AQTIVATE Conference - *September 29 - 2026*

# AI for Physical Dynamics: From Neural Operators to Foundation Models

Patrick Gallinari, https://pages.isir.upmc.fr/gallinari/
patrick.gallinari@sorbonne-universite.fr,
p.gallinari@criteo.com

CRITEO

SORBONNE UNIVERSITÉ

# Outline

- Context
  - Modeling spatio-temporal physical dynamics
  - Motivating example: weather forecasting
- Neural operators
- Foundation models
- Scaling neural operators & Datasets

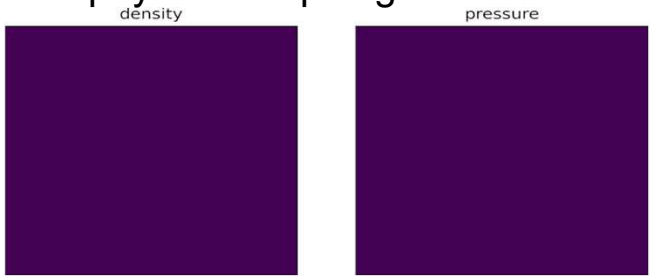
# Context: Modeling spatio-temporal physical dynamics

Data-driven Neural PDE solvers

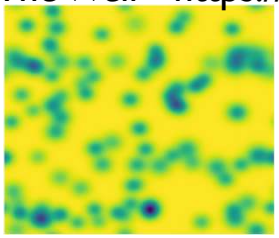
# Modeling spatio-temporal physical dynamics

- Examples: spatio temporal dynamics

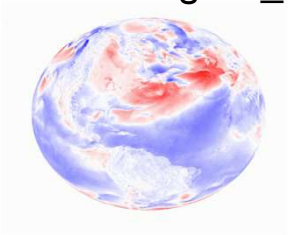
Astrophysics - https://github.com/leo1200/astronomix




The Well - https://polymathic-ai.org/the_well/



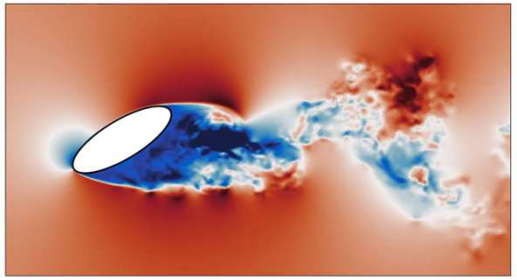
Chemical reaction



Weather Forecast

Fluid

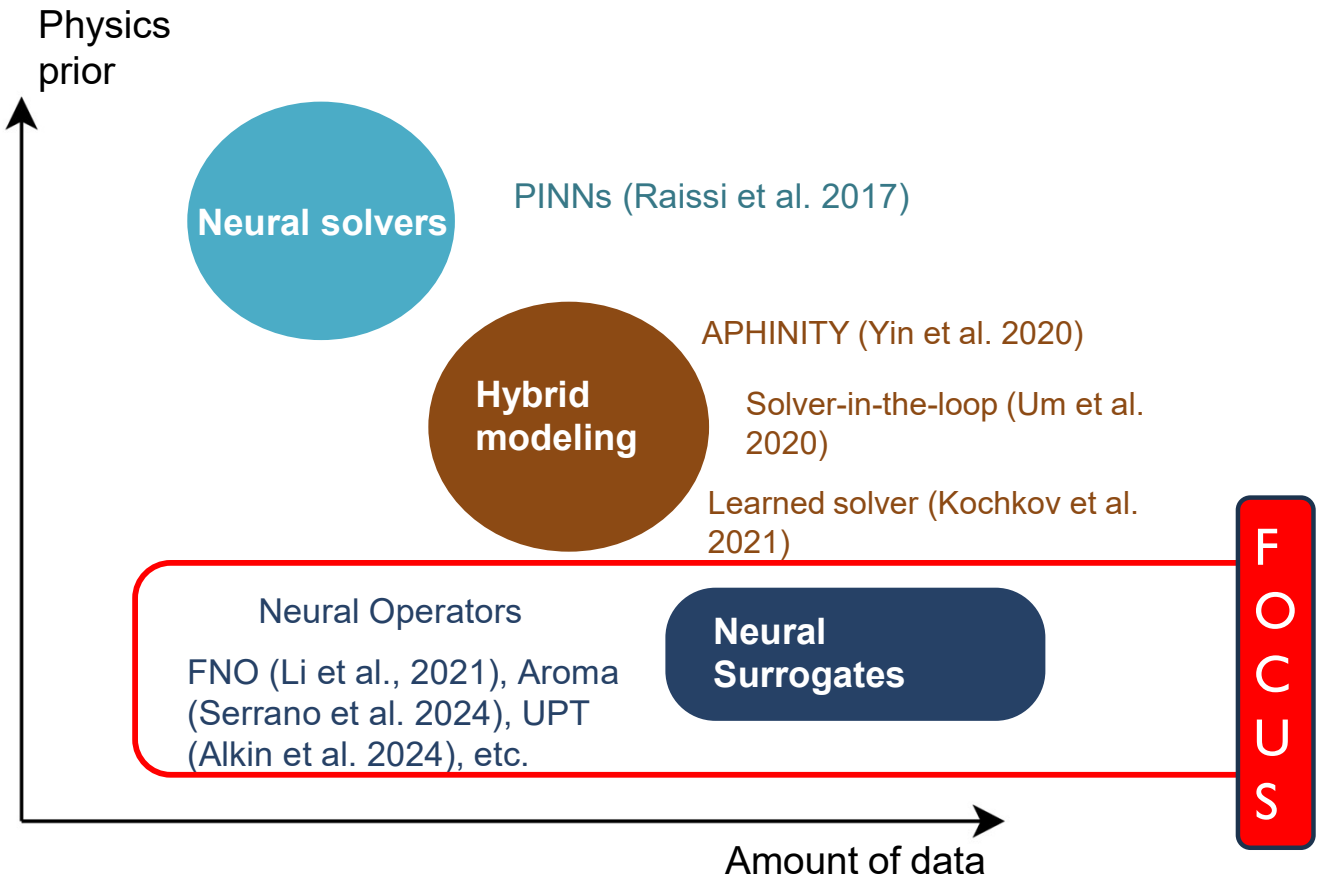




# Modeling spatio-temporal physical dynamics
# AI frameworks

## Deep learning landscape for modeling dynamics and solving PDEs

Focus of the presentation: data-driven approaches



# Modeling spatio-temporal physical dynamics
# Scientific challenges

- Physical Consistency: Incorporating constraints such as conservation laws, symmetries, and invariants into learning models.
- Generalization and Extrapolation: Ensuring robust predictions beyond training regimes (unseen physical regimes, new domains, new sensors, etc).
- Data Scarcity & Heterogeneity: Leveraging limited, noisy, or multi-fidelity data typical of scientific domains.
- Integration with Simulation & Theory: Building hybrid frameworks that combine data-driven and physics-based modeling.
- Scaling: Training large-scale scientific foundation models while managing cost and sustainability.
- Uncertainty Quantification…

# Modeling spatio-temporal physical dynamics Illustrative example

## Weather forecasting

# Weather forecasting
# Physical framework

- Objective: forecast the spatio-temporal evolution of atmospheric variables
  - Variables: temperature, pressure, wind speed, humidity,…, over the globe

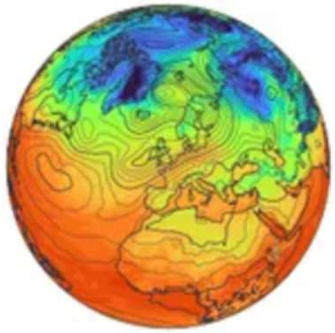
Physical framework

Observations | Physical model | Simulation



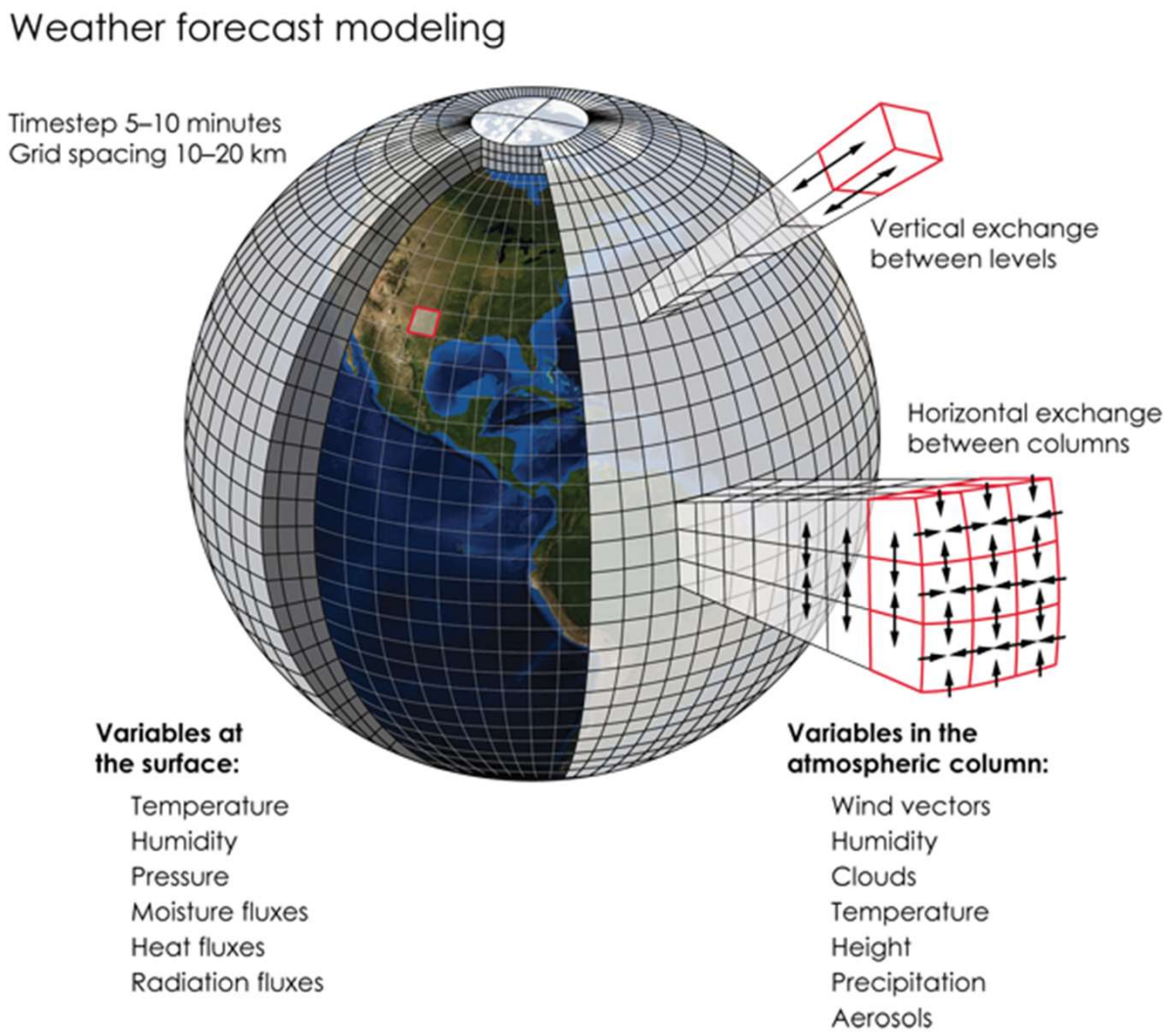
Variables: $\mathrm{x} = (\mathrm{u}, \rho, \mathrm{T}, \mathrm{p}, \mathrm{q})$

$Dynamics: \frac{\partial x}{\partial t} = F(x)$



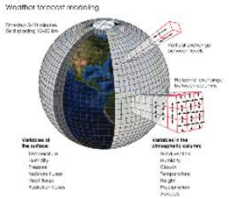
# Weather forecasting
# Physical simulation

