ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

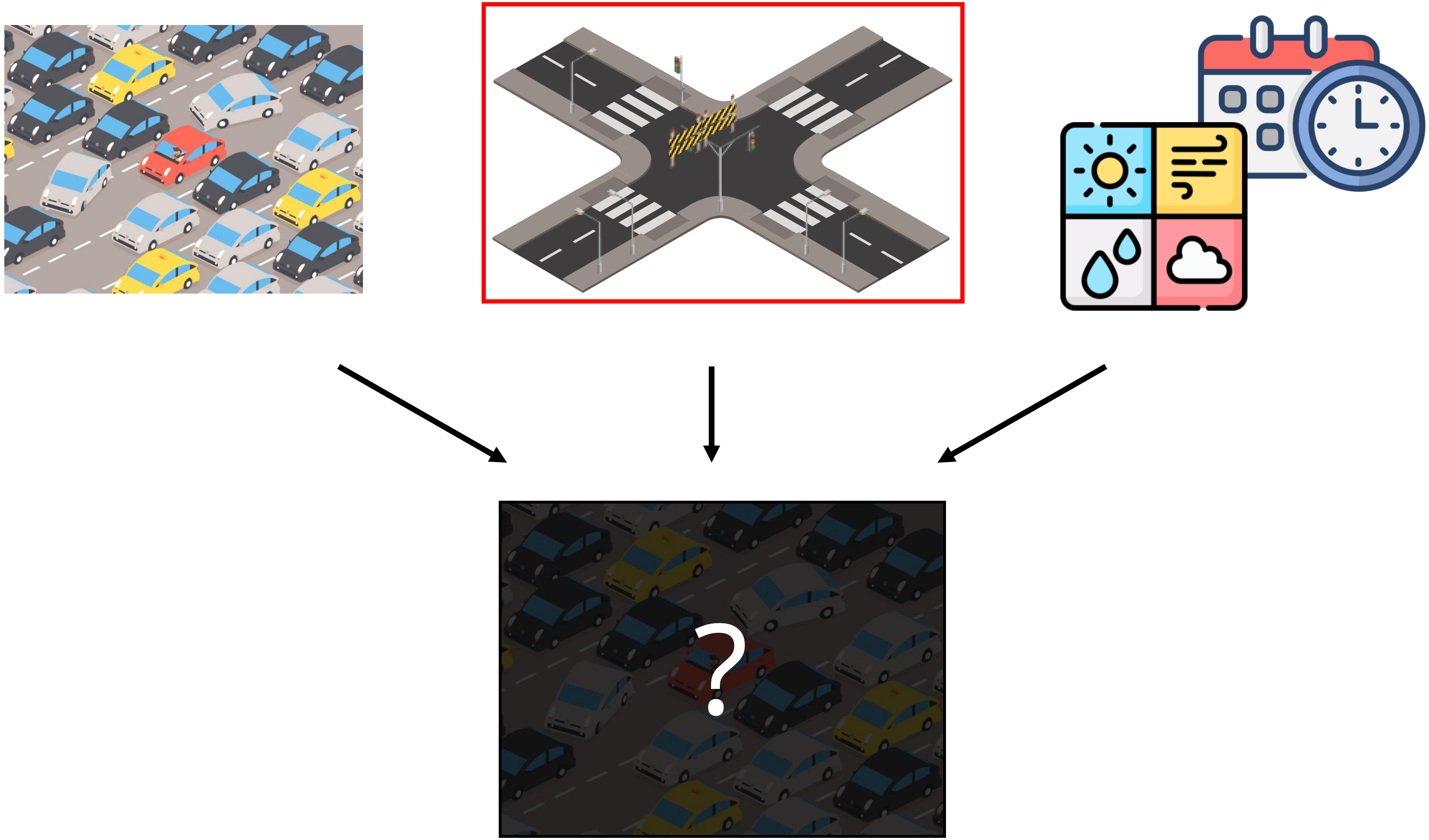
TESTAM: Time-Enhanced Spatio-Temporal Attention Network with Mixture of Experts

Hyunwook Lee and Sungahn Ko

Ulsan National Institute of Science and Technology

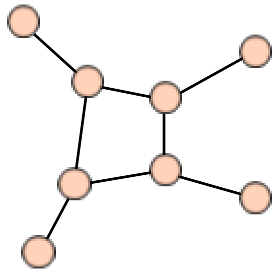
ICLR

Background: Traffic Forecasting



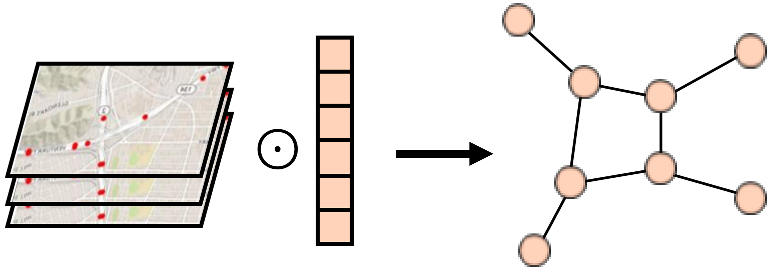
Main research themes on spatial modeling methods

