

A MORE METHODOLOGY DETAILS

A.1 DECODER

The dimension of the feature map is $H/16$ and $W/16$. After going through the Decoder operation, the output dimension becomes H and W , which is the target dimension. The detailed process is as follows:

$$X_0 \in \mathbb{R}^{B \times \text{embed_dim} \times \frac{H}{16} \times \frac{W}{16}} \quad (14)$$

Here, X_0 is the tensor input to this module.

$$X_1 = \text{Tanh}(\text{ConvTranspose2d}(X_0)) \in \mathbb{R}^{B \times \text{out_channels} \times \frac{H}{8} \times \frac{W}{8}} \quad (15)$$

In this step, we use a transposed convolution layer followed by a Tanh activation function to obtain the tensor X_1 .

$$X_2 = \text{Tanh}(\text{ConvTranspose2d}(X_1)) \in \mathbb{R}^{B \times \text{out_channels} \times \frac{H}{4} \times \frac{W}{4}} \quad (16)$$

Next, we again use a transposed convolution layer and Tanh activation function to get the tensor X_2 .

$$X_3 = \text{ConvTranspose2d}(X_2) \in \mathbb{R}^{B \times \text{out_channels} \times H \times W} \quad (17)$$

Finally, we apply another transposed convolution layer to get the output tensor X_3 with dimensions H and W .

A.2 LOSS FUNCTION DETAILS

Our loss function is as follows, inside the main text, we have described \mathcal{L}^{MSE} and \mathcal{L}^{ADV} in detail, next, we describe \mathcal{L}^{SSIM} in detail.

$$\mathcal{L} = \mathcal{L}^{MSE} + \mathcal{L}^{SSIM} + \mathcal{L}^{ADV}. \quad (18)$$

SSIM (Structural Similarity Index) is a metric used to assess the structural similarity of two images. It is proposed to better reflect the human eye's subjective perception of image quality, and provides a more intuitive and accurate assessment of image quality than the traditional mean square error (MSE) or peak signal-to-noise ratio (PSNR).

Specifically, two four-dimensional tensors of dimension $[T \times C \times H \times W]$ are given: predicted data P and real labeled data G , where T stands for the time dimension or batch size, C is the number of channels, and H and W are the height and width of the tensor, respectively. To compute the SSIM loss, a window of fixed size w (e.g., a Gaussian window of 11×11) and two constants for stabilizing the denominator c_1 and c_2 are first chosen.

For each sample t and each channel c , we define the following calculations:

1. mean value:

$$\mu_{P_{t,c}} = w \cdot P_{t,c} \quad (19)$$

$$\mu_{G_{t,c}} = w \cdot G_{t,c} \quad (20)$$

2. variance:

$$\sigma_{P_{t,c}}^2 = w \cdot P_{t,c}^2 - \mu_{P_{t,c}}^2 \quad (21)$$

$$\sigma_{G_{t,c}}^2 = w \cdot G_{t,c}^2 - \mu_{G_{t,c}}^2 \quad (22)$$

3. covariance:

$$\sigma_{P_{t,c}G_{t,c}} = w \cdot P_{t,c} \cdot G_{t,c} - \mu_{P_{t,c}} \cdot \mu_{G_{t,c}} \quad (23)$$

Then, the SSIM value for each position is calculated using the SSIM formula:

$$\text{SSIM}_{t,c} = \frac{(2\mu_{P_{t,c}}\mu_{G_{t,c}} + c_1)(2\sigma_{P_{t,c}G_{t,c}} + c_2)}{(\mu_{P_{t,c}}^2 + \mu_{G_{t,c}}^2 + c_1)(\text{sigma}_{P_{t,c}}^2 + \sigma_{G_{t,c}}^2 + c_2)} \quad (24)$$

Average over all positions to obtain the SSIM value for that channel. Finally, average the SSIM values over all samples and channels and subtract this value from 1 to get the SSIM loss:

$$\mathcal{L}^{SSIM} = 1 - \text{mean}(\text{SSIM}_{t,c}) \quad (25)$$

For efficiency, the computation of mean, variance and covariance can be realized by convolution operation.

A.3 ALGORITHMIC PROCESS

The proposed algorithm, named STAC, offers a novel approach to model the spatio-temporal evolution in dynamical systems. At its core, the algorithm integrates a twin spatio-temporal encoder, which captures both spatial and frequency domain semantics, and a cache-based recurrent propagator, which leverages historical data to enhance long-term dynamics prediction. The encoder consists of a Frequency-enhanced Spatial Module (FSM) and an ODE-enhanced Temporal Module (OTM), which are then fused together. The recurrent propagator utilizes a cache mechanism to store and update previous representations. Finally, a decoder transforms the updated feature maps into predicted trajectories, which are then optimized using a combination of loss functions. The algorithm aims to provide accurate and reliable long-term predictions for dynamical systems.

Algorithm 1 The STAC Approach

```

0: function METHOD(Input: system states in interval  $[0, T^{obs}]$ ) {Twin Spatio-temporal Encoder}
0:    $I^{in} \leftarrow \text{Input}$ 
0:    $I^{out} \leftarrow \text{FSM}(I^{in})$  {Frequency-enhanced Spatial Module}
0:    $F_0 \leftarrow \text{OTM}(I^{in})$  {ODE-enhanced Temporal Module}
0:    $X \leftarrow \text{IFTM}(I^{out}, F_0)$  {Information Fusion between Twin Modules}
0:   {Cache-based Recurrent Propagator}
0:   Initialize cache  $\mathcal{M}$  with size  $R$ 
0:   for  $m = 1$  to  $M$  do
0:      $Q_m \leftarrow \text{UpdateCache}(Q_{m-1}, X_m, \mathcal{M})$ 
0:     Add  $X_{m-1}$  to cache  $\mathcal{M}$ 
0:   end for
0:   {Decoder and Optimization}
0:    $Y_{\text{hat}} \leftarrow \text{Decode}(Q_m)$ 
0:   Loss  $\leftarrow \text{CalculateLoss}(Y_{\text{hat}}, \text{GroundTruth})$ 
0:   UpdateModelParameters(Loss)
0:   return  $Y_{\text{hat}}$ 
0: end function=0

```

B DETAILED DESCRIPTION OF BENCHMARKS

We summarize the benchmark configurations in Tab. Here are the details of the dataset.

B.1 TURBULENCE DATASET

This dataset (Khojasteh et al., 2022) contains Eulerian velocity fields and pressure fields. An open-source direct numerical simulation (DNS) flow solver named Incompact3d was used to compute the Eulerian fields around the cylinder. Following the original thesis setup, highly resolved direct numerical simulations (DNS) of the flow over a smooth cylinder at a subcritical Reynolds number of 3900 (based on the diameter D of the cylinder and the diameter D of the freestream velocity) were performed to generate the data. Double-precision Eulerian and Lagrangian fields were collected for both subdomains as shown in Fig. 7. Due to online cloud storage limitations, every 10 DNS time steps

were saved every 10 DNS time steps (saving each time step would require about 30 TB of storage space per vortex shed). The 1000 snapshots are also used for smaller subdomains with dimensions of $4D \times 2D \times 2D$ (i.e., per DNS time step). Subdomain 2 is suitable for studies that require the highest possible temporal resolution. Detailed information on the two subdomains can be found in Table 2. An Eulerian snapshot of the current tail stream is shown in Fig. 2. For both subdomains, Lagrangian trajectories are provided for about 200000 synthetic particles. Three main categories are provided in the data repository: subdomain 1, subdomain 2, and software. Snapshots are in text format (.txt) and are collected in compressed files (.zip). There are no special requirements for reading and opening the data. Euler 3D snapshots are saved in vector format. Therefore, they need to be extracted within three internal loops in the xyz direction.

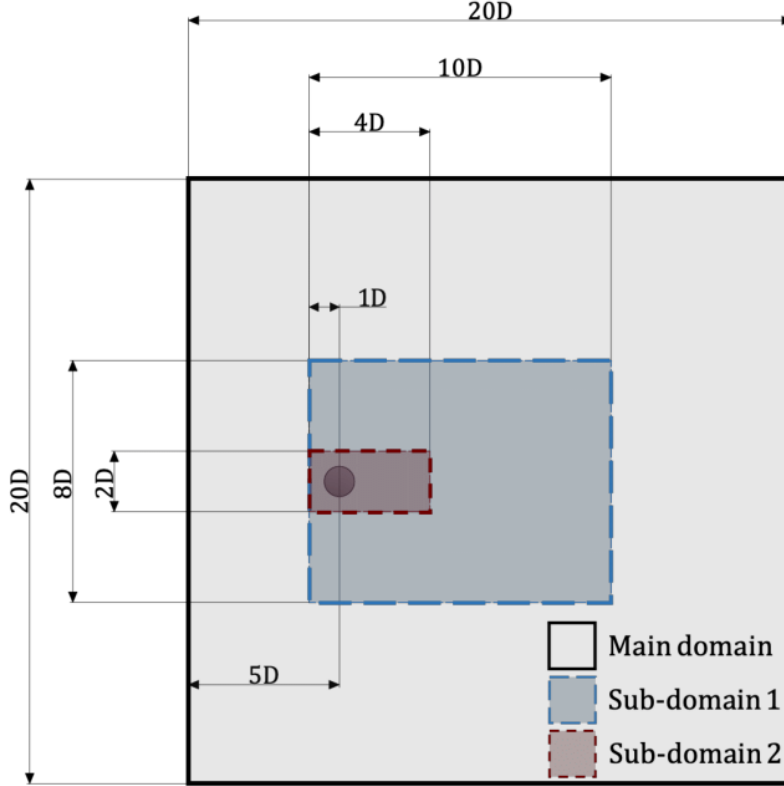


Figure 7: The flow around a smooth cylinder at a subcritical Reynolds number of 3900, with dimensions of two computational subdomains. (Khojasteh et al., 2022)

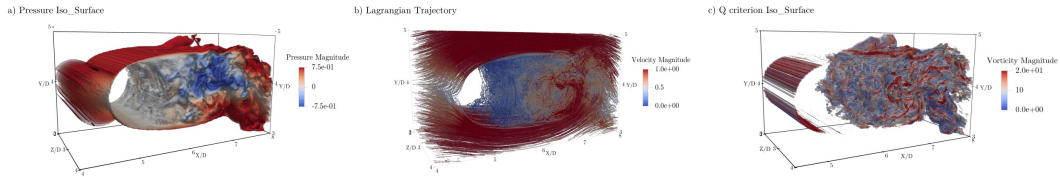


Figure 8: Snapshot Overview of Sub-domain 2: (a) Pressure iso-surface highlighted by the intensity of the pressure. (b) Lagrangian trajectories of 20,000 particles, visualized after 1,000 DNS time steps, color-coded by velocity magnitude. (c) Q-criterion representation depicting Eulerian flow structures, color-graded by the magnitude of vorticity.

B.2 ERA5 DATASET

ERA5 is the latest global reanalysis product released by the European Centre for Medium-Range Weather Forecasts (ECMWF). It provides researchers and meteorologists with high-resolution meteorological and climatic data from 1979 to the present. The spatial resolution of ERA5 data is 31 kilometers, with a temporal resolution of hourly, representing a significant improvement over previous reanalysis products. It encompasses observations from the atmosphere, land, and oceans, offering invaluable data resources for global climate change, weather forecasting, and other related studies. The data quality and accuracy of ERA5 have been widely recognized by researchers, making it an essential tool in climate research and meteorological forecasting.

We have selected global temperature field data and local velocity field data as our training set. The former has a resolution of 1440x720. In our experiments, we downsampled it to a quarter of its original size, i.e., 360x180. The velocity field has a resolution of 64x64. For the temperature data, we employed autoregressive training, using 5 days of data as input and 10 days of data as output. For the velocity field, we chose 8 hours of data as input and another 8 hours of data as output.

B.3 SEVIR DATASET

The Storm Event Imagery (SEVIR) dataset presents a meticulously curated set of spatiotemporally synchronized images, capturing meteorological phenomena via the GOES-16 geostationary satellite and the NEXRAD weather radar system. Encompassing in excess of 10,000 distinct weather events, each individual event in this collection showcases an image sequence persisting for a duration of 4 hours, spanning a geographical expanse of 384 km x 384 km. Delving into this dataset can expedite advancements in the realms of weather sensing, hazard avoidance, near-term forecasting, and other pertinent meteorological applications.

We follow the same setup as Earthformer, using 13 frames as input and 12 frames as output.

B.4 FLAME FLOW FIELD DATASET

In this study, we select a typical highway tunnel for simulation, with dimensions of 50 meters in length, 10 meters in width, and 5 meters in height. The fire source has dimensions of 4.6 meters in length, 1.8 meters in width, and 2.4 meters in height. The top surface of the truck is set as a "burner" type. To simulate a realistic scenario, the maximum heat release rate (HRR) of the fire source is set at 20 MW, a value recommended by the standard for the maximum HRR of tunnel fires in the event of a truck fire. The fire source is modeled as a propane gas fire, with its HRR growing at a t^2 rate. Four operating conditions are designed. In all four scenarios, the power of the fire source is consistently 20 MW. In the first scenario, the fire source growth factor is 0.0029 kW/s^2 , with the time to reach steady state being 2626 seconds and another steady state time being 2700 seconds. In the second scenario, the fire source growth factor is 0.0117 kW/s^2 , with the times to reach steady state being 1307 seconds and 1400 seconds, respectively. In the third scenario, the fire source growth factor is 0.0469 kW/s^2 , with the steady state times being 653 seconds and 700 seconds. Lastly, in the fourth scenario, the fire source growth factor is 0.1876 kW/s^2 , with the times to reach steady state being 326 seconds and 400 seconds. The choice of actual tunnel dimensions, fire source size, and HRR values ensures the validity and relevance of the simulation results, providing a solid foundation for the proposed artificial intelligence fire prediction method.

In this study, the input dimensions are set at [10,2,80,480], while the output dimensions are [90,2,80,480]. Here, the input duration of 10 seconds represents the observation time, and the value of 2 corresponds to the temperature field and smoke field, both of which have a resolution of 80x480. The output duration of 90 seconds is used for extended time-range predictions. To achieve this long-term forecasting, we employ a rollout strategy. Moreover, the caching mechanism introduced in this paper plays a pivotal role in enhancing the accuracy and efficiency of long-term predictions.

B.5 KTH DATASET

The KTH dataset stands as a benchmark in the domain of human activity recognition, stemming from the esteemed KTH Royal Institute of Technology in Sweden. This collection distinctly captures six

human activities: walking, jogging, running, boxing, hand waving, and hand clapping. Across diverse scenarios—ranging from outdoor settings (s1, s4) and scaled outdoor variations (s2, s3) to indoor environments (s5, s6)—25 participants, donning varied attire, repetitively perform these actions. Each video in this dataset is recorded at a clarity of 128x128 pixel resolution and maintains a consistent frame rate of 25 frames per second.

B.6 DYNAMIC SYSTEM DATASETS RECORDED BY VIDEO.

We have provided 9 datasets of dynamic system, recorded in the form of videos. Both the input and output dimensions are [10,3,128,128], indicating an input length of 10 time steps and an output length of 10 time steps. Since they are recorded as videos, there are 3 channels, representing RGB, with a resolution of 128x128 for each image.

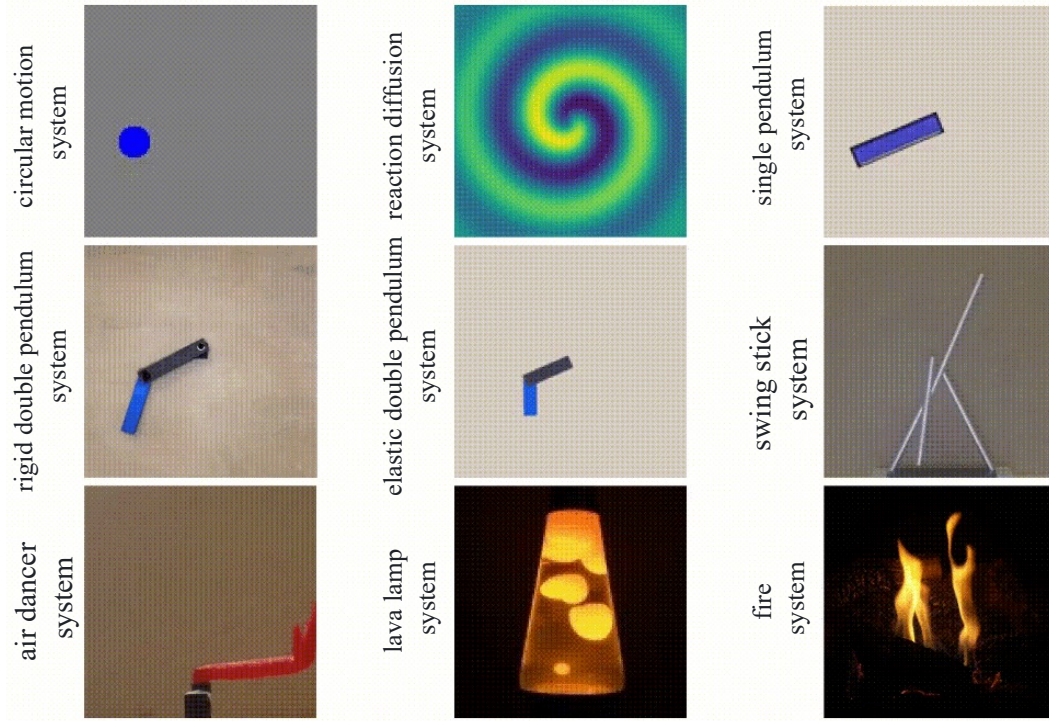


Figure 9: A case study of a dataset of dynamical systems recorded by video

1. Circular motion system. (CMS) An object moves uniformly along a fixed radius. Formula as follows:

$$v = r\omega, \quad (26)$$

$$a = r\omega^2, \quad (27)$$

where v is the linear speed, r is the radius of the circle, ω is the angular speed, and a is the centripetal acceleration.

2. Reaction diffusion system. (RDS) A system describing how the concentration of substances changes over time due to reactions and diffusion. Formula as follows:

$$\frac{\partial u}{\partial t} = D_u \nabla^2 u - uv^2 + F(1 - u), \quad (28)$$

$$\frac{\partial v}{\partial t} = D_v \nabla^2 v + uv^2 - (F + k)v, \quad (29)$$

where u and v are concentrations, D_u and D_v are their diffusion coefficients, and F and k are system parameters.

3. Single pendulum system. (SPS) A point mass hung from a fixed point swings due to gravity. Formula as follows:

$$\frac{d^2\theta}{dt^2} = -\frac{g}{l} \sin(\theta), \quad (30)$$

where θ is the pendulum angle, l is the length of the pendulum, and g is gravitational acceleration.

4. Rigid double pendulum system. (RDPS) A complex pendulum system consisting of two pendulums, with one pendulum attached to the end of another. The equations of a double pendulum are usually relatively complex and involve multiple variables. But the basic idea is to use the Lagrange equation. For a simplified description:

$$\frac{d^2\theta_1}{dt^2} = \frac{g(\sin\theta_2 \cos(\theta_1 - \theta_2) - \sin\theta_1) - (\ell_2\ddot{\theta}_2 + \ell_1\dot{\theta}_1^2 \sin(\theta_1 - \theta_2)) \cos(\theta_1 - \theta_2)}{\ell_1(\cos^2(\theta_1 - \theta_2) - 1)}, \quad (31)$$

$$\frac{d^2\theta_2}{dt^2} = \frac{g \sin\theta_1 \cos(\theta_1 - \theta_2) - \ell_1\dot{\theta}_1^2 \sin(\theta_1 - \theta_2) - g \sin\theta_2}{\ell_2(\cos^2(\theta_1 - \theta_2) - 1)}, \quad (32)$$

5. Elastic double pendulum system. (EDPS) Similar to the double pendulum, but the connecting component between the pendulums is elastic. The basic mathematical description of an elastic pendulum involves Hooke's law of springs and the motion of a pendulum. A simplified description is:

$$\frac{d^2x}{dt^2} = -kx/m - g \sin\theta, \quad (33)$$

$$\frac{d^2\theta}{dt^2} = -g/x \cos\theta, \quad (34)$$

where x is the displacement from the equilibrium position, k is the spring constant, and m is the mass.

6. Swing stick system. (SSS) A long stick with a fixed endpoint that swings under the influence of gravity and other possible external forces. The pendulum system is equivalent to a long pendulum. The basic description is similar to a simple pendulum, but requires consideration of the mass distribution and length of the rod. The simplest description is:

$$\frac{d^2\theta}{dt^2} = -\frac{3g}{2L} \sin\theta, \quad (35)$$

where L is half the length of the rod.

7. Air Dancer System. (ADS) For the air dancer, it is crucial to consider the influence of the gas flow. This can be described by the incompressible Navier-Stokes equation:

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla) \mathbf{u} = -\frac{1}{\rho} \nabla p + \nu \nabla^2 \mathbf{u}, \quad (36)$$

$$\nabla \cdot \mathbf{u} = 0, \quad (37)$$

where \mathbf{u} is the velocity, p is the pressure, ρ is the density, and ν is the kinematic viscosity.

8. Lava Lamp System. (LLS) At the heart of the lava lamp is the fluid flow caused by density changes due to temperature. This can be described using the Navier-Stokes equation and the Boussinesq approximation:

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla) \mathbf{u} = -\frac{1}{\rho_0} \nabla p + \nu \nabla^2 \mathbf{u} - \alpha(T - T_0) \mathbf{g}, \quad (38)$$

$$\nabla \cdot \mathbf{u} = 0, \quad (39)$$

$$\frac{\partial T}{\partial t} + \mathbf{u} \cdot \nabla T = \kappa \nabla^2 T, \quad (40)$$

where T is the temperature, α is the thermal expansion coefficient, κ is the thermal diffusivity, and \mathbf{g} is the acceleration due to gravity.

9. Fire System. (FS) The fire system involves chemical reactions, heat transfer, and fluid dynamics. A common description is:

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla) \mathbf{u} = -\frac{1}{\rho} \nabla p + \nu \nabla^2 \mathbf{u} + \text{source terms due to combustion}, \quad (41)$$

$$\nabla \cdot \mathbf{u} = 0, \quad (42)$$

$$\frac{\partial T}{\partial t} + \mathbf{u} \cdot \nabla T = \kappa \nabla^2 T + \text{source terms due to combustion}, \quad (43)$$

$$\frac{\partial Y_i}{\partial t} + \mathbf{u} \cdot \nabla Y_i = D \nabla^2 Y_i + \text{reaction rate of species } i, \quad (44)$$

where Y_i is the mass fraction of the i -th chemical species, and D is the diffusion coefficient.

C EXPERIMENTAL DETAILS

C.1 EVALUATION METRICS

We utilize the following metrics:

Mean Squared Error (MSE) Given the predicted data dimension $Y_{\text{pred}} \in \mathbb{R}^{T \times C \times H \times W}$ and the label data dimension $Y_{\text{label}} \in \mathbb{R}^{T \times C \times H \times W}$, the MSE is computed as:

$$\text{MSE}(Y_{\text{pred}}, Y_{\text{label}}) = \frac{1}{T \times C \times H \times W} \sum_{t=1}^T \sum_{c=1}^C \sum_{h=1}^H \sum_{w=1}^W (Y_{\text{pred}}^{tchw} - Y_{\text{label}}^{tchw})^2 \quad (45)$$

Mean Absolute Error (MAE) The MAE is given by:

$$\text{MAE}(Y_{\text{pred}}, Y_{\text{label}}) = \frac{1}{T \times C \times H \times W} \sum_{t=1}^T \sum_{c=1}^C \sum_{h=1}^H \sum_{w=1}^W |Y_{\text{pred}}^{tchw} - Y_{\text{label}}^{tchw}| \quad (46)$$

Anomaly Correlation Coefficient (ACC) The ACC, often used in meteorology, is defined as:

$$\text{ACC} = \frac{\sum (Y_{\text{pred}} - \bar{Y}_{\text{pred}})(Y_{\text{label}} - \bar{Y}_{\text{label}})}{\sqrt{\sum (Y_{\text{pred}} - \bar{Y}_{\text{pred}})^2 \sum (Y_{\text{label}} - \bar{Y}_{\text{label}})^2}} \quad (47)$$

where \bar{Y}_{pred} and \bar{Y}_{label} represent the means of Y_{pred} and Y_{label} , respectively.

Structural Similarity Index (SSIM) For each local window or region, the SSIM is calculated as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x \mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (48)$$

where x and y are pixel values within two windows or regions, μ_x and μ_y are their means, σ_x^2 and σ_y^2 are their variances, and σ_{xy} is their covariance. C_1 and C_2 are small constants to prevent division by zero.

Peak Signal-to-Noise Ratio (PSNR) The PSNR is given by:

$$\text{PSNR} = 10 \times \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right) \quad (49)$$

where MAX represents the maximum possible pixel value. For an 8-bit image, MAX = 255.

C.2 HYPERPARAMETERS

In the experimental settings, various hyperparameters are set for different datasets. For the attention head, the Turbulence, KTH, and Video DS datasets are set to 2, while the ERA5, SEVIR, and FLAME FLOW datasets are set to 4. The Fourier Transform Layers are configured as follows: 6 for Turbulence, ERA5, and SEVIR; 4 for FLAME FLOW; 10 for KTH; and 12 for Video DS. The hidden layer dimension in both the Feature Selection Module (FSM) and Other Training Module (OTM) is set to 64 for Turbulence, ERA5, and KTH, and 128 for SEVIR, FLAME FLOW, and Video DS. Across all datasets, the learning rate is consistently set at 0.01. In terms of the number of epochs, Turbulence, ERA5, SEVIR, and KTH have 500 epochs, FLAME FLOW has 300, and Video DS has 100. Lastly, the batch sizes vary: 2 for Turbulence, 6 for ERA5, 10 for both SEVIR and KTH, 4 for FLAME FLOW, and 20 for Video DS.

Table 3: Hyperparameters for Different Datasets

Hyperparameter	Turbulence	ERA5	SEVIR	FLAME FLOW	KTH	Video DS
Attention head	2	4	4	4	2	2
Fourier Transform Layers	6	6	6	4	10	12
Hidden layer dimension in FSM	64	64	128	128	64	128
Hidden layer dimension in OTM	64	64	128	128	64	128
Learning rate	0.01	0.01	0.01	0.01	0.01	0.01
Number of epochs	500	500	500	300	500	100
Batch size	2	6	10	4	10	20

D ADDITIONAL EXPERIMENTS

Table 4: Performance of Models on Various Datasets

MODEL	Datasets								
	CMS	RDS	SPS	RDPS	EDPS	SSS	ADS	LLS	FS
ADFV	87.24	90.93	94.78	93.43	91.32	87.74	88.65	82.41	92.87
STAC	88.64	92.14	96.34	95.26	93.14	88.96	90.12	83.90	94.54

Our dataset is derived from (Chen et al., 2022), nine dynamics system datasets, recorded in video format. Comparison experiment with the model of the original text, we named the original text model as ADFV, because it is a video recording, we use SSIM as the evaluation metrics, and the experimental results are shown in the Table 4.

According to the experimental results, the STAC model performs better on all datasets compared to the original ADFV model. Specifically, the SSIM scores of the STAC model are higher than those of the ADFV model both on the CMS, RDS, SPS, RDPS, EDPS, SSS, ADS, LLS, and FS datasets, which clearly highlights the advantages and efficiency of the STAC model. These results demonstrate the strong performance and reliability of the STAC model in dealing with dynamic systems of video recordings.

E LONG-TERM PREDICTION RESULTS OF STAC

In this section, we present the complete visualization results of STAC on the long-term prediction benchmark. We observe that on the Flame benchmark, our model is capable of excellently reconstructing the forecast results over an extended time frame, nearly encapsulating detailed contour information of the fire dynamics as well as the flow velocity. The astonishing consistency between the ground-truth and prediction at 210 seconds further substantiates our model’s prowess in long-term forecasting.

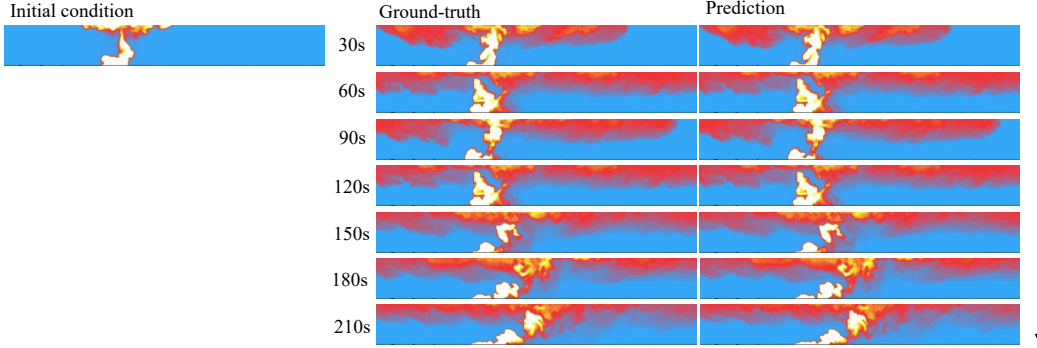


Figure 10: Case study of predicting the fire temperature field for the next 210 seconds based on the dynamic of the past 30 seconds.

F ABLATION STUDY

F.1 DATASET DETIALS

We conduct a comprehensive ablation study on STAC, with the dataset sourced from (Yin et al., 2022). The dataset is based on the shallow water equation, and our primary focus is on modeling its velocity field. The dataset has a basic dimensionality of $[160, 1, 128, 256]$, which signifies 160 consecutive time snapshots. The 1 in the dimension stands for the channel variable, representing velocity. The resolution of the velocity field is captured by the 128×256 dimension. We slice the dataset into two sections: $[20, 1, 128, 256]$ and $[140, 1, 128, 256]$. The former serves as the input, while the latter acts as the ground truth. Subsequently, these slices are fed into various model variants for ablation experiments.

F.2 EXPERIMENTAL RESULTS

To gain deeper insights into the significance and impact of each component within the *Spatio-temporal Twins with A Cache* (STAC) model, we conducted a series of ablation experiments. Firstly, we set up a basic model for comparison, termed as *Basic STAC*. Subsequently, we removed specific components from the STAC model one at a time to evaluate their contributions to the overall performance. Specifically, we examined the following model variants:

1. **STAC w/o FESM** - without the Frequency-enhanced Spatial Module.
2. **STAC w/o ODETM** - excluding the ODE-enhanced Temporal Module.
3. **STAC w/o CA** - devoid of the Channel Attention module.
4. **STAC w/o CRP** - lacking the Cache-based Recursive Propagator.
5. **STAC w/o TF&M** - omitting Teacher Forcing and Mixup.
6. **STAC w/o SSAL** - without the Semi-supervised Adversarial Learning.

To thoroughly evaluate the influence of each component within the STAC model on predictions, we have devised the following ablation strategy: Initially, the "Basic STAC" serves as our benchmark. Then, by eliminating the Frequency-enhanced Spatial Module (STAC w/o FESM) and the ODE-enhanced Temporal Module (STAC w/o ODETM), we assess their impact on prediction intricacies. Lastly, delving deeper into the pivotal components for long-term predictive robustness, we evaluate the variant devoid of the Cache-based Recursive Propagator (STAC w/o CRP) and the variant excluding Teacher Forcing and Mixup (STAC w/o TF&M). These experiments are geared towards revealing the specific contributions of each component to short-term prediction accuracy and long-term predictive stability.

All these model variants were trained and tested on the same dataset to ensure the consistency of results. Through these ablation experiments, our goal is to pinpoint which components play a pivotal role in the model's performance and which might be auxiliary.

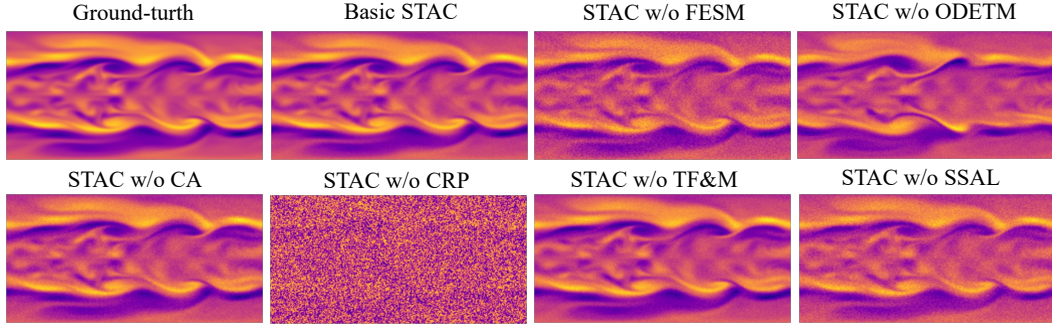


Figure 11: Visualization of the ablation experiment results.

Table 5: Ablation study on the Spherical Shallow Water dataset using RMSE as the evaluation metric.

Model Variant	RMSE
Basic STAC	0.0287
STAC w/o FESM	0.0485
STAC w/o ODETM	0.0464
STAC w/o CA	0.0344
STAC w/o CRP	0.0795
STAC w/o TF&M	0.0311
STAC w/o SSAL	0.0421

The Basic STAC model, with an RMSE of 0.0287, serves as our benchmark. The absence of the Frequency-enhanced Spatial Module elevates the RMSE to 0.0485, highlighting its significance. Similarly, excluding the ODE-enhanced Temporal Module pushes RMSE to 0.0464, underscoring its importance. The omission of the Channel Attention leads to a modest rise in RMSE to 0.0344. Notably, removing the Cache-based Recursive Propagator causes the most dramatic increase, with RMSE soaring to 0.0795, emphasizing its pivotal role in long-term predictions. Minor shifts are observed when removing Teacher Forcing and Mixup, and the Semi-supervised Adversarial Learning, with RMSE values of 0.0311 and 0.0421 respectively. In essence, the CRP stands out as the most critical component, followed closely by FESM and ODETM. To optimize the model further, it's recommended to delve deeper into enhancing the CRP, FESM, and ODETM functionalities.