Handling Spatial-Temporal IoT Data

Presenters: Jiayi Xiao, Kai-Siang Wang, Madhav Khirwar
Date: 03/03/2026

UNIVERSITY OF
ILLINOIS
URBANA-CHAMPAIGN

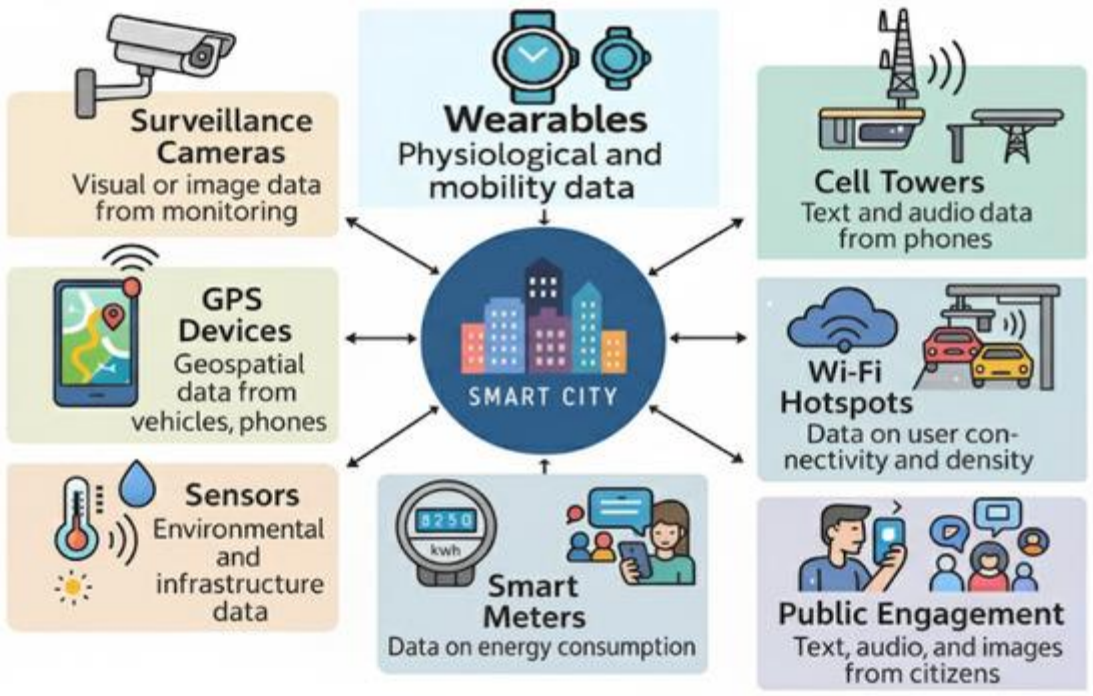
Importance of Spatial-Temporal IoT Data

IoT data are not **isolated readings**

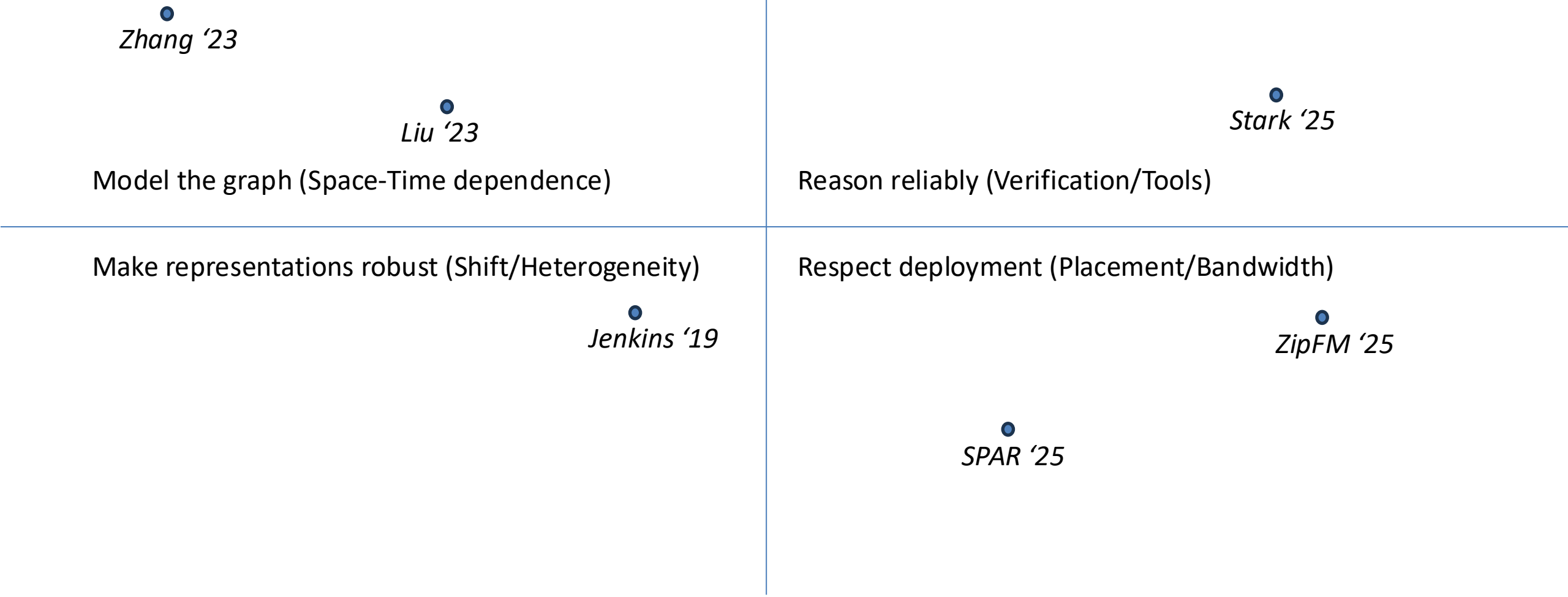
- Spatial: Physical distribution & topological correlation
- Temporal: continuous measurements

Spatial-temporal analysis needed for tasks:

- Forecasting
- Anomaly detection
- Localization
- Etc.