Static Graph Modeling



$\mathcal{G}$ =softmax(relu($\mathbf{E}^{\top}\mathbf{E}$))

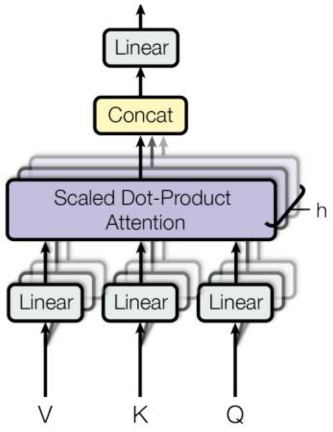
Dynamic Graph Modeling

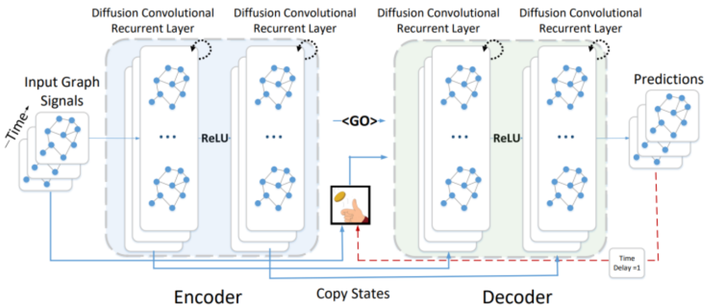


$\boldsymbol{H} = \boldsymbol{X} \odot \boldsymbol{W}$

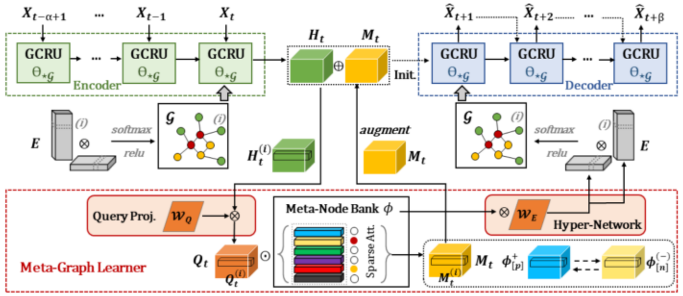
$\mathcal{G}$ =softmax(relu($\boldsymbol{H}^{\top}\boldsymbol{H}$)),

Attention Modeling*

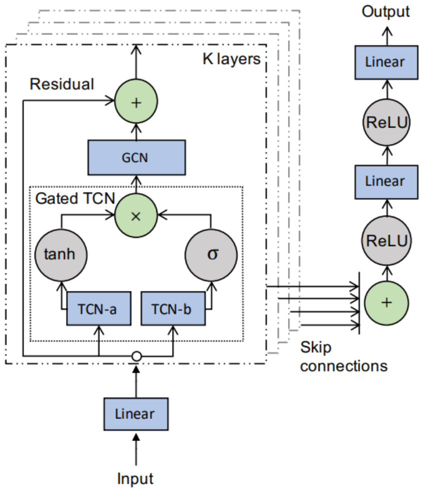



Diffusion Convolutional Recurrent Layer
Diffusion Convolutional Recurrent Layer
Diffusion Convolutional Recurrent Layer
Diffusion Convolutional Recurrent Layer
Input Graph Signals
Time
ReLU
<GO>
ReLU
Predictions
Time Delay =1
Encoder
Copy States
Decoder

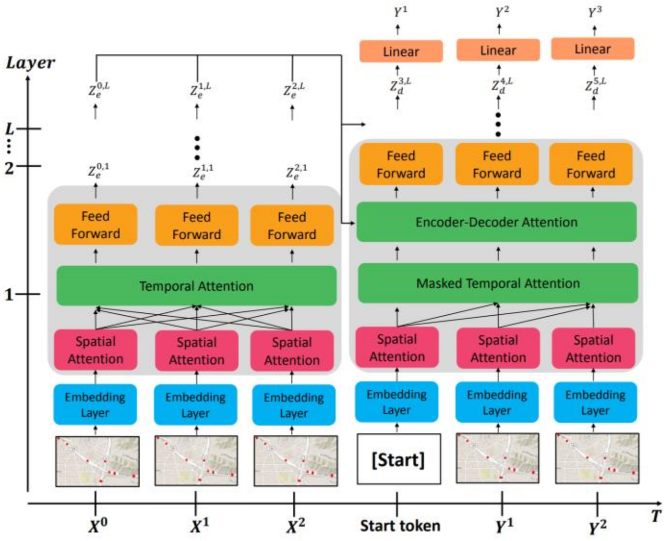
DCRNN (Li et al., 2018)



MegaCRN (Jiang et al., 2023)



Graph-WaveNet (Wu et al., 2019)

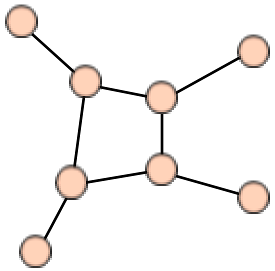


ST-GRAT (Park et al., 2020)

* Figure reproduced from Vaswani et al. 2017

Most of the research focused on finding *all-in-one* best solution – Q: does it really exist in the traffic forecasting?

