Big Data Management & Analytics Master

Differential Equations Solver using NFTM for Burger Equations

Samuel CHAPUIS
samuel.chapuis@student-cs.fr

Lucia Victoria FERNANDEZ SANCHEZ
lucia-victoria.fernandez@student-cs.fr

Alexandra PERRUCHOT-TRIBOULET RODRIGUEZ
alexandra.perruchot-triboulet-rodriguez@student-cs.fr

Github Link

Advisor: Nacéra SEGHOUANI - Nacera.Seghouani@centralesupelec.fr
Advisor: Akash MALHOTRA - akash.malhotra@centralesupelec.fr

Contents

Chapter 1

Introduction

1.1 Context and motivation

1.1.1 Project goals

Our objective is to build a generic neural architecture that can internalize the governing rules of physical systems and then advance their fields over time. The model is trained on challenging PDEs with a focus on fluid mechanics (Navier–Stokes) to stress-test its ability to learn complex, multi-scale, nonlinear dynamics. Beyond matching the training horizon, the same architecture should simulate and extrapolate trajectories past the seen time window, remaining stable and accurate while generalizing across PDE families. Comparing these predictions to reference simulations allows us to identify which ingredients are essential for robustness and transfer.

1.1.2 Minimal mathematical frame

Many PDEs of interest can be written in a (semi-)invariant form:

$$\partial_t u(x,t) = \mathcal{F}\big(u(x,t), \nabla u(x,t), \nabla^2 u(x,t); \theta\big), \quad x \in \Omega,\ t > 0, \tag{1.1}$$

with initial/boundary conditions. Here u is a scalar or vector field; $\mathcal{F}$ encodes advection, diffusion, reaction, constraints; and θ collects physical parameters (viscosity ν, conductivity α, permeability $a(x)$, etc.). Our network seeks to learn a local time-advance operator that approximates $\mathcal{F}$ across multiple PDE families.

1.1.3 Historical milestones: AI and PDEs

- **1990s–2000s**: early universal approximators for simple PDEs; occasional RBF/MLP surrogates.
- **2017–2019**: *Physics-Informed Neural Networks* (PINNs) enforce PDE residuals in the loss; good on simple geometries, sensitive to stiffness and turbulent regimes.
- **2020**: *Fourier Neural Operator* (FNO) and *Neural Operators* learn the operator between function spaces; reference benchmarks on Burgers, incompressible Navier–Stokes, Darcy.
- **2021–2024**: rise of spatio-temporal architectures (Transformers, ConvNeXt) for continuous fields; work on numerical stability, spectral regularization, and conservation of physical quantities.

- **NFTM (2025)**: Neural Field Turing Machine (Malhotra & Seghouani) combines **continuous spatial memory + read/write heads + a neural controller** (CNN/RNN/Transformer). It resembles an explicit scheme: local read, compute an update, local write. Our work builds on this idea to generalize across PDEs.

1.2 Fluid Mechanics Overview

1.2.1 Why Navier–Stokes?

Navier–Stokes models is a canonical set of nonlinear PDEs governing fluid flow, encompassing a wide range of phenomena from laminar to turbulent regimes. Their complexity and multi-scale nature make them an ideal testbed for evaluating the capabilities of neural architectures in learning physical dynamics. Successfully modeling Navier–Stokes would demonstrate the potential of AI-driven approaches in computational fluid dynamics (CFD) and beyond.

1.2.2 Governing equations

The Navier–Stokes equations describe the motion of fluid substances and are derived from fundamental conservation laws: mass, momentum, and energy. They can be expressed as follows:

$$\text{Mass Conervation}: \quad \frac{\partial \rho}{\partial t} + \nabla\cdot(\rho\mathbf{u}) = 0 \tag{1.2}$$

$$\text{Momentum Conervation}: \quad \frac{\partial (\rho\mathbf{u})}{\partial t} + \nabla\cdot(\rho\mathbf{u}\otimes\mathbf{u}) = -\nabla p + \nabla\cdot\boldsymbol{\Sigma} + \rho\mathbf{g} \tag{1.3}$$

$$\text{Energy Conervation}: \quad \frac{\partial E}{\partial t} + \nabla\cdot((E+p)\mathbf{u}) = \nabla\cdot(\boldsymbol{\tau}\cdot\mathbf{u}) - \nabla\cdot\mathbf{q} + \rho\mathbf{u}\cdot\mathbf{g} \tag{1.4}$$

Where all the variables are defined as:

- ρ: fluid density [kg·m^{-3}]
- $\mathbf{u} = (u, v, w)$: velocity field [m·s^{-1}]
- $\nabla\cdot(\rho\mathbf{u})$: divergence of mass flux
- $\partial\rho/\partial t$: local time rate of change of density
- $\rho\mathbf{u}$: momentum density [kg·m^{-2}·s^{-1}]
- $\nabla\cdot(\rho\mathbf{u}\otimes\mathbf{u})$: convective momentum transport
- $-\nabla p$: pressure gradient force per unit volume