Bottleneck Landscape



Agenda

1. Supervised Spatiotemporal Representation Learning (Inter-Sensor)

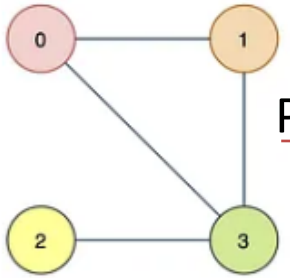
2. Self-supervised/Unsupervised Learning for Spatial Representation (City-Scale)

3. Self-supervised Distributed Sensing (Placement-Aware & Network Efficiency)

4. Zero-shot Reasoning on Representation

Graphs for Mobile Sensor Networks

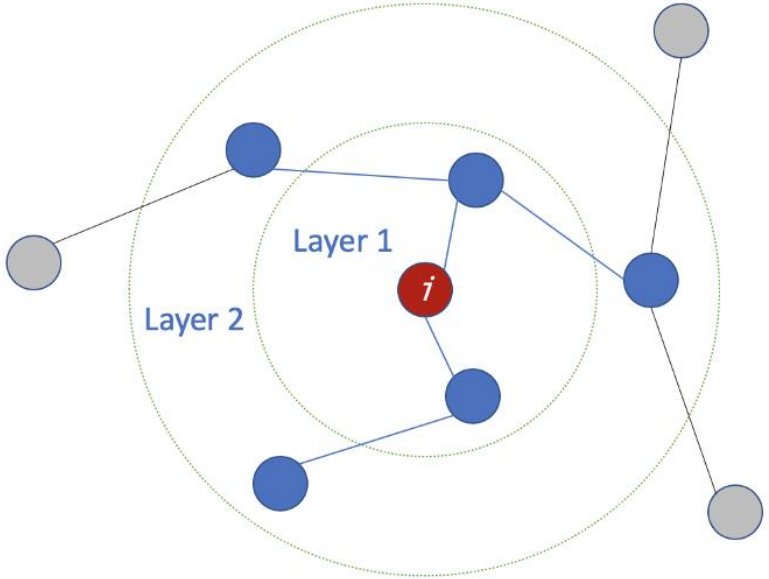
- Monitoring toxic gas (CO) using distributed, mobile sensors
- Traditional Graph Convolutional Network (GCN)



Pre-defined Adjacency Matrix A

	0	1	2	3
0	1	1	0	1
1	1	1	0	1
2	0	0	1	1
3	1	1	1	1

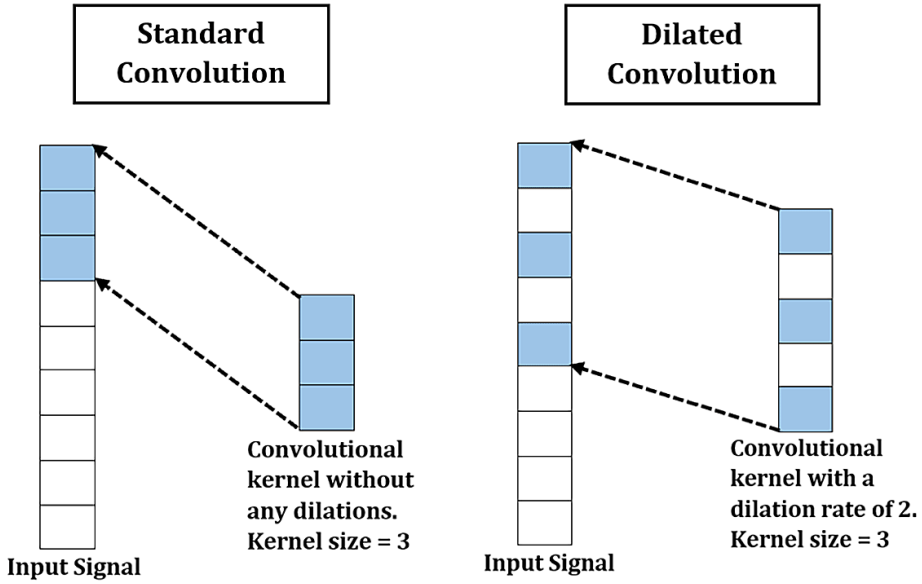
Aggregate
Update



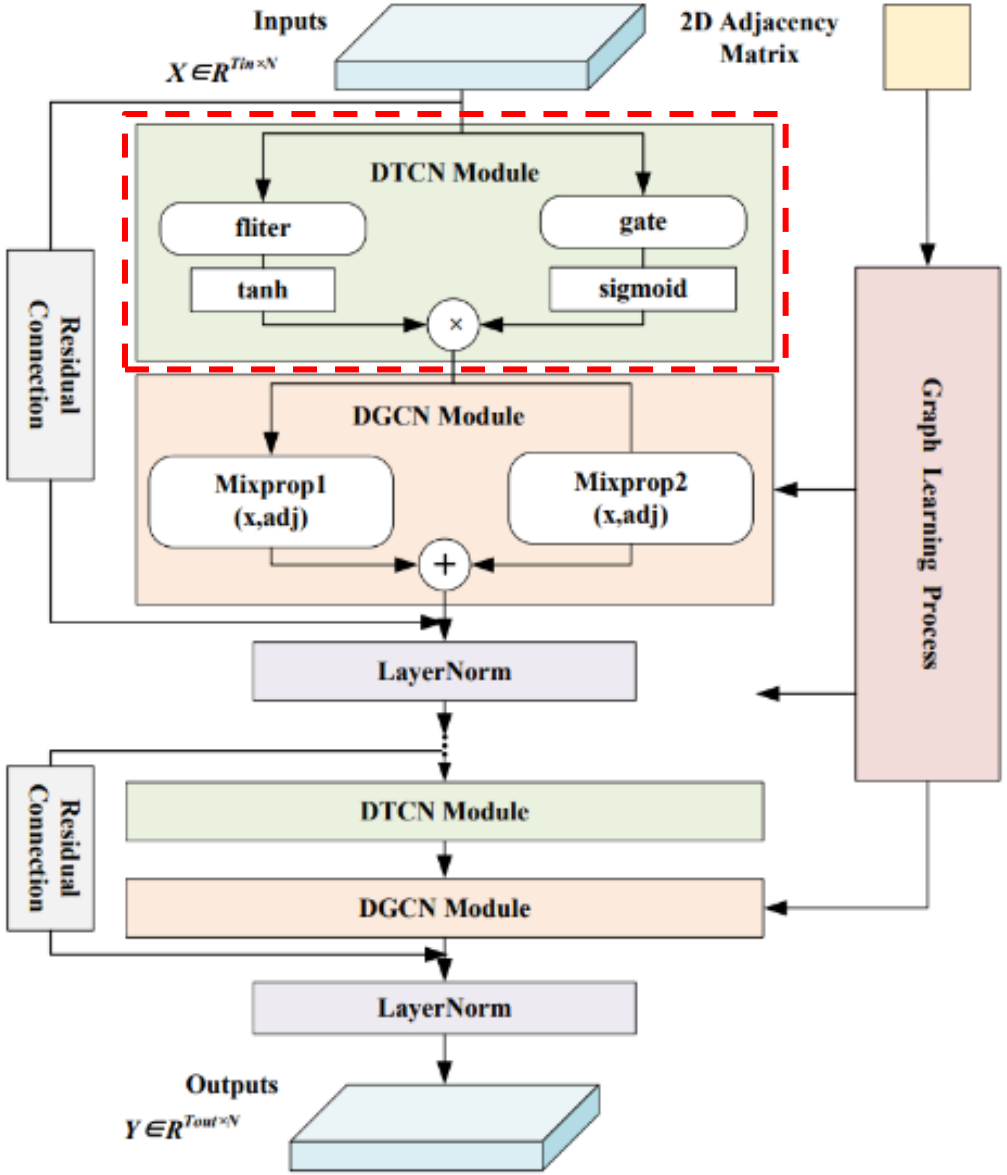
- Each node: a feature vector of gas concentration
- Adjacency matrix: connections between nodes
- Limitation of traditional GCN: fixed pre-defined graph & over smoothing

How do we build a dynamic graph for moving sensors without going too deep?