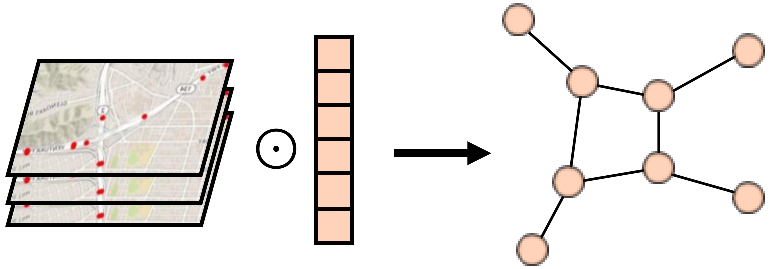
Static Graph Modeling



$\mathcal{G}$ =softmax(relu($\mathbf{E}^\top \mathbf{E}$))

- Rely on static graph
- Robustness on outliers or noises
- Vulnerable to domain shifts
- Hard to achieve in-situ dependency modeling

Dynamic Graph Modeling

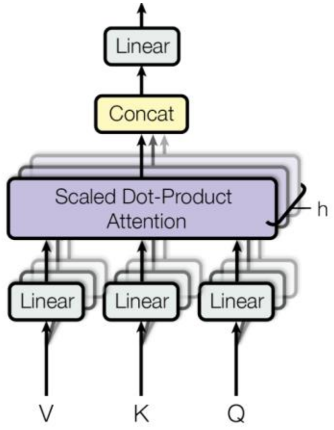


$\boldsymbol{H} = \boldsymbol{X} \odot \boldsymbol{W}$

$\mathcal{G}$ =softmax(relu($\boldsymbol{H}^\top \boldsymbol{H}$)),

- Rely on input-conditioned dynamic graph
- Relatively more robust for domain shift
- Vulnerable to input noises or outliers

Attention Modeling*



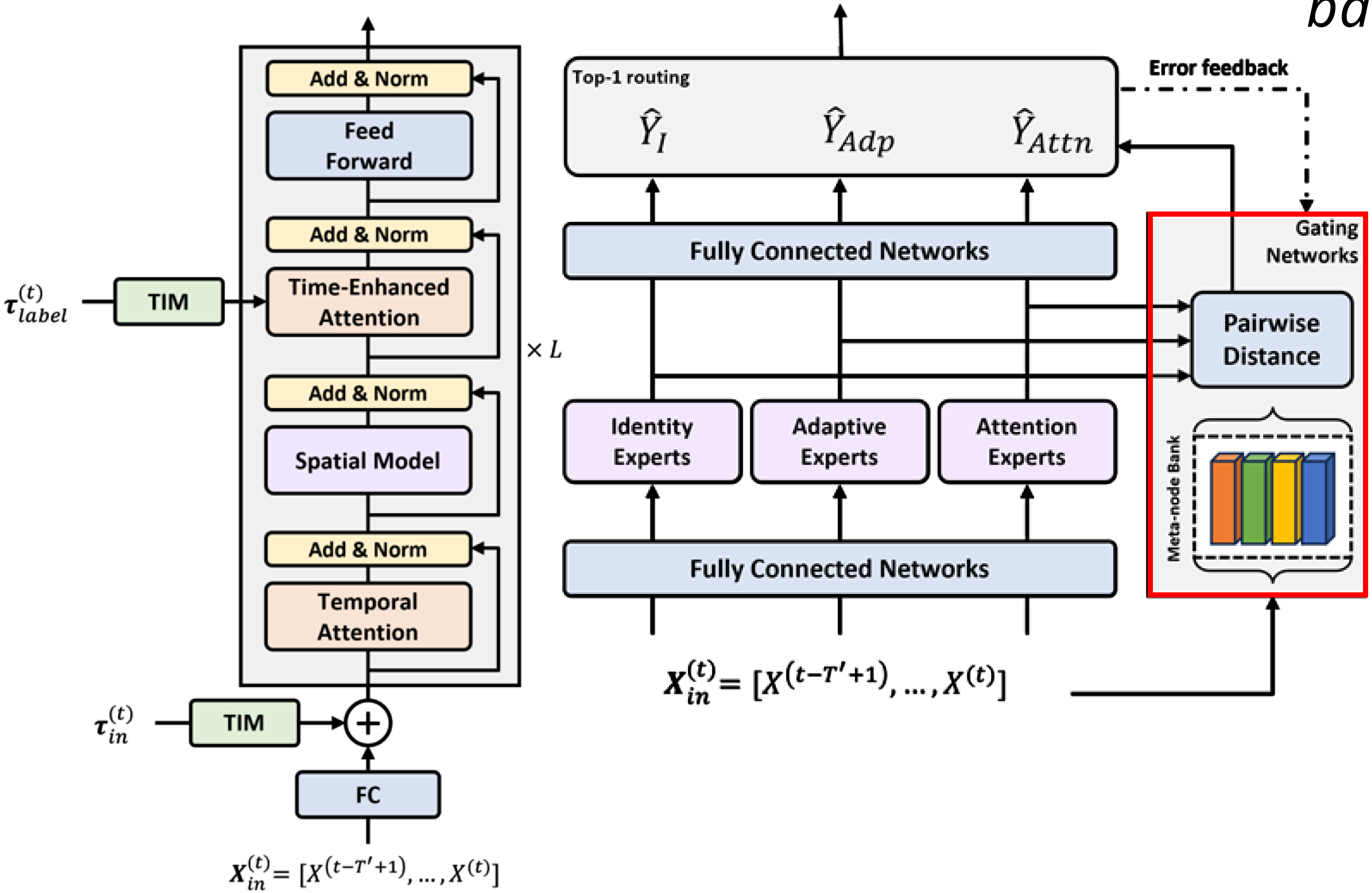
- No restriction on spatial dependency
- Take whole roads in account
- May need auxiliary connectivity information
- Often generate non-informative, blurry attention (Jin et al., 2023)