- $\mathbf{\Sigma}$: viscous stress tensor [Pa]
- $\rho\mathbf{g}$: body force density (e.g. gravity) [N·m^{-3}]
- $E = \rho\big(e + \frac{1}{2}|\mathbf{u}|^2\big)$: total energy density (internal + kinetic)
- p: pressure [Pa]
- $\boldsymbol{\tau}$: stress tensor (viscous + pressure)
- $\mathbf{q}$: heat flux vector [W·m^{-2}]
- $\rho\mathbf{u}\cdot\mathbf{g}$: work done by body forces

These equations require a viscosity model (e.g., Newtonian fluid) and an equation of state (e.g., ideal gas law or van der Waals equation) to close the system. They form the foundation for simulating fluid dynamics in various applications, from aerodynamics to weather forecasting.

$$\text{Newtonian fluid model :} \quad \begin{cases} \mathbf{\Sigma} = \mu\left(\nabla\mathbf{u} + (\nabla\mathbf{u})^T\right) + \lambda(\nabla\cdot\mathbf{u})\mathbf{I}, \\ \boldsymbol{\tau} = \mathbf{\Sigma} - p\mathbf{I} \end{cases} \tag{1.5}$$

$$\text{Ideal gas :} \quad p = \rho R T \tag{1.6}$$

$$\text{van der Waals :} \quad \left(p + a\left(\frac{n}{V}\right)^2\right)(V - nb) = nRT \tag{1.7}$$

1.2.3 Burger's equation as a starting point

As the project starts, we focus the goal was to use a model simple enough to be learned with limited data and computational resources, yet rich enough to exhibit complex dynamics. Thus, we begin with the burger s equation, a simplified 1D version of the Navier–Stokes equations that captures some features of fluid dynamics, such as shock formation and nonlinear advection. The viscous Burger's equation in 1D is given by:

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \nu\frac{\partial^2 u}{\partial x^2}, \quad x \in [0, L],\ t > 0, \tag{1.8}$$

Chapter 2

Related Work

It is expected to find:

- The research articles related to your defined problem/objective
- Try to present them by category, the main ideas behind and their limitations
- Justify your directions/choices. This part should make a link with the next chapter.

Chapter 3
Background

The objective here is to detail the main concepts/definitions existing algorithms needed to understand your work to be detailed later and to introduce the notations to be used.

Use examples

Don't forget to cite again these existing approaches

Chapter 4

Our Methodology and Approach

This chapter presents our comprehensive methodology for solving the one-dimensional Burgers equation using Neural Field Turing Machines (NFTMs). We detail the complete pipeline from data generation through model architecture, training procedures, and evaluation metrics. Our approach builds upon the NFTM framework while introducing several key innovations to improve stability, physical consistency, and generalization capabilities.

4.1 Overview of the Proposed Pipeline

Our neural PDE solver pipeline consists of four main components, illustrated in Figure **??**:

1. **Physics-Based Data Generation**: High-fidelity numerical solutions of the Burgers equation using stable finite difference methods
2. **Neural Architecture**: A causal temporal attention mechanism combined with convolutional decoders
3. **Training Strategy**: Multi-stage curriculum learning with physics-informed regularization
4. **Evaluation Framework**: Comprehensive metrics assessing accuracy, stability, and physical consistency

The complete workflow operates as follows: training trajectories are generated by solving the Burgers equation numerically for various initial conditions and viscosity parameters. These trajectories are then used to train a neural network that learns to predict future states given a window of historical states. During inference, the model performs autoregressive rollout to generate long-term predictions. Finally, predictions are evaluated against ground truth using multiple metrics to assess both numerical accuracy and physical plausibility.

4.2 Data Generation: Stable Numerical Solver

4.2.1 Finite Difference Scheme

To generate high-quality training data, we implement a stable numerical solver for the one-dimensional Burgers equation:

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \nu\frac{\partial^2 u}{\partial x^2} \tag{4.1}$$

We discretize the spatial domain $x \in [0, 2\pi]$ with periodic boundary conditions using $N = 128$ uniformly spaced grid points:

$$x_i = \frac{2\pi i}{N}, \quad i = 0, 1, \ldots, N-1 \tag{4.2}$$

The spatial step size is:

$$\Delta x = \frac{2\pi}{N} \approx 0.049 \tag{4.3}$$

Diffusion Term. The viscous diffusion term is discretized using central differences, which is second-order accurate in space:

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_{x_i} \approx \frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta x^2} = D_i \tag{4.4}$$

This stencil is applied with periodic boundary conditions: $u_{-1} = u_{N-1}$ and $u_N = u_0$.