Dilated Temporal Convolutional Network (DTCN)



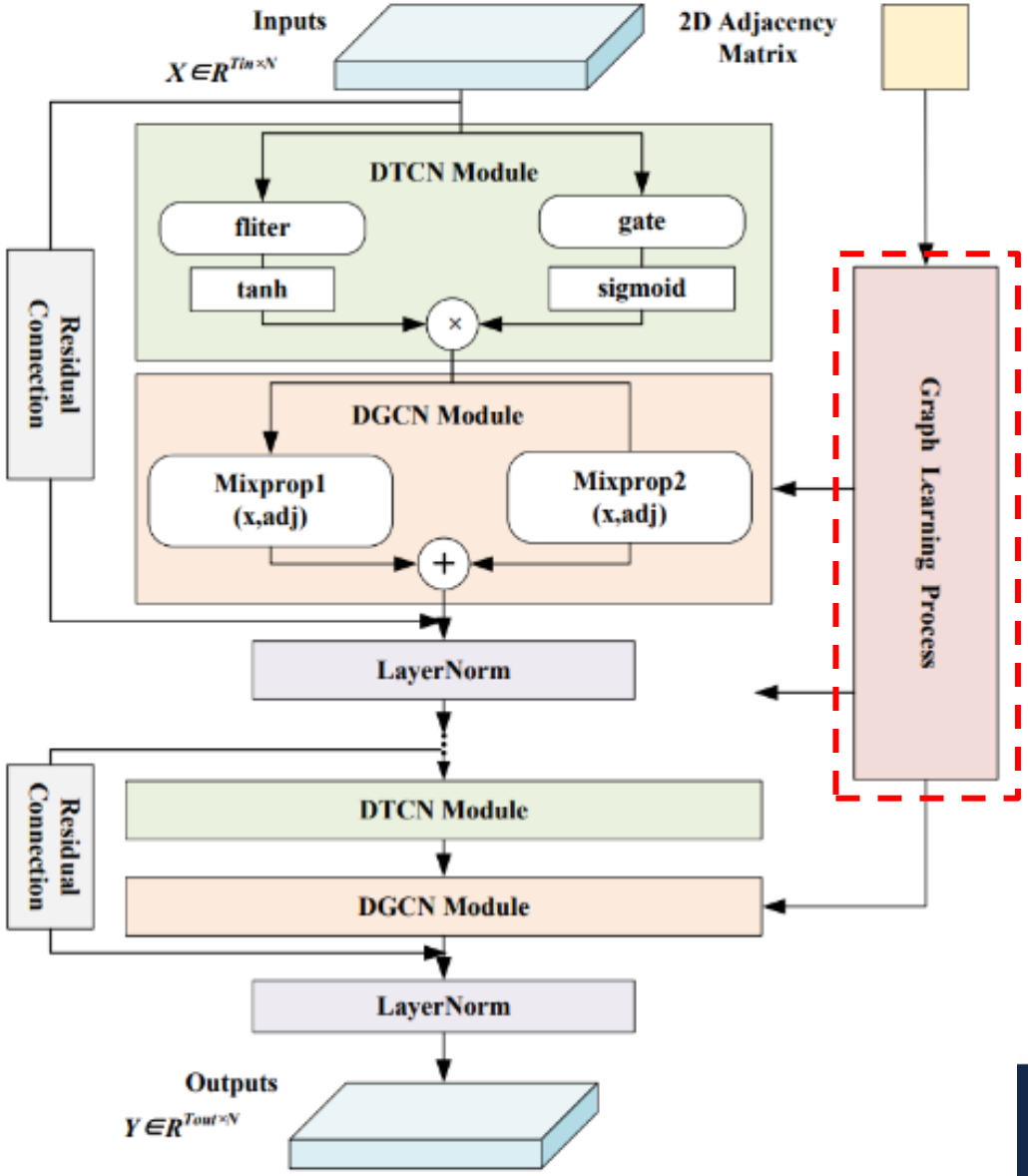
- Dilation: enlarge temporal receptive field
- 2 expanded receptive layers
  - Filter
  - Gate



Graph Learning Process

- Randomly initialized node embeddings

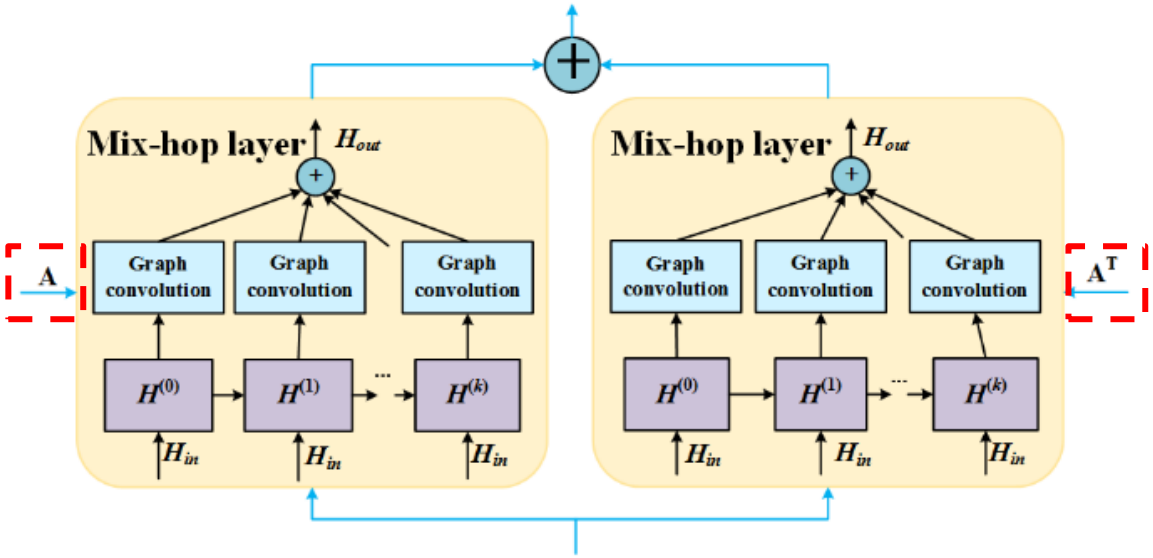
  (learnable parameters ($E1$,$E2$,$\vartheta 1$,$\vartheta 2$) representing the latent identity of each node)
- Compute pairwise relation scores
- Keep only top-k neighbors
- Output: adaptive asymmetric adjacency matrix