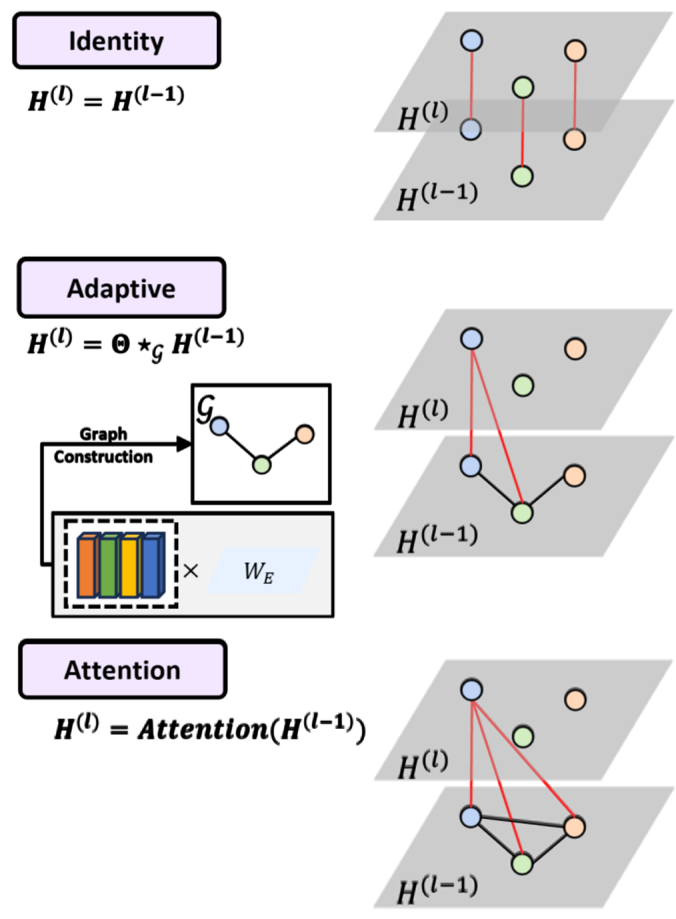
There are no *all-in-one* solution – Each method has its own pros and cons!

* Figure reproduced from Vaswani et al. 2017

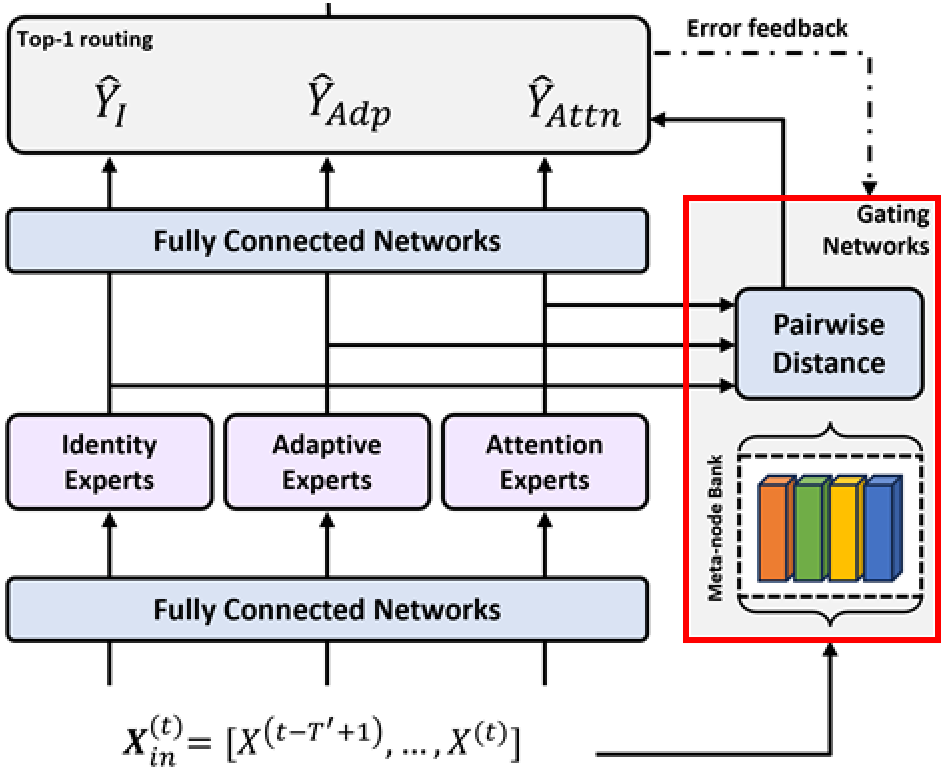
There are no *all-in-one* solution – Each method has its own pros and cons!