Advection Term. For the nonlinear advection term $u\frac{\partial u}{\partial x}$, we employ an upwind differencing scheme to maintain numerical stability:

$$\left.\frac{\partial u}{\partial x}\right|_{x_i} \approx \begin{cases} \frac{u_i - u_{i-1}}{\Delta x} & \text{if } u_i > 0 \text{ (backward difference)} \\ \frac{u_{i+1} - u_i}{\Delta x} & \text{if } u_i \leqslant 0 \text{ (forward difference)} \end{cases} \tag{4.5}$$

The upwind scheme selects the differencing direction based on the local flow direction, which prevents spurious oscillations in regions with steep gradients or shocks.

Time Integration. We use the forward Euler method for time integration with timestep $\Delta t = 0.01$:

$$u_i^{n+1} = u_i^n - \Delta t \cdot u_i^n \left(\frac{\partial u}{\partial x}\right)_i^n + \Delta t \cdot \nu \cdot D_i^n \tag{4.6}$$

To ensure stability, the timestep satisfies the CFL (Courant-Friedrichs-Lewy) condition:

$$\Delta t \leqslant \min\left(\frac{\Delta x}{|u_{\max}|}, \frac{\Delta x^2}{2\nu}\right) \tag{4.7}$$

For our parameters ($\Delta x \approx 0.049$, $\nu \in [0.01, 0.1]$, $|u_{\max}| \lesssim 1$), $\Delta t = 0.01$ satisfies this stability criterion.

4.2.2 Dataset Construction

Sampling Strategy. For each trajectory in our dataset, we randomly sample:

- **Viscosity parameter**: $\nu \sim \mathcal{U}(0.01, 0.1)$, where $\mathcal{U}(a, b)$ denotes the uniform distribution
- **Initial condition**: $u_0(x) = \sin(x)$ (smooth sinusoidal profile)

The sinusoidal initial condition is chosen because:

1. It is smooth and differentiable (realistic physical scenario)

2. It produces rich dynamics including wave steepening and eventual shock formation for small ν
3. It has zero mean: $\int_0^{2\pi} \sin(x)\, dx = 0$, which tests mass conservation
4. It provides analytical benchmarks via the Cole-Hopf transformation

Trajectory Length. Each trajectory consists of $T = 100$ timesteps, spanning total time $T_{\text{total}} = 1.0$ seconds. The state at each timestep is stored:

$$\mathcal{T}_{\text{trajectory}} = \{u^0, u^1, \ldots, u^{99}\}, \quad u^t \in \mathbb{R}^{128} \tag{4.8}$$

Dataset Split. We generate:

- **Training set**: 1000 trajectories (100,000 state snapshots)
- **Test set**: 200 trajectories (20,000 state snapshots)

No trajectory in the test set has the same viscosity parameter ν as any training trajectory, ensuring evaluation on truly unseen parameter regimes.

4.2.3 Quality Assurance

We verify the quality of generated data through:

Mass Conservation. For periodic boundary conditions, the total mass $M(t) = \int_0^{2\pi} u(x,t)\, dx$ should be conserved. Numerically:

$$M^n = \Delta x \sum_{i=0}^{N-1} u_i^n \tag{4.9}$$

We verify:

$$\left| \frac{M^{99} - M^0}{M^0 + \varepsilon} \right| < 10^{-3} \tag{4.10}$$

where $\varepsilon = 10^{-6}$ prevents division by zero for the sinusoidal IC where $M^0 \approx 0$.

Energy Dissipation. The kinetic energy:

$$E^n = \frac{\Delta x}{2} \sum_{i=0}^{N-1} (u_i^n)^2 \tag{4.11}$$

must monotonically decrease due to viscous dissipation:

$$E^{n+1} \leqslant E^n + \varepsilon_{\text{tol}} \tag{4.12}$$

where $\varepsilon_{\text{tol}} = 10^{-6}$ accounts for numerical roundoff.

Stability Check. We reject any trajectory where:

$$\max_{t,i} |u_i^t| > 5 \quad \text{or} \quad \exists\,(t,i) : \text{isnan}(u_i^t) \vee \text{isinf}(u_i^t) \tag{4.13}$$

These checks ensure only physically plausible and numerically stable trajectories enter the training dataset.

4.3 Neural Architecture: Improved NFTM with Temporal Attention

Our neural architecture enhances the basic NFTM framework with causal temporal attention and multi-scale processing.

4.3.1 Input Representation and Window-Based Prediction

Historical Window. Rather than predicting the next state from a single previous state, we use a sliding window of length W:

$$\mathbf{U}_{\text{hist}} = [u^{t-W+1}, u^{t-W+2}, \ldots, u^t] \in \mathbb{R}^{W \times N} \tag{4.14}$$

In our implementation, $W = 20$, providing the model with approximately 0.2 seconds of history to capture temporal dynamics.

