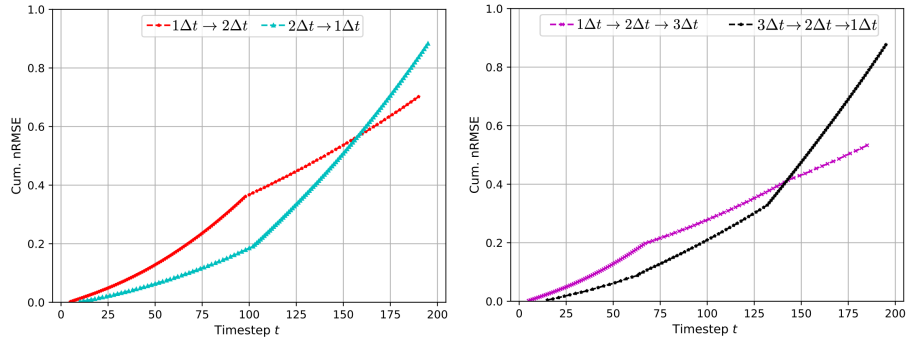

Figure 11: Cumulative error growth during the autoregressive rollout for the Advection equation. The near-linear accumulation of error across all models reflects the system's non-dissipative nature, in which local prediction inaccuracies are advected forward without damping.

H VISCOUS BURGERS

The Viscous Burgers' equation introduces non-linear shock formation and dissipative dynamics, providing a more rigorous test for the temporal conditioning mechanisms. Unlike the pure translation seen in advection, the model must account for changing gradients and the smoothing effect of viscosity over time.

The **MLP-based Attention (MLP-At)** architecture emerged as the top performer. As shown in Table 12, the MLP-At model using a $\mathbf{3\Delta t}$ ($67\ t_s$) constant-timestep or $\mathbf{3\Delta t} \rightarrow \mathbf{2\Delta t} \rightarrow \mathbf{1\Delta t}$ ($125 t_s$) variable-sampling strategy achieved the lowest total error across the entire rollout.

For constant-timestep sampling, the total error decreases as the timestep increases or the number of rollouts decreases, so only 67 steps are required for a $3\Delta t$ interval. However, the best performing variable-timestep sampling (3-2-1) demonstrates superior resilience; it can achieve an equivalent error level while utilising almost twice as many rollout steps. This highlights the unique flexibility of the variable-timestep approach, which maintains high predictive accuracy even as temporal sampling increases, a task that typically increases error propagation.

A key trend across all tables is that **Attention (At)** consistently outperforms Concatenation (Co) and Addition (Ad), suggesting that the attention mechanism better weights temporal information against the complex spatial features of the shock front. While variable-timestep strategies remain competitive, constant-timestep strategies (specifically $3\Delta t$) showed high robustness in final-time accuracy (Table 13).

H.1 ANALYSIS OF CONSTANT-TIME SAMPLING AND NONLINEAR ERROR

Figure 12 displays a nonlinear error profile with different early and late behaviour influenced by the PDE dynamics:

- **Nonlinear Error and Early-Stage Difficulty:** Unlike the linear error growth seen in advection, the error here is nonlinear and significantly higher during the initial timesteps. This corresponds to the time phase where the solution forms a sharp shock front. The high-frequency spectral components generated during this steepening phase pose a significant challenge for the FNO's spectral layers. As time progresses, the viscous term (νu_{xx}) dominates, causing the shock to dissipate into a smooth, low-gradient solution. This transition to lower-frequency dynamics, which is easier for the FNO model to predict, explains the plateauing and eventual stabilisation of the error in later stages.
- **Early error propagation:** The constant $1\Delta t$ strategy performs best during the first 25 timesteps. In this early, fast-evolution phase, rapid temporal change requires fine temporal resolution to be accurately resolved, especially in the presence of high spatial gradients. However, as the solution dynamics slow down for smooth solutions ($t > 25$), larger time steps become more effective.
- **Later Error propagation:** The constant $1\Delta t$ strategy exhibits significantly higher errors in the later stages compared to $2\Delta t$ and $3\Delta t$. This is a clear manifestation of exposure bias in autoregressive modelling. Because the $1\Delta t$ model must perform three times as many rollout steps as the $3\Delta t$ model, the accumulation of small, per-step errors eventually outweighs the accuracy gained from a finer temporal mesh. By contrast, the $3\Delta t$ strategy maintains the lowest error at the end of the rollout by reaching the target state in fewer rollouts, thereby minimising error propagation. However, beyond $3\Delta t$ (i.e., $4\Delta t$ and $5\Delta t$), the error increases, indicating that $3\Delta t$ is the optimal timestep.

H.2 ANALYSIS OF VARIABLE-TIME SAMPLING AND DIRECTIONALITY

The results for variable-timestep sampling reveal the benefit of variable-timestep sampling.

- **Increasing timestep sampling (Figure 13)** : This figure illustrates the impact of transitioning from small-to-large during the later stages of the evolution, where the dynamics are dominated by a smooth, low-gradient, and low-magnitude dissipative solution. Starting with a $1\Delta t$ baseline (top-left), the 1-2 and 1-2-3 variable-timestep sampling significantly

reduce the final error compared to a constant $1\Delta t$ rollout. For samples starting with $2\Delta t$ (top-right) and $3\Delta t$ (bottom-middle), while the variable-timestep strategies do not outperform their respective constant-timestep counterparts, they achieve comparable accuracy with significantly fewer total rollouts. This demonstrates that once the initial shock has dissipated, the model can remain stable with coarser temporal resolution.