… How about making the model specialized to "choose" the best one based on current situation?





… How about making the model specialized to "choose" the best one based on current situation?



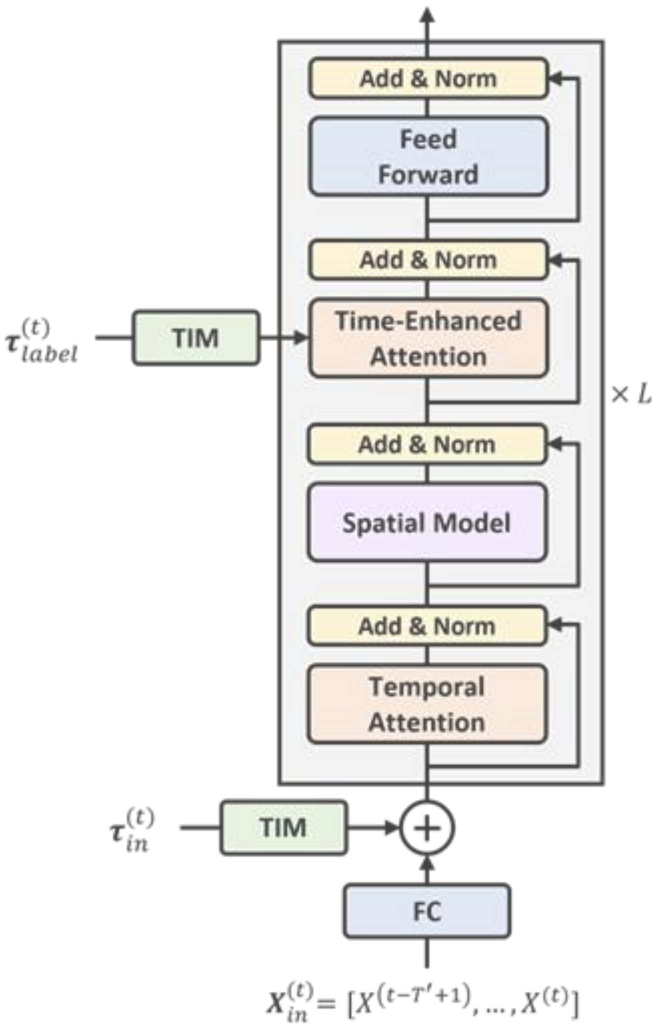
Input-queried memory unit representation

$$Q_i^{(t)} = X_i^{(t)} W_q + b_q$$

$$\begin{cases} a_j = \dfrac{\exp(Q_i^{(t)} M[j]^\top)}{\sum_{j=1}^m \exp(Q_i^{(t)} M[j]^\top)} \\ O_i^{(t)} = \sum_{j=1}^m a_j M[j] \end{cases}$$

Routing probability

$$r_e = g(z_e, O_i^{(t)}); \quad p_e = \frac{r_e}{\sum_{e \in [e_1, \dots, e_E]} r_e},$$



