SpoT-Mamba:
Learning Long-Range Dependency on Spatio-Temporal Graphs with Selective State Spaces

Jinhyeok Choi, Heehyeon Kim, Minhyeong An and Joyce Jiyoung Whang*
School of Computing, KAIST
* Corresponding Author

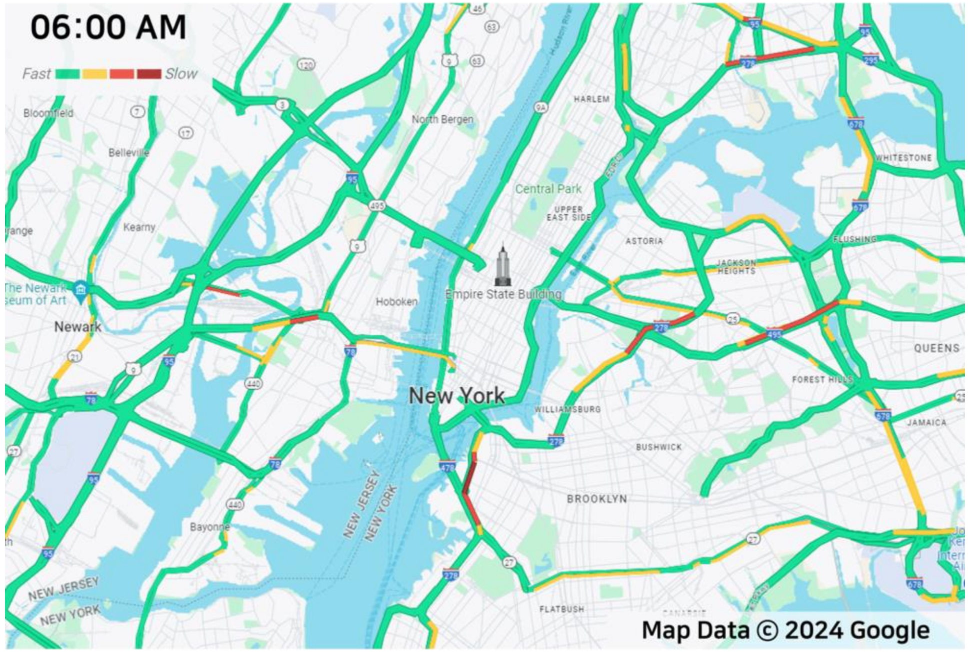
Spatio-Temporal Reasoning and Learning (STRL) Workshop at
the 33rd International Joint Conference on Artificial Intelligence (IJCAI 2024)

KAIST BDILab
BIG DATA INTELLIGENCE

01 STG Forecasting

What is a spatio-temporal graph (STG)?

- A Spatio-Temporal Graph (STG) is a type of data structure designed to represent and analyze data which varies across both spatial and temporal dimensions.
  - Urban traffic networks, weather data, skeleton-based human actions, etc.

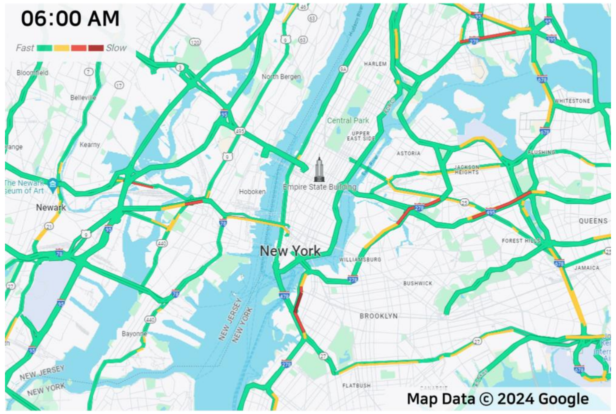


Traffic flow of road networks

01 STG Forecasting

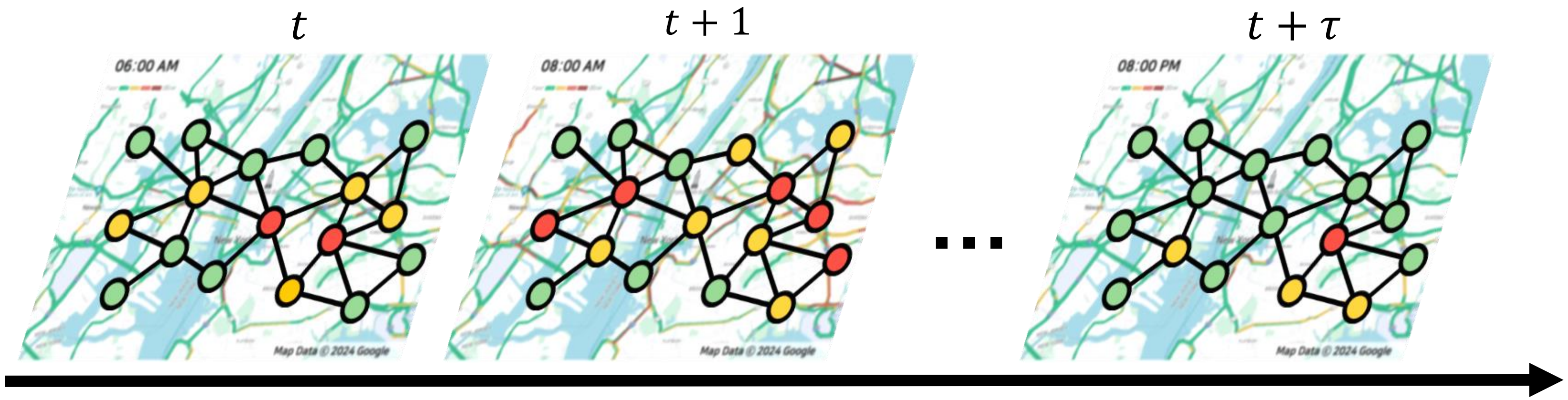
What is a spatio-temporal graph (STG)?