Prediction Task. Given $\mathbf{U}_{\text{hist}}$, the model predicts the next state:

$$\hat{u}^{t+1} = f_\theta(\mathbf{U}_{\text{hist}}) \tag{4.15}$$

For autoregressive rollout over K steps:

$$\begin{aligned}
\hat{u}^{t+1} &= f_\theta([u^{t-W+2}, \ldots, u^t]) \\
\hat{u}^{t+2} &= f_\theta([u^{t-W+3}, \ldots, u^t, \hat{u}^{t+1}]) \\
&\vdots \\
\hat{u}^{t+K} &= f_\theta([\hat{u}^{t+K-W}, \ldots, \hat{u}^{t+K-1}])
\end{aligned} \tag{4.16}$$

This autoregressive formulation allows arbitrary-length predictions but accumulates errors over time.

4.3.2 Causal Temporal Attention Mechanism

Motivation. The temporal attention mechanism allows the model to selectively weight different historical timesteps based on their relevance to the current prediction. This is particularly important for PDEs where recent states may be more informative than distant past states, but long-range temporal dependencies still matter.

Architecture Overview. The causal temporal attention module processes the history window $\mathbf{U}_{\text{hist}} \in \mathbb{R}^{B\times W\times N}$ (batch size B, window size W, spatial points N) in three stages:

1. **Feature Embedding**: Each timestep is embedded into a high-dimensional feature space
2. **Attention Computation**: Query-Key-Value attention focuses on relevant historical information
3. **Context Aggregation**: Weighted temporal features are aggregated into a context vector

Feature Embedding. First, we embed each spatial field into a feature space using a 1D convolutional layer:

$$\mathbf{H}^t = \text{GELU}(\text{Conv1D}(u^t; \theta_{\text{emb}})) \tag{4.17}$$

where $\mathbf{H}^t \in \mathbb{R}^{d\times N}$ and $d = 32$ is the embedding dimension.

For the full window:

$$\mathbf{H} = [\mathbf{H}^{t-W+1}, \dots, \mathbf{H}^t] \in \mathbb{R}^{W\times d\times N} \tag{4.18}$$

Query-Key-Value Projections. We use the most recent frame $\mathbf{H}^t$ to generate queries $\mathbf{Q}$, while all historical frames generate keys $\mathbf{K}$ and values $\mathbf{V}$:

$$\mathbf{Q} = \text{Conv1D}(\mathbf{H}^t; \theta_Q) \in \mathbb{R}^{d\times N} \tag{4.19}$$

$$\mathbf{K}^\tau = \text{Conv1D}(\mathbf{H}^\tau; \theta_K) \in \mathbb{R}^{d\times N}, \quad \tau \in [t-W+1, t] \tag{4.20}$$

$$\mathbf{V}^\tau = \text{Conv1D}(\mathbf{H}^\tau; \theta_V) \in \mathbb{R}^{d\times N} \tag{4.21}$$

Attention Scores. The attention scores measure similarity between the current query and past keys:

$$\alpha_i^\tau = \frac{\mathbf{Q}_i^T \mathbf{K}_i^\tau}{\sqrt{d}}, \quad i = 1, \dots, N \tag{4.22}$$

Normalization via softmax across the temporal dimension yields attention weights:

$$a_i^\tau = \frac{\exp(\alpha_i^\tau)}{\sum_{\tau'=t-W+1}^{t} \exp(\alpha_i^{\tau'})} \tag{4.23}$$

This ensures:

$$\sum_{\tau=t-W+1}^{t} a_i^\tau = 1, \quad a_i^\tau \geqslant 0 \tag{4.24}$$

Context Aggregation. The weighted sum of value vectors produces the context representation:

$$\mathbf{C}_i = \sum_{\tau=t-W+1}^{t} a_i^\tau \mathbf{V}_i^\tau \tag{4.25}$$

Output Projection and Residual Connection. The context is projected back to the feature dimension and combined with the last frame via residual connection:

$$\mathbf{F} = \text{GroupNorm}(\text{Conv1D}(\mathbf{C}; \theta_{\text{out}}) + \mathbf{H}^t) \tag{4.26}$$

The GroupNorm stabilizes training by normalizing activations across channel groups.

4.3.3 Convolutional Decoder

The decoder transforms the attention-enriched features $\mathbf{F} \in \mathbb{R}^{d \times N}$ into spatial corrections $\Delta u \in \mathbb{R}^{1 \times N}$:

$$\begin{aligned} \mathbf{D}_1 &= \text{GELU}(\text{BatchNorm}(\text{Conv1D}(\mathbf{F}; \text{kernel} = 5))) \\ \Delta u &= \text{Conv1D}(\mathbf{D}_1; \text{kernel} = 5, \text{out\_channels} = 1) \end{aligned} \tag{4.27}$$

Multi-Scale Receptive Field. The two convolutional layers with kernel size 5 provide an effective receptive field:

$$R_{\text{eff}} = 1 + 2 \times (5 - 1) = 9 \tag{4.28}$$

This covers approximately $\frac{9}{128} \times 2\pi \approx 0.44$ spatial units, sufficient to capture local gradients and shock structures.