Time-Enhanced Attention

$$\alpha_{i,j} = \frac{\exp(e_{i,j})}{\sum_{k=t+1}^{T} \exp(e_{i,k})},$$

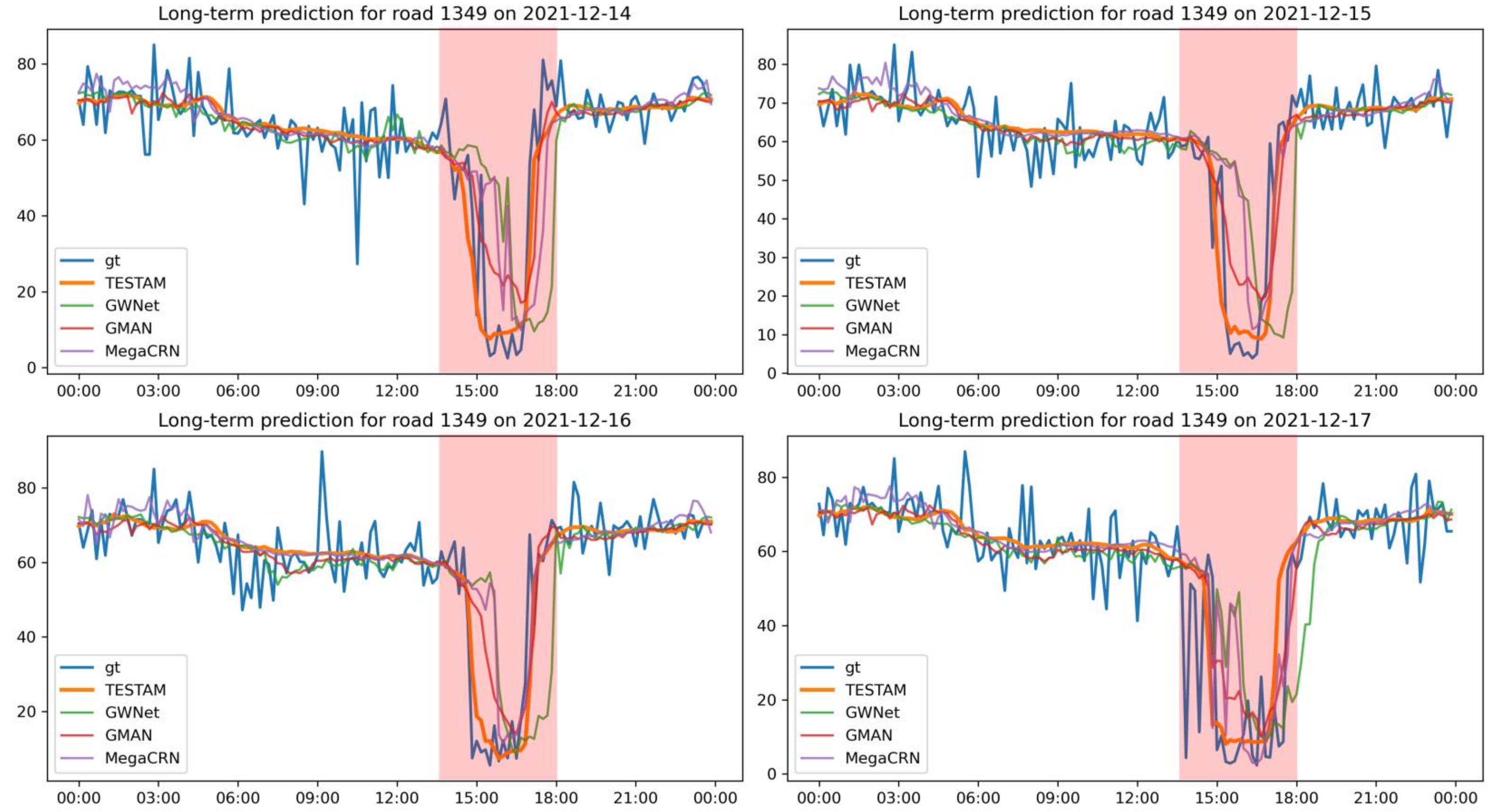
$$e_{i,j} = \frac{(H^{(i)} W_q^{(k)})(\mathbf{TIM}(\boxed{\tau^{(j)}})W_k^{(k)})^\top}{\sqrt{d_k}},$$

Temporal Information Embedding

$$TIM(\tau)[i] = \begin{cases} w_i v(\tau)[i] + \phi_i, & \text{if } i = 0 \\ \mathcal{F}(w_i v(\tau)[i] + \phi_i) & \text{if } 1 \leq i \leq h-1, \end{cases}$$

$$\boldsymbol{X}_{out}^{(t)} = Y^{(t)} = [X^{(t+1)}, \dots, X^{(t+T)}] \longrightarrow \boxed{\boldsymbol{\tau}_{label}^{(t)}}$$

Showcase: highway entrance near Tokyo station

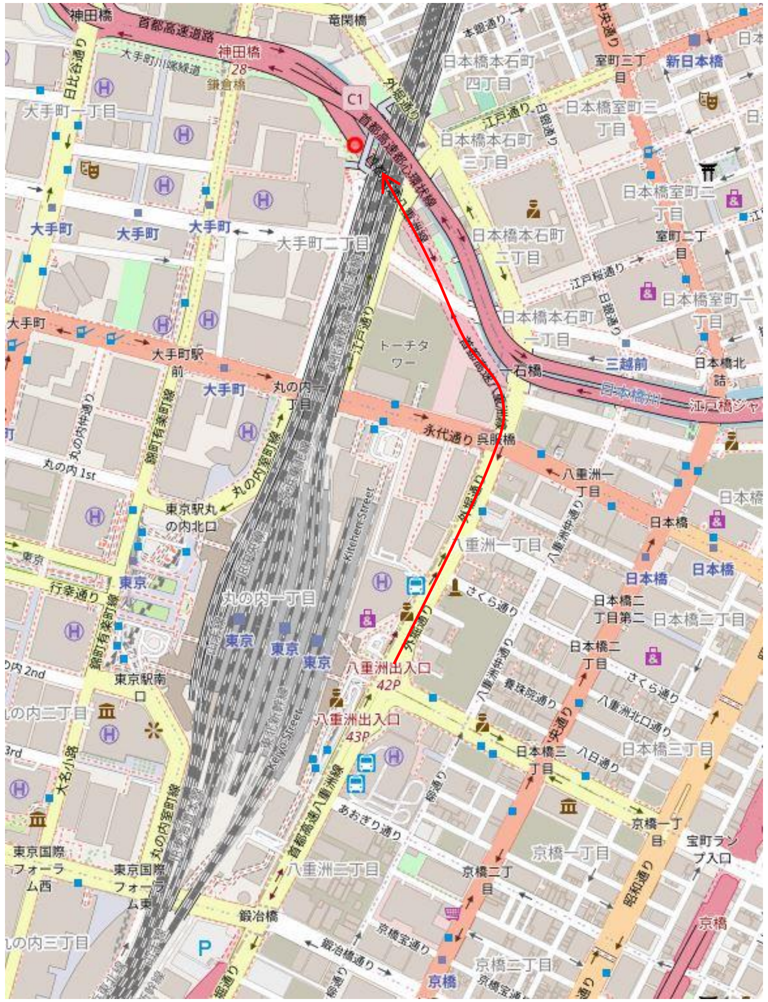


Long-term prediction for road 1349 on 2021-12-14

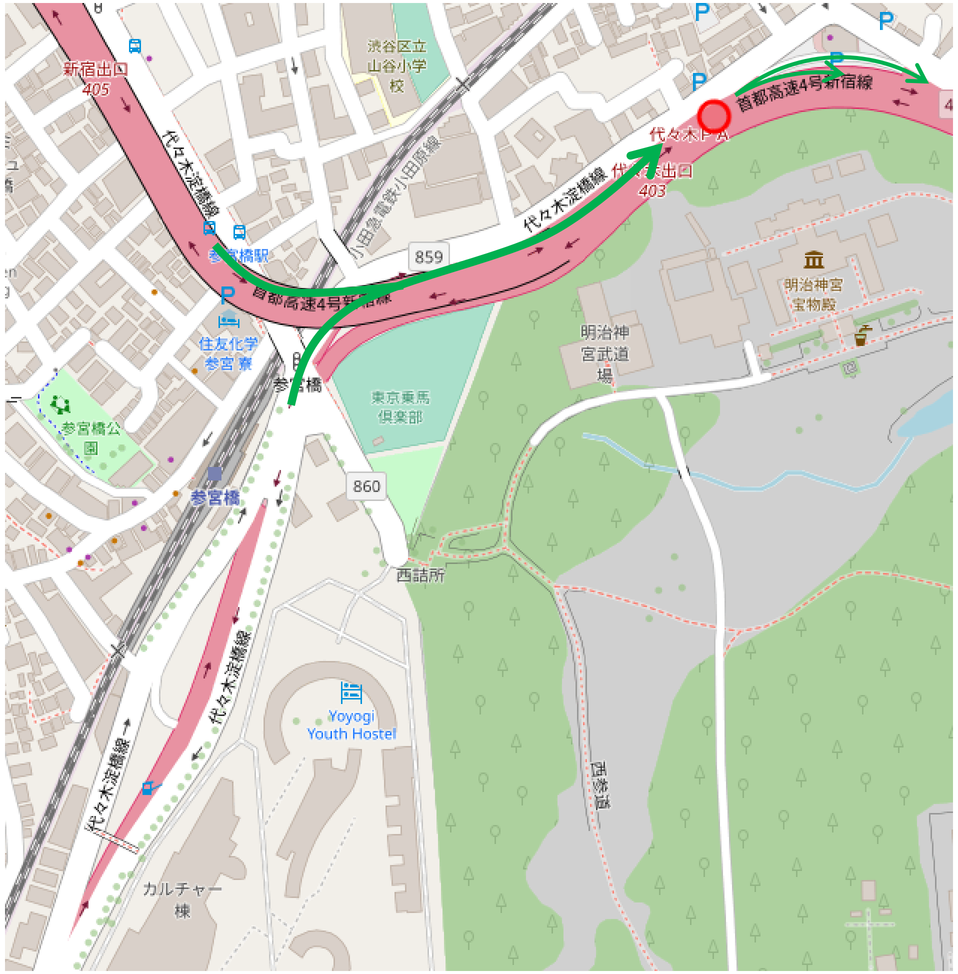
Long-term prediction for road 1349 on 2021-12-15

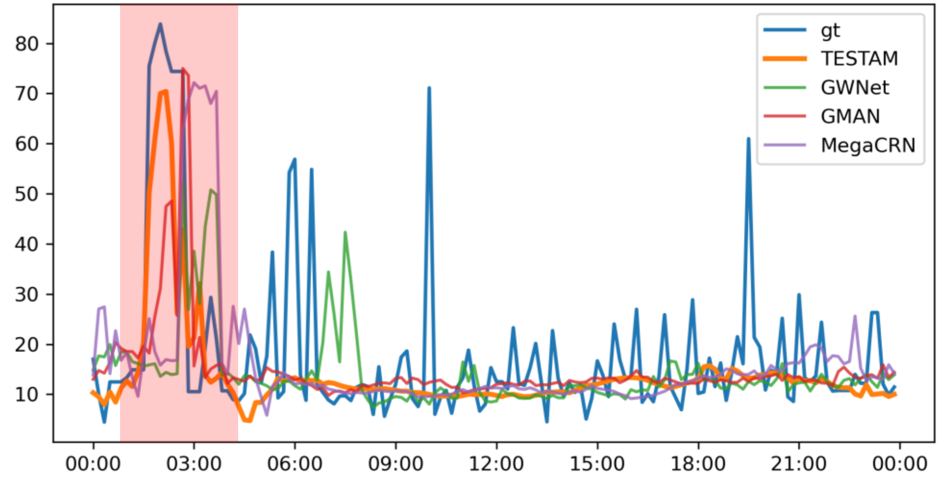
Long-term prediction for road 1349 on 2021-12-16