- **Nodes** represent **entities with spatial properties**, such as locations in a geographic area.
    - **Node features** evolve **over time**.
- **Edges** represent **relationships or interactions** between **entities**.



Traffic flow of road networks

Spatio-Temporal Graph (STG)



Time steps

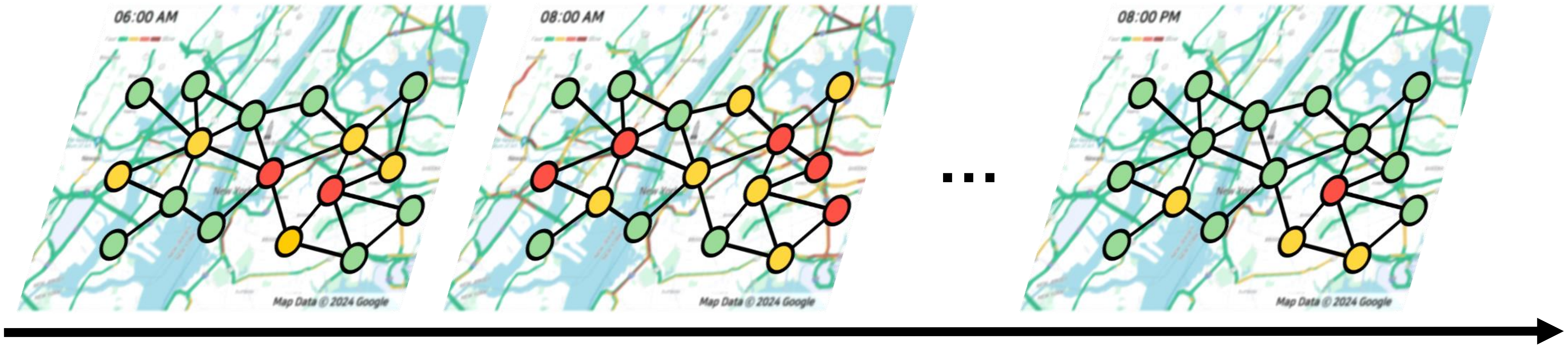
01 STG Forecasting

Definition of STG: $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{X})$

- $\mathcal{V}$ is a set of N nodes and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is a set of edges.
- $\mathcal{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_\tau]$ is **a sequence of observed data** for all nodes at each time step.
    - $\mathbf{X}_t \in \mathbb{R}^{N \times D_{\text{in}}}$ represents **the node features** at time step t.

t $\quad$ $t+1$ $\quad$ $t+\tau$

Traffic flow of road networks



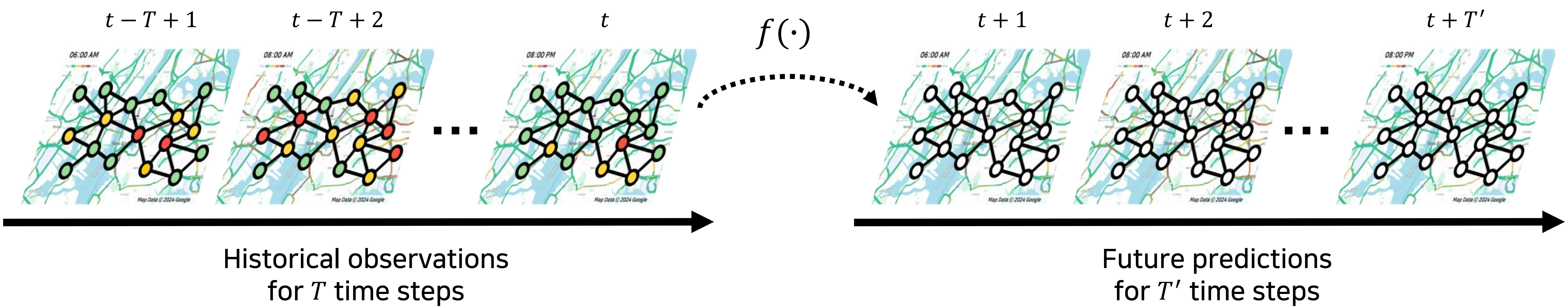
Time steps

$\mathbf{X}_t$ $\quad$ $\mathbf{X}_{t+1}$ $\quad$ $\mathbf{X}_{t+\tau}$

01 STG Forecasting

Problem definition

- STG forecasting aims to predict future observations for T' time steps, given historical observations for the previous T time steps with $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{X})$.
  - This is formulated as $[\mathbf{X}_{t-T+1}, \dots, \mathbf{X}_t] \xrightarrow{f(\cdot)} [\mathbf{X}_{t+1}, \dots, \mathbf{X}_{t+T'}]$.
  - $f(\cdot)$ represents the STG forecasting model.