Bounded Correction via Tanh. To prevent instabilities, corrections are bounded:

$$\Delta u_{\text{bounded}} = \tanh(\Delta u) \cdot c_{\text{clip}} \tag{4.29}$$

where $c_{\text{clip}} = 0.1$. This ensures:

$$|\Delta u_{\text{bounded}}| \leqslant 0.1 \tag{4.30}$$

preventing single-step divergence during autoregressive rollout.

4.3.4 Final Prediction

The next state is computed via residual update:

$$\hat{u}^{t+1} = u^t + \Delta u_{\text{bounded}} \tag{4.31}$$

This formulation has two advantages:

1. **Easier Optimization**: Learning perturbations Δu rather than absolute states u^{t+1} simplifies the learning task
2. **Physical Alignment**: For small timesteps, the true dynamics satisfy $u^{t+1} \approx u^t + \Delta t \cdot f(u^t)$, matching our residual formulation

4.4 Training Strategy

4.4.1 Loss Function Design

Our loss function combines multiple terms to enforce both accuracy and physical consistency.

Mean Squared Error (MSE). The primary loss is the L2 prediction error:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{BN} \sum_{b=1}^{B} \sum_{i=1}^{N} (\hat{u}_i^b - u_i^b)^2 \tag{4.32}$$

Gradient Matching Loss. To preserve spatial structures and shocks, we match spatial gradients:

$$\mathcal{L}_{\text{grad}} = \frac{1}{BN} \sum_{b=1}^{B} \sum_{i=1}^{N} \left(\frac{\partial \hat{u}_i^b}{\partial x} - \frac{\partial u_i^b}{\partial x} \right)^2 \tag{4.33}$$

Gradients are computed via central differences:

$$\frac{\partial u_i}{\partial x} \approx \frac{u_{i+1} - u_{i-1}}{2\Delta x} \tag{4.34}$$

Energy Dissipation Constraint. The Burgers equation dissipates energy due to viscosity:

$$E(t + \Delta t) \leqslant E(t) \tag{4.35}$$

We enforce this via a penalty:

$$\mathcal{L}_{\text{energy}} = \frac{1}{B} \sum_{b=1}^{B} \text{ReLU} \left(\frac{E(\hat{u}^b) - E(u^{t,b})}{E(u^{t,b}) + \varepsilon} \right) \tag{4.36}$$

where:

$$E(u) = \frac{\Delta x}{2} \sum_{i=1}^{N} u_i^2 \tag{4.37}$$

The ReLU ensures we only penalize energy *increases*, not decreases. Normalization by $E(u^t)$ makes the penalty scale-invariant.

Combined Loss. The total loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}} + \lambda_{\text{grad}} \mathcal{L}_{\text{grad}} + \lambda_{\text{energy}} \mathcal{L}_{\text{energy}} \tag{4.38}$$

with weights $\lambda_{\text{grad}} = 0.1$ and $\lambda_{\text{energy}} = 0.05$.

4.4.2 Curriculum Learning Strategy

To improve long-term stability, we employ curriculum learning that gradually increases task difficulty.

Rollout Depth Scheduling. Early in training, we use short rollouts; later, we extend to longer predictions:

$$K_{\text{rollout}}(\text{epoch}) = \begin{cases} 8 & \text{epoch} < 10 \\ 16 & 10 \leqslant \text{epoch} < 30 \\ 32 & \text{epoch} \geqslant 30 \end{cases} \tag{4.39}$$

During each training iteration, we unroll for K_{rollout} steps and backpropagate through the entire sequence:

$$\mathcal{L}_{\text{rollout}} = \frac{1}{K_{\text{rollout}}} \sum_{k=1}^{K_{\text{rollout}}} \mathcal{L}_{\text{total}}(\hat{u}^{t+k}, u^{t+k}) \tag{4.40}$$

Teacher Forcing. To prevent error accumulation during training, we occasionally use ground truth states instead of predictions:

$$u_{\text{next}}^{t+k} = \begin{cases} u_{\text{true}}^{t+k} & \text{with probability } p_{\text{TF}} \\ \hat{u}^{t+k} & \text{with probability } 1 - p_{\text{TF}} \end{cases} \tag{4.41}$$

The teacher forcing rate decays over training:

$$p_{\text{TF}}(\text{epoch}) = \begin{cases} 0.5 & \text{epoch} < 10 \\ 0.2 & 10 \leqslant \text{epoch} < 30 \\ 0.05 & \text{epoch} \geqslant 30 \end{cases} \tag{4.42}$$

This provides stability early (when the model is weak) while ensuring the model learns to handle its own predictions later.

4.4.3 Noise Injection for Robustness

After epoch 20, we inject Gaussian noise into the input history to improve robustness:

$$\tilde{\mathbf{U}}_{\text{hist}} = \mathbf{U}_{\text{hist}} + \varepsilon \mathcal{N}(0, I) \tag{4.43}$$

where:

$$\varepsilon = 0.01 \times \min\left(1.0, \frac{\text{epoch}}{50}\right) \tag{4.44}$$

This prevents overfitting to exact trajectories and improves generalization to perturbed initial conditions.