- **Decreasing timestep sampling (Figure 14)**: In contrast, Figure 14 shows that reducing the timestep over time (e.g., 2-1 or 3-2-1) has a negligible or even negative impact on the final error. Because the late-stage solution is smooth and physically less demanding, the larger steps ($2\Delta t$ and $3\Delta t$) are already highly effective. Transitioning to a finer $1\Delta t$ resolution late in the rollout increases the total number of autoregressive steps, thereby allowing more opportunities for error propagation.

H.3 Cumulative Error and Rollout Stability

Figure 15 provides a view of these dynamics by plotting the cumulative error sum. A key observation is that 1-2 sampling results in higher cumulative error than 2-1 sampling. This is because the 1-2 strategy requires more frequent sampling during the high-error, high-gradient initial phase. Although a larger timestep is suitable for the later stages, the error accumulated during the early shock-formation phase dominates. Consequently, as shown in the total error summary (Table 12), the **Large-to-Small** (3-2-1) strategy consistently outperforms the **Small-to-Large** (1-2-3) approach. The 1-2-3 strategy suffers from higher *exposure bias* because it relies on a greater number of small-step evaluations in the high-error phase. By taking fewer, larger steps ($3\Delta t$) in the early stages, the model bypasses the region of highest instability, resulting in lower total error. This confirms that, for the Viscous Burgers' equation, the optimal variable-timestep strategy minimises the number of rollouts during periods of high physical complexity.

	TIME	**VARIABLE-TIMESTEP MODEL**						
	CONDITION-EMBED	**CONSTANT-TIMESTEP**			**VARIABLE-TIMESTEP**			
	STRATEGY	1 (201 t_s)	2 (101 t_s)	3 (67 t_s)	1-2 (150 t_s)	2-1 (150 t_s)	1-2-3 (125 t_s)	3-2-1 (125 t_s)
Viscous Burgers (vB)	Ad	6.01	**3.15**	4.45	5.85	3.16	5.40	3.52
	Co	3.09	1.30	1.68	3.12	**1.29**	3.16	1.53
	At (**ours**)	2.50	1.21	**1.08**	2.35	1.26	2.25	1.09
	STE-Ad	7.96	4.30	**3.97**	7.64	4.68	7.43	4.26
	STE-Co	6.04	**1.64**	2.41	5.87	1.86	5.74	2.15
	STE-At (**ours**)	2.24	1.97	**1.24**	2.27	1.80	2.34	1.33
	EL-Ad	10.2	**5.5**	11.3	10.6	6.20	10.7	10.5
	EL-Co	3.9	1.7	2.3	3.9	**1.6**	3.9	10.5
	EL-At (**ours**)	1.6	1.0	1.0	1.5	1.1	1.5	1.0
	MLP-Ad	16.2	**9.0**	10.3	16.1	9.5	16.2	10.4
	MLP-Co	2.2	**1.2**	1.6	2.2	1.3	2.2	1.5
	MLP-At (**ours**)	1.8	0.93	**0.84**	1.7	0.96	1.6	**0.84**

Table 12: Total nRMSE for the Viscous Burgers' equation: The values in **bold** and underline are the smallest error per row and column, respectively. The smallest error is highlighted in blue. MLP-At with a 3-2-1 sampling strategy achieves the lowest error, highlighting the advantage of decreasing the timestep when dealing with non-linear shock formation.

	TIME	**VARIABLE-TIMESTEP MODEL**						
	CONDITION-EMBED	**CONSTANT-TIMESTEP**			**VARIABLE-TIMESTEP**			
	STRATEGY	1 (201 t_s)	2 (101 t_s)	3 (67 t_s)	1-2 (150 t_s)	2-1 (150 t_s)	1-2-3 (125 t_s)	3-2-1 (125 t_s)
Viscous Burgers (vB)	Ad	0.030	0.017	0.044	0.034	0.015	0.028	**0.013**
	Co	0.0060	0.0027	0.0035	0.0061	**0.0026**	0.0063	0.0029
	At (**ours**)	0.012	0.0047	**0.0031**	0.0096	0.0053	0.0084	0.0042
	STE-Ad	0.036	0.018	**0.018**	0.034	0.027	0.033	0.023
	STE-Co	0.021	**0.0047**	0.0062	0.016	0.0087	0.017	0.0055
	STE-At (**ours**)	0.0066	0.0081	0.0039	0.0069	0.0051	0.0065	**0.0038**
	EL-Ad	0.027	0.046	0.036	0.037	0.026	**0.019**	0.035
	EL-Co	0.0087	0.0030	0.0045	0.0081	**0.0030**	0.0079	0.0359
	EL-At (**ours**)	0.0059	0.0034	**0.0034**	0.0043	0.0048	0.0043	0.0039
	MLP-Ad	0.068	0.03	**0.03**	0.05	0.05	0.05	0.05
	MLP-Co	0.0077	0.0041	**0.0035**	0.0067	0.0041	0.0067	0.0041
	MLP-At (**ours**)	0.01	0.0037	**0.0024**	0.0085	0.0042	0.0071	0.0031

Table 13: Final time error for Viscous Burgers': The values in **bold** and underline are the smallest error per row and column, respectively. The smallest error is highlighted in blue. The constant $3\Delta t$ strategy performs well, as it minimises the number of autoregressive steps required to reach the final state.