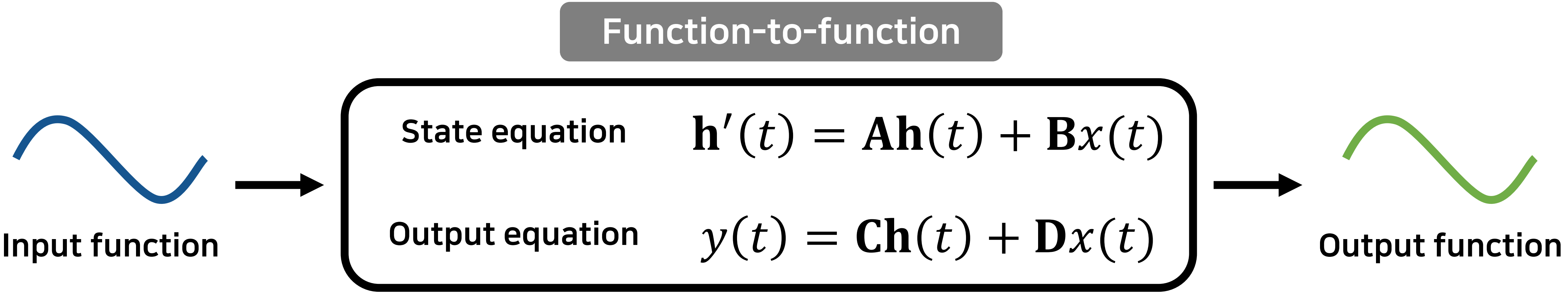
A New Spatio-Temporal Graph Forecasting Framework with Mamba-based Sequence Modeling

SpoT-Mamba

03 Preliminaries

State Space Model (SSM)*

- SSM defines the evolution of a dynamic system's state with two equations.
- Given the input signal $x(t) \in \mathbb{R}$, SSM transforms $x(t)$ into the latent state $\mathbf{h}(t) \in \mathbb{R}^D$.
  - $\mathbf{A} \in \mathbb{R}^{D \times D}$, $\mathbf{B} \in \mathbb{R}^{D \times 1}$, $\mathbf{C} \in \mathbb{R}^{1 \times D}$, and $\mathbf{D} \in \mathbb{R}$ are learnable parameters.
  - $y(t) \in \mathbb{R}$ is the output signal.



* A. Gu et al., "Efficiently modeling long sequences with structured state spaces", ICLR, 2022

Preliminaries

Discretized SSM*

- The discretized SSM is defined in two forms: a recurrent form and a convolutional form.
  - $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$ are approximated learnable parameters with a step size $\mathbf{\Delta}$.
  - L denotes the sequence length and $*$ indicates the convolution operation.

Function-to-function

Continuous-time SSM

$$\mathbf{h}'(t) = \mathbf{A}\mathbf{h}(t) + \mathbf{B}x(t)$$
$$y(t) = \mathbf{C}\mathbf{h}(t) + \mathbf{D}x(t)$$

Sequence-to-sequence

Recurrent form

$$\mathbf{h}_t = \bar{\mathbf{A}}\mathbf{h}_{t-1} + \bar{\mathbf{B}}x_t$$
$$y_t = \bar{\mathbf{A}}\mathbf{h}_t$$

Efficient inference for SSM

or

Convolutional form

$$\mathbf{y} = \bar{\mathbf{K}} * \mathbf{x}$$
$$\bar{\mathbf{K}} \in \mathbb{R}^L = (\overline{\mathbf{CB}}, \overline{\mathbf{CAB}}, \ldots, \overline{\mathbf{CA}^{L-1}\mathbf{B}})$$

Parallelizable training of SSM

* A. Gu et al., "Efficiently modeling long sequences with structured state spaces", ICLR, 2022

Mamba**

- Mamba removes the linear time-invariant (LTI) constraint of SSMs with selection mechanisms.
  - Selection mechanisms allow learnable parameters of SSMs to interact with the input sequence.
  - $\mathbf{B}$, $\mathbf{C}$, and the step size $\boldsymbol{\Delta}$ become functions of the input sequence.

Recurrent form

$$\mathbf{h}_t = \bar{\mathbf{A}}\mathbf{h}_{t-1} + \bar{\mathbf{B}}x_t,\ y_t = \bar{\mathbf{A}}\mathbf{h}_t$$

Efficient inference for SSM

Convolutional form

$$\mathbf{y} = \bar{\mathbf{K}} * \mathbf{x}$$

$$\bar{\mathbf{K}} \in \mathbb{R}^L = (\bar{\mathbf{C}}\bar{\mathbf{B}}, \overline{\mathbf{CAB}}, \dots, \bar{\mathbf{C}}\bar{\mathbf{A}}^{L-1}\bar{\mathbf{B}})$$