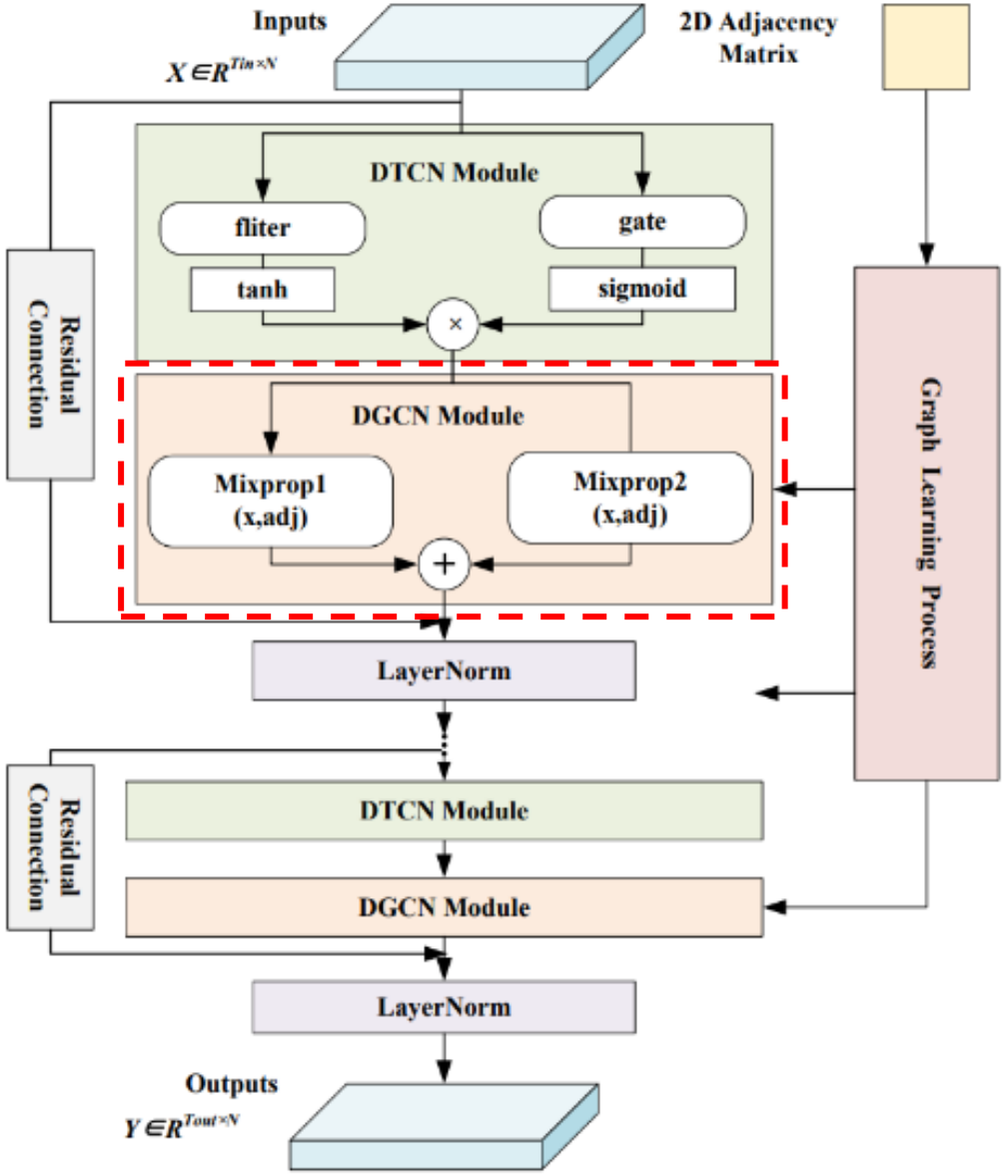


Dynamic Graph Convolutional Network (DGCN)