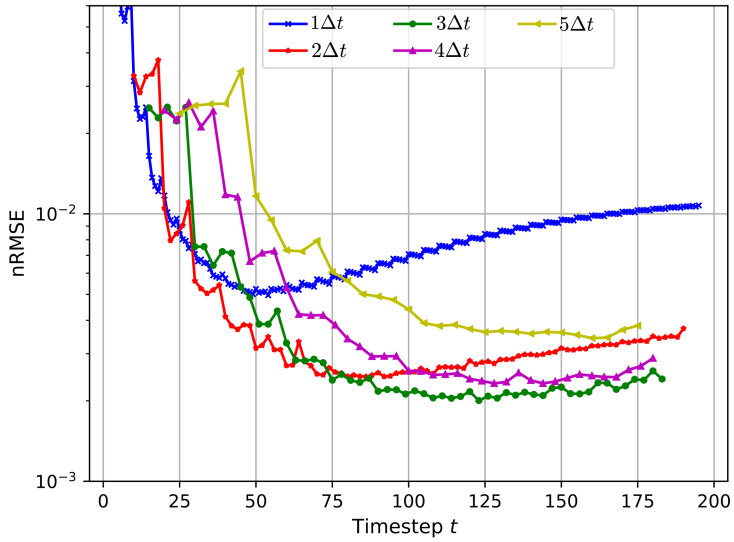

Figure 12: MLP-At temporal error propagation for constant timesteps. The non-linear error curves reflect the physical difficulty of capturing early-stage shock formation, with larger timesteps ($3\Delta t$) showing better long-term stability due to reduced rollouts.

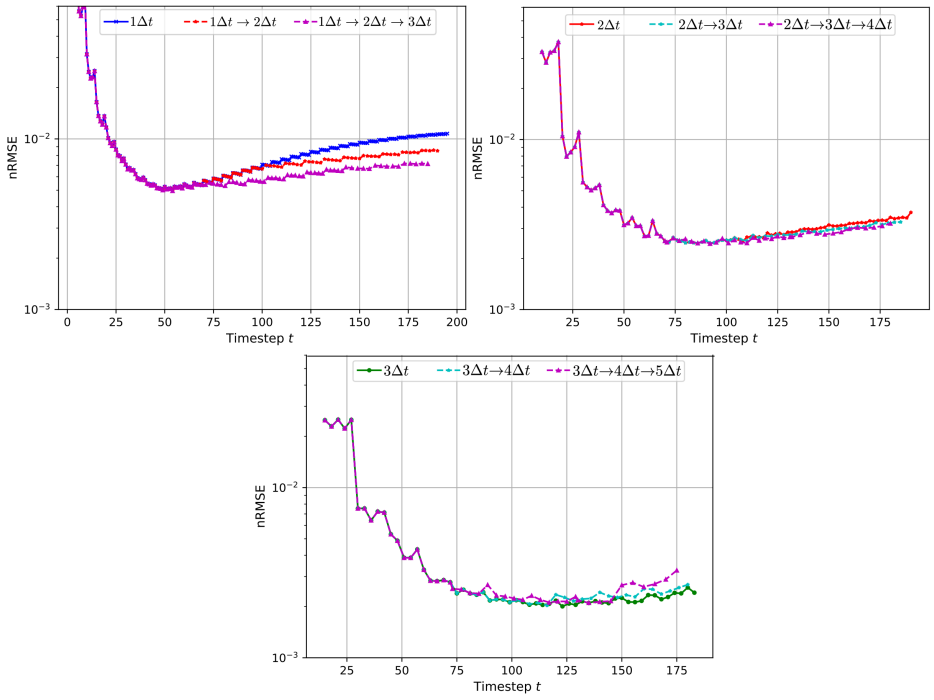

Figure 13: Performance of **increasing** variable-timestep steps. While 1-2 and 1-2-3 improve upon the constant $1\Delta t$ baseline, they are still hindered by high error in the early timesteps.

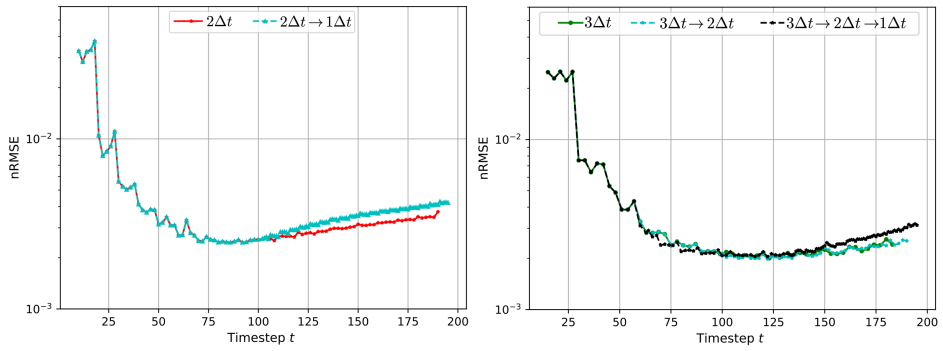

Figure 14: Performance of **decreasing** variable-timestep steps. Decreasing timesteps leads to worse performance because the smooth solutions are more susceptible to error propagation.