$3D$ grid




Credit: A. Cantner AGI

# Weather forecasting
# Physical simulation



- Millions coupled equations
- Simulation
  - For each cell, one computes the values of atmospheric variables
    - i.e. solve the equations on each grid point
  - Interactions with neighbor cells are used to forecast the future values of the variables
- Costly: requires fine spatial and temporal resolution to operate
  - IFS-HRES model: horizontal 10 km x 10 km, vertical 137 layers, temporal 7.5 mn,

# Weather forecasting
# The rise of AI models




ECMWF, Dueben et al. 2018 , Challenges … for weather & Climate based ML
Scientific exploration

- GraphCast 2022 – DeepMind
- ClimaX 2023 – Msoft
- Pangu-W. 2023 – Huawei
- FourCastNet – NVIDIA
- NGCM 2023– Google
- …

- Operational forecast models
- ECMWF AIFS, AIFS-ENS, NOAA AIGFS, NVIDIA Earth2, …

# Weather forecasting
# Operational deployment by the main agencies - USA and Europ

National Oceanic and Atmospheric Administration
U.S. Department of Commerce

December 17, 2025

Search NOAA sites

Home / News & Features

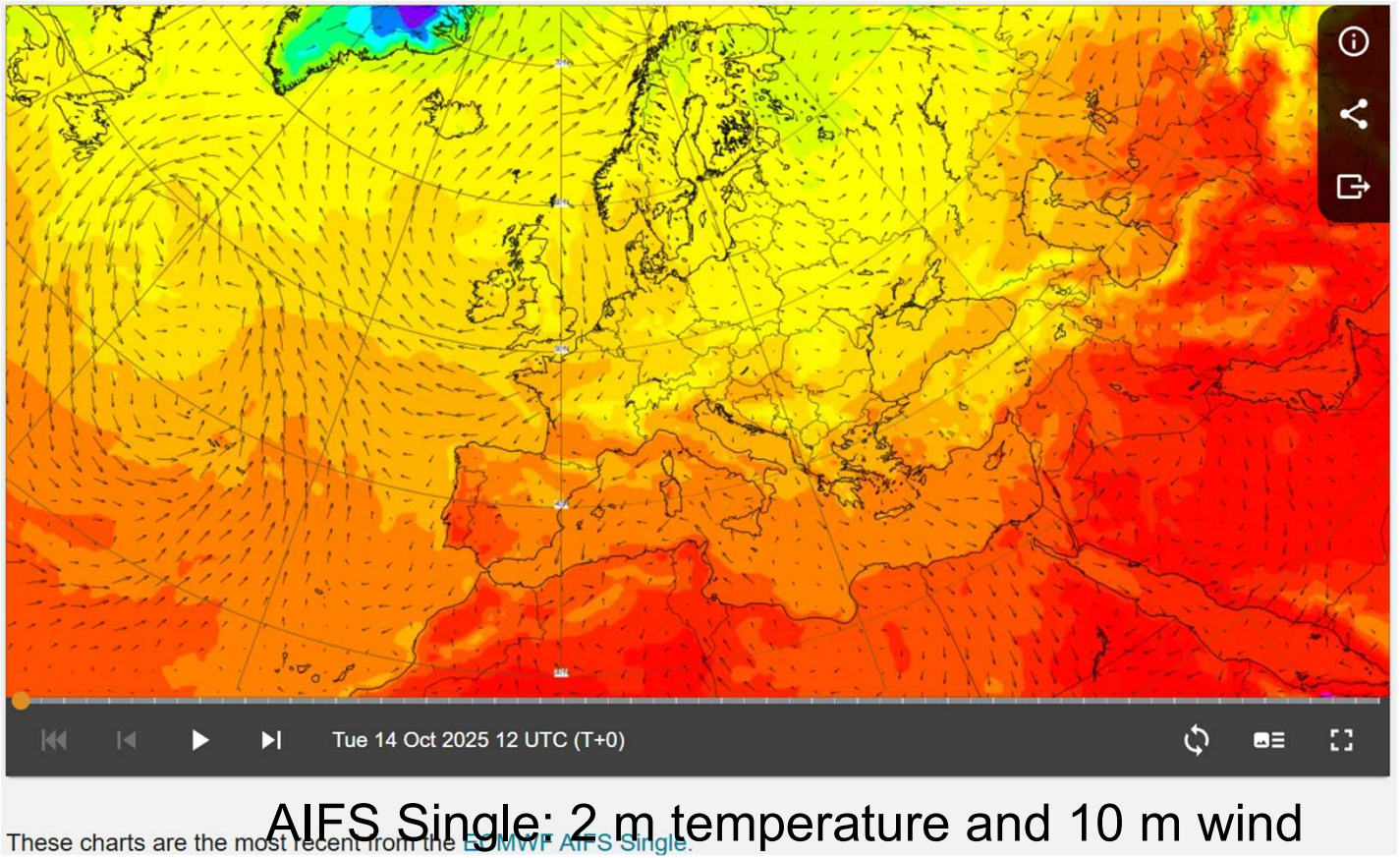
**Performance:** improved forecasting performance compared to traditional systems across most evaluation metrics.
**Efficiency:** approximately 1 minute to generate a single 16-day forecast on a GPU, compared to about 30 minutes on 50 high-end CPUs for ECMWF-IFS.




https://www.ecmwf.int/en/about/media-centre/news/2025/ecmwfs-ensemble-ai-forecasts-become-operational

# Weather forecasting
# ECMWF site dedicated to AI models

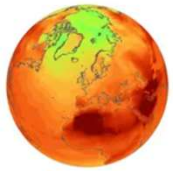
- AIFS model – ECMWF Operational February 2025: [AIFS charts]




AIFS Single: 2 m temperature and 10 m wind

# Weather forecasting
# Lessons learned from the deployment

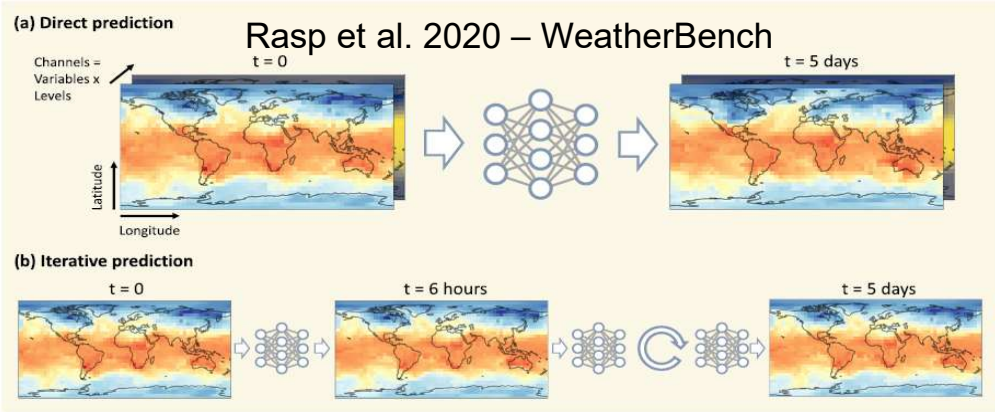
- A sweet spot for AI models
  - Large training datasets available



**ERA5 post-processed daily statistics on pressure levels from 1940 to present**

ERA5 is the fifth generation ECMWF reanalysis for the global climate and weather for the past 8 decades. Data is available from 1940 onwards. ERA5 replaces the ERA-Interim reanalysis. Reanalysis combines model data with observations from across the world into a globally complete and consistent datas...

  - Problem that formulates well for AI: spatio temporal forecasting




- Why is it faster?
  - AI operates on $10+$ pressure levels instead of $100+$ and does not require the « fine » $3D$ + Time integration of simulations

# Neural operators as surrogate models for solving PDEs

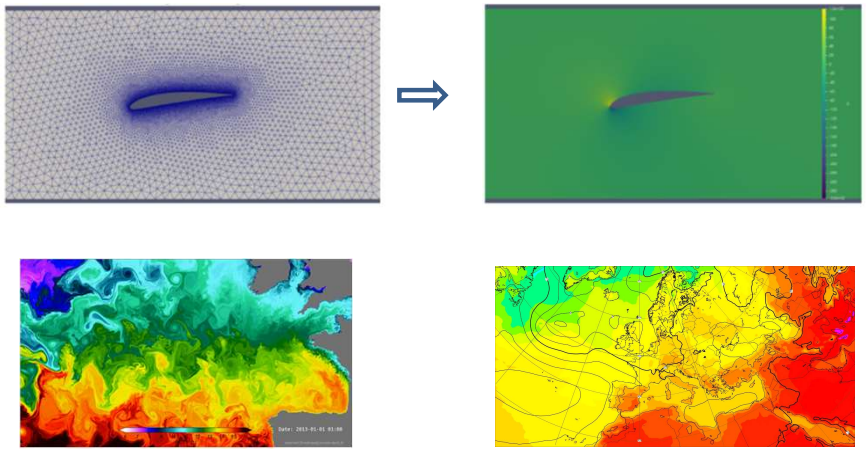
Fourier Neural Operators

AROMA: Attentive Reduced Order Model with Attention

ENMA Tokenwise Autoregression for Generative Neural PDE Operators

# Neural operators

Classical numerical solvers operate on grids or meshes (finite differences, finite elements, finite volumes)



2018 + Early efforts leverage standard NN (Convolutional NN and Graph NN) to learn functions between input and output discrete spaces

2020 + Neural operators formulate the problem as learning mappings between function spaces opening the way to meshless formulations

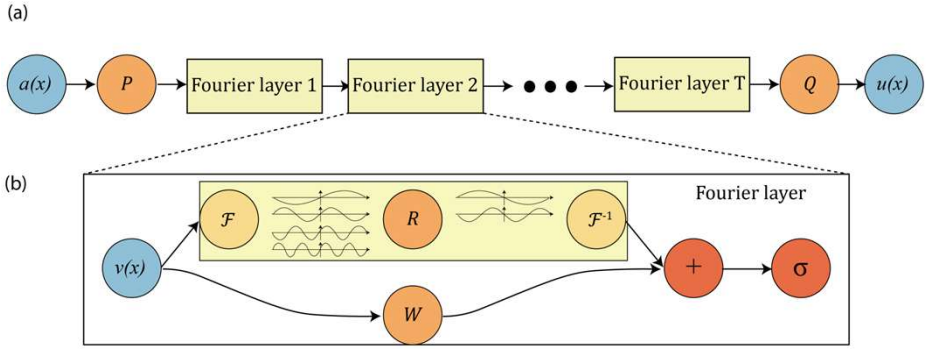
- e.g. images are considered as continuous functions

Potential benefits

- Functions and operators are resolution invariant
- Handle irregular geometries: meshes, point sets, grids
- Query at any space-time coordinate in the output space

# Neural operators for solving PDEs

- Pionneer work 2020+
  - Fourier Neural Operator (Li et al. 2020)




Initial FNO: Limited ability to handle multiple resolutions and continuous inputs & queries

Multiple extensions

- Recent approaches 2023+
  - Leverage developments in vision e.g. ViTs, diffusion/ flow matching
  - Illustrations
    - Full meshless approach
      - AROMA (Serrano et al. 2024)
    - Solving parametric PDEs
      - Incontext learning
        - Meshless: ENMA (Koupai et al. 2025)

# Neural Operators – Meshless approaches

- ✓ AROMA: Attentive Reduced Order Model with Attention (Serrano et al., NeurIPS 2024)