a parameter used to retain a fraction of the node's original state

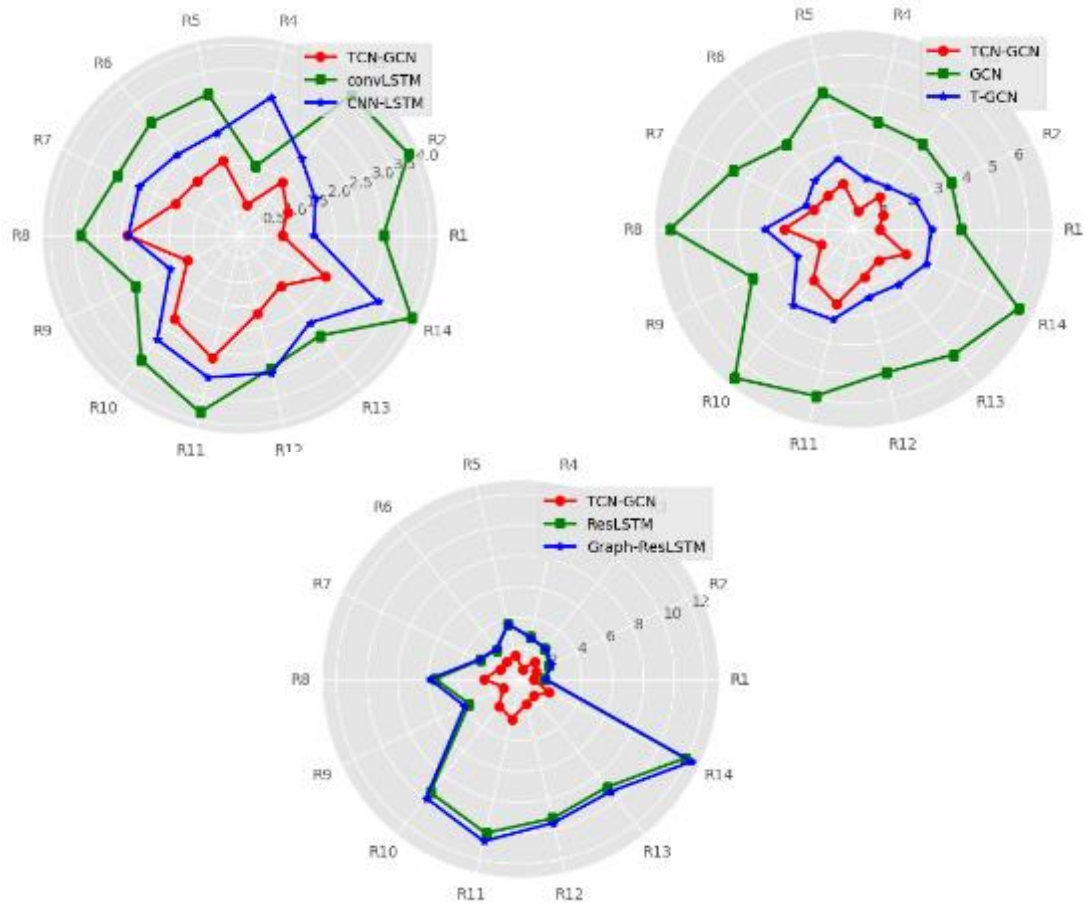
To avoid over-smoothing





Prediction Results

14 gas sensors to detect CO



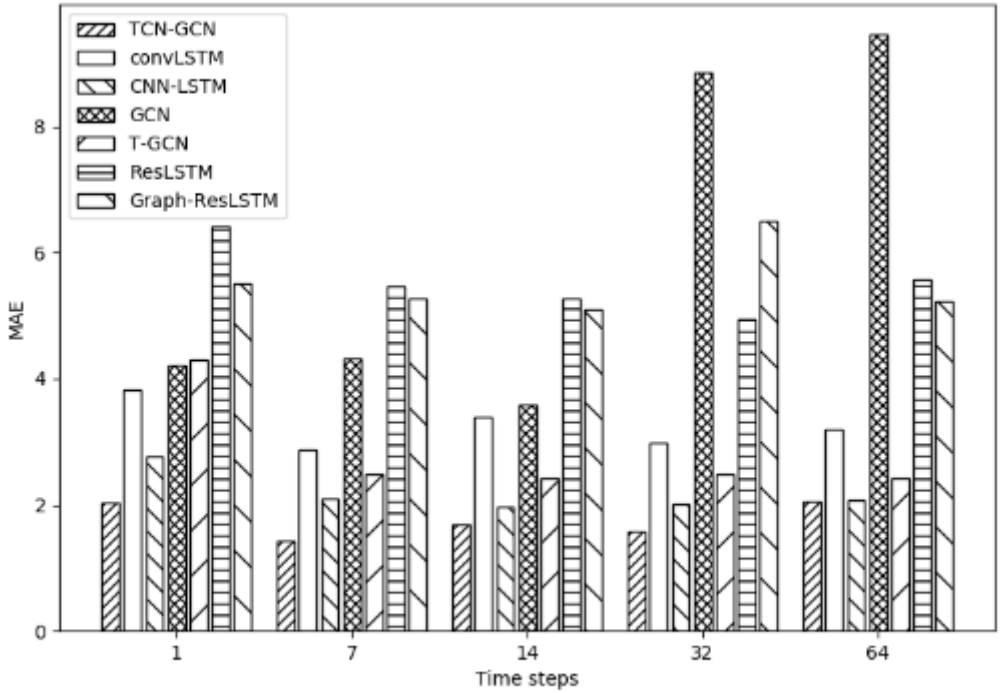
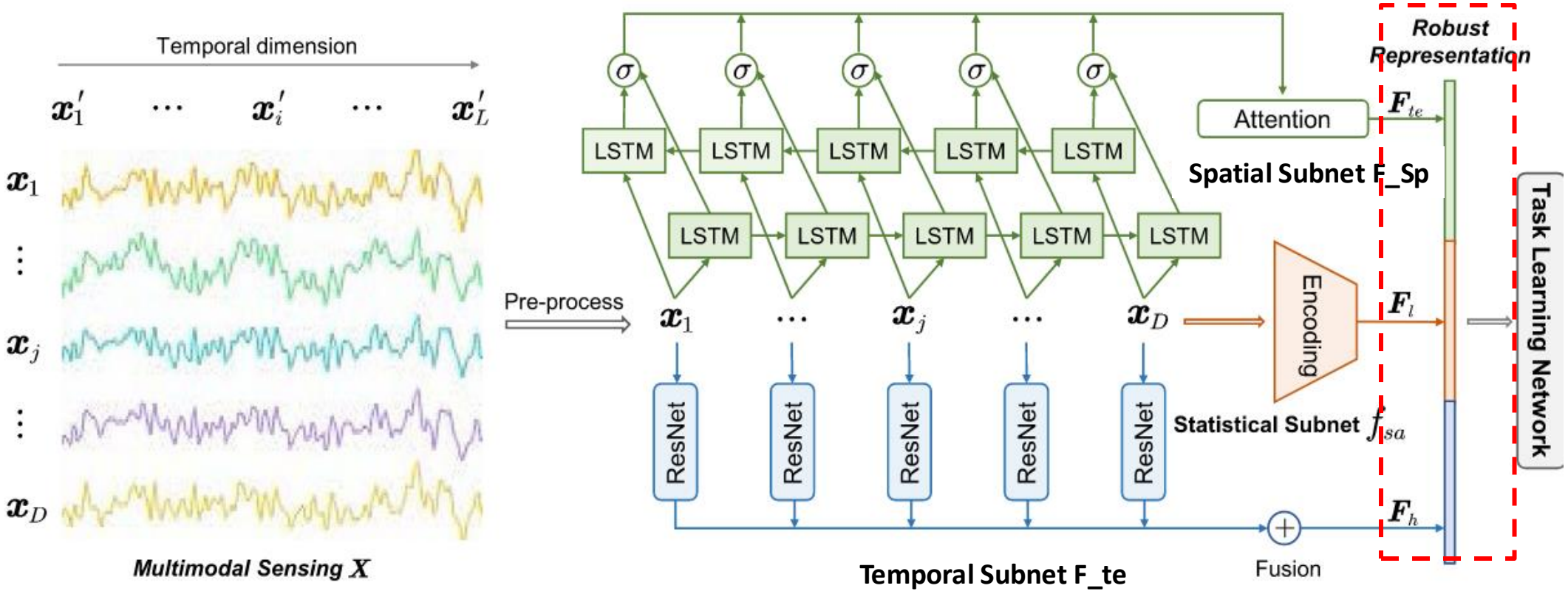


Fig. 13. Comparison of MAE at different time steps.