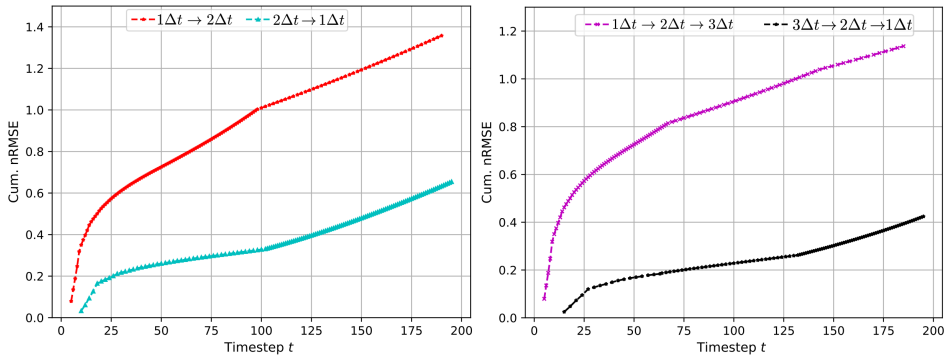

Figure 15: Cumulative error sum for Burgers' equation. The higher error in the 1-2 sampling compared to the 2-1 sampling demonstrates the impact of the sampling strategy.

I Diffusion-Reaction

The Diffusion-Reaction equation presents a case in which temporal stability is challenged by potentially exponential growth driven by the reaction term.

The **Embedding Layer with Attention (EL-At)** architecture provided the most accurate predictions. As shown in Tables 14 and 15, the constant $3\Delta t$ sampling strategy achieved the lowest total and final-time errors. A consistent trend across all evaluations is the superiority of the **Attention (At)** mechanism over Concatenation and Addition, reinforcing its ability to dynamically weigh temporal features. Notably, in this system, constant-timestep sampling with $3\Delta t$ outperformed variable-timestep strategies, suggesting that the benefits of a fixed, optimised temporal step outweigh the flexibility of variable-timestep sampling for this specific PDE.

I.1 Analysis of Constant-Time Sampling and Late-Stage Complexity

Figure 16 illustrates the temporal evolution of error for fixed timesteps:

- **Late-Stage Error Growth:** Unlike the Burgers' equation, the error for Diffusion-Reaction is lowest during the initial phases and grows non-linearly over time. This indicates that while the model effectively captures the initial diffusive spreading, the cumulative effect of the reaction term introduces complexities that are harder to resolve as the rollout progresses.
- **Optimal Sampling ($3\Delta t$):** The $1\Delta t$ strategy exhibits the highest error at later timesteps. This is attributed to the high frequency of rollouts: in a system where errors accumulate

over time, increasing the number of autoregressive steps exacerbates overall divergence from the ground truth. The $3\Delta t$ strategy achieves the best balance, providing enough temporal density to capture the reaction dynamics while minimising the exposure bias inherent in long rollouts.

I.2 Variable-Time Sampling and Cumulative Error Analysis

The performance of variable-timestep strategies is heavily influenced by the sampling region in reference to error propagation:

- **Increasing Vs Decreasing strategy (Figures 17 & 18):** Increasing the timestep (e.g., 1-2-3) show improvement over the constant $1\Delta t$ baseline. Conversely, decreasing strategies like the 3-2-1 sampling, significantly increases error upon constant $3\Delta t$ result. This is because the $3\Delta t$ timestep is already highly optimised for this equation; adding smaller, more error-prone steps ($1\Delta t$) into the rollout degrades the overall performance.
- **Cumulative Error (Figure 19):** While the small-to-large strategy (1-2-3) starts with higher error than 3-2-1, it ends with a lower total cumulative error. This is because 1-2-3 utilises larger timesteps ($3\Delta t$) during the final stages of the rollout, when the system is most prone to error. By reducing the number of samples taken during this high-error late phase, the 1-2-3 strategy effectively limits the total error compared to 3-2-1.

	TIME	VARIABLE-TIMESTEP MODEL						
	CONDITION-EMBED	CONSTANT-TIMESTEP			VARIABLE-TIMESTEP			
	STRATEGY	1 (201 t_s)	2 (101 t_s)	3 (67 t_s)	1-2 (150 t_s)	2-1 (150 t_s)	1-2-3 (125 t_s)	3-2-1 (125 t_s)
Diffusion-Reaction (DR)	Ad	13.2	12.4	12.0	12.2	**11.2**	11.8	13.3
	Co	20.3	14.2	**11.3**	18.8	14.5	18.3	12.3
	At (**ours**)	10.5	5.4	**4.4**	10.1	5.7	9.8	4.7
	STE-Ad	194.7	188.3	**181.4**	193.5	189.6	192.2	184.5
	STE-Co	24.0	17.0	**10.4**	26.86	15.0	22.6	13.8
	STE-At (**ours**)	13.9	9.1	**7.9**	12.0	10.7	10.8	9.3
	EL-Ad	161.38	126.0	123.8	155.7	135.8	147.9	**123.7**
	EL-Co	11.2	9.4	**9.3**	10.8	9.6	10.4	9.5
	EL-At (**ours**)	9.4	4.8	**3.4**	7.8	6.4	6.8	5.1
	MLP-Ad	27.9	18.22	**18.1**	26.9	19.2	27.9	18.4
	MLP-Co	15.1	10.2	**9.3**	12.7	12.0	11.4	11.2
	MLP-At (**ours**)	11.4	8.0	**7.1**	10.2	9.0	9.4	8.3

Table 14: Total nRMSE for the Diffusion-Reaction equation: The EL-At model with a constant $3\Delta t$ strategy achieves the best performance, suggesting that minimising rollouts is vital for stability in reactive systems. For variable-timestep sampling, the large-to-small strategy performs better.