4.4.4 Optimization Details

Optimizer. We use AdamW optimizer with:

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} - \lambda_{\text{WD}} \theta_t \tag{4.45}$$

where:

- Learning rate: $\eta = 0.001$
- Weight decay: $\lambda_{\text{WD}} = 10^{-4}$
- $\beta_1 = 0.9$, $\beta_2 = 0.999$
- $\varepsilon = 10^{-8}$

Learning Rate Schedule. Cosine annealing reduces the learning rate:

$$\eta_t = \eta_{\min} + \frac{\eta_{\max} - \eta_{\min}}{2}\left(1 + \cos\left(\frac{t}{T_{\max}}\pi\right)\right) \tag{4.46}$$

with $\eta_{\max} = 0.001$, $\eta_{\min} = 0$, and $T_{\max} = 50$ epochs.

Gradient Clipping. To prevent gradient explosion during backpropagation through time:

$$\|\nabla_\theta \mathcal{L}\| > 1.0 \implies \nabla_\theta \mathcal{L} \leftarrow \frac{\nabla_\theta \mathcal{L}}{\|\nabla_\theta \mathcal{L}\|} \tag{4.47}$$

Batch Size and Epochs.

- Batch size: $B = 32$
- Total epochs: 50
- Training samples per epoch: $\frac{1000}{32} \approx 31$ batches

4.5 Evaluation Metrics

We employ comprehensive metrics to assess model performance across multiple dimensions.

4.5.1 Accuracy Metrics

Mean Squared Error (MSE).

$$\text{MSE} = \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}(\hat{u}_i^t - u_i^t)^2 \tag{4.48}$$

Relative L2 Error. Scale-invariant error:

$$E_{L^2}^{\text{rel}} = \frac{\|\hat{\mathbf{u}} - \mathbf{u}\|_2}{\|\mathbf{u}\|_2 + \varepsilon} \tag{4.49}$$

Per-Timestep Relative L2.

$$E_{L^2}^{\text{rel}}(t) = \frac{\sqrt{\sum_{i=1}^{N}(\hat{u}_i^t - u_i^t)^2}}{\sqrt{\sum_{i=1}^{N}(u_i^t)^2} + \varepsilon} \tag{4.50}$$

This reveals when predictions diverge during rollout.

Peak Signal-to-Noise Ratio (PSNR).

$$\text{PSNR} = 20\log_{10}\left(\frac{\text{RANGE}}{\sqrt{\text{MSE}}}\right) \tag{4.51}$$

where $\text{RANGE} = \max(u) - \min(u)$.

Higher PSNR indicates better reconstruction quality.

Structural Similarity Index (SSIM). Measures perceptual similarity:

$$\mathrm{SSIM}(\mathbf{u}, \hat{\mathbf{u}}) = \frac{(2\mu_u\mu_{\hat{u}} + C_1)(2\sigma_{u\hat{u}} + C_2)}{(\mu_u^2 + \mu_{\hat{u}}^2 + C_1)(\sigma_u^2 + \sigma_{\hat{u}}^2 + C_2)} \tag{4.52}$$

SSIM $\in [-1, 1]$, with 1 indicating perfect structural match.

4.5.2 Physical Consistency Metrics

Mass Conservation Error. For sinusoidal IC where $M_0 \approx 0$:

$$E_{\text{mass}}(t) = |M(t) - M_0| \tag{4.53}$$

For non-zero initial mass:

$$E_{\text{mass}}^{\text{rel}}(t) = \frac{|M(t) - M_0|}{|M_0| + \varepsilon} \tag{4.54}$$

Energy Dissipation Consistency. Verify that predicted energy decreases:

$$\text{Monotonicity} = \frac{1}{T-1}\sum_{t=1}^{T-1} \nVdash[E(t+1) \leqslant E(t) + \varepsilon_{\text{tol}}] \tag{4.55}$$

Perfect monotonicity gives 1.0; violations reduce this score.

4.5.3 Correlation and Spectral Metrics

Pearson Correlation per Timestep.

$$\rho(t) = \frac{\mathrm{Cov}(\hat{u}^t, u^t)}{\sigma_{\hat{u}^t}\sigma_{u^t}} \tag{4.56}$$

High correlation ($\rho \approx 1$) indicates the model captures spatial patterns accurately.

Energy Spectrum Error. Fourier analysis reveals whether the model captures correct wavenumber content:

$$\hat{U}^t(k) = \mathrm{FFT}(u^t), \quad S(k) = |\hat{U}(k)|^2 \tag{4.57}$$

Spectrum error:

$$E_{\text{spec}} = \frac{1}{K}\sum_{k=1}^{K/2} |\log_{10} S_{\text{pred}}(k) - \log_{10} S_{\text{true}}(k)| \tag{4.58}$$

4.6 Implementation Details

4.6.1 Software and Hardware

Framework. All models are implemented in PyTorch 2.0+ with CUDA support for GPU acceleration.