DSTRR: Micro-level Inter-Sensor Correlation (Transportation Activity Recognition via Smartphone Sensors)

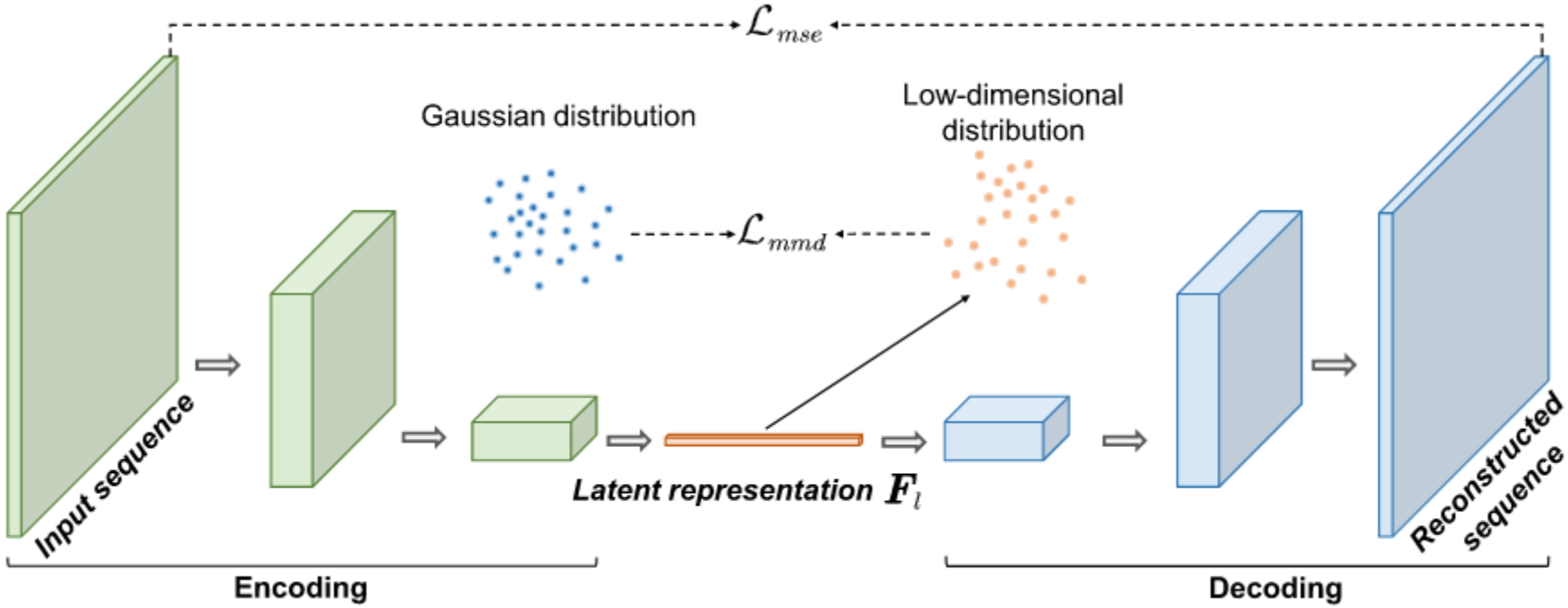
- Statistical + Spatial + Temporal representations -> Concatenate -> a robust representation F
- L timesteps, D dimensions (multiple sensors on a same device)