	TIME	VARIABLE-TIMESTEP MODEL						
	CONDITION-EMBED	CONSTANT-TIMESTEP			VARIABLE-TIMESTEP			
	STRATEGY	1 (201 t_s)	2 (101 t_s)	3 (67 t_s)	1-2 (150 t_s)	2-1 (150 t_s)	1-2-3 (125 t_s)	3-2-1 (125 t_s)
Diffusion-Reaction (DR)	Ad	0.13	0.15	0.13	0.10	**0.09**	0.11	0.12
	Co	0.21	0.13	**0.10**	0.18	0.14	0.18	0.11
	At (**ours**)	0.10	0.06	**0.04**	0.10	0.06	0.09	0.05
	STE-Ad	1.0	1.0	1.03	1.01	1.01	1.01	1.01
	STE-Co	0.21	0.20	**0.08**	0.31	0.17	0.21	0.15
	STE-At (**ours**)	0.14	0.087	**0.081**	0.11	0.12	0.098	0.10
	EL-Ad	0.95	**0.72**	0.74	0.88	0.88	0.80	0.81
	EL-Co	0.09	**0.07**	0.13	0.12	0.12	0.12	0.11
	EL-At (**ours**)	0.10	0.052	**0.031**	0.077	0.077	0.066	0.065
	MLP-Ad	0.18	0.14	**0.12**	0.17	0.18	0.15	0.17
	MLP-Co	0.19	0.13	0.11	0.12	0.15	**0.11**	0.14
	MLP-At (**ours**)	0.12	0.094	**0.066**	0.10	0.10	0.090	0.093

Table 15: Final time error for Diffusion-Reaction: Constant $3\Delta t$ remains the most robust strategy, while in the variable-timestep sampling, there is no clear superior sampling strategy for our Attention-based options.

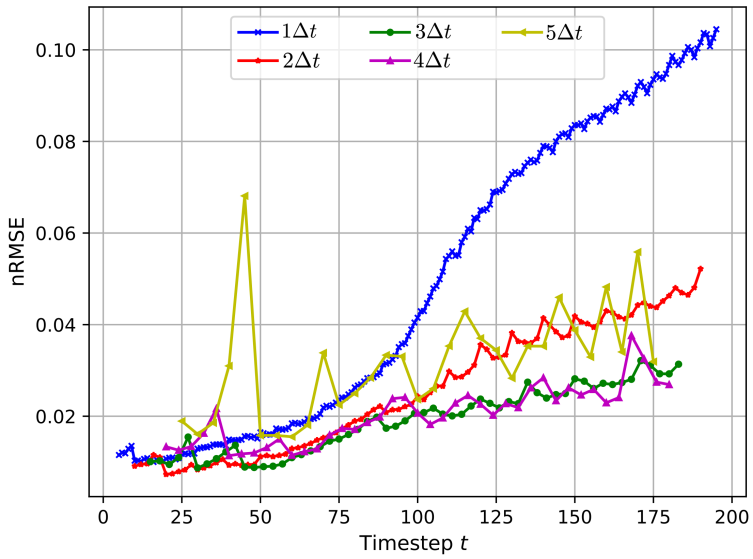

Figure 16: Temporal error propagation for constant-timestep sampling. The error increases over time as reaction-driven complexities accumulate, with the $1\Delta t$ model suffering the most from autoregressive error propagation.

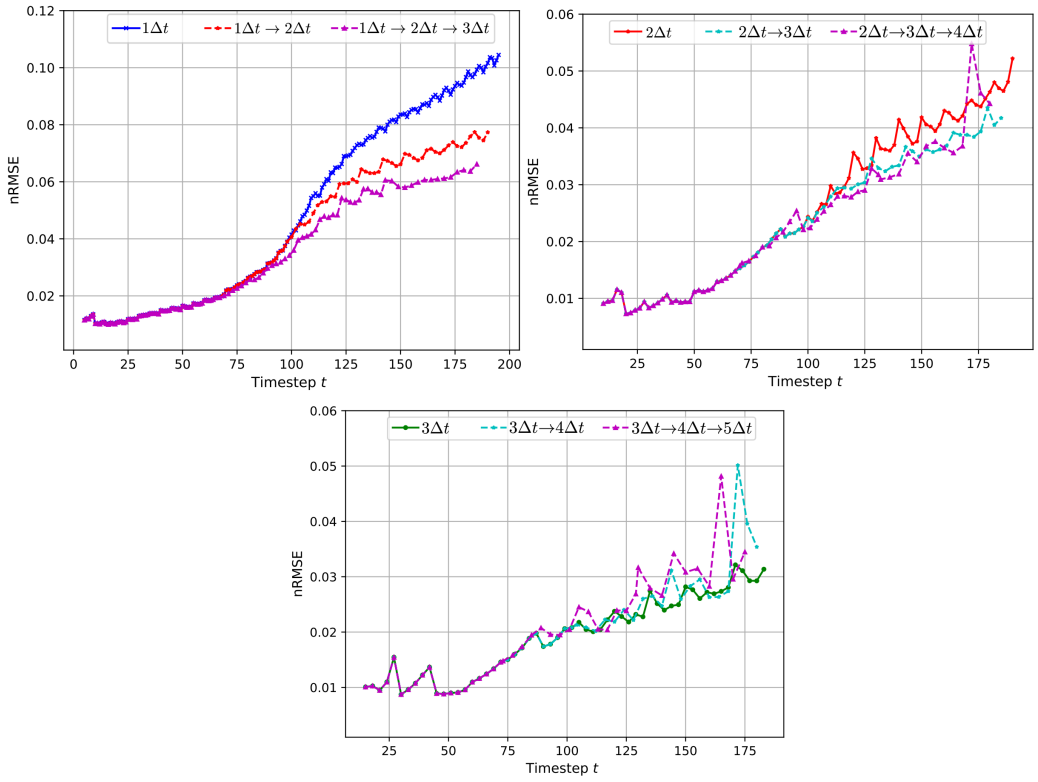

Figure 17: Increasing variable-timestep sampling: Strategies like 1-2 and 1-2-3 improve upon the $1\Delta t$ baseline by transitioning to fewer, larger steps as the physical system's error sensitivity increases.

