ICLR
International Conference On Learning Representations

ConvT3: Structured State Kernels for Convolutional State Space Models

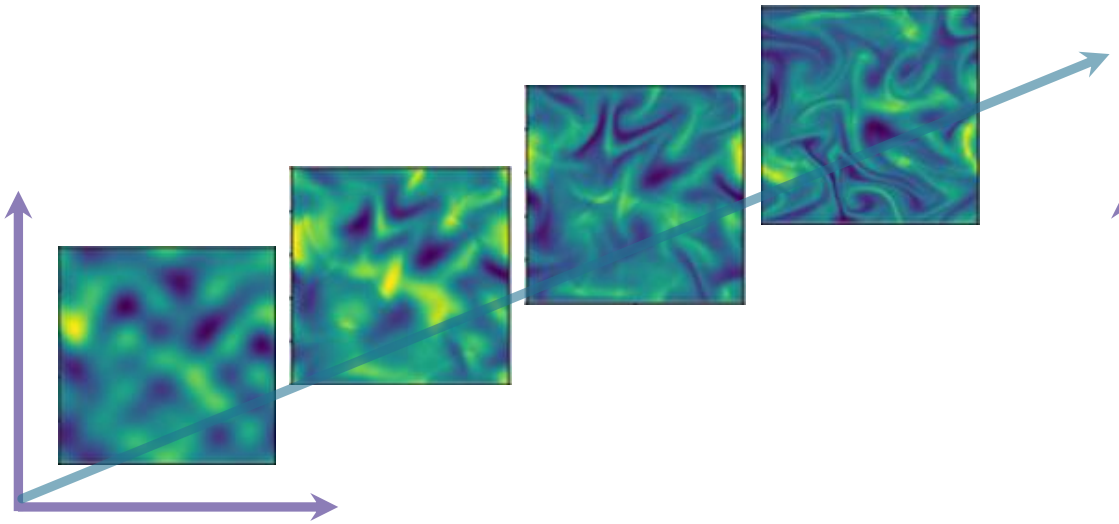
Jaeyoung Hong[*,1,2] Yun Young Choi[*,1] Joohwan Ko[3] Minseon Gwak[†,4]

[1]SolverX [2]Seoul National University
[3]University of Massachusetts Amherst [4]Pohang University of Science and Technology

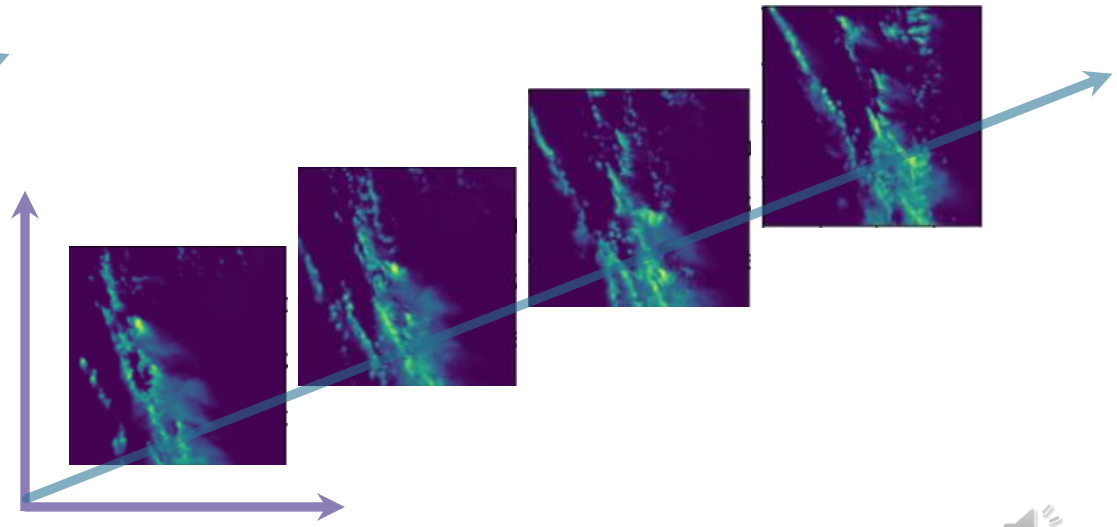
SolverX

Background: Spatiotemporal Modeling

- Spatiotemporal modeling
  - Simultaneously model spatial correlations and temporal dependencies



PDE Simulation



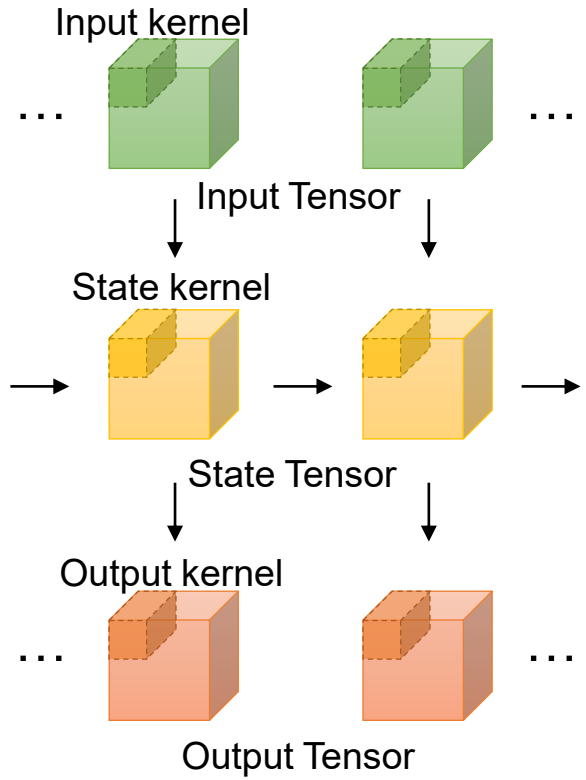
Weather Forecasting

Background: Convolutional State Space Models

- Convolutional State Space Model (ConvSSM)

$$\mathcal{X}'(t) = \boldsymbol{\mathcal{A}} * \mathcal{X}(t) + \boldsymbol{\mathcal{B}} * \mathcal{U}(t)$$
$$\mathcal{Y}(t) = \boldsymbol{\mathcal{C}} * \mathcal{X}(t) + \boldsymbol{\mathcal{D}} * \mathcal{U}(t)$$

where $*$ denotes the convolution operation; $\boldsymbol{\mathcal{A}} \in \mathbb{C}^{P \times P \times k_A \times k_A}$, $\boldsymbol{\mathcal{B}} \in \mathbb{C}^{P \times U \times k_B \times k_B}$, $\boldsymbol{\mathcal{C}} \in \mathbb{C}^{U \times P \times k_C \times k_C}$, and $\boldsymbol{\mathcal{D}} \in \mathbb{C}^{U \times U \times k_D \times k_D}$ are the state, input, output, and feedthrough convolutional kernels, respectively; and $\mathcal{X}(t) \in \mathbb{C}^{H \times W \times P}$, $\mathcal{U}(t) \in \mathbb{C}^{H \times W \times U}$, and $\mathcal{Y}(t) \in \mathbb{C}^{H \times W \times U}$ represent the state, input, and output tensors, respectively.



Background: The 1×1 Bottleneck in ConvS5

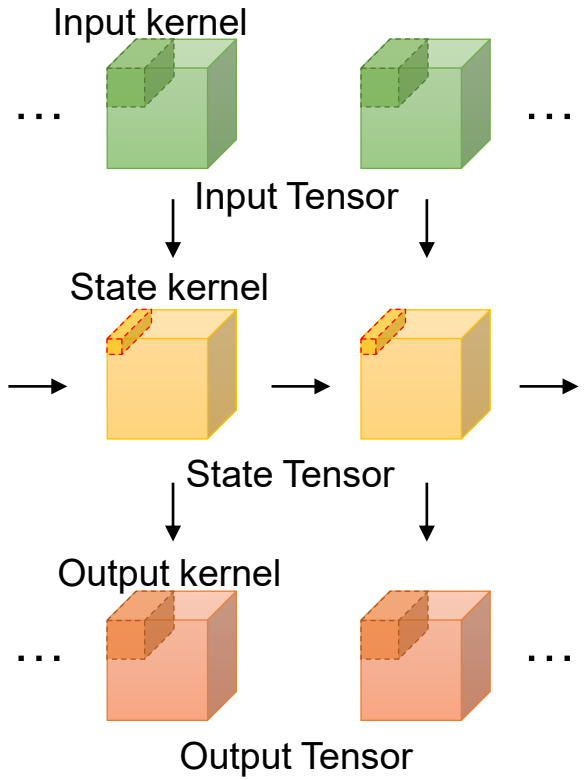
- ConvS5 (Efficient Implementation)

$$\mathcal{X}'(t) = \boldsymbol{\mathcal{A}} * \mathcal{X}(t) + \boldsymbol{\mathcal{B}} * \mathcal{U}(t)$$
$$\mathcal{Y}(t) = \boldsymbol{\mathcal{C}} * \mathcal{X}(t) + \boldsymbol{\mathcal{D}} * \mathcal{U}(t)$$

where $\boldsymbol{\mathcal{A}} \in \mathbb{C}^{P \times P \times 1 \times 1}$, $\boldsymbol{\mathcal{B}} \in \mathbb{C}^{P \times U \times k_B \times k_B}$, $\boldsymbol{\mathcal{C}} \in \mathbb{C}^{U \times P \times k_C \times k_C}$, and $\boldsymbol{\mathcal{D}} \in \mathbb{C}^{U \times U \times k_D \times k_D}$.

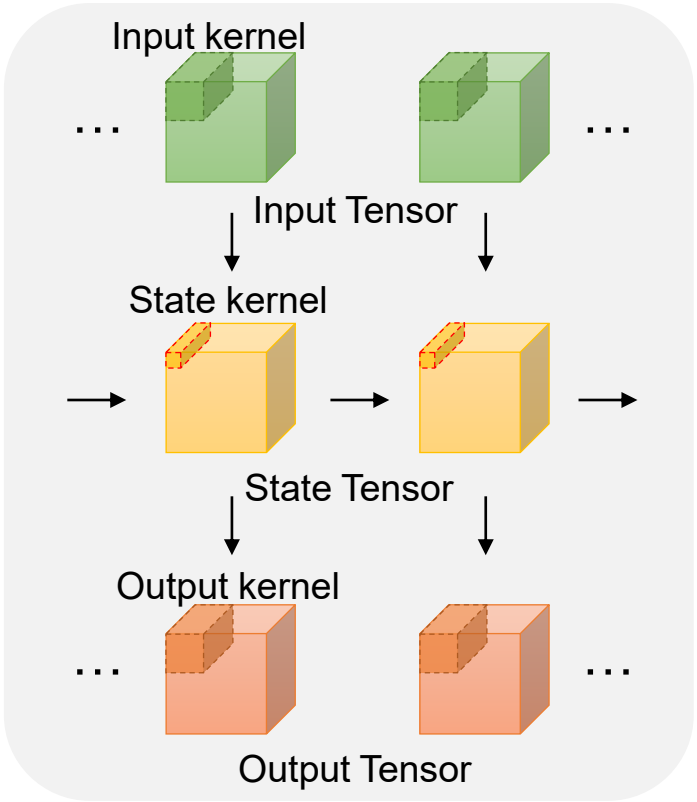
➔ Parallel scan

	Transformer	ConvRNNs	ConvS5
Inference	$\mathcal{O}(L)$	$\boldsymbol{\mathcal{O}(1)}$	$\boldsymbol{\mathcal{O}(1)}$
Training	$\mathcal{O}(L^2)$	$\boldsymbol{\mathcal{O}(L)}$	$\boldsymbol{\mathcal{O}(L)}$
Parallelizable	**Yes**	No	**Yes**

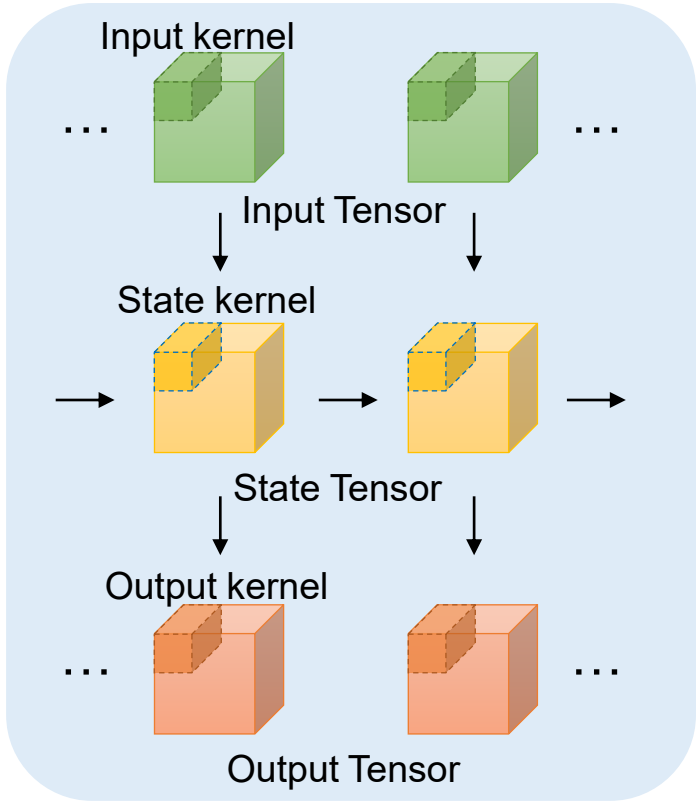


[1] Jimmy Smith, Shalini De Mello, Jan Kautz, Scott Linderman, and Wonmin Byeon. Convolutional state space models for long-range spatiotemporal modeling. NeurIPS 2023.

Can we use larger state kernels without breaking efficiency?



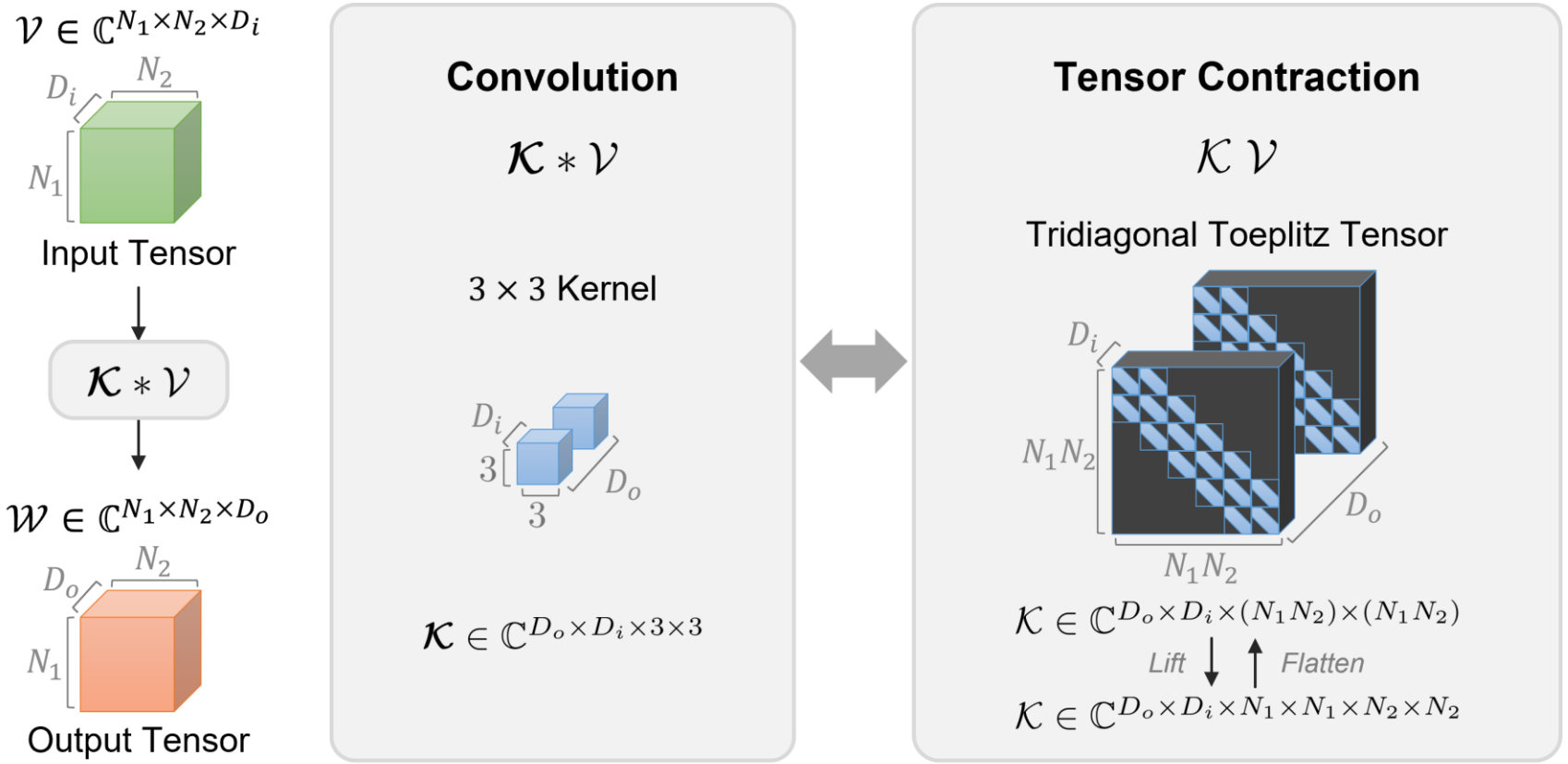
$$\mathcal{A} \in \mathbb{C}^{P \times P \times 1 \times 1}$$



$$\mathcal{A} \in \mathbb{C}^{P \times P \times 3 \times 3}$$

Key Idea: Diagonalizable Structuring

Convolution with **3×3 kernels** can be expressed as contraction with **tridiagonal Toeplitz (TT) tensors.**



Key Idea: Diagonalizable Structuring

A TT matrix has a **closed-form eigendecomposition**, determined by the lower, diagonal, and upper entities.

$$T = \text{tridiag}(l_T, d_T, u_T) = \begin{bmatrix} d_T & u_T & & & \\ l_T & d_T & u_T & & \\ & \ddots & \ddots & \ddots & \\ & & l_T & d_T & u_T \\ & & & l_T & d_T \end{bmatrix}$$

For *i*th eigenvalue/eigenvector:

$$\lambda_i = d_T + 2\sqrt{l_T u_T}\cos\left(\tfrac{i\pi}{N+1}\right),$$
$$x_{ij} = (l_T/u_T)^{j/2}\sin\left(\tfrac{ij\pi}{N+1}\right).$$

Key Idea: Diagonalizable Structuring

A ConvSSM with a **3×3 state kernel** has a **TT state tensor**.

Convolution Form

$$\mathcal{X}'(t) = \boldsymbol{\mathcal{A}} * \mathcal{X}(t) + \boldsymbol{\mathcal{B}} * \mathcal{U}(t)$$
$$\mathcal{Y}(t) = \boldsymbol{\mathcal{C}} * \mathcal{X}(t) + \boldsymbol{\mathcal{D}} * \mathcal{U}(t)$$

$$\boldsymbol{\mathcal{A}} \in \mathbb{C}^{P \times P \times 3 \times 3}$$

Tensor Contraction Form

$$\mathcal{X}'(t) = \mathcal{A}\, \mathcal{X}(t) + \mathcal{B}\, \mathcal{U}(t)$$
$$\mathcal{Y}(t) = \mathcal{C}\, \mathcal{X}(t) + \mathcal{D}\, \mathcal{U}(t)$$

Key Idea: Diagonalizable Structuring

We construct the TT state tensor with a **diagonalizable matrix R** and a **proportionality-constrained TT (PTT) tensor *S***.

Convolution Form

$$\mathcal{X}'(t) = \boldsymbol{\mathcal{A}} * \mathcal{X}(t) + \boldsymbol{\mathcal{B}} * \mathcal{U}(t)$$
$$\mathcal{Y}(t) = \boldsymbol{\mathcal{C}} * \mathcal{X}(t) + \boldsymbol{\mathcal{D}} * \mathcal{U}(t)$$

$$\boldsymbol{\mathcal{A}} \in \mathbb{C}^{P \times P \times 3 \times 3}$$

Tensor Contraction Form

$$\mathcal{X}'(t) = \mathcal{A}\,\mathcal{X}(t) + \mathcal{B}\,\mathcal{U}(t)$$
$$\mathcal{Y}(t) = \mathcal{C}\,\mathcal{X}(t) + \mathcal{D}\,\mathcal{U}(t)$$

$$\mathcal{A} := f(R, \mathcal{S}) \in \mathbb{C}^{P \times P \times H \times H \times W \times W}$$

Specifically, the PTT tensor $\mathcal{S}$ satisfies two proportionality conditions with some nonzero ratios $\alpha_H, \alpha_W \in \mathbb{C}$:

$$\text{(i)} \quad l_{\mathcal{S}_{q,r,:,:,i_w,j_w}} = \alpha_H u_{\mathcal{S}_{q,r,:,:,i_w,j_w}}, \qquad \text{(ii)} \quad l_{\mathcal{S}_{q,r,i_h,j_h,:,:}} = \alpha_W u_{\mathcal{S}_{q,r,i_h,j_h,:,:}}, \tag{11}$$

for all $q, r \in \{1, \ldots, P\}$, $i_h, j_h \in \{1, \ldots, H\}$, $i_w, j_w \in \{1, \ldots, W\}$ such that $|i_h - j_h| \leq 1$, $|i_w - j_w| \leq 1$.

Key Idea: Diagonalizable Structuring

Then, the original form can be transformed in to a **diagonalized form** using eigenvector matrices.

Convolution Form

$$\mathcal{X}'(t) = \boldsymbol{\mathcal{A}} * \mathcal{X}(t) + \boldsymbol{\mathcal{B}} * \mathcal{U}(t)$$
$$\mathcal{Y}(t) = \boldsymbol{\mathcal{C}} * \mathcal{X}(t) + \boldsymbol{\mathcal{D}} * \mathcal{U}(t)$$

$$\boldsymbol{\mathcal{A}} \in \mathbb{C}^{P \times P \times 3 \times 3}$$

Tensor Contraction Form

$$\mathcal{X}'(t) = \mathcal{A}\, \mathcal{X}(t) + \mathcal{B}\, \mathcal{U}(t)$$
$$\mathcal{Y}(t) = \mathcal{C}\, \mathcal{X}(t) + \mathcal{D}\, \mathcal{U}(t)$$

$$\mathcal{A} := f(R, \mathcal{S}) \in \mathbb{C}^{P \times P \times H \times H \times W \times W}$$

Diagonalized Form

$$\mathcal{X}'_T(t) = \mathcal{A}_T\, \mathcal{X}_T(t) + \mathcal{B}_T\, \mathcal{U}(t)$$
$$\mathcal{Y}(t) = \mathcal{C}_T\, \mathcal{X}_T(t) + \mathcal{D}\, \mathcal{U}(t)$$

✓ **Parallel scans**

Theorem 2. A ConvT3 can be diagonalized as

$$\mathcal{X}'_T(t) = \mathcal{A}_T \mathcal{X}_T(t) + \mathcal{B}_T \mathcal{U}(t),$$
$$\mathcal{Y}(t) = \mathcal{C}_T \mathcal{X}_T(t) + \mathcal{D}\mathcal{U}(t), \tag{17}$$

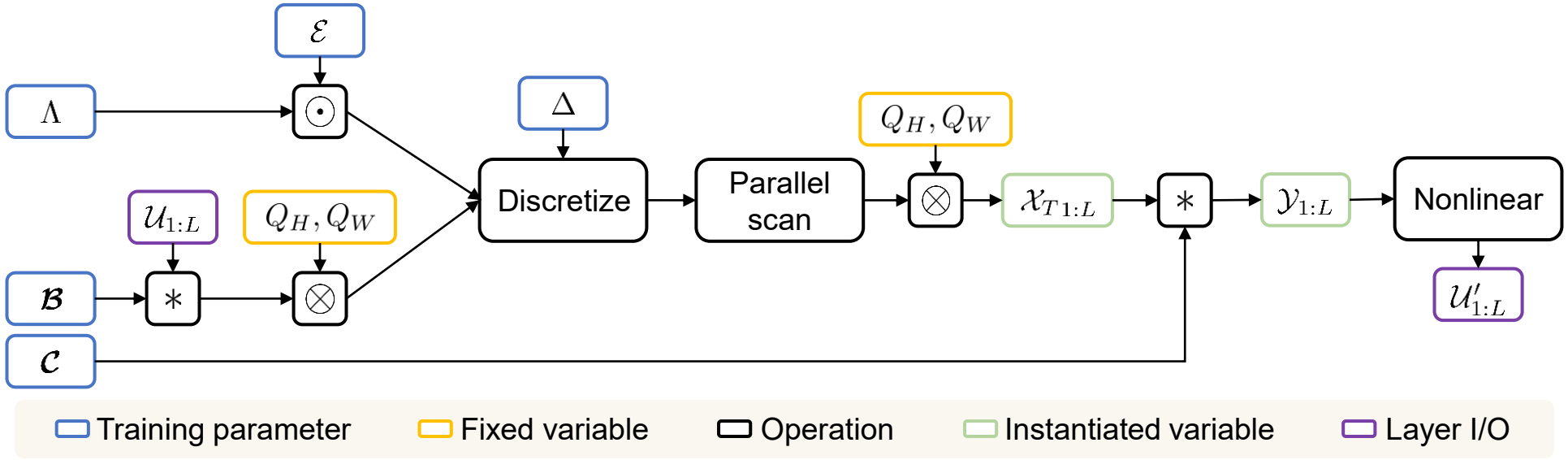
where

$$\mathcal{A}_T = (\Lambda \otimes I_H \otimes I_W) \odot \mathcal{E}, \qquad \mathcal{B}_T = \mathcal{Q}^{-1} B, \qquad \mathcal{C}_T = C\mathcal{Q}, \tag{18}$$

under the change of state $\mathcal{X}_T(t) = \mathcal{Q}^{-1}\mathcal{X}(t)$, with $\mathcal{Q} := Q_P \otimes Q_H \otimes Q_W$. The contraction dimensions of tensor contraction are stated by Einstein notation in Appendix A.1.

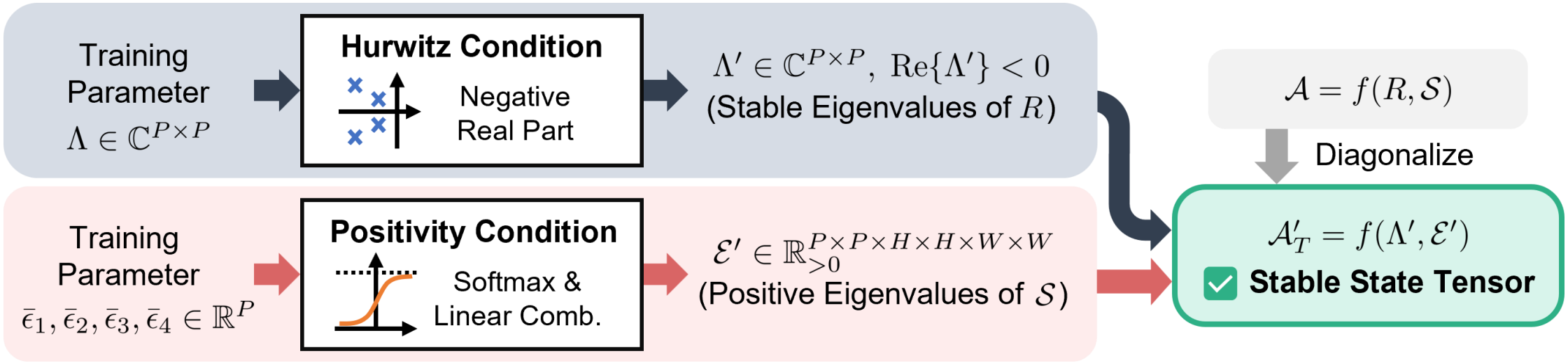
Algorithmic Flow of ConvT3

The overall computation including the transformation is **efficient** since the eigenvector matrices are **fixed**.



Training Stability of ConvT3

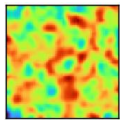
The training stability is ensured by conditioning the state tensor.



Results (Video Prediction)

Method	100 → 800 FVD ↓	PSNR ↑	SSIM ↑	LPIPS ↓	100 → 1200 FVD ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Trained on 300 frames								
Transformer (Vaswani et al., 2017)	159	12.6	0.609	0.287	265	12.4	0.591	0.321
Performer (Choromanski et al., 2021)	234	13.4	0.652	0.379	275	13.2	0.592	0.393
CW-VAE (Saxena et al., 2021)	104	12.4	0.592	0.277	**117**	12.3	0.585	0.286
ConvLSTM (Shi et al., 2015)	128	15.0	0.737	0.169	187	14.1	0.706	0.203
ConvS5 (Smith et al., 2023)	**72**	16.0	0.761	0.156	187	14.5	0.678	0.230
ConvT3	79	**16.1**	**0.776**	**0.146**	118	**15.2**	**0.746**	**0.179**
Trained on 600 frames								
Transformer	42	13.7	0.672	0.207	91	13.1	0.631	0.252
Performer	93	12.4	0.616	0.274	243	12.2	0.608	0.312
CW-VAE	94	12.5	0.598	0.269	107	12.3	0.590	0.280
ConvLSTM	91	15.5	0.757	0.149	137	14.6	0.727	0.180
ConvS5	47	16.4	0.788	0.134	71	15.6	0.763	0.162
ConvT3	**36**	**17.7**	**0.823**	**0.104**	**56**	**16.7**	**0.795**	**0.131**

Results (PDE Simulation)



Model	#Params	NRMSE ↓		Time (s)	
		Shallow-Water	Diffusion-Reaction	Train Step	Evaluation Step
AViT-B	116M	0.00047	0.0110	-	-
FNO-B	115M	0.00246	0.0599	-	-
UNet	7M	0.083–	0.84–	-	-
FNO	927K	0.0044	0.12–	-	-
AViT-Ti	7M	0.00053	<u>0.0090</u>	303 (2.31×)	2.74 (2.06×)
ConvS5	6M	<u>0.00035</u>	0.0106	131 (1.00×)	1.33 (1.00×)
ConvT3	6M	**0.00033**	**0.0087**	151 (1.15×)	1.51 (1.14×)

Conclusion

1. We extends the 1×1 state kernel limit by enabling 3×3 state kernels.
2. We preserve both efficiency and stability of ConvSSMs.
3. ConvT3 achieves superior performance in video and PDE tasks.