Statistical Subnet

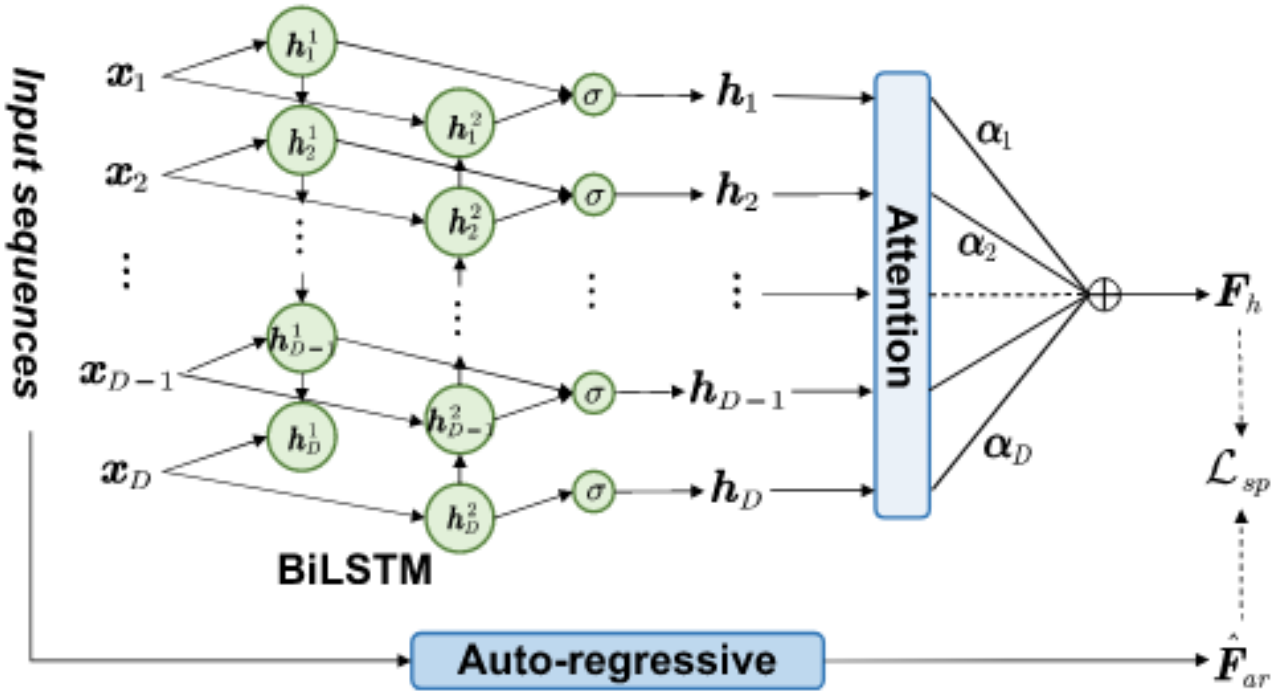
Input: 2D matrix

Loss = Reconstruction loss + MMD loss



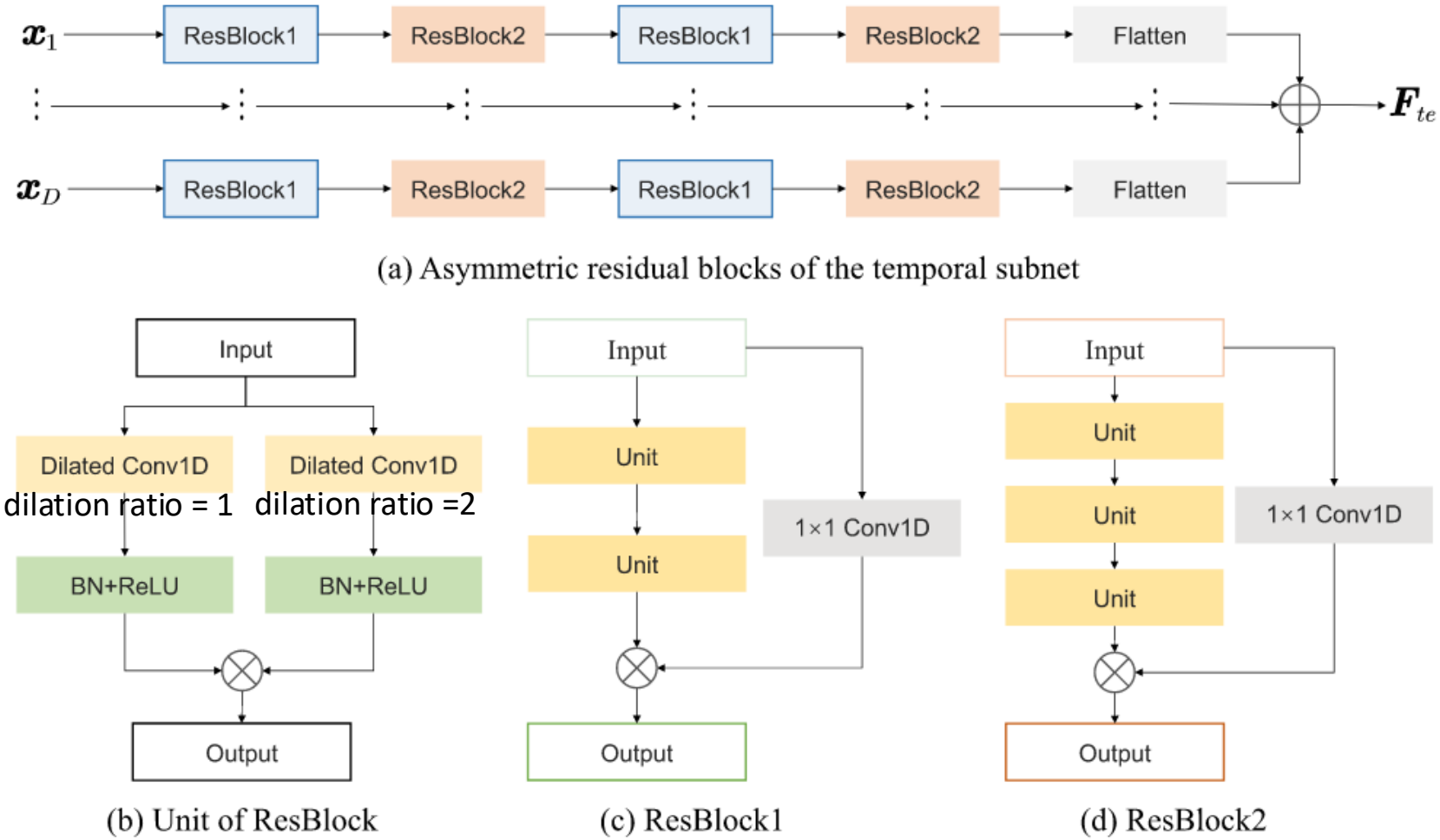
Spatial Subnet

- Input: a set of col vectors (different sensor readings across the entire time window)
- Spatial loss = difference between 2 representations



Temporal Subnet

- Input: Matrix consisting of a set of D column vectors
- Dilation
- ResBlock 1 & 2: capture features at different scales



Results

Description of the datasets. A, G, and M denote the accelerometer, gyroscope, and magnetometer, respectively.

Dataset	# Subject	# Class	# Sample	Frequency	Sensor
TMD	13	5	735,621	20 Hz	A, G, M
SHL	3	8	878,746	100 Hz	A, G, M, GPS
UAH	6	3	359,903	10 Hz	A, G, GPS

Ground truth labels: Still/Walking/Running/Car, Train/Bus

Method	TMD		SHL		UAH	
	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
Random Forest [41]	93.00	92.54	92.24	91.77	93.73	94.32
Decision Tree [27]	91.84	90.02	86.12	83.69	68.12	69.09
SVM [38]	71.12	72.46	87.64	84.08	83.34	81.46
kNN [39]	88.73	90.62	86.06	83.45	82.02	84.56
XGBoost [43]	82.38	85.64	90.20	90.00	75.31	74.72
CNN_Ito [47]	90.81	88.54	91.72	93.00	67.57	66.76
LSTM [44]	90.00	89.45	91.95	91.08	90.33	90.18
DeepConvLSTM [45]	91.57	92.40	92.45	92.61	94.34	94.83
LRNL [2]	91.24	91.61	92.53	92.61	-	-
DSTRR-f_{st}	93.09	92.25	93.65	93.54	95.03	95.97
DSTRR-f_{sp}	93.85	93.68	94.77	94.17	96.23	96.78
DSTRR-f_{te}	93.43	92.84	93.82	94.01	95.85	96.30
DSTRR-f_{sa}	94.37	93.65	95.03	94.85	96.75	97.17
DSTRR	**95.14**	**94.57**	**95.62**	**95.49**	**97.34**	**97.81**

Limitations

- Learned spatial representation is for prediction/classification (specific tasks)

Agenda

1. Supervised Spatiotemporal Representation Learning (Inter-Sensor)

2. Self-supervised/Unsupervised Learning for Spatial Representation (City-Scale)

3. Self-supervised Distributed Sensing (Placement-Aware & Network Efficiency)

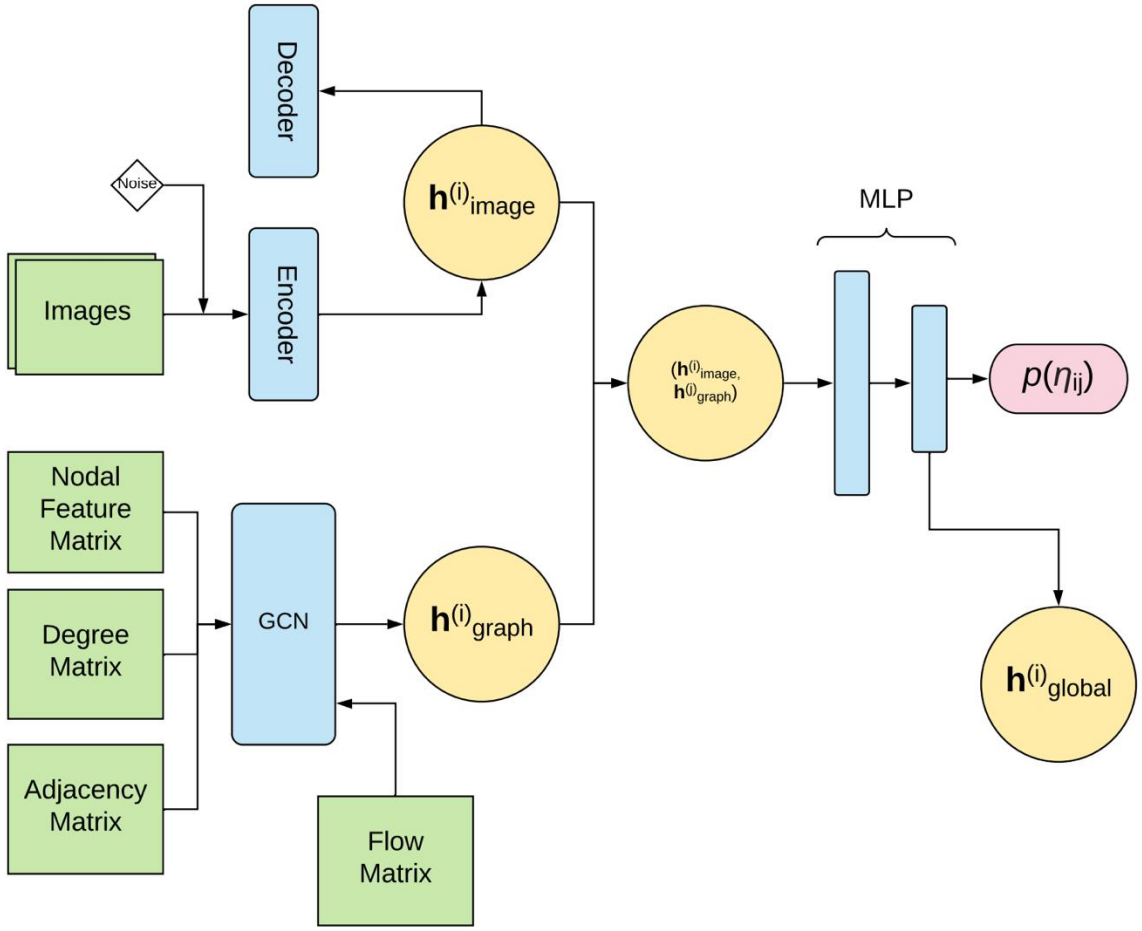
4. Zero-shot Reasoning on Representation

Problem: Regions aren’t just geometry

Same distance ≠ same meaning

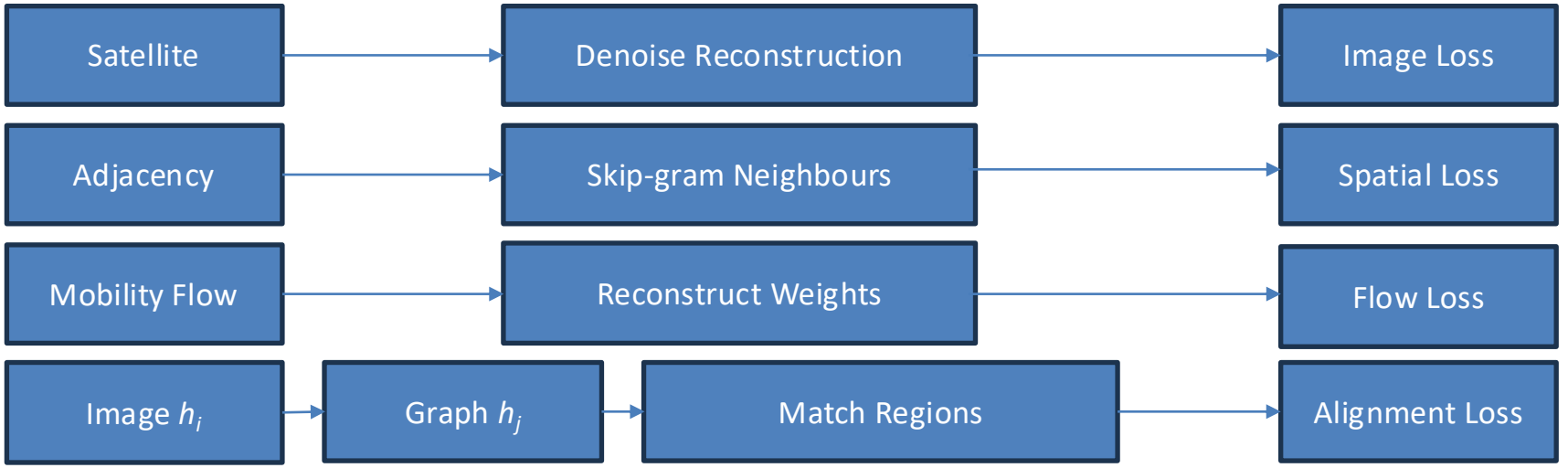
Key idea: “RegionEncoder = multimodal fusion”

A region encoder that enables multimodal fusion…



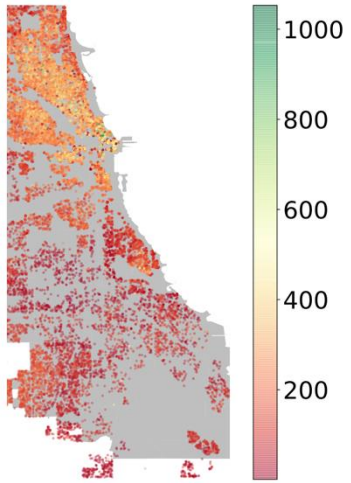
4 losses → one embedding

One embedding vector with many semantics…

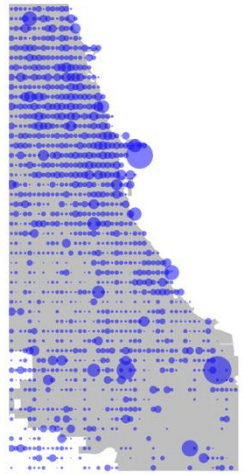


$$L_{total} = \lambda_I L_{img} + \lambda_s L_{spatial} + \lambda_f L_{flow} + \lambda_a L_{align}$$

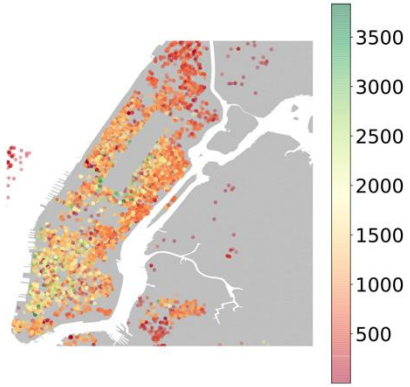
Downstream tasks as probes



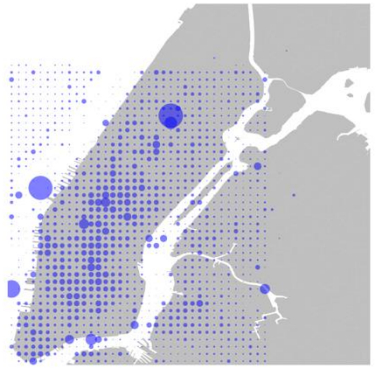
(a) Chicago House Prices



(b) Chicago Regions by POI Popularity



(c) NYC House Prices