Hardware. Training is performed on:

- GPU: NVIDIA A100 (40GB) or equivalent
- Training time: ∼30 minutes for 50 epochs
- Inference time: ∼10ms per rollout step

4.6.2 Reproducibility

To ensure reproducibility:

- Random seeds fixed: PyTorch, NumPy, Python (`seed=42`)
- Deterministic algorithms enabled where available
- Complete hyperparameter logging
- Model checkpoints saved every 10 epochs

4.6.3 Code Organization

Our implementation follows modular design:

```
burgers_solver/
|-- data_generation.py      # Finite difference solver
|-- dataset.py              # PyTorch Dataset/DataLoader
|-- model.py                # Neural architecture
|   |-- CausalTemporalAttention
|   +-- ImprovedBurgersNet
|-- training.py             # Training loop
|   |-- Loss functions
|   +-- Curriculum learning
|-- metrics.py              # Evaluation metrics
+-- visualization.py        # Plotting utilities
```

The complete source code and pre-trained models are available in our GitHub repository[1].

4.7 Summary

This chapter has presented our comprehensive methodology for neural PDE solving:

1. **High-quality data generation** using stable finite difference methods with upwind advection and central difference diffusion
2. **Novel architecture** combining causal temporal attention with convolutional decoding for effective spatiotemporal modeling

[1] https://github.com/Samuel-Chapuis/ML_Differential_Solver

3. **Physics-informed training** via energy dissipation constraints and gradient matching losses
4. **Curriculum learning** with scheduled rollout depth and teacher forcing for improved stability
5. **Comprehensive evaluation** across accuracy, physical consistency, and spectral metrics

The next chapter presents experimental results demonstrating the effectiveness of these design choices and comparing performance against baseline methods.

Transition to Chapter 5. Having established our complete methodological framework, we now turn to empirical validation. Chapter **??** presents comprehensive experiments evaluating our approach across multiple metrics, analyzing the contribution of each architectural component through ablation studies, and comparing performance against established baseline methods.

Chapter 5

Approaches for Neural PDE Simulators

5.1 Approach 1: Samuel's Model

5.1.1 Model Architecture

5.1.2

5.2 Approach 2: Causal Temporal Convolutional Attention Network

5.2.1 Model Architecture

The Causal Temporal Convolutional Attention Network (TCAN) is a feed-forward autoregressive model that predicts next state of a PDE given a sliding window of past states. At each time step, the model receives a history window u_{history} of shape (B, W, P)—where B is the batch size, W is the history length (e.g., 20), and P is the number of spatial grid points—and outputs the next field u_{t+1} of shape $(B, 1, P)$. The model operates as a one-step neural PDE surrogate, repeatedly applied in an autoregressive rollout to generate full trajectories.

The architecture consists of three main components:
1. A temporal attention encoder that aggregates information across the history window.
2. A convolutional decoder that produces a bounded correction.
3. A residual update that adds the correction to the last frame.

The full data flow for a single prediction step is illustrated in Figure 5.1.

5.3 Temporal Attention Encoder

The `CausalTemporalAttention` module processes the history window through the following steps:

1. **Frame-wise embedding**: Reshape u_{history} to $(B \cdot W, 1, P)$ and apply a 1D convolution with kernel size 3 followed by GELU activation, producing features of shape $(B \cdot W, C, P)$ where $C = 32$ is the embedding dimension.

2. **Temporal restructuring**: Reshape features back to (B, W, C, P) to expose the temporal dimension.

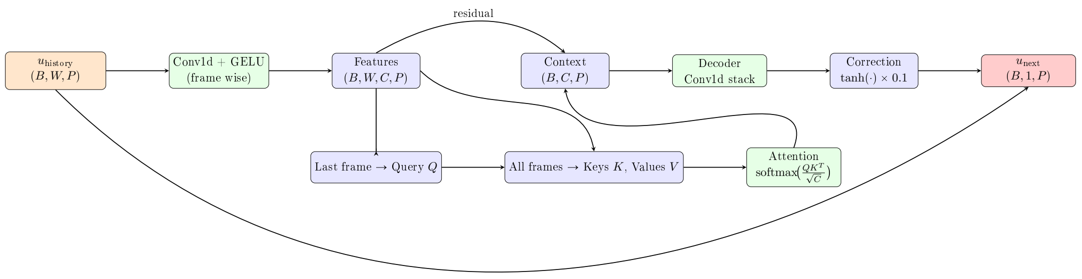
residual
u_{history} (B, W, P)
Conv1d + GELU (frame wise)
Features (B, W, C, P)
Context (B, C, P)
Decoder Conv1d stack
Correction $\tanh(\cdot) \times 0.1$
u_{next} $(B, 1, P)$
Last frame → Query Q
All frames → Keys K, Values V
Attention softmax$(\frac{QK^T}{\sqrt{C}})$

Figure 5.1: Data flow through the Causal Temporal Attention Network during one prediction step.