Parallelizable training of SSM

→

Mamba

$$\mathbf{B} \coloneqq S_B(\mathrm{x}) = \mathrm{Linear}_N(\mathrm{x})$$

$$\mathbf{C} \coloneqq S_C(\mathrm{x}) = \mathrm{Linear}_N(\mathrm{x})$$

$$\boldsymbol{\Delta} \coloneqq \mathrm{Broadcast}_D(\mathrm{Linear}_N(\mathrm{x}))$$

$$\bar{\mathbf{A}}, \bar{\mathbf{B}} = \mathrm{Discretize}(\boldsymbol{\Delta}, \mathbf{A}, \mathbf{B})$$

$$y = \mathrm{SSM}(\bar{\mathbf{A}}, \bar{\mathbf{B}}, \mathbf{C})(\mathrm{x})$$

Selectivity for SSM
Hardware-aware parallel algorithm

** A. Gu et al., "Mamba: Linear-time sequence modeling with selective state spaces", arXiv, 2023

SpoT-Mamba: A New STG Forecasting Framework

Multi-way walk sequence

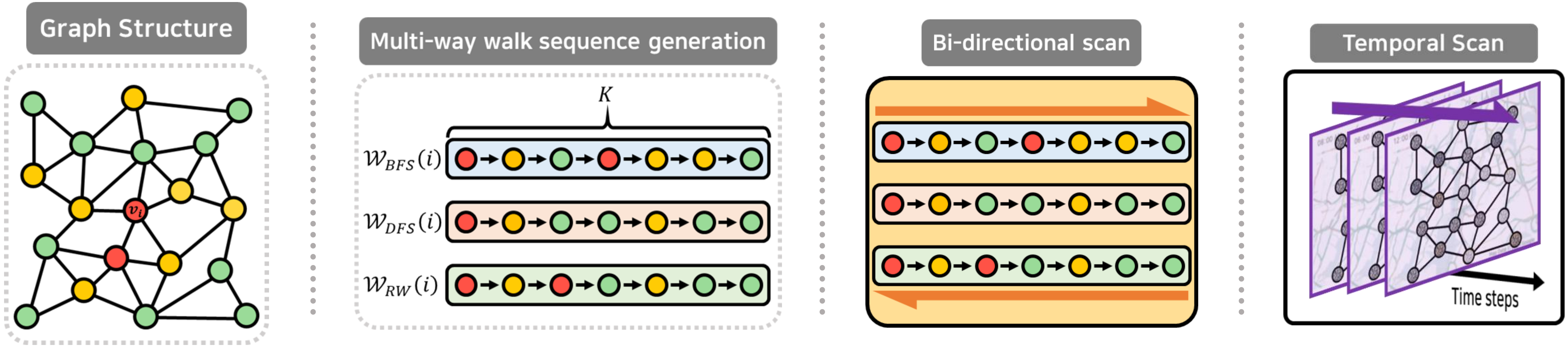
- Spatial sequences of nodes with three walk algorithms.

Walk sequence embedding

- Node embedding from node-specific walk sequences with Mamba blocks.

Temporal scan with Mamba blocks

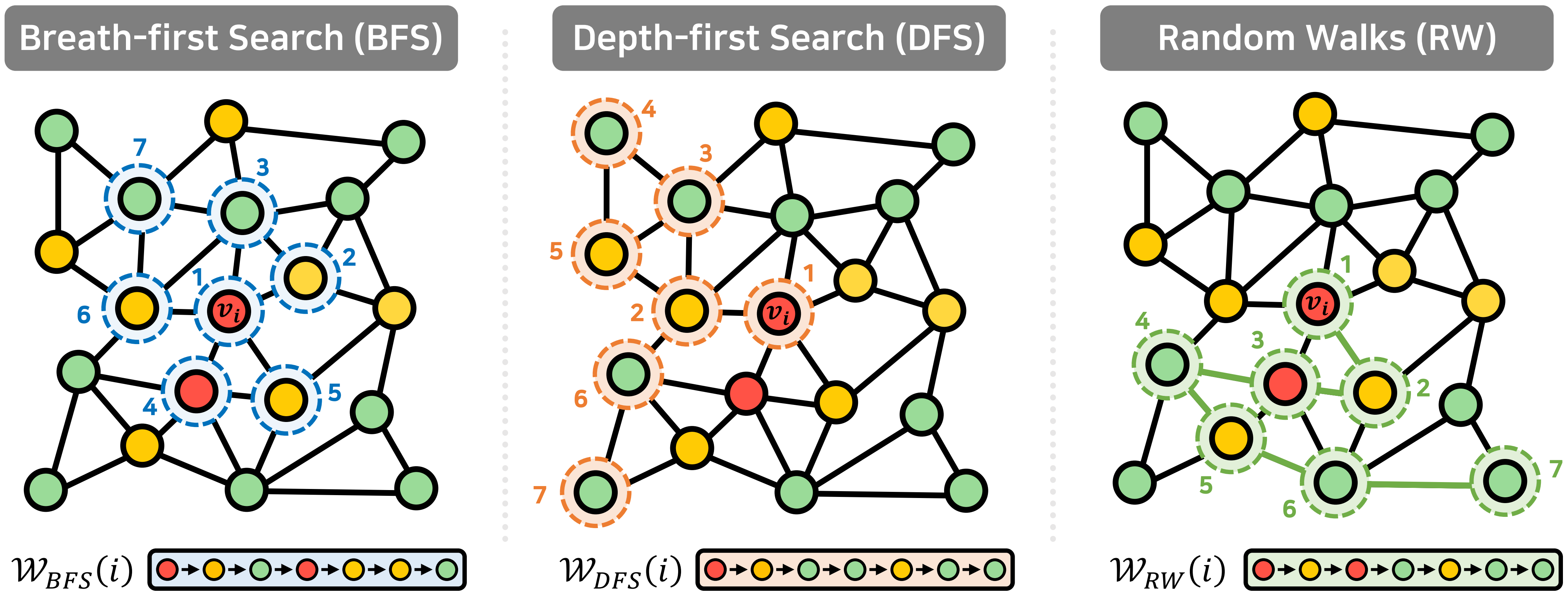
- Capturing temporal dynamics with selective mechanisms.