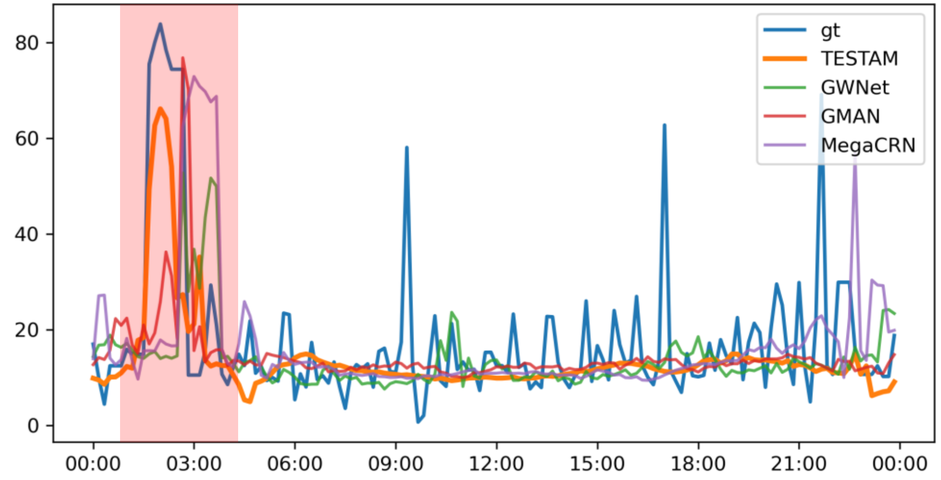
Long-term prediction for road 1349 on 2021-12-17



Showcase: complex intersection near Yoyogi park



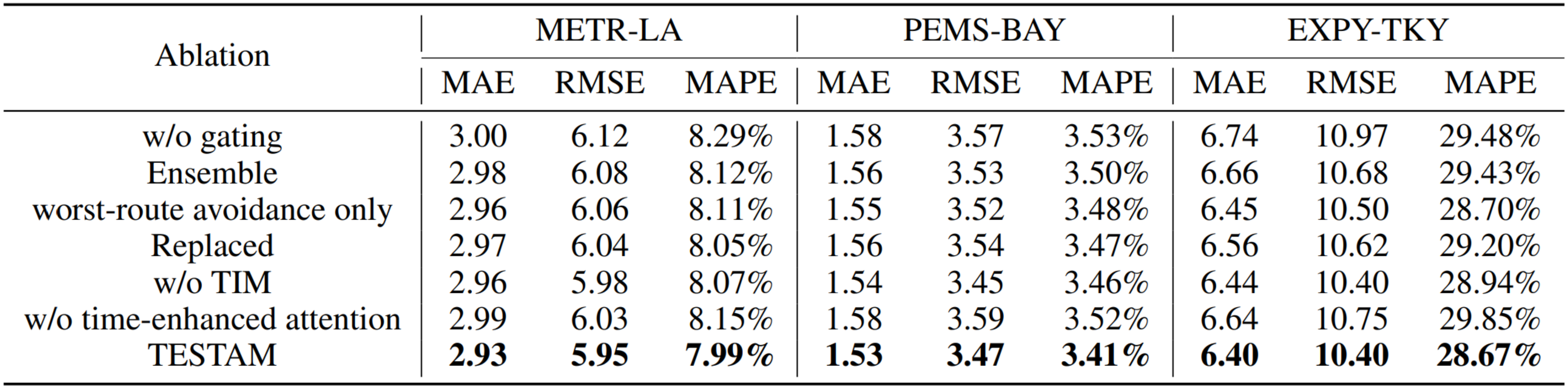

Long-term prediction for road 1111 on 2021-12-9


Long-term prediction for road 1111 on 2021-12-10

Ablation Study Results

- Removing components makes degradation in performance
- Mixture-of-Experts, diversity of spatial modeling methods, and time-enhanced attention take an important role in the model

Table 2: Ablation study results across all prediction windows (i.e., average performance)



Ablation	METR-LA			PEMS-BAY			EXPY-TKY		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
w/o gating	3.00	6.12	8.29%	1.58	3.57	3.53%	6.74	10.97	29.48%
Ensemble	2.98	6.08	8.12%	1.56	3.53	3.50%	6.66	10.68	29.43%
worst-route avoidance only	2.96	6.06	8.11%	1.55	3.52	3.48%	6.45	10.50	28.70%
Replaced	2.97	6.04	8.05%	1.56	3.54	3.47%	6.56	10.62	29.20%
w/o TIM	2.96	5.98	8.07%	1.54	3.45	3.46%	6.44	10.40	28.94%
w/o time-enhanced attention	2.99	6.03	8.15%	1.58	3.59	3.52%	6.64	10.75	29.85%
TESTAM	**2.93**	**5.95**	**7.99%**	**1.53**	**3.47**	**3.41%**	**6.40**	**10.40**	**28.67%**

Summary

- Diversifying spatial modeling methods is beneficial
- Temporal information could be an indicator to guide attention domain
- Gating mechanism and spatial modeling have rooms for improvements

Hyunwook Lee
UNIST

Sungahn Ko
UNIST

Paper

Official Code

Our in-person poster session is on Wednesday 4:30 pm!