# NNs as surrogate models for solving PDEs – Operators

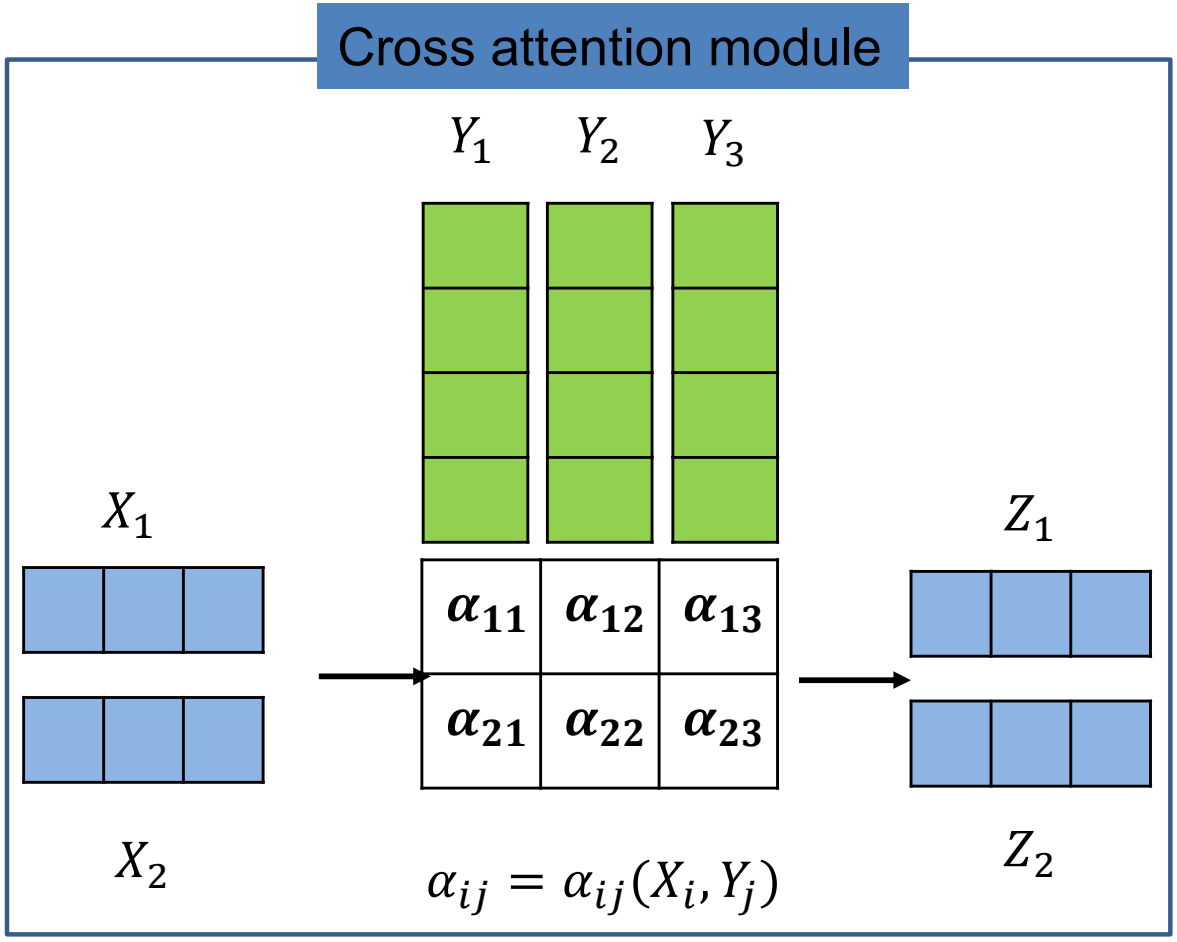
AROMA: Attentive Reduced Order Model with Attention (Serrano et al. 2024) - https://arxiv.org/abs/2406.02176

- Principled Framework:
  - Properties
    - Handle diverse geometries: inputs and outputs may consist in point sets, grids, irregular meshes
    - Captures local spatial information
    - Can be queried at any spatial position
  - Demonstrates how modern NN components allow building versatile PDE solvers
    - Encode/ Process/ Decode framework
      - Encoding: cross-attention maps variable-size inputs to a fixed-size compact latent token space encoding local spatial information
      - Processing: a diffusion transformer architecture to model dynamics and exploit spatial relations locally and globally via self-attention + diffusion
      - Decoding: uses a conditional neural field + cross-attention to query forecast values at any spatial point within the equation's domain

# NNs as surrogate models for solving PDEs – Operators

AROMA: Attentive Reduced Order Model with Attention (Serrano et al. 2024) - https://arxiv.org/abs/2406.02176 - Cross attention

Cross attention maps a sequence of vectors ($Y$) of variable sequence size $N$ into a sequence of vectors ($Z'$) of fixed sequence size $M$




Attention weights

$$Z_1 = \alpha_{11} W Y_1 + \alpha_{12} W Y_2 + \alpha_{13} W Y_3$$

$$Z_2 = \alpha_{21} W Y_1 + \alpha_{22} W Y_2 + \alpha_{13}\, W Y_3$$

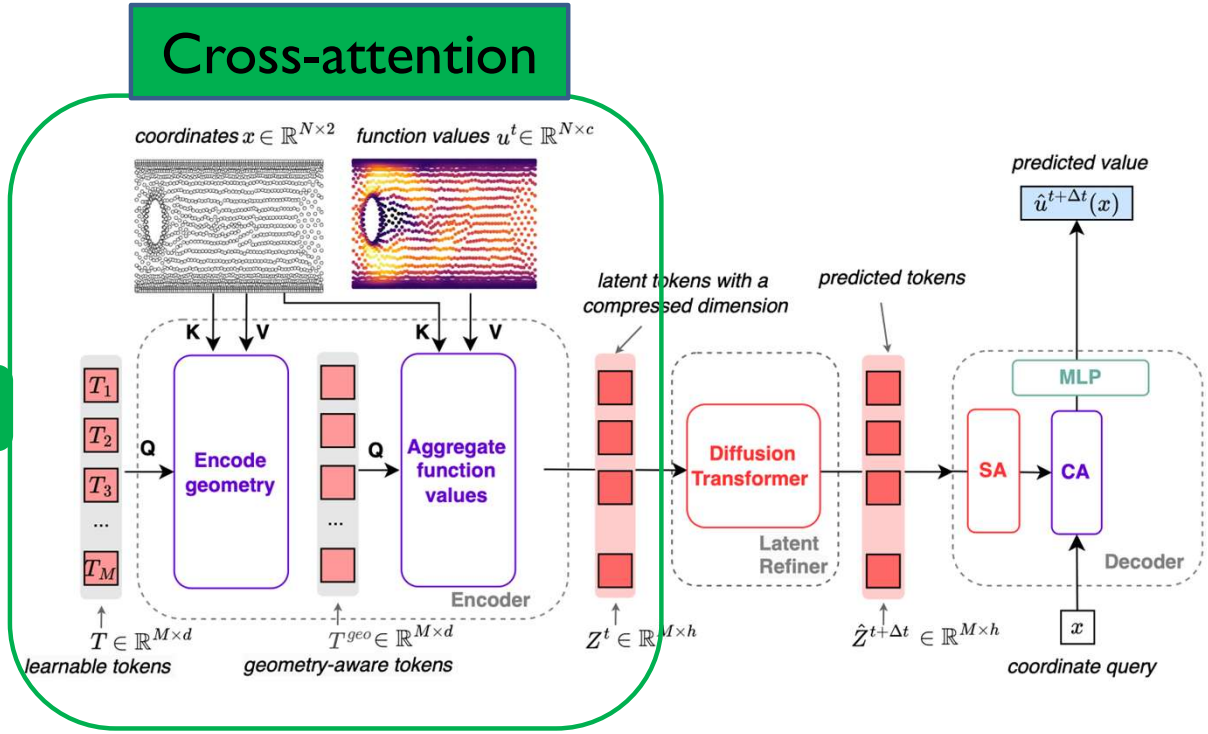
$W$ projection matrix

Complexity: $O(N(Y)N(X))$

# NNs as surrogate models for solving PDEs – Operators

AROMA: Attentive Reduced Order Model with Attention (Serrano et al. 2024) - https://arxiv.org/abs/2406.02176 - General framework

Encoder module




Cross-attention encoder: $u^t \rightarrow Z^t$

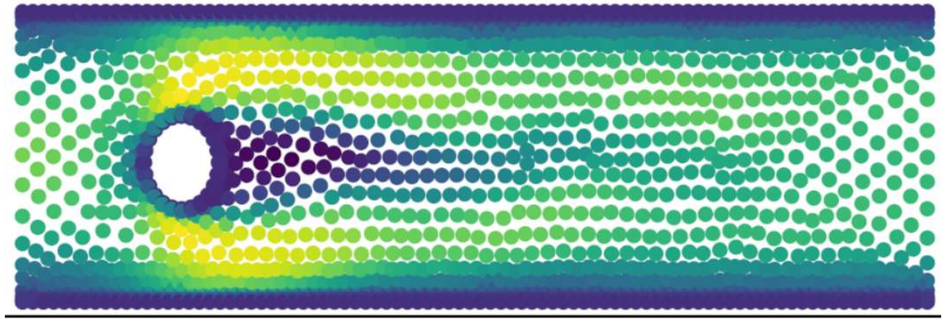
- Encodes variable size discretized input $u$ (point set, mesh, regular grid) into a fixed size & small dimensional sequence of latent embedding tokens $Z$
- Latent space encodes local spatial information on problem geometry + variable values

# NNs as surrogate models for solving PDEs – Operators

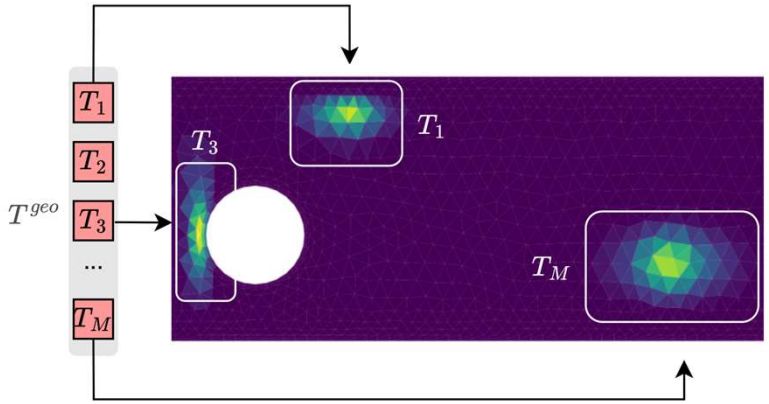
## AROMA: Attentive Reduced Order Model with Attention (Serrano et al. 2024) - https://arxiv.org/abs/2406.02176 - Cross-attention encoder captures spatial attention