04 Multi-Way Walk Sequence

Spatial sequences of nodes with three well-known walk algorithms.

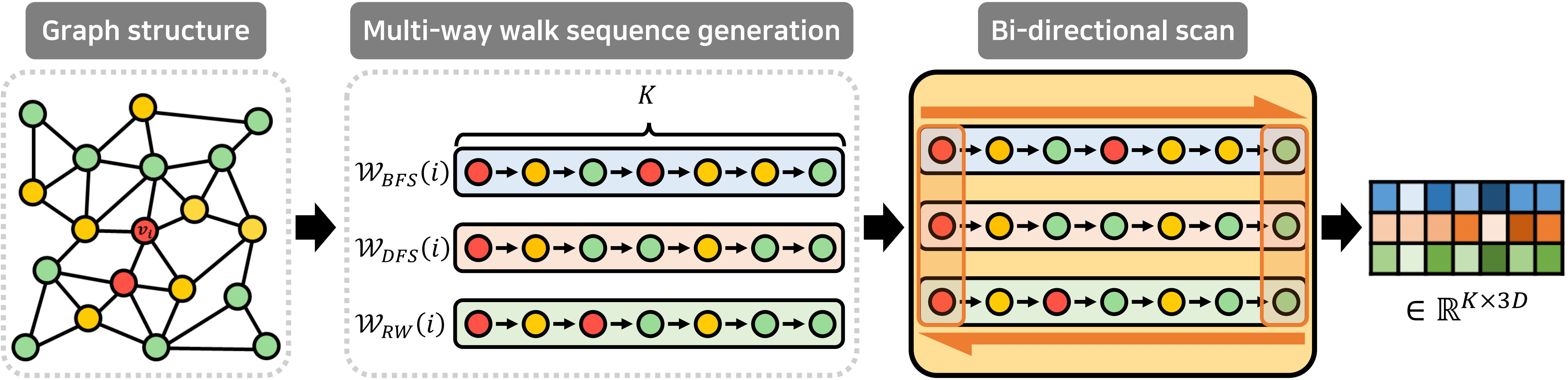
- SpoT-Mamba extracts diverse local and global structural information.



Walk Sequence Embedding

Bi-directional scan with Mamba blocks

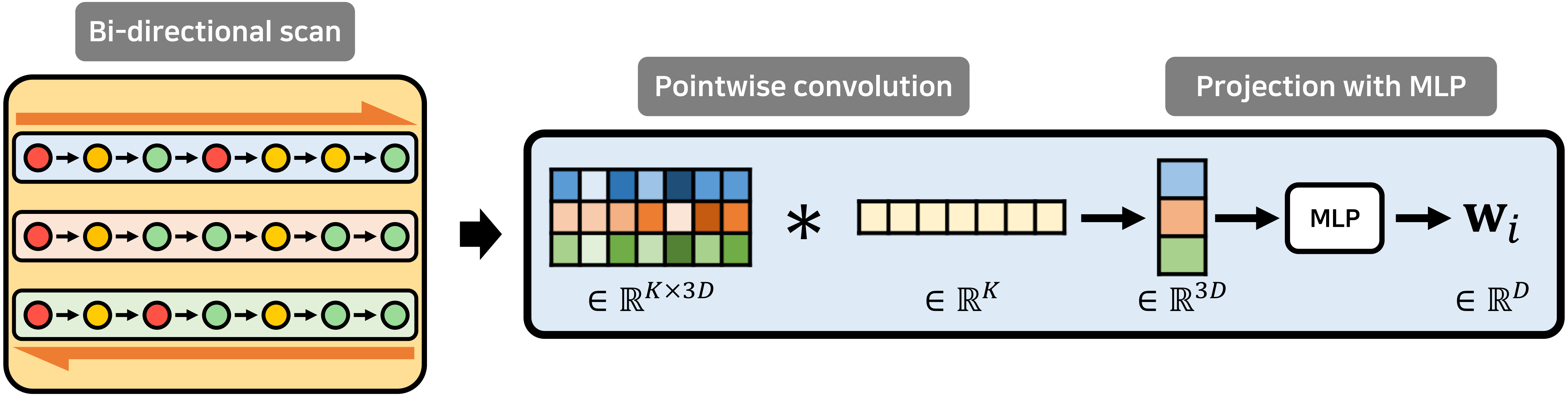
- Embeddings for node-specific walk sequences by scanning each sequence.
  - Capturing both short and long-range structural information from each node's neighborhood.
  - K indicates the length of the walk sequence, and D denotes the embedding dimension.