3. **Query computation**: Extract features of the last frame features$[:, -1, :, :] \in \mathbb{R}^{B \times C \times P}$ and apply a 1×1 convolution to obtain queries $Q \in \mathbb{R}^{B \times C \times P}$.
4. **Key and value projection**: Apply 1×1 convolutions to all W frames (flattened to $(B \cdot W, C, P)$ then reshaped) to obtain keys K and values V, both of shape (B, W, C, P).
5. **Causal attention scores**: Compute scaled dot-product attention over the temporal dimension:

$$\text{scores}_{b,w,p} = \frac{\langle Q_b, K_{b,w} \rangle}{\sqrt{C}}, \quad \alpha_{b,w,p} = \text{softmax}_w(\text{scores}_{b,w,p})$$

6. **Context aggregation**: Compute the attended context:

$$\text{context}_b = \sum_{w=1}^{W} \alpha_{b,w} \cdot V_{b,w} \in \mathbb{R}^{B \times C \times P}$$

7. **Residual projection**: Apply a 1×1 convolution to the context, add the residual connection to the last frame features, and normalize with GroupNorm.

This mechanism allows the last frame to dynamically query relevant past frames at each spatial location, producing spatially-aware temporal context features.

5.4 Decoder and Residual Correction

The `ImprovedBurgersNet` decoder processes the attention output context $\in \mathbb{R}^{B \times C \times P}$ as follows:

1. **Convolutional decoder stack**:
   - Conv1d: $C \to 32$ channels, kernel size 5, BatchNorm1d, GELU.
   - Conv1d: $32 \to 1$ channel, kernel size 5, producing raw_correction $\in \mathbb{R}^{B \times 1 \times P}$.

The final layer is zero-initialized to ensure near-zero corrections during early training.

2. **Bounded correction**: Apply hyperbolic tangent clipping scaled by corr_clip = 0.1:

$$\Delta u = \tanh(\text{raw\_correction}) \cdot 0.1$$

3. **Residual update**: Add the correction to the last input frame:

$$u_{\text{next}} = u_{\text{history}}[:, -1 :, :] + \Delta u$$

This design enforces incremental updates, preventing instability during long autoregressive rollouts.

5.5 Autoregressive Training Procedure

Training employs curriculum learning with multi-step rollouts:

1. Initialize current_window $= u_{\text{gt}}[:, : W, :]$ from ground truth trajectories.
2. For each rollout step $k \in \{0, \ldots, D-1\}$ where D is the rollout depth (curriculum: 8→16→64):
   - Predict $\hat{u}_{W+k} = \text{TCAN}(\text{current\_window})$.
   - Compute composite loss:
   $$\mathcal{L} = \text{MSE}(\hat{u}, u_{\text{gt}}) + 0.1 \cdot \|\nabla_x \hat{u} - \nabla_x u_{\text{gt}}\|_2^2 + 0.05 \cdot \mathbb{E}[\max(0, E(\hat{u}) - E(u_t))]$$
   where $E(u) = \frac{1}{2}\int u^2\, dx$ enforces energy dissipation.
   - Update window: current_window $\leftarrow$ [current_window$[:, 1 :, :], \hat{u}$] (with occasional teacher forcing).
3. Average loss over rollout, backpropagate through unrolled computation graph, apply gradient clipping, and optimize with AdamW + cosine annealing.

5.6 Evaluation Metrics

Full-trajectory predictions are evaluated using:

- Standard: MSE, relative L^2, PSNR, SSIM, temporal correlation.
- Physics-informed: mass conservation error, energy monotonicity fraction, mean PDE residual, spectral error, max gradient error.

This comprehensive evaluation ensures both accuracy and physical fidelity across long rollouts.

Chapter 6

Conclusion and Perspectives

A summary of your work. More focused on the results

The limitations of the work -> which perspectives/clues to deal with limitations, to improve your work

the last paragraph must be dedicated to the work in team

6.1 Gl remarks

GENERAL :

Each table, figure must be cited and explained in the text.

The references must be complete

Each chapter must start with a paragraph to introduce its content (no need to have a separated for that), except the introduction and the conclusion. In the same manner each chapter must finish with a paragraph to conclude and to make a link with the next one, except the introduction and the conclusion.

6.2 Gl remarks about the presentation

The slides must be numbered

The presentation follows more and less the structure of the report

No too much blabla about the the gl context you need to define the objectives of the project (with examples if possible) ...

Then how your work fits into existing works (some main related works), the overall pipeline, your main contributions in this pipeline

also your main contributions in terms of implementation

the main results

Conclusions and next ...

Then the overall

Appendix A

Appendix

The progress draft must be included in the appendix

Bibliography

Abstract:

Keywords: