南京航空航天大學
NANJING UNIVERSITY OF

CVIT: CONTINUOUS VISION TRANSFORMER FOR OPERATOR LEARNING

Sifan Wang[1], Jacob H. Seidman[2], Shyam Sankaran[3], Hanwen Wang[2], George J. Pappas[4] Paris Perdikaris[3]

ICLR 2025

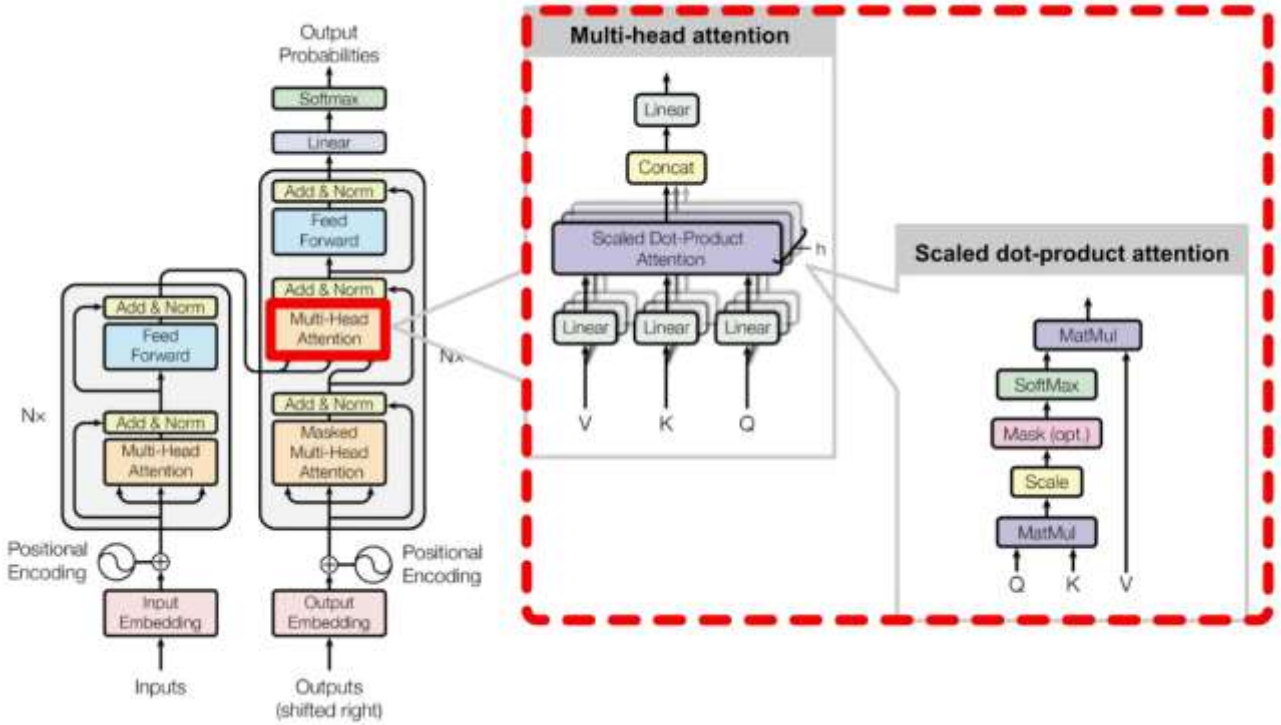
Background

The emergence of neural operators has driven advancements in solving partial differential equations (PDEs) and modeling physical systems. Recent developments have focused on enhancing the expressiveness and efficiency of these models.For instance, extensions of the Fourier Neural Operator (FNO) (Li et al., 2021) have explored factorized representations (Tran et al., 2021) to reduce computational complexity while maintaining performance. Similarly, wavelet-based approaches (Gupta et al., 2021; Tripura & Chakraborty, 2022) have been proposed to capture multi-scale features more effectively.

Transformer-Based Operator Learning. The self-attention mechanism central to transformers offers a powerful tool for capturing long-range dependencies in spatial and temporal data, a crucial aspect in many PDE-governed systems. For example, OFormer (Li et al., 2022) introduced a novel way to embed continuous input functions and output queries into a transformer architecture. This approach demonstrated the potential of transformers in handling the inherent continuity of physical problems. GNOT (Hao et al., 2023) further extended this idea by incorporating graph structures, allowing for more flexible representation of input functions and improved handling of irregular domains.

Background

Transformer：

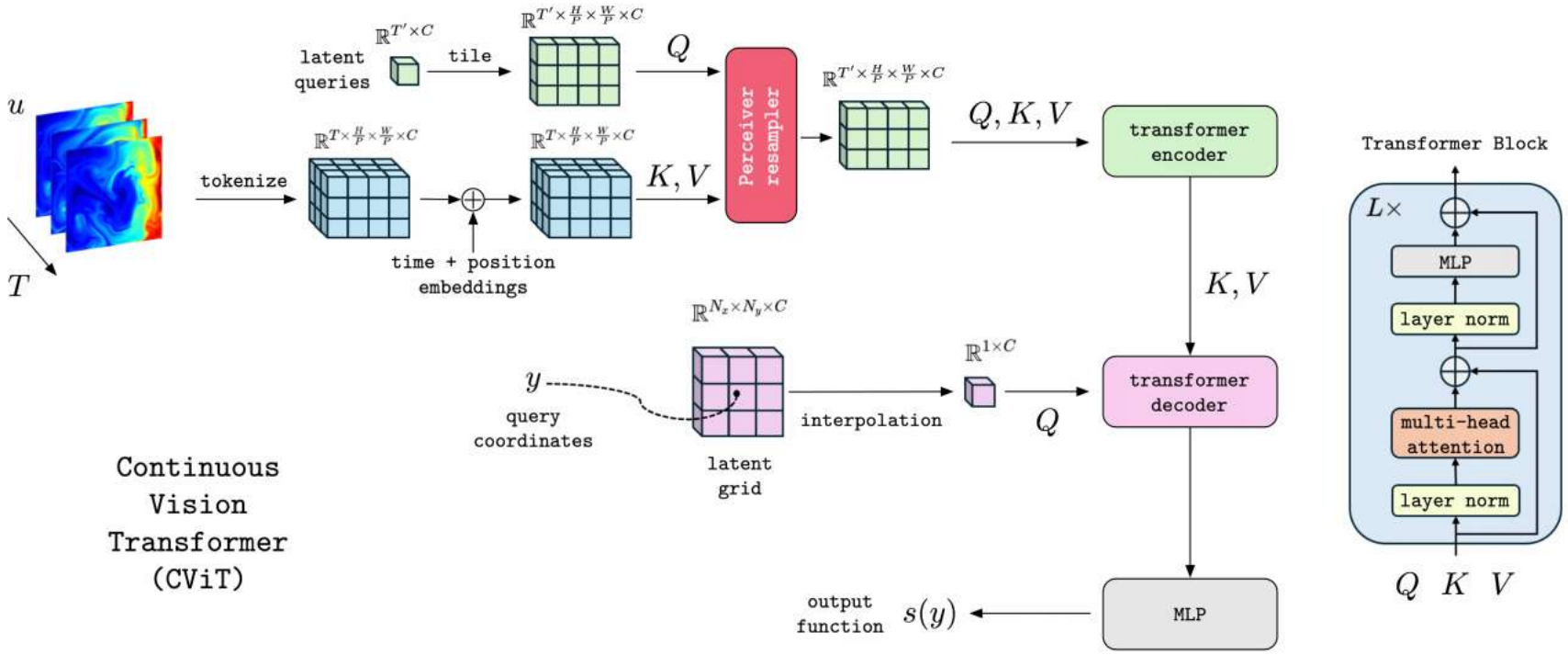


$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- $Q \in \mathbb{R}^{N \times d_k}$：查询矩阵（Query），由查询序列 X_q 经过投影得到。
- $K \in \mathbb{R}^{M \times d_k}$：键矩阵（Key），由输入序列 X_k 经过投影得到。
- $V \in \mathbb{R}^{M \times d_v}$：值矩阵（Value），由输入序列 X_v 经过投影得到。
- d_k：键向量的维度。
- N：查询序列的长度（查询点个数）。
- M：输入序列的长度（键/值的个数）。

Method

CVIT:


u
T
tokenize
latent queries
tile
Q
K, V
Perceiver resampler
time + position embeddings
Q, K, V
transformer encoder
K, V
transformer decoder
MLP
output function
s(y)
y
query coordinates
interpolation
latent grid
Continuous Vision Transformer (CViT)
Transformer Block
L×
MLP
layer norm
multi-head attention
layer norm
Q K V

Method

Continuous Vision Transformer (CViT) Architecture: CViT consists of the following components:

(1) Spatio-temporal patch embeddings to extract localized features.

(2) A temporal aggregation module based on the Perceiver architecture, which captures temporal correlations to compresses tokens along the time axis.

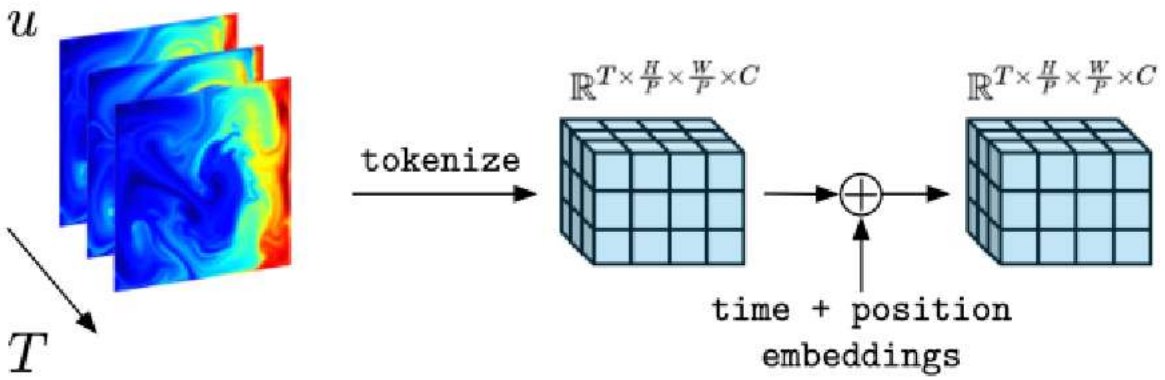
(3) A Transformer encoder that captures multi-scale spatial dependencies via self-attention layers.

(4) A novel grid-based positional encoding scheme for query coordinates, allowing for flexible output representation and interpolation.

(5) A cross-attention decoder that integrates information from the input function with query coordinates.

Method

(1) Spatio-temporal patch embeddings to extract localized features.

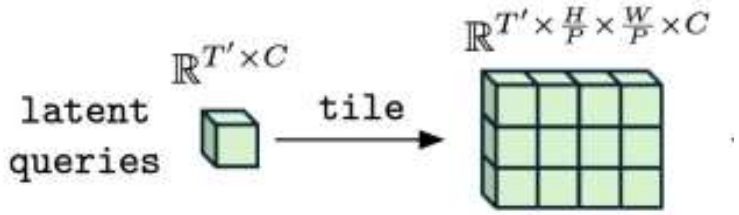


The vision transformer encoder takes as input a gridded representation of the input function u, yielding a spatio-temporal data tensor $u \in R^{T\times H\times W\times D}$ with D channels. We patchify our inputs into 3D tokens $u_p \in R^{T\times\frac{H}{P}\times\frac{W}{P}\times C}$ by tokenizing each 2D spatial frame independently, following the process used in standard Vision Transformers. We then add trainable 1D temporal and 2D spatial positional embeddings to each token:

$$u_{pe} = u_p + PE_t + PE_s, \qquad PE_t \in R^{T\times 1\times 1\times C}, \qquad PE_s \in R^{1\times\frac{H}{P}\times\frac{W}{P}\times C}$$

Method

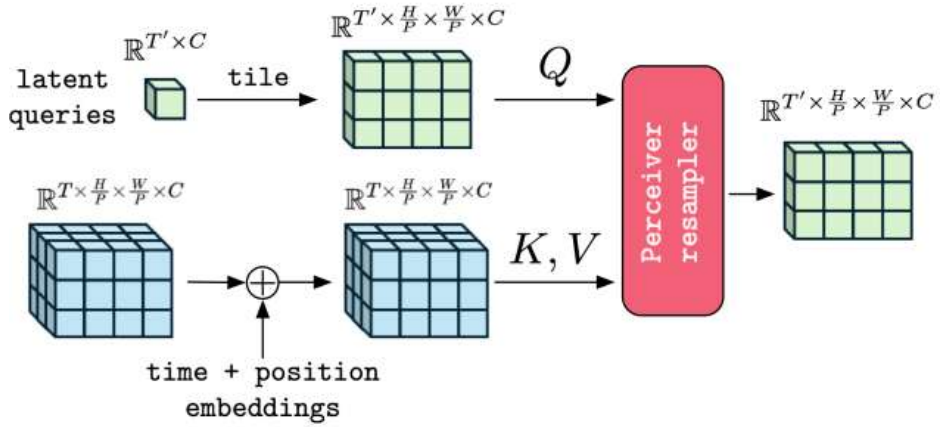
(2) A temporal aggregation module based on the Perceiver architecture.



To reduce computational cost, we use a temporal aggregation layer based on the Perceiver architecture. Specifically, we learn a pre-defined number of latent input queries $z \in R^{T' \times C}$. These queries serve as learnable parameters in a cross-attention module, which processes the visual features of our input.

Method

(2) A temporal aggregation module based on the Perceiver architecture.



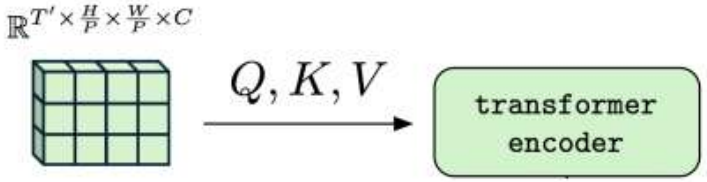
The Perceiver module operates on flattened visual features $u_f \in R^{(\frac{H}{P}\times\frac{W}{P})\times T\times C}$ obtained by flattening the positionally encoded features u_{pe}. Through this cross-attention mechanism, we aggregate the temporal information into a compact latent representation $z_{agg} \in R^{(\frac{H}{P}\times\frac{W}{P})\times T'\times C}$ as:

$$z^{'} = \hat{z} + MHA(LN(\hat{z}), LN(u_f), LN(u_f)),$$
$$z_{agg} = z^{'} + MLP(LN(z^{'})).$$

Here, z is initialized by a unit Gaussian distribution and $\hat{z} \in R^{(\frac{H}{P}\times\frac{W}{P})\times T'\times C}$ is obtained by tiling the latent query z $(\frac{H}{P}\times\frac{W}{P})$ times. Besides compressing tokens over the time axis, this module also enables the model to handle inputs with a variable number of time steps.

Method

(3) A Transformer encoder that captures multi-scale spatial dependencies via self-attention layers.

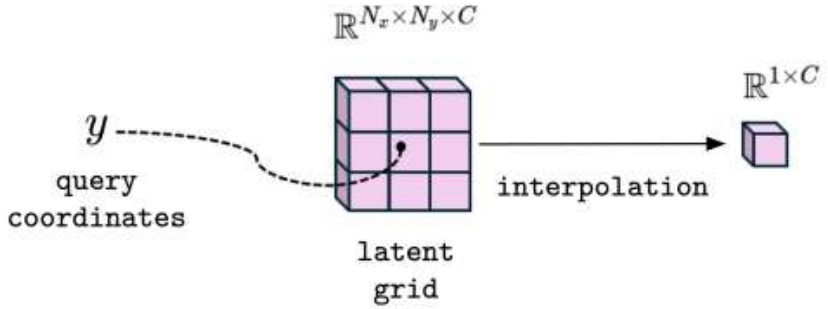


We then process the aggregated tokens z_{agg} using a sequence of L pre-norm Transformer blocks.

$$z_0 = LN(z_{agg}),$$
$$z_l^{'} = MSA(LN(z_{l-1})) + z_{l-1}, \quad l = 1,2,\ldots,L$$
$$z_l = MLP(LN(z_l^{'})) + z_l^{'}, \qquad l = 1,2,\ldots,L$$

Method

(4) A novel grid-based positional encoding scheme for query coordinates.



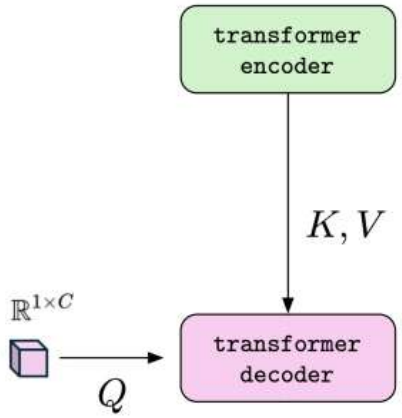
To enable continuous evaluation of outputs, we design a novel and efficient coordinate embedding to capture the fine-scale features of the target functions. Specifically, we create a uniform grid $\{y_{ij}\} \subset [0,1]^2$, for i $= 1, \ldots, N_x$ and j $= 1, \ldots, N_y$, along with associated trainable latent grid features $x \in R^{N_x \times N_y \times C}$. For a query point y $\in R^2$, we compute a NadarayaWatson interpolant over grid latent features:

$$\mathbf{x}' = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} w_{ij} \mathbf{x}_{ij}, \quad w_{ij} = \frac{\exp\left(-\beta \|y - \mathbf{y}_{ij}\|^2\right)}{\sum_{ij} \exp\left(-\beta \|y - \mathbf{y}_{ij}\|^2\right)}.$$

Here β > 0 is a hyperparameter that determines the locality of the interpolated features. Specifically, larger values of β yield more localized weight distributions ω_{ij} , resulting in a higher-frequency interpolant that captures finer-scale variations. Conversely, smaller β values produce a smoother interpolant by averaging over a broader neighborhood of points.

Method

(5) A cross-attention decoder that integrates information from the input function with query coordinates.



The interpolated grid feature x_0 = x′ ∈ $R^{1\times C}$ is used as query input to a transformer decoder. Thisdecoder uses the output of the vision transformer encoder z_L as keys and values in a cross-attention mechanism:

$$\mathbf{x}'_k = \mathbf{x}_{k-1} + \mathrm{MHA}\left(\mathrm{LN}\left(\mathbf{x}_{k-1}\right), \mathrm{LN}\left(\mathbf{z}_L\right), \mathrm{LN}\left(\mathbf{z}_L\right)\right), \qquad k = 1 \ldots K,$$
$$\mathbf{x}_k = \mathbf{x}'_k + \mathrm{MLP}\left(\mathrm{LN}\left(\mathbf{x}'_k\right)\right), \qquad k = 1 \ldots K.$$

Experiment

For all experiments, unless otherwise stated, we use a patch size of 8×8 for tokenizing inputs.

Table 1: Details of Continuous Vision Transformer model variants.

Model	Encoder layers	Embedding dim	MLP width	Heads	# Params
CViT-S	5	384	384	6	13 M
CViT-B	10	512	512	8	30 M
CViT-L	15	768	1536	12	92 M

Model	# Params	NS	CNS	DR
FNO	0.5 M	9.12 %	9.60 %	12.00 %
FFNO	1.3 M	8.39 %	5.20 %	5.71 %
GK-T	1.6 M	9.52%	3.77%	3.59 %
GNOT	1.8 M	17.20 %	4.20 %	3.11 %
Oformer	1.9 M	13.50 %	6.25 %	1.92 %
DPOT-Ti	7 M	12.50 %	3.97 %	3.21 %
MPP-S	30M	-	-	1.12 %
DPOT-S	30 M	9.91 %	3.37 %	3.79 %
MPP-L (Pre-trained)	400 M	-	-	0.98 %
DPOT-L (Pre-trained)	500 M	7.98 %	2.16 %	2.32 %
DPOT-L (Fine-tuned)	500 M	2.78 %	1.31 %	0.73 %
DPOT-H (Pre-trained)	1.03 B	3.79 %	1.80 %	1.91 %
CViT-S	13 M	**3.75 %**	**2.71 %**	**1.13 %**
CViT-B	30 M	**3.18 %**	**1.99 %**	**1.11 %**
CViT-L	92 M	**2.35 %**	**1.29 %**	**0.68 %**

Experiment

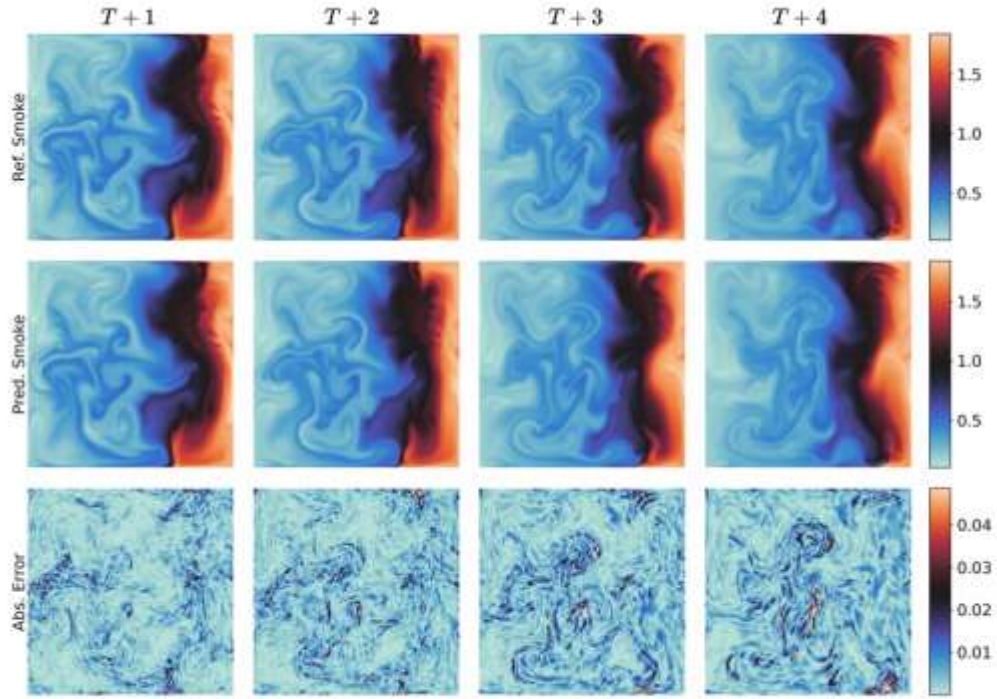


Figure 5: *Incompressible Navier-Stokes benchmark (NS).* Representative CViT rollout predictions of the passive scalar field, and point-wise error against the ground truth.

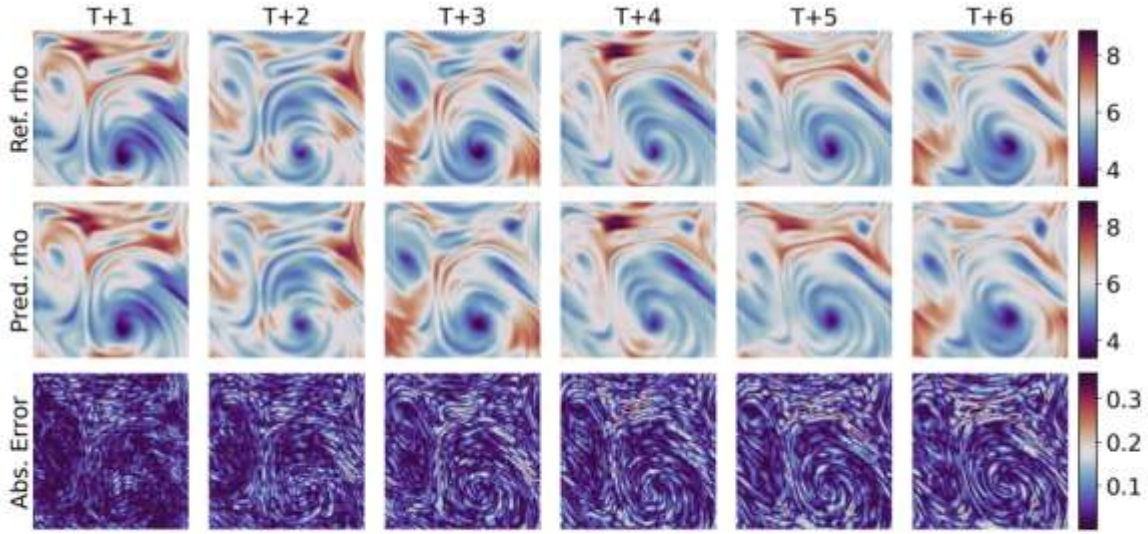


Figure 6: *Compressible Navier-Stokes Benchmark (CNS).* Representative CViT rollout prediction of the density field ρ, and point-wise error against the ground truth.

Thanks