04 Walk Sequence Embedding

Node-specific walk embeddings to node embeddings

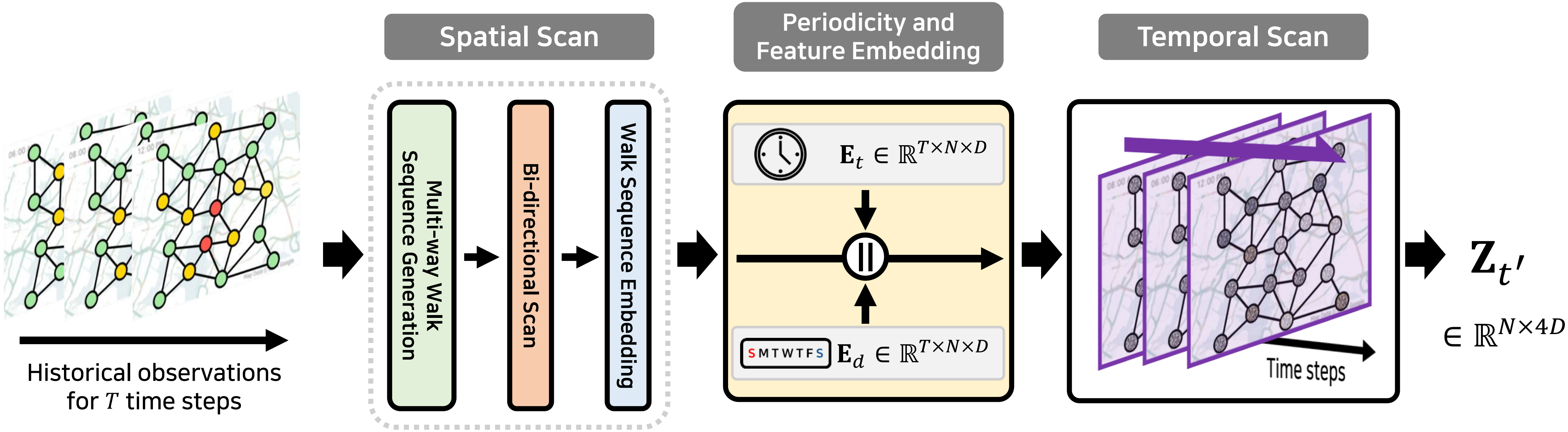
- **Pointwise convolution** allows for **incorporating representations of neighboring nodes** in sequences.
- **SpoT-Mamba** integrates **representations of sequences for a target node** with MLP.



STG Forecasting of SpoT-Mamba

Temporal scan with Mamba blocks

- **Learnable embeddings** are adopted to capture **the repetitive patterns over time**.
- **SpoT-Mamba** performs **selective scans** across **the sequences of node embeddings** with time axis.

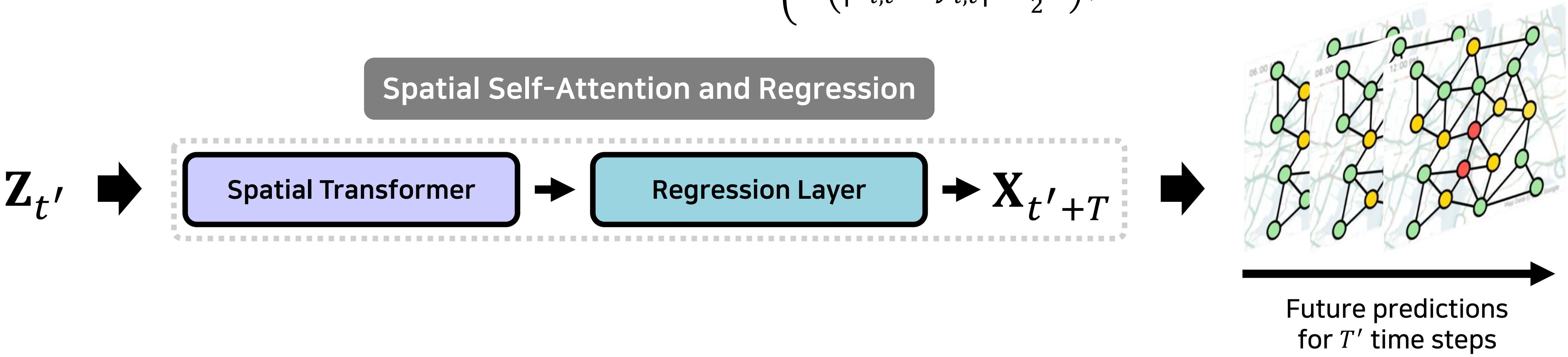


STG Forecasting of SpoT-Mamba

Spatial self-attention with Transformers and regression with MLP

- Incorporating **global information** from the entire graph at each time step through **Transformers**.
- **MLP** is applied to **forecast the attributes of each node** for future time steps.
- To ensure **robustness** to **outliers**, SpoT-Mamba is trained with **Huber Loss**.

$$\mathcal{L}_{\text{Huber}} = \sum_{i,t} l_{i,t} \quad \text{where} \quad l_{i,t} = \begin{cases} \frac{1}{2}\left(x_{i,t} - y_{i,t}\right)^2, & \text{if } \left|x_{i,t} - y_{i,t}\right| < \delta \\ \delta\left(\left|x_{i,t} - y_{i,t}\right| - \frac{1}{2}\delta\right), & \text{otherwise} \end{cases}$$



05 Experiments

- **Dataset**
  - *PEMS04*: A real-world **traffic flow forecasting** benchmark.

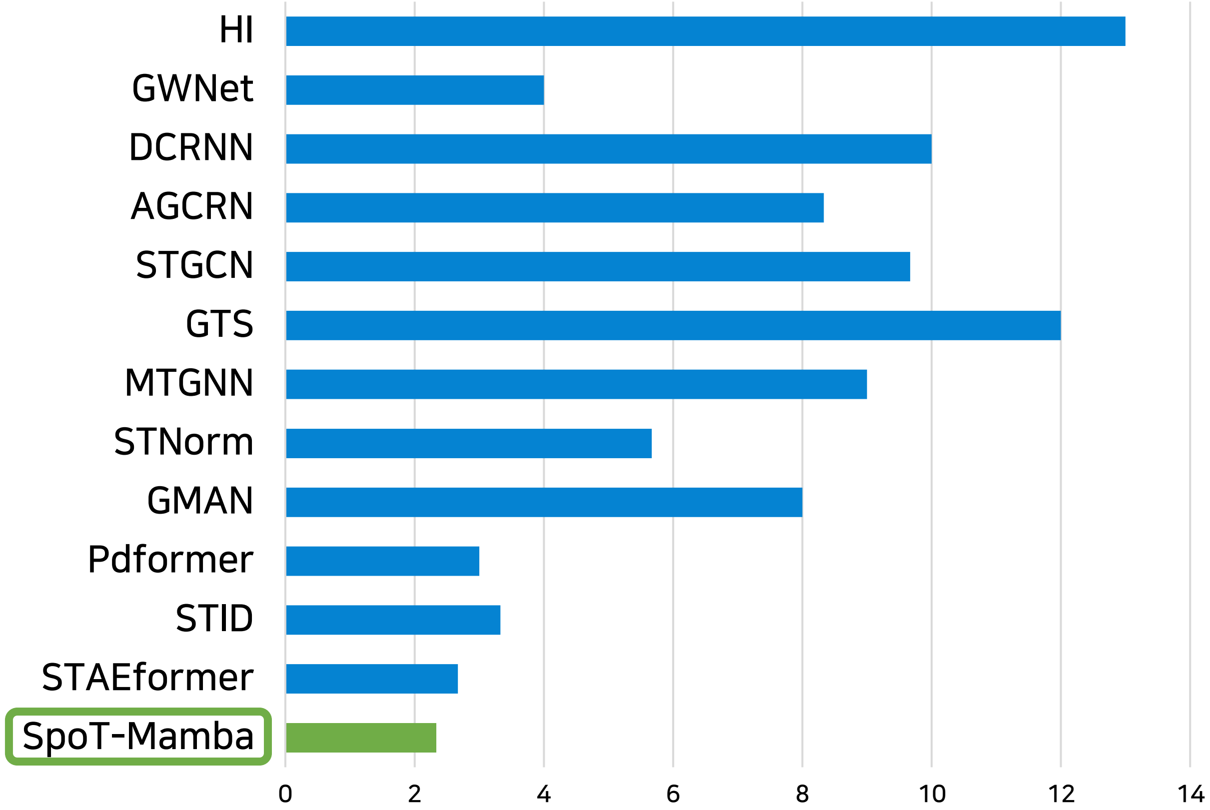
$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	**#Time Steps**	**Time Interval**	**Time Range**
307	338	16,992	5 min.	01/2018 – 02/2018

- **Baselines**
  - **GNN-based**: DCRNN (ICLR 2018), GWNet (IJCAI 2019), MTGNN (KDD 2020)
    AGCRN (NeurIPS 2020), GTS (ICLR 2021)
  - **Transformer-based**: GMAN (AAAI 2020), STAEformer (CIKM 2023), PDformer (AAAI 2023)
  - **Others**: STNorm (KDD 2021), HI (CIKM 2021), STID (CIKM 2022)

05 Experiments

Traffic forecasting performance on *PEMS04*

PEMS04	MAE(↓)	RMSE(↓)	MAPE(↓)
HI	42.35	61.66	29.92
GWNet	18.53	29.92	12.89
DCRNN	19.63	31.26	13.59
AGCRN	19.38	31.25	13.40
STGCN	19.57	31.38	13.44
GTS	20.96	32.95	14.66
MTGNN	19.17	31.70	13.37
STNorm	18.96	30.98	12.69
GMAN	19.14	31.60	13.19
PDformer	18.36	30.03	12.00
STID	18.38	29.95	12.04
STAEformer	18.22	30.18	11.98
SpoT-Mamba	18.31	30.11	11.86