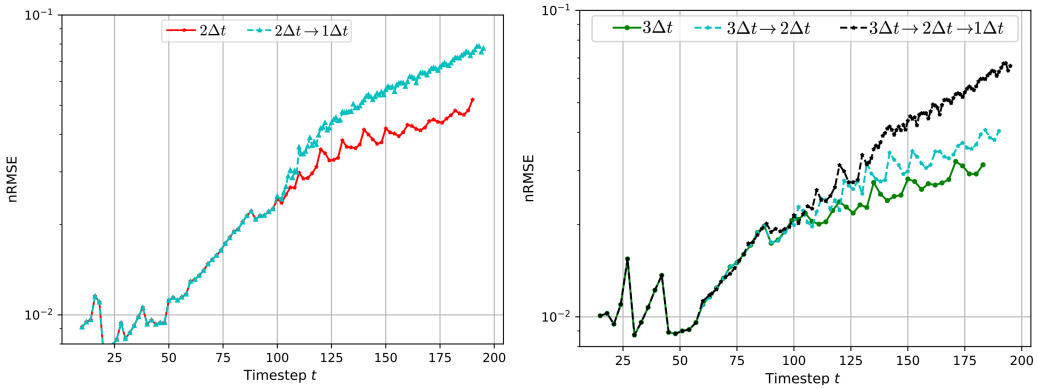

Figure 18: Decreasing variable-timestep sampling: The 3-2-1 and 2-1 sampling increase errors over the constant $3\Delta t$ baseline, as the introduction of smaller timesteps increases the total number of rollout steps and the associated exposure bias.

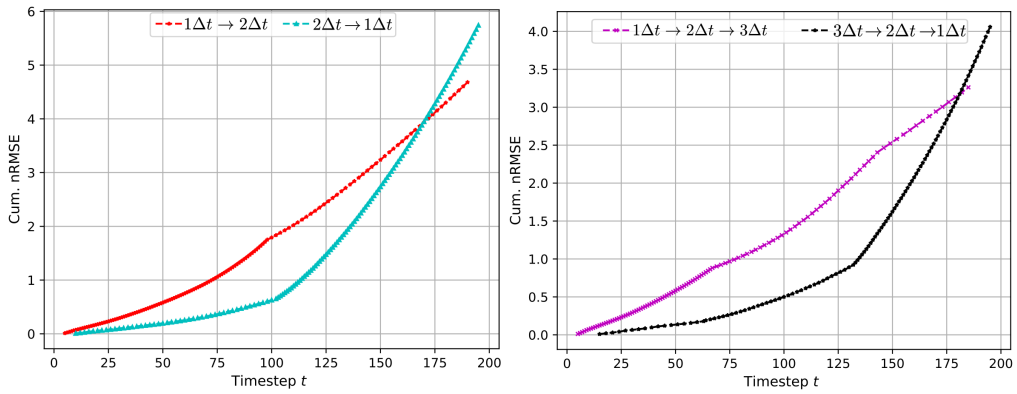

Figure 19: Cumulative error sum. Although 3-2-1 begins with lower error, the 1-2-3 strategy achieves a lower final cumulative total by using larger timesteps to bypass frequent sampling during the high-error later stages of the PDE evolution.

J DIFFUSION-SORPTION

The Diffusion-Sorption equation models the transport of solutes in porous media, where interactions between the liquid and solid phases (sorption) introduce non-linear retardation. This process typically creates high-gradient ronts early in the simulation that stabilise as the media reaches equilibrium

Analysis of the results shows that the **Embedding Layer with Attention (EL-At)** architecture is the most effective for this system. As detailed in Table 16, the constant $3\Delta t$ strategy achieved the lowest total error. While constant-timestep sampling generally demonstrates higher accuracy in total error metrics, the distinction between constant and variable-timestep strategies becomes negligible when evaluating the final-time error (Table 17). Across all tests, the **Attention (At)** mechanism consistently outperformed Concatenation and Addition, highlighting its ability to adapt to the sorption process's shifting gradients.

An observation in this PDE is the relationship between sampling density and model stability. In constant-timestep sampling, accuracy depends on the number of rollout steps; i.e., increasing the timestep (decreasing the number of rollouts) reduces error because there is less error propagation.

Variable-timestep sampling, however, offers a structural advantage that enables the model to remain competitive in large rollouts without suffering significant error propagation.

J.1 Analysis of Constant-Time Sampling and Initial Dynamics

Figure 20 illustrates the temporal error propagation for fixed timestep and reveals that the early-time phase is the primary driver of total error:

- **Early dynamics:** There is a clear separation in performance during the initial timesteps. Smaller constant timesteps achieve higher accuracy during the early phase of the rollout ($t < 20$) because they better resolve discontinuities and sharp gradients caused by rapid chemical interactions when the solute first encounters the solid matrix.
- **Late-Stage Convergence:** As the rollout progresses, the differences between various constant timesteps ($1\Delta t$ through $5\Delta t$) diminish. Once the initial sorption shock has passed and the system enters a more linear, diffusion-dominant regime, the model's error profiles converge, suggesting that temporal resolution becomes less critical in the later stages of evolution, allowing for larger timesteps.