Example: Cross attention on cylinder flow

- Cylinder flow ground truth



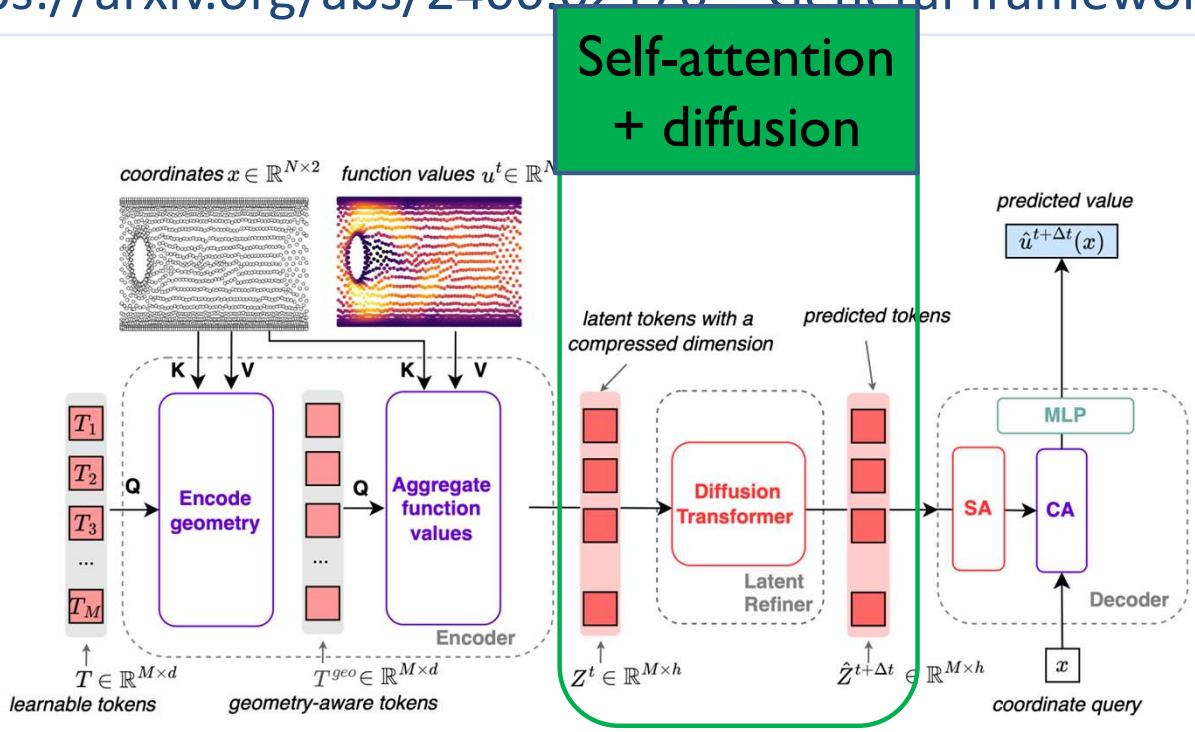
- Tokens encode local spatial information – cross attention between $T^{geo}$ tokens and "$x$"



# NNs as surrogate models for solving PDEs – Operators

AROMA: Attentive Reduced Order Model with Attention Attention (Serrano et al. 2024) - https://arxiv.org/abs/2406.02176 - General framework

Processor




Dynamics is modeled by a time stepping diffusion transformer: $Z^t \rightarrow Z^{t+\Delta t}$

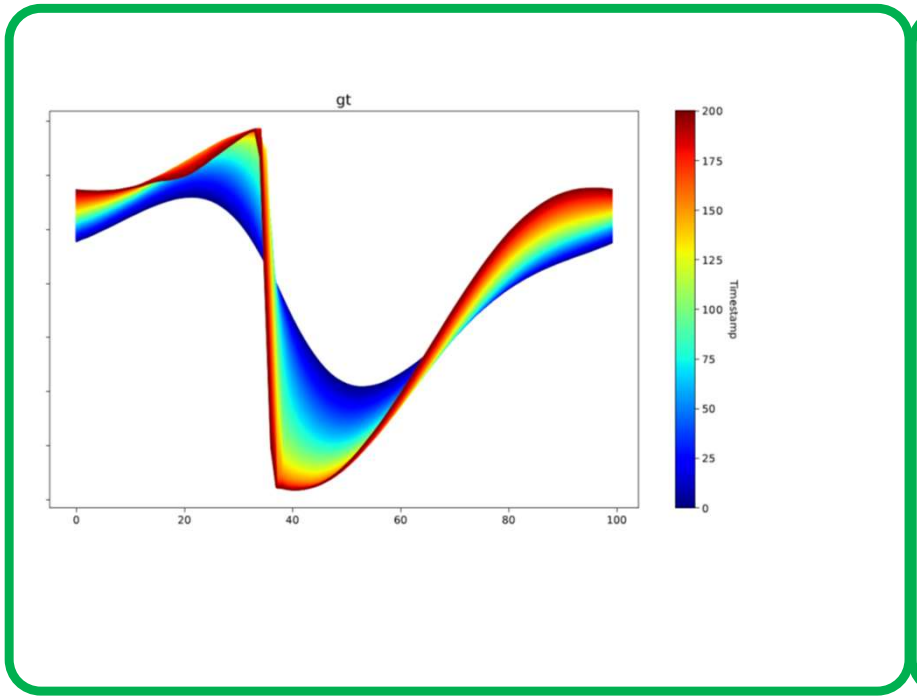
- Learns the dynamics in the small dimensional latent space
- Self attention models relations between spatial latent tokens
- Inference: dynamics is enrolled in the latent space starting from an initial condition– low complexity
- Diffusion: introduces a stochastic component

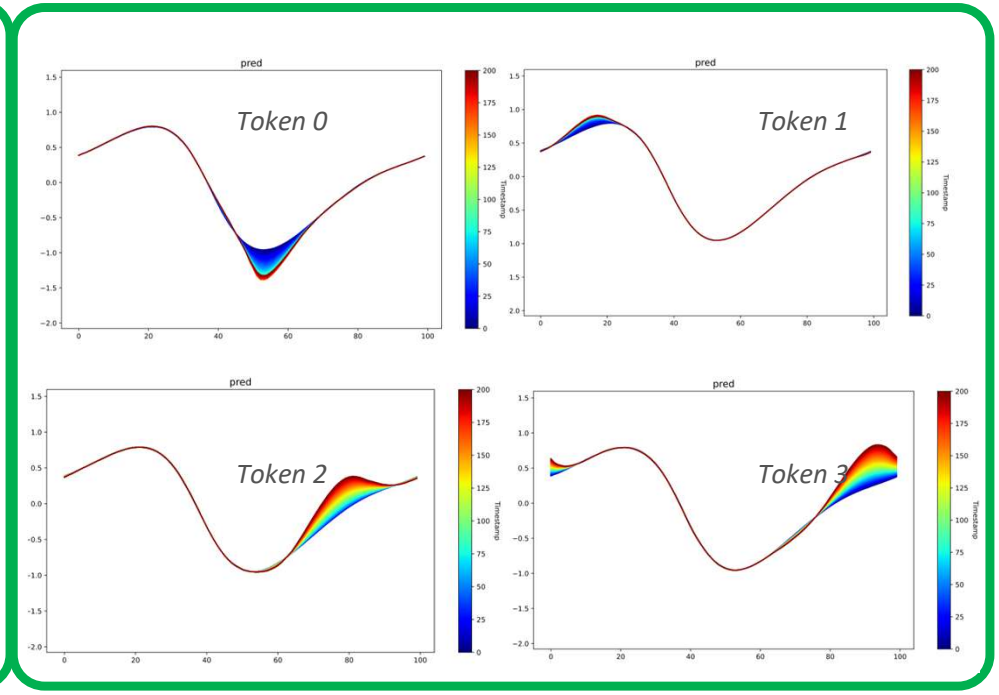
# NNs as surrogate models for solving PDEs – Operators

## AROMA: Attentive Reduced Order Model with Attention - (Serrano et al. 2024) - https://arxiv.org/abs/2406.02176

Example: Burgers equation – perturbation analysis on the tokens

- Burgers equation ground truth
- Tokens encode local spatial dynamics

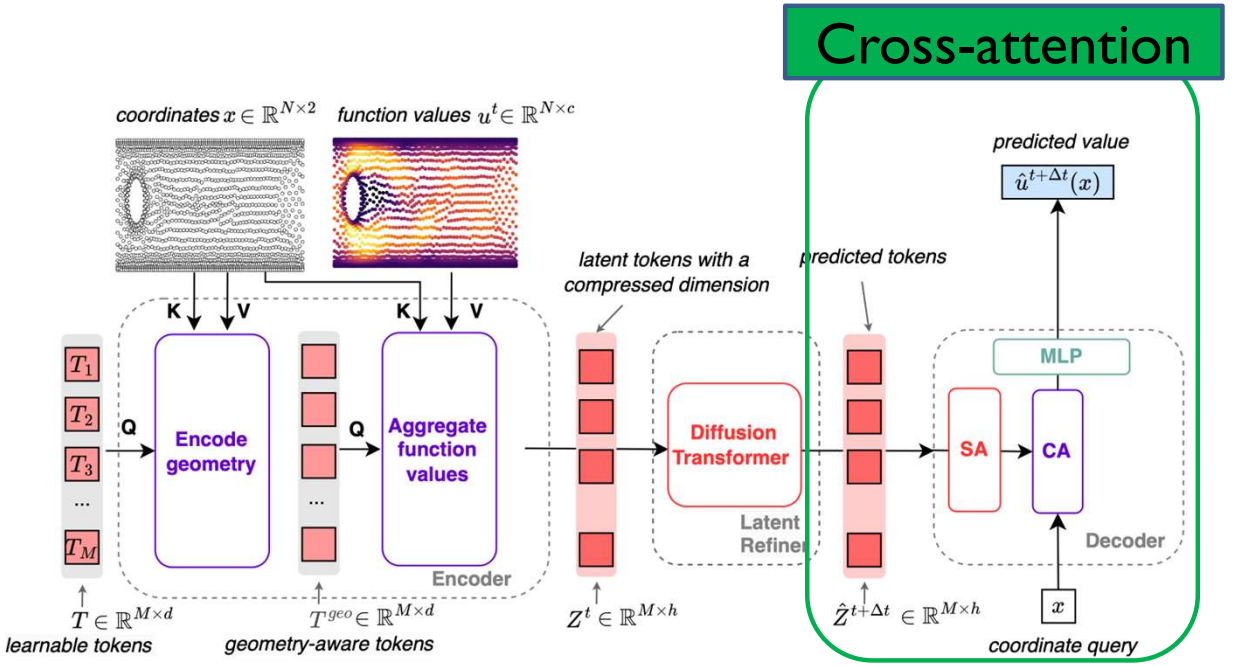


# NNs as surrogate models for solving PDEs – Operators

## AROMA: Attentive Reduced Order Model with Attention Attention (Serrano et al. 2024)- https://arxiv.org/abs/2406.02176 - General framework

Decoder module




Cross-attention + neural fields decoder: $Z^{t+\Delta t} \rightarrow u^{t+\Delta t}$

- Maps the latent representation $Z^{t+\Delta t}$ to the original physical space
- Can be queried at any position $x$ of the physical space

## NNs as surrogate models for solving PDEs – Operators

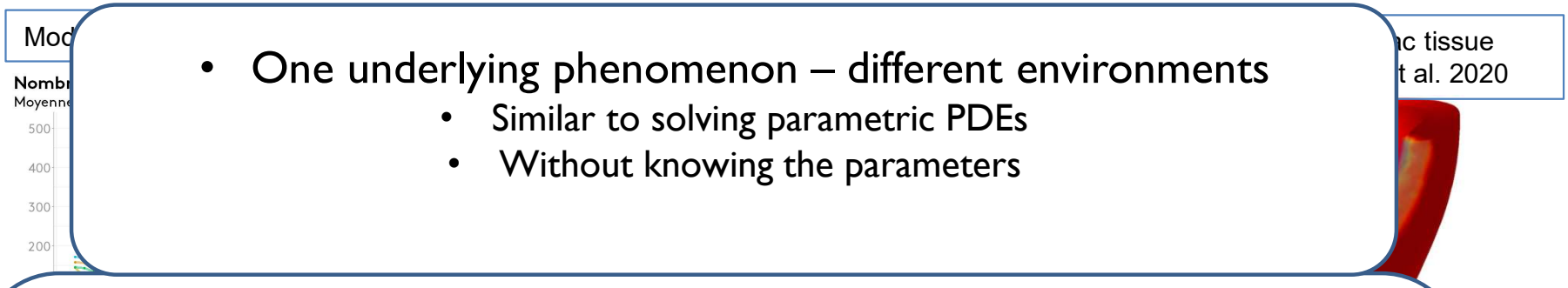
AROMA: Attentive Reduced Order Model with Attention - (Serrano et al. 2024) - https://arxiv.org/abs/2406.02176 - Summary