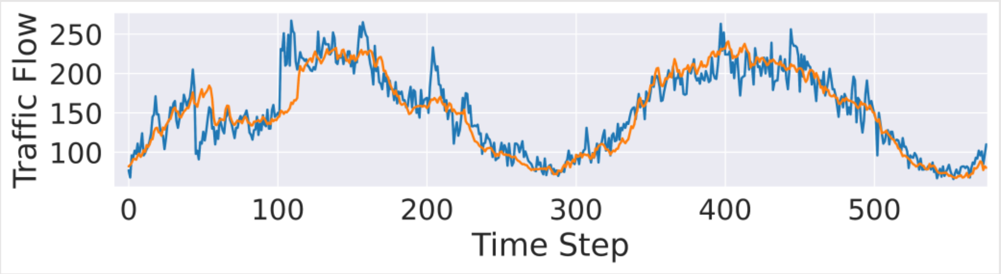
Average rank over the three metrics (↓)

05 Experiments

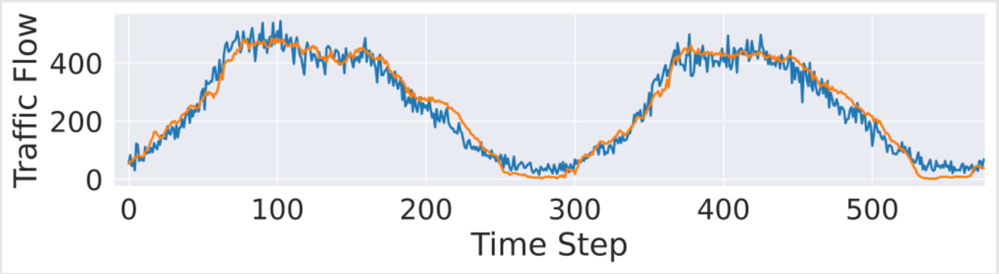
Visualization of SpoT-Mamba's predictions

- The blue line represents the ground truth, and the orange line denotes predicted traffic data.
  - Four nodes are randomly selected in *PEMS04*.

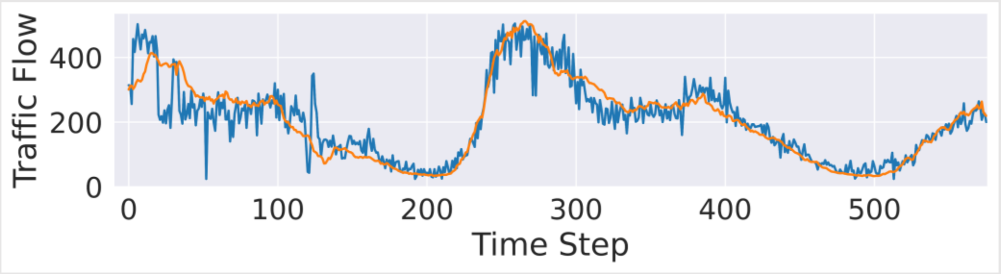
Case 1



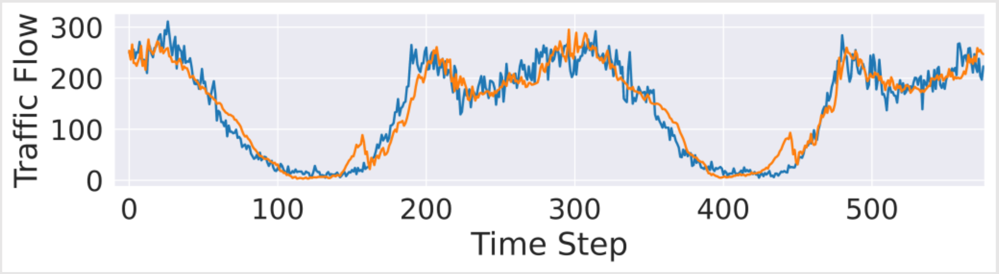
Case 2



Case 3



Case 4



Ablation study of SpoT-Mamba

- Mamba blocks of SpoT-Mamba are replaced with transformer encoders.
  - Transformer encoders do not recognize input sequence order by themselves.
    - SpoT-Mamba does not apply learnable embeddings for walk sequences.
  - As a result, transformer encoders struggle to perceive the order in walk sequences.

Walk Scan	Temporal Scan	MAE(↓)	RMSE(↓)	MAPE(↓)
Transformer	Transformer	18.41	30.32	12.12
Transformer	Mamba	18.69	30.17	12.28
Mamba	Transformer	18.29	30.06	11.93
Mamba	Mamba	18.31	30.11	11.86

06 Conclusion

- Propose a new **Spatio-Temporal** graph forecasting framework with a **Mamba**-based sequence modeling architecture, **SpoT-Mamba**.

- Effectively capturing **the long-range spatio-temporal dependencies** in STGs.
  - Extracting diverse **local and global structures** by utilizing **BFS**, **DFS**, and **random walks**.

- Promising results on the real-world traffic forecasting benchmark *PEMS04*.

Thank You!

▲ GitHub

▲ BDILab

Our datasets and codes are available at:

https://github.com/bdi-lab/SpoT-Mamba

You can find us at:

{cjh0507, heehyeon, amh0360, jjwhang}@kaist.ac.kr

https://bdi-lab.kaist.ac.kr

KAIST BDILab

BIG DATA INTELLIGENCE