J.2 Analysis of Variable-Time Sampling and Cumulative Error

The variable-timestep sampling results (Figures 21 & 22) show negligible difference and the cumulative error (Figure 23) reveals a familiar pattern:

- **Increasing and Decreasing (Figures 21 and 21)**: For increasing variable-timestep sampling (21), the figure shows the effect of transitioning to larger timesteps during the later stages of the evolution, where the dynamics are dominated by a smooth, low-gradient solution. Starting with a $1\Delta t$ baseline (left) and $2\Delta t$ (top-right), while the variable-timestep strategies do not outperform their respective constant-timestep counterparts, they achieve comparable accuracy with significantly fewer total rollouts. This demonstrates that once the initial discontinuity has dissipated, the model can remain stable at coarser temporal resolutions.

  For decreasing variable-timestep sampling (22), we observe a similar behaviour: decreasing variable-timestep sampling does not significantly affect the rollout accuracy, since all constant timesteps perform similarly in the later phase (figure 20).
- **Cumulative Error:** Figure 23 shows that starting with larger timesteps ($2\Delta t$ and $3\Delta t$) and moving to smaller ones ($1\Delta t$ and $2\Delta t \rightarrow 1\Delta t$), respectively, results in a lower cumulative error than the reverse case, i.e., stating with smaller timesteps. Although smaller timesteps are more accurate in the early phase, the higher number of rollouts in the high-error region leads to larger cumulative errors. This suggests that the early dynamics of the PDE are the most difficult for the model to capture. By using larger timesteps during the high-error early phase and smaller timesteps in the later stage, the model can reduce the total cumulative error.

	TIME	**VARIABLE-TIMESTEP MODEL**						
	CONDITION-EMBED	**CONSTANT-TIMESTEP**			**VARIABLE-TIMESTEP**			
	STRATEGY	1 (201 t_s)	2 (101 t_s)	3 (67 t_s)	1-2 (150 t_s)	2-1 (150 t_s)	1-2-3 (125 t_s)	3-2-1 (125 t_s)
Diffusion-Sorption (DS)	Ad	1.14	**1.00**	1.20	1.12	1.03	1.10	1.22
	Co	0.17	**0.15**	0.16	0.16	0.16	0.16	0.16
	At (**ours**)	0.13	0.12	0.12	0.13	0.12	0.12	0.12
	STE-Ad	1.0	**0.72**	1.46	0.86	0.86	0.75	1.47
	STE-Co	0.70	0.29	0.33	0.68	**0.28**	0.64	0.33
	STE-At (**ours**)	0.20	0.19	**0.16**	0.22	0.22	0.19	0.18
	EL-Ad	0.37	**0.29**	0.37	0.31	0.32	0.30	0.39
	EL-Co	0.19	**0.16**	0.17	0.18	0.17	0.17	0.17
	EL-At (**ours**)	0.14	0.12	**0.11**	0.13	0.12	0.13	0.12
	MLP-Ad	0.46	0.25	**0.18**	0.42	0.29	0.37	0.23
	MLP-Co	0.20	0.14	**0.13**	0.19	0.15	0.17	0.14
	MLP-At (**ours**)	0.13	0.12	0.12	0.13	0.12	0.13	0.12

Table 16: Total nRMSE for the Diffusion-Sorption equation. The EL-At architecture with constant $3\Delta t$ sampling achieves the lowest error.

	TIME CONDITION-EMBED STRATEGY	VARIABLE-TIMESTEP MODEL						
		CONSTANT-TIMESTEP			VARIABLE-TIMESTEP			
		1 (201 t_s)	2 (101 t_s)	3 (67 t_s)	1-2 (150 t_s)	2-1 (150 t_s)	1-2-3 (125 t_s)	3-2-1 (125 t_s)
Diffusion-Sorption (DS)	Ad	0.0051	0.0027	**0.0021**	0.0042	0.0048	0.0035	0.0056
	Co	0.0015	0.0010	**0.0009**	0.0012	0.0013	0.0011	0.0011
	At (**ours**)	0.0010	0.0008	0.0008	0.0008	0.0009	0.0008	0.0008
	STE-Ad	0.010	**0.0037**	0.010	0.0052	0.0091	0.0043	0.012
	STE-Co	0.0038	0.0013	0.0013	0.0092	0.0013	0.0027	0.0014
	STE-At (**ours**)	0.0013	0.0014	0.0010	0.0020	0.0027	**0.0011**	0.0019
	EL-Ad	0.0039	0.0022	0.0013	**0.0010**	0.0020	0.0011	0.0024
	EL-Co	0.0014	0.0010	0.0010	0.0011	0.0012	0.0010	0.0012
	EL-At (**ours**)	0.0013	0.0011	**0.0008**	0.0011	0.0011	0.0009	0.0010
	MLP-Ad	0.0067	0.0036	**0.0022**	0.0050	0.0049	0.0042	0.0041
	MLP-Co	0.0020	**0.0009**	0.0010	0.0012	0.0015	0.0096	0.0014
	MLP-At (**ours**)	0.0009	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008

Table 17: Prediction error at the final timestep for Diffusion-Sorption: There is no conclusive winner between constant and variable strategies, indicating that multiple temporal paths can converge to an accurate final state.