- Demonstrates how recent deep learning developments could be adapted to develop neural operators
- Advantages
  - One the very first fully meshless N.O.
  - Generative framework
    - uncertainty estimates
  - Competitive for in and out-distribution extrapolation
- Drawbacks
  - Does not scale to dense mesh points

# Neural operators – Solving Parametric PDEs

✓ ENMA Tokenwise Autoregression for Generative Neural PDE Operators (Kassai et al., NeurIPS 2025)

# Tackling the generalization problem for physical dynamics Examples




- One underlying phenomenon – different environments
  - Similar to solving parametric PDEs
  - Without knowing the parameters

Modeling : Meta-Learning

- Learn from a sample of environments, generalize to new ones
- How: learn to condition on environments

→ Infer the unknown dynamics from a few samples of the new environment

# In-context generative pretraining for PDEs (Serrano et al. 2025)

https://arxiv.org/abs/2410.03437

- Objective: solve parametric PDES for a parameter distribution
- Conditioning inspired by In-context learning in NLP decoders)

**LLMs**

In-Context examples ($u_{n+1} = 3u_n - 1$)

**Prompt**

*Sequence 1: 1, 2, 5, 14, …*
*Sequence 2: 2, 5, 14, 41, …*
*Sequence 3: 3, 8, 23, 68, …*
*Sequence 4: 4, ?*

Query new initial condition

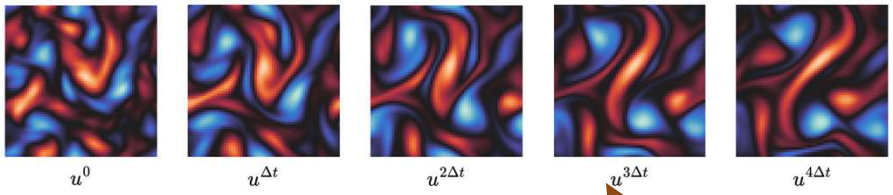
**Answer**

*Sequence 4: 4, 11, 32, 95, etc..*

*Without optimization LLMs are somehow capable of inferring the correct pattern in the sequences, i.e the underlying parameters:* $u_{n+1} = 3u_n - 1$

**PDE**

In-Context example




Same PDE parameters

Query new initial condition

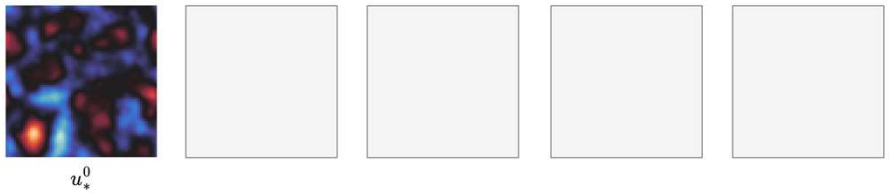


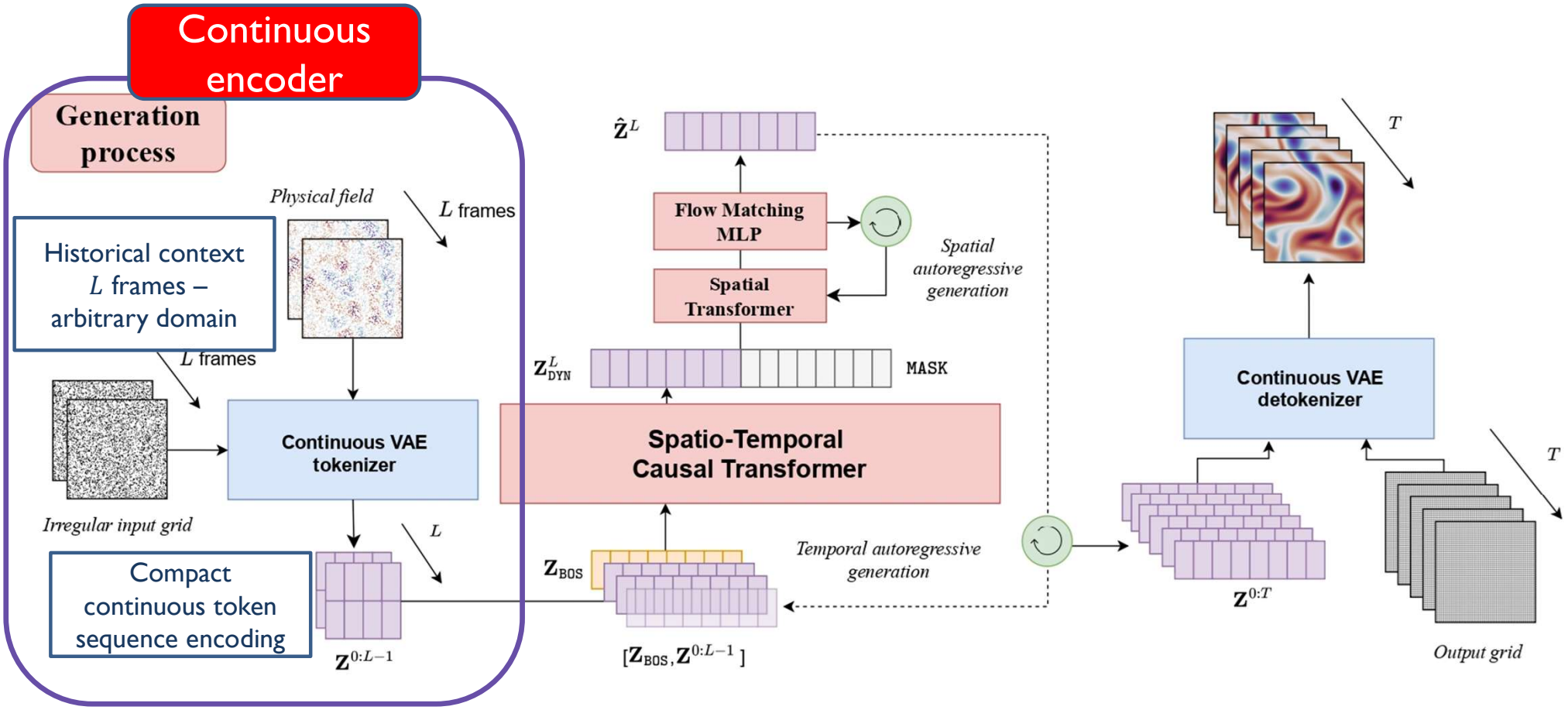

Fig. Serrano 2025

No gradient update – only context

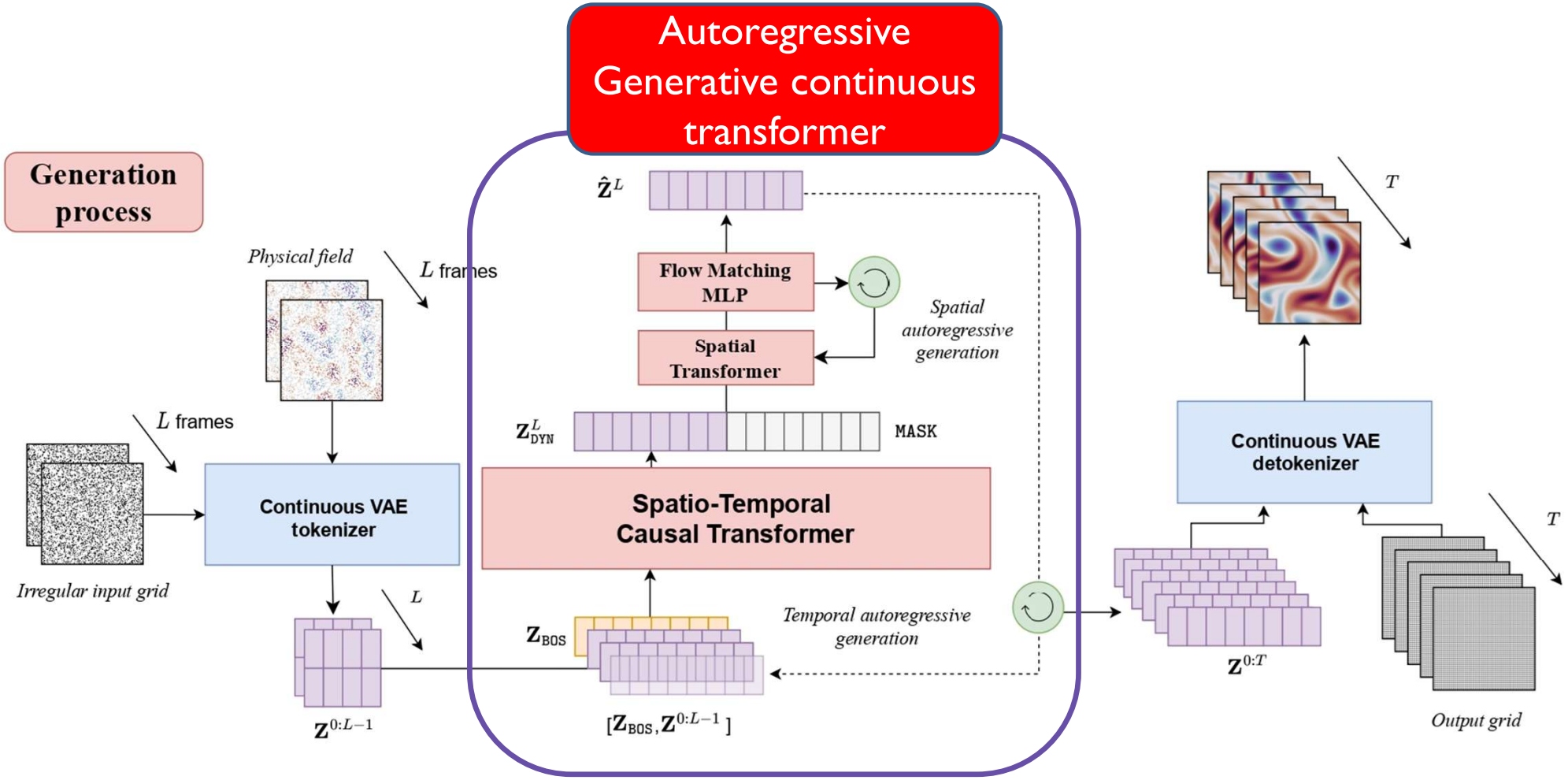
ENMA: Tokenwise Autoregression for Generative Neural PDE Operators (Kassai et al. 2025)
https://arxiv.org/abs/2506.06158

- Follows an encode-process-decode framework
- Extends the **in context learning** ideas to a fully continuous token space instead of a discrete space in LLMs
  - More natural for physical phenomena
  - Trained as a generative model using a flow matching loss
  - Properties
    - Handle irregularly positioned and possibly faulty sensors
    - Performs in context learning and well as forecasts from past history
    - Adapt to changes in the discretisation grid
    - Models uncertainty