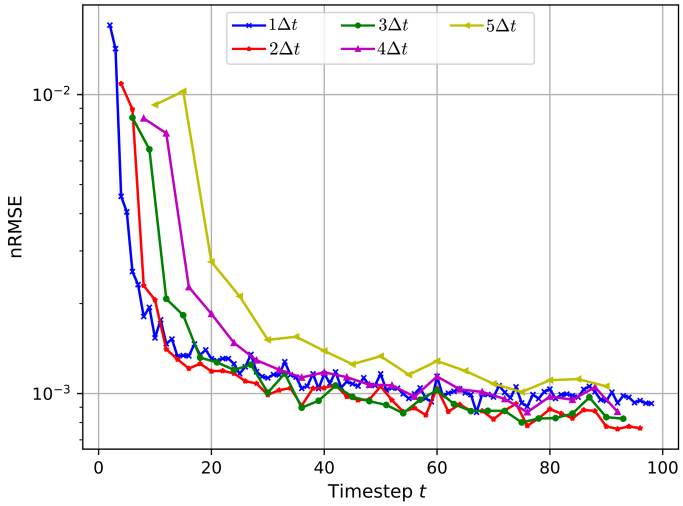

Figure 20: Temporal error propagation for constant timesteps. The high error at early timesteps highlights the difficulty of capturing initial sorption dynamics, with error profiles converging as the system stabilises.

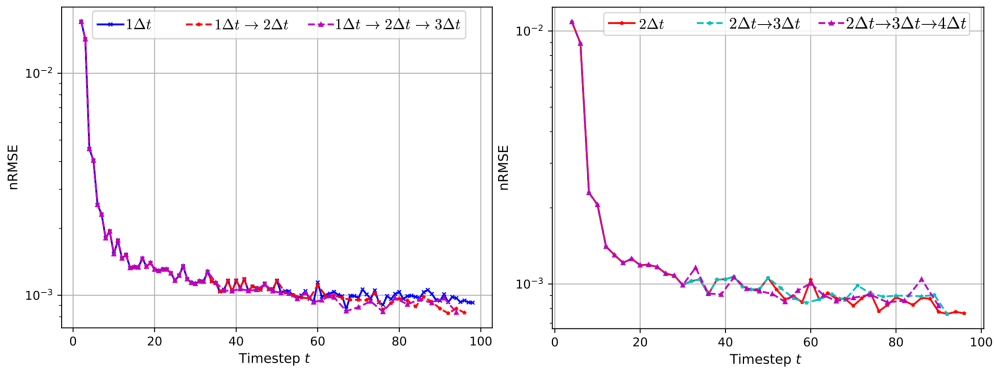

Figure 21: Performance of **increasing** variable-timestep sampling. Variable-timestep sampling does not improve over constant-timestep sampling.

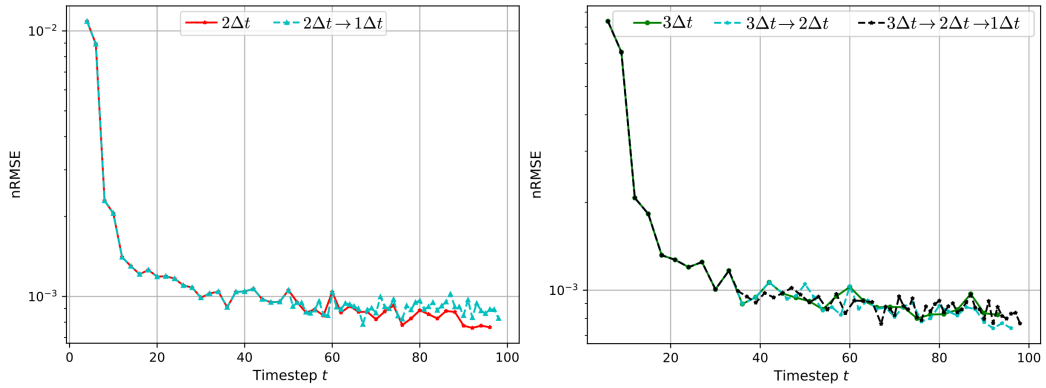

Figure 22: Performance of **decreasing** variable-timestep sampling. Variable-timestep sampling does not improve over constant-timestep sampling.

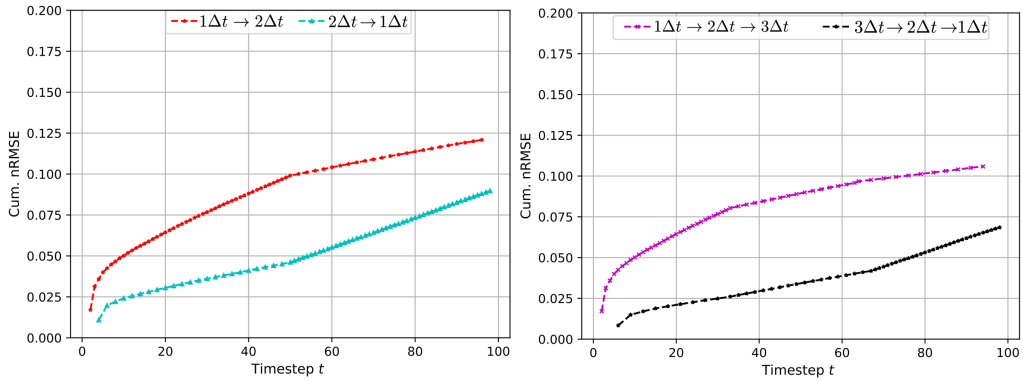

Figure 23: Cumulative error for Diffusion-Sorption. The higher error in the 1-2 1-2-3 sampling compared to 2-1 and 3-2-1 illustrates how a larger timestep during the high-error early phase can reduce overall error propagation.