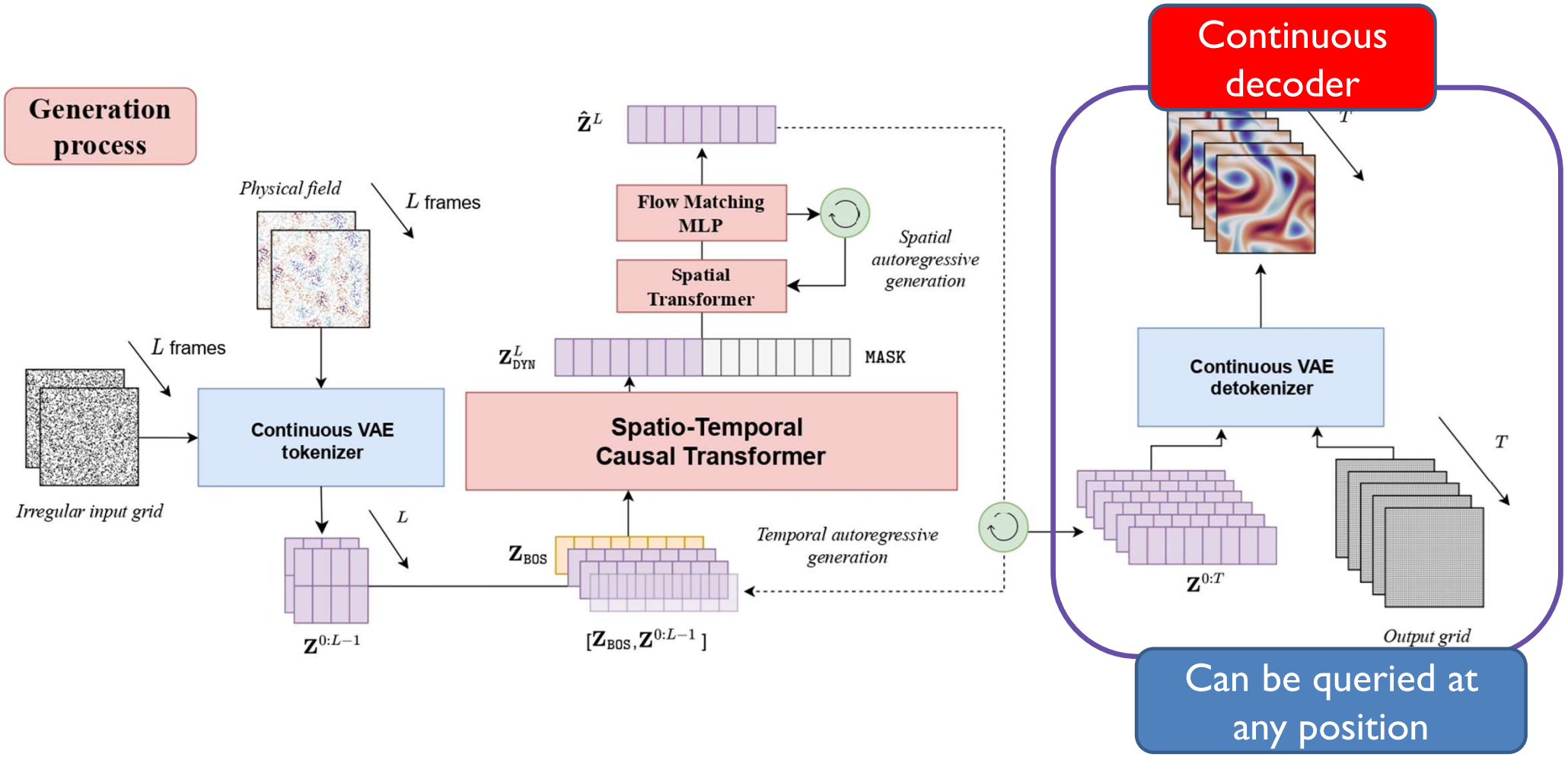
# ENMA: Tokenwise Autoregression for Generative Neural PDE Operators (Kassai et al. 2025) - https://arxiv.org/abs/2506.06158



# ENMA: Tokenwise Autoregression for Generative Neural PDE Operators (Kassai et al. 2025) - https://arxiv.org/abs/2506.06158



# ENMA: Tokenwise Autoregression for Generative Neural PDE Operators (Kassai et al. 2025) - https://arxiv.org/abs/2506.06158



# ENMA: Tokenwise Autoregression for Generative Neural PDE Operators (Kassai et al. 2025) - https://arxiv.org/abs/2506.06158

## Examples: Gray Scott

In-distribution

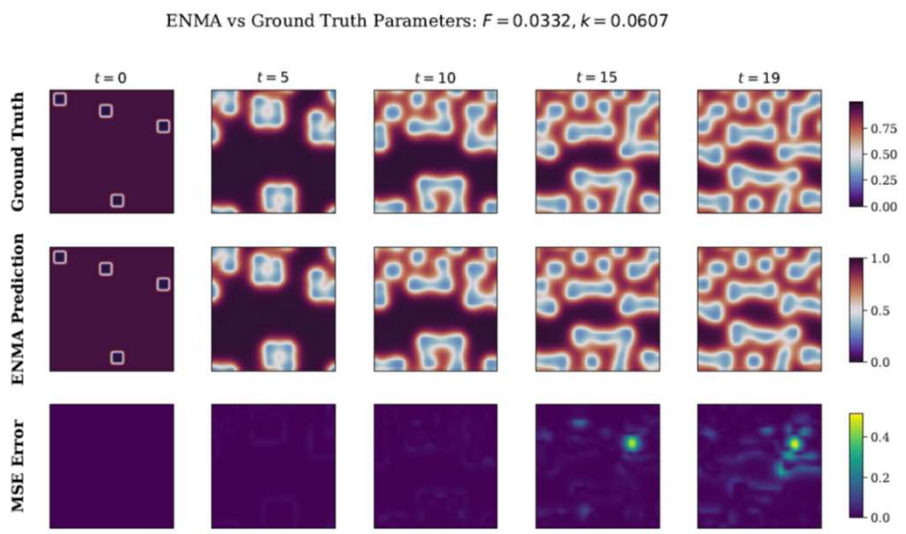



Figure 37: Qualitative comparison between ENMA prediction and ground truth for an in-distribution sample from the Gray-Scott dataset ($F = 0.0323$, $k = 0.0606$).

Out-of-distribution

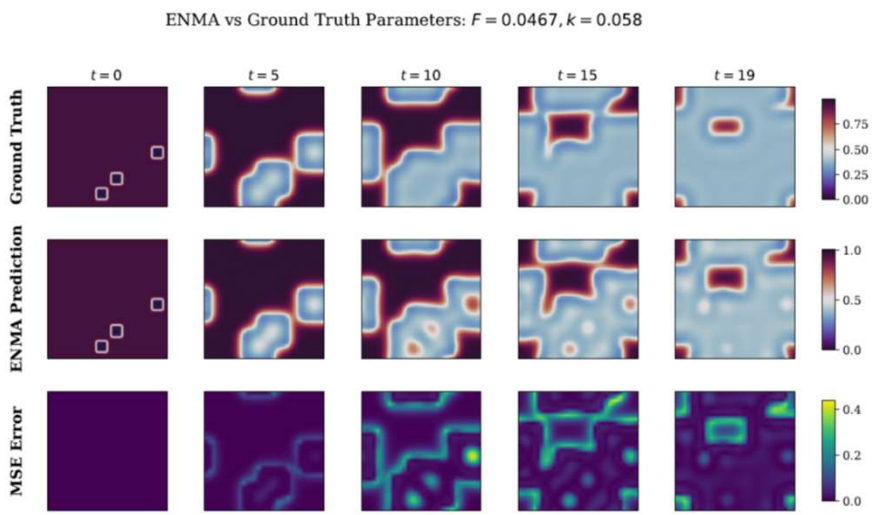



Figure 39: Out-of-distribution (OOD) generalization for the Gray-Scott equation. ENMA prediction remains consistent despite being evaluated at unseen parameters ($F = 0.0467$, $k = 0.058$).

# ENMA: Tokenwise Autoregression for Generative Neural PDE Operators (Kassai et al. 2025) - https://arxiv.org/abs/2506.06158

Examples: vorticity at different viscosities

In-distribution

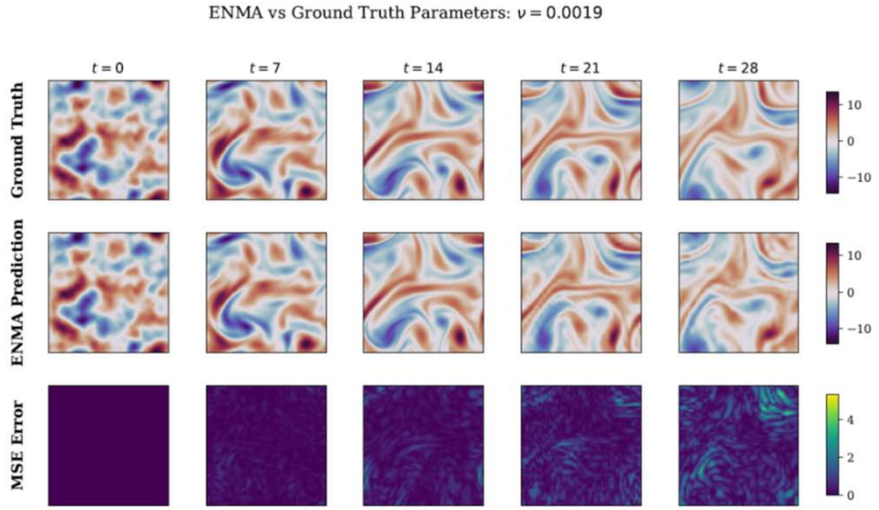



Figure 43: Qualitative comparison between ENMA prediction and ground truth for an in-distribution sample from the Vorticity dataset ($\nu = 0.0019$).

Out-of-distribution

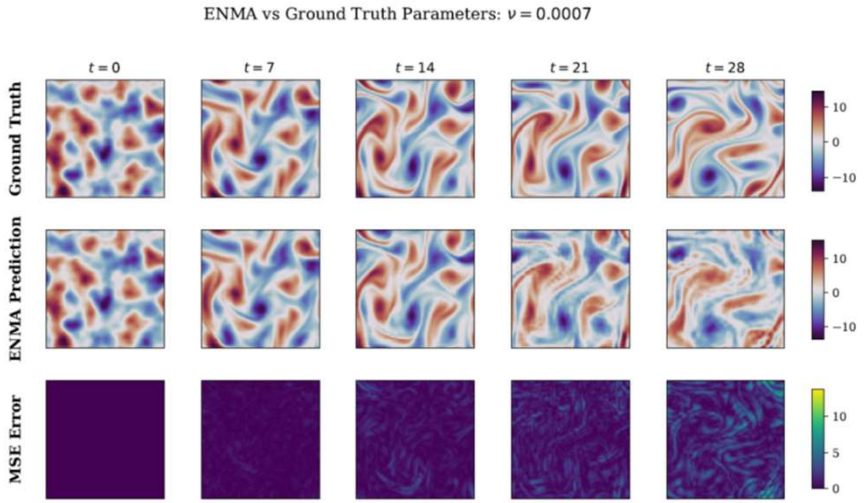



Figure 45: Out-of-distribution example from the Vorticity dataset ($\nu = 0.0007$), highlighting ENMA's robustness in extrapolating vortex dynamics.

## ENMA: Tokenwise Autoregression for Generative Neural PDE Operators (Kassai et al. 2025) - https://arxiv.org/abs/2506.06158 Summary

- Advantages
  - Full meshless neural operator
  - Adaptation of in-context to continuous domains
  - In-context adaptation for solving parametric PDEs
  - Generative framework
- Drawbacks
  - Does not scale to dense mesh points
  - Slow inference compared to fully deterministic

# Foundation models for PDEs

# Foundation models for PDEs
# Motivations & objectives

- Analogy with language and vision
  - Self training on large corpora allows us learning useful features
- Objectives: from task-specific surrogates to general-purpose models
  - Initiated in 2023-2024
  - Learn across many PDEs, geometries, regimes
  - Enable few-shot generalization to new situations
- Motivations: Data efficiency and amortization
  - PDE simulations are expensive
    - Current surrogates require large datasets per problem
  - Foundation models trained on large & diverse corpora
    - May adapt cheaply to new PDEs/ Boundary condition - Many situations where data is scarce

# Foundation models for PDEs
# Models & Challenges

- Methods
  - Largely based on Vision transformer adaptations
  - Trained by forecasting on multiple PDEs
  - Evaluated in few shot setting/ supervised fine tuning
  - Often limited to regular grids
- Challenges
  - Heterogeneity of the physical problems
    - Variables at different numerical magnitudes (temperature, speed, pression, etc)
    - Multiple resolutions, multiple scales, channels
    - Multiple domains (space, time, parameters), e.g. $1D, 2D, 3D$
  - Technical challenges
    - Compared to vision, requires backbone architectures that can adapt to different types of PDEs, domains, boundary conditions, scales, etc
    - With limited complexity
      - Key issue for transformers: PDE FMs all introduce specific attention mechanisms to limit the complexity

# Foundation models for PDEs
# Learning Task-agnostic representations for physical dynamics (Castano Segade et al. 2026)

- Objective
  - Learn reusable representations of dynamics that can be used for different dowsntream tasks
    - Leverage self-supervised learning - SSL
    - Disentengles dynamical system representation from state evolution
      - Forecasting is only one possible downstream task
- Why?
  - Successful idea in Vision for still images (DINO models) or Videos (JEPA models)
  - Comes with potential advantages
    - One representation for multiple tasks
    - Limited complexity
    - 0 shot inference – generalizes to different contexts

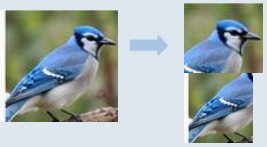
# Foundation models for PDEs
# Learning Task-agnostic representations for physical dynamics
# JEPA: Predicting Target Representations from Context

## Self-supervised learning principle

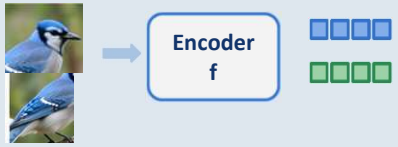
1 **No labels**



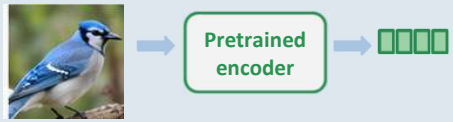

2 **Two related observations or partial views**



3 **Train a latent representation mapping**




4 **Use the pretrained encoder at inference**




**Learn one reusable encoder from structure in the data itself.**

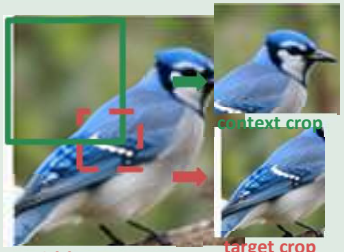
full image → representation

## JEPA

**Training**

Predict target representations from context

Visible context (observed)




Context encoder fθ

Target encoder fψ (EMA copy)

Context representation

Predictor gϕ

Predicted target representation

keep

Target representation

Latent prediction loss L = ||p(z_context) − z_target||²

stop gradient on target

Training view: two branches are used, but the target encoder is just an EMA copy used to provide stable targets.

**Objective: learn a single good encoder, not two different encoders.**

**After pretraining / inference**



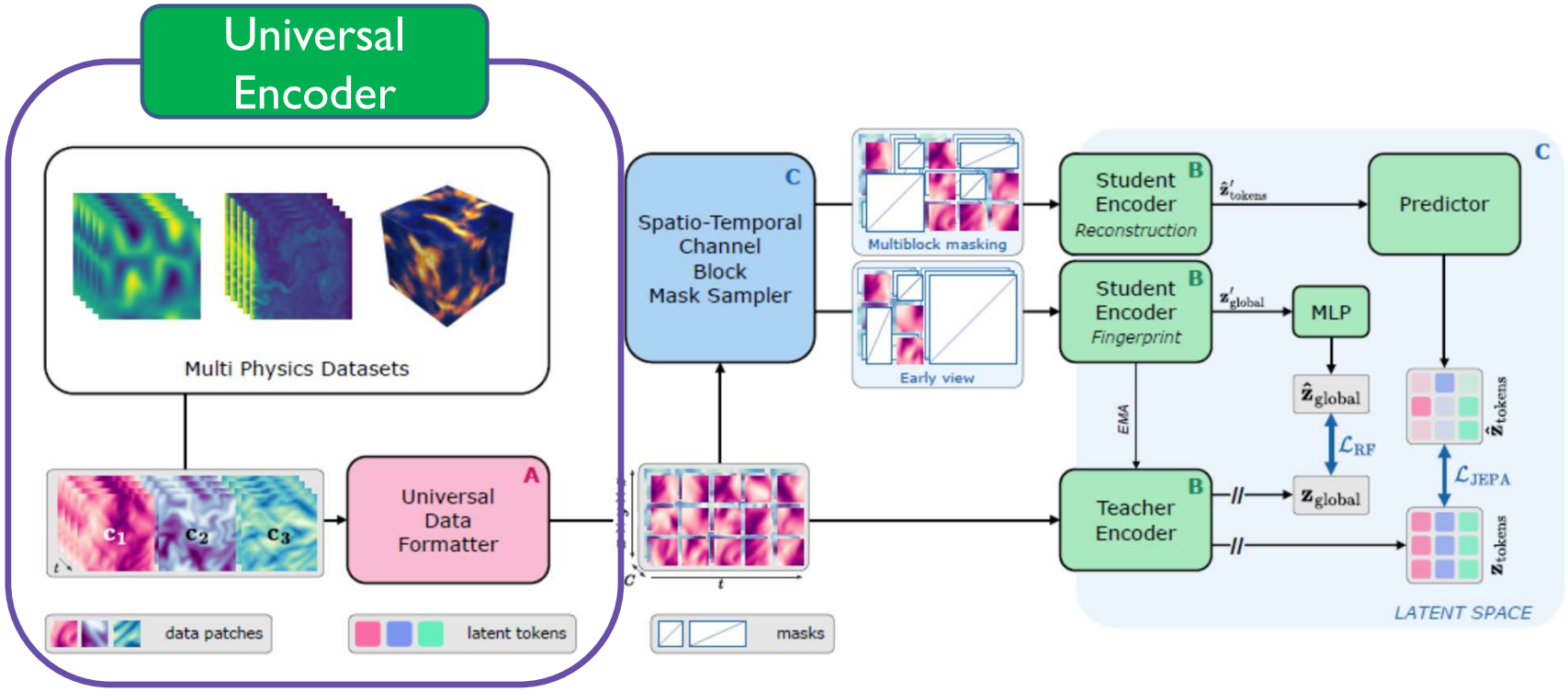
full input

Keep EMA target encoder fψ

Representation

At inference, discard the predictor and use a single pretrained encoder: the EMA target encoder.

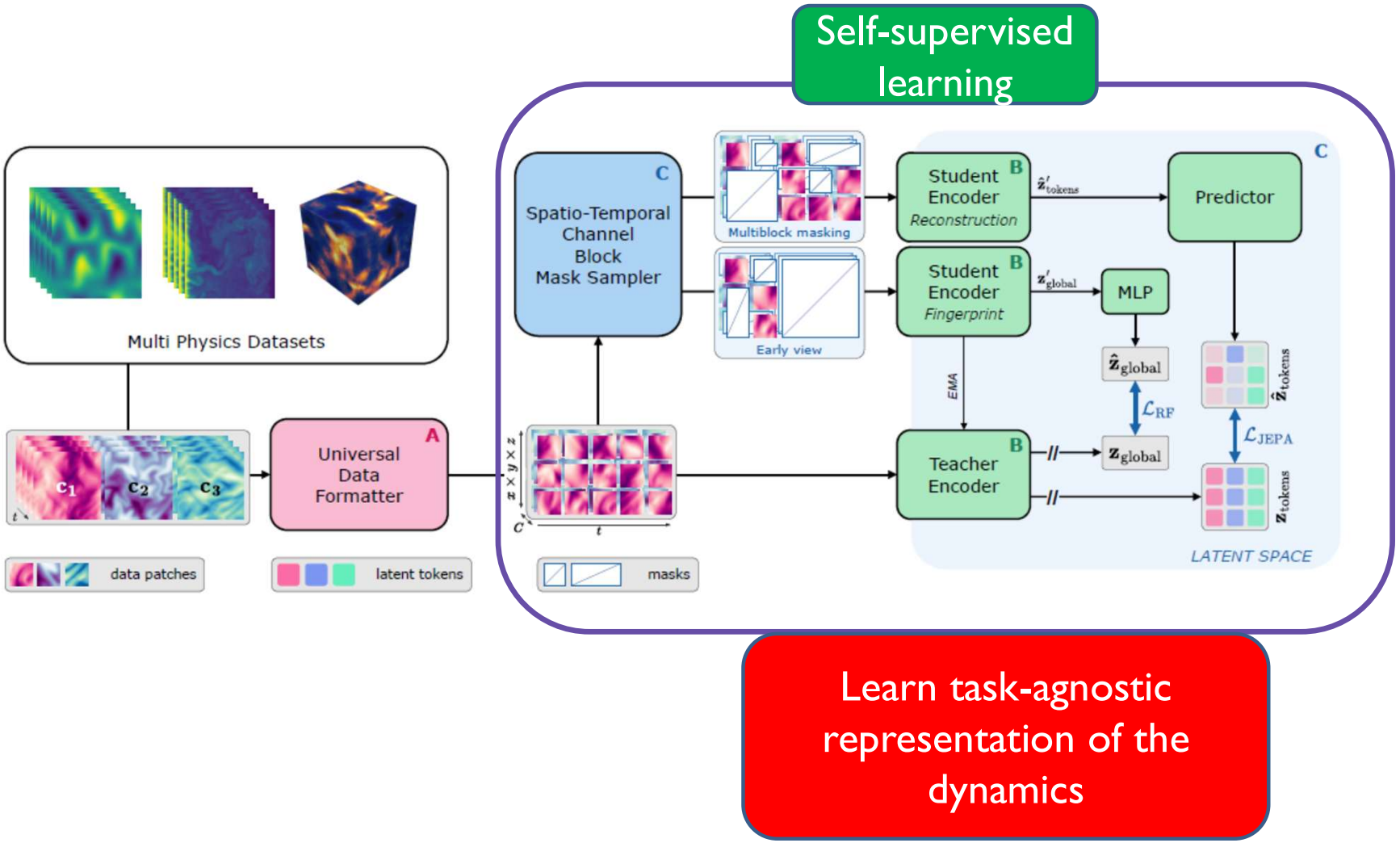
# Foundation models for PDEs
# Learning Task-agnostic representations for physical dynamics (Castano Segade et al. 2026)




Encodes heterogenous inputs into a uniform format: $1D, 2D, 3D$, multiple fields

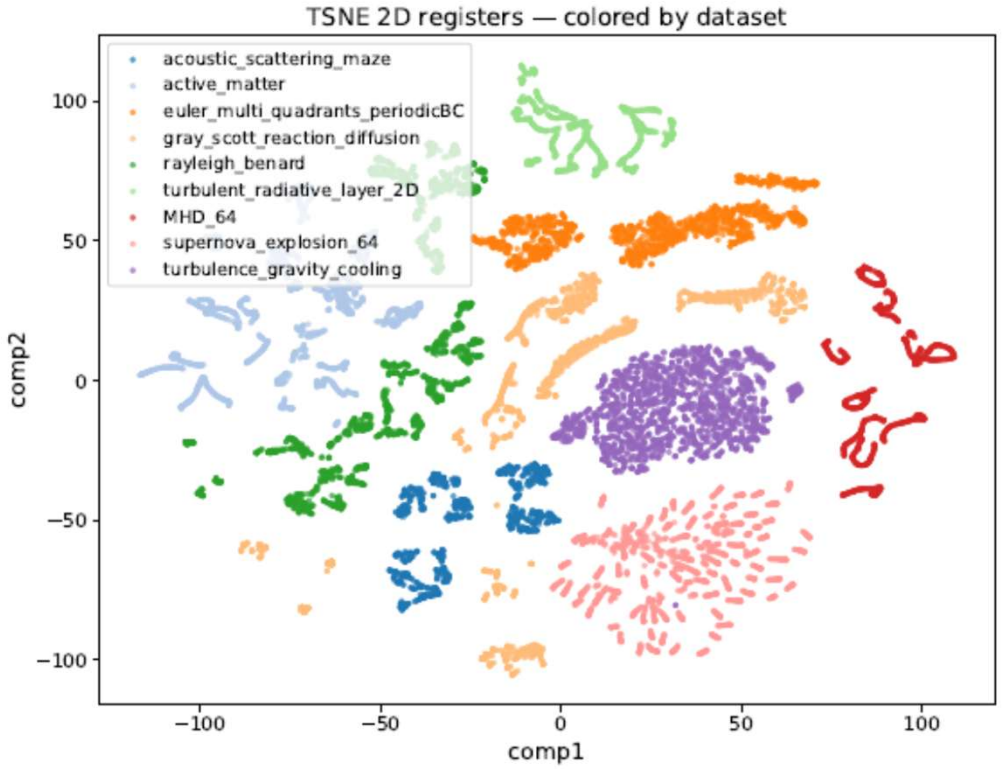
- Normalize variables
- Embed fields in a shared space
- Axial attention – limit complexity

# Foundation models for PDEs
# Learning Task-agnostic representations for physical dynamics (Castano Segade et al. 2026)



# Foundation models for PDEs
# Learning Task-agnostic representations for physical dynamics (Castano Segade et al. 2026)

- Learned dynamics visualisation – 9 physics




- One point: one trajectory
- Each color corresponds to a different physics

# Foundation models for PDEs
# Summary

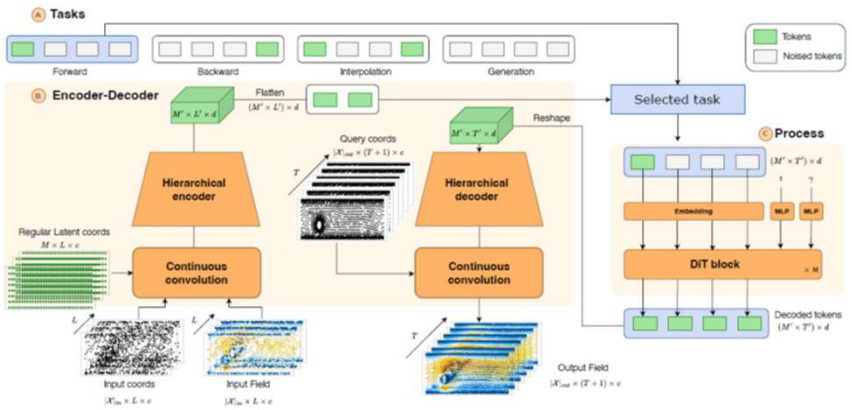
- Direction started in 2023-2024
  - Attempts to learn multiple physics simultanously
  - Objective: capture physics « primitives » that help for few shot and 0-shot
  - Largely based on ViT/ DiT adaptations
- Limitations
  - Mainly limited to simplified settings, e.g. regular grids $1D$ or $2D$
  - Scaling and datasets availability
  - Interpretabilty - what do we learn: what/ how/ where dynamics are captured?

# Scaling

## Scaling

- While neural operators & foundation models introduce new concepts for meshless operators, they do not scale to real world operational requirements
- 2025 +, witnesses initial efforts for extending neural operators to large size problems
- They are all transformer based architectures
  - Most models target steady state prediction, e.g. automotive applications
    - i.e. no explicit modeling of the temporal dynamics
  - A few models targeting full spatio-temporal forecasting

# ECHO: Efficient Generative Transformer Operators for Million-Point PDEs (Koupai et al. NeurIPS 2026)




| Dataset | Dim. | Trajectory size | Total spatio-temporal values |
|---|---|---|---|
| Vorticity | 2D | 20 × 1024 × 1024 × 1 | **20M** |
| Active Matter | 2D | 20 × 65536 × 11 | **14.4M** |
| MHD | 3D | 16 × 262144 × 7 | **29M** |
| TGC | 3D | 16 × 262144 × 6 | **25M** |

ECHO handles trajectories with several million points

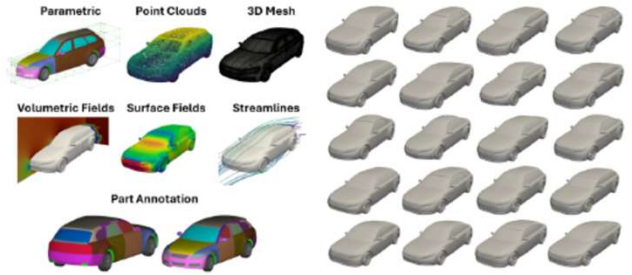
# Datasets

Datasets for foundation models, for large scale training static and dynamics

# Datasets

- Datasets are key to the development of neural PDE solvers
  - Recent attempts (2024 +) to build realistic/ complex datasets
    - Steady state RANS, multiple physics (Foundations), full dynamics
  - Difficulties
    - High fidelity simulations, e.g. for CFD, require important compute ressources + huge storage capacities

**DrivAerNet++: A Large-Scale Multimodal Car Dataset with Computational Fluid Dynamics Simulations and Deep Learning Benchmarks**




(a) Data modalities of DrivAerNet++. Top row: different data representations; middle row: CFD simulation results; bottom row: annotated car components.

(b) Selected samples from DrivAerNet++ showing diversity in shape with different car designs (fastback, estateback, and notchback), wheels configurations, and underbody configurations.

Figure 1: Data modalities and shape variations in the DrivAerNet++ dataset.

**The Well: a Large-Scale Collection of Diverse Physics Simulations for Machine Learning**

| Dataset | CS | Resolution (pixels) | n_steps | n_traj |
|---|---|---|---|---|
| acoustic_scattering | $(x,y)$ | $256\times256$ | 100 | 8000 |
| active_matter | $(x,y)$ | $256\times256$ | 81 | 360 |
| convective_envelope_rsg | $(r,\theta,\phi)$ | $256\times128\times256$ | 100 | 29 |
| euler_multi_quadrants | $(x,y)$ | $512\times512$ | 100 | 10000 |
| gray_scott_reaction_diffusion | $(x,y)$ | $128\times128$ | 1001 | 1200 |
| helmholtz_staircase | $(x,y)$ | $1024\times256$ | 50 | 512 |
| MHD | $(x,y,z)$ | $64^3$ and $256^3$ | 100 | 100 |
| planetswe | $(\theta,\phi)$ | $256\times512$ | 1008 | 120 |
| post_neutron_star_merger | $(\log r,\theta,\phi)$ | $192\times128\times66$ | 181 | 8 |
| rayleigh_benard | $(x,y)$ | $512\times128$ | 200 | 1750 |
| rayleigh_taylor_instability | $(x,y,z)$ | $128\times128\times128$ | 120 | 45 |
| shear_flow | $(x,y)$ | $128\times256$ | 200 | 1120 |
| supernova_explosion | $(x,y,z)$ | $64^3$ and $128^3$ | 59 | 1000 |
| turbulence_gravity_cooling | $(x,y,z)$ | $64\times64\times64$ | 50 | 2700 |
| turbulent_radiative_layer_2D | $(x,y)$ | $128\times384$ | 101 | 90 |
| turbulent_radiative_layer_3D | $(x,y,z)$ | $128\times128\times256$ | 101 | 90 |
| viscoelastic_instability | $(x,y)$ | $512\times512$ | variable | 260 |

Table 1: Dataset description: coordinate system (CS), resolution of snapshots, `n_steps` (number of time-steps per trajectory), `n_traj` (total number of trajectories in the dataset).

**APEBench: A Benchmark for Autoregressive Neural Emulators of PDEs**


Felix Koehler
Technical University of Munich
Munich Center for Machine Learning
f.koehler@tum.de

Simon Niedermayr
Technical University of Munich
simon.niedermayr@tum.de

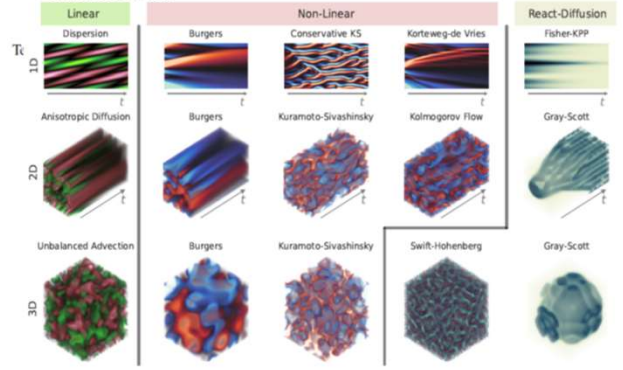
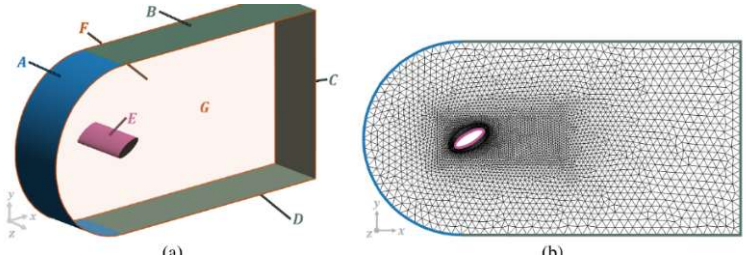
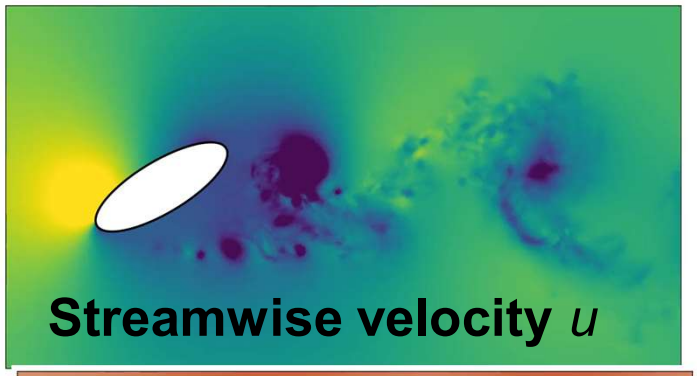



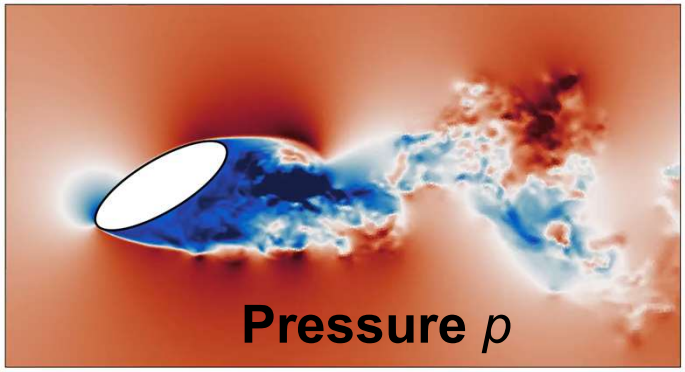

Figure 1: APEBench provides an efficient pseudo-spectral solver to simulate 46 PDE dynamics across one to three spatial dimensions. Shown are examples visualized with APEBench's custom volume renderer.

# Datasets
# TurtLES a dataset for large scale CFD (Platenburg et al. NeurIPS 2026)

- 3D turbulence dynamics dataset of LES simulations
  - Dense irregular discretization reaching millions points per frame
  - Long horizon evaluation on hundreds of time steps
  - Support physically grounded simulations
  - 15 TB data
  - Vary the cylinder cross section, Reynolds number, angle of attack



# Conclusion

- Survey recent evolution of data driven modeling physical dynamics
  - Benefits from advances in different ML fields (Vision, NLP, …)
  - Large size applications deployed in some domains e.g. weather forecast + recent new CFD datasets
  - Still gap to real world applications in many domains e.g. CFD
  - Recent developments for building foundation models and scaling neural

- Note
  - Data is essential – recent efforts to generate/ gather large sieze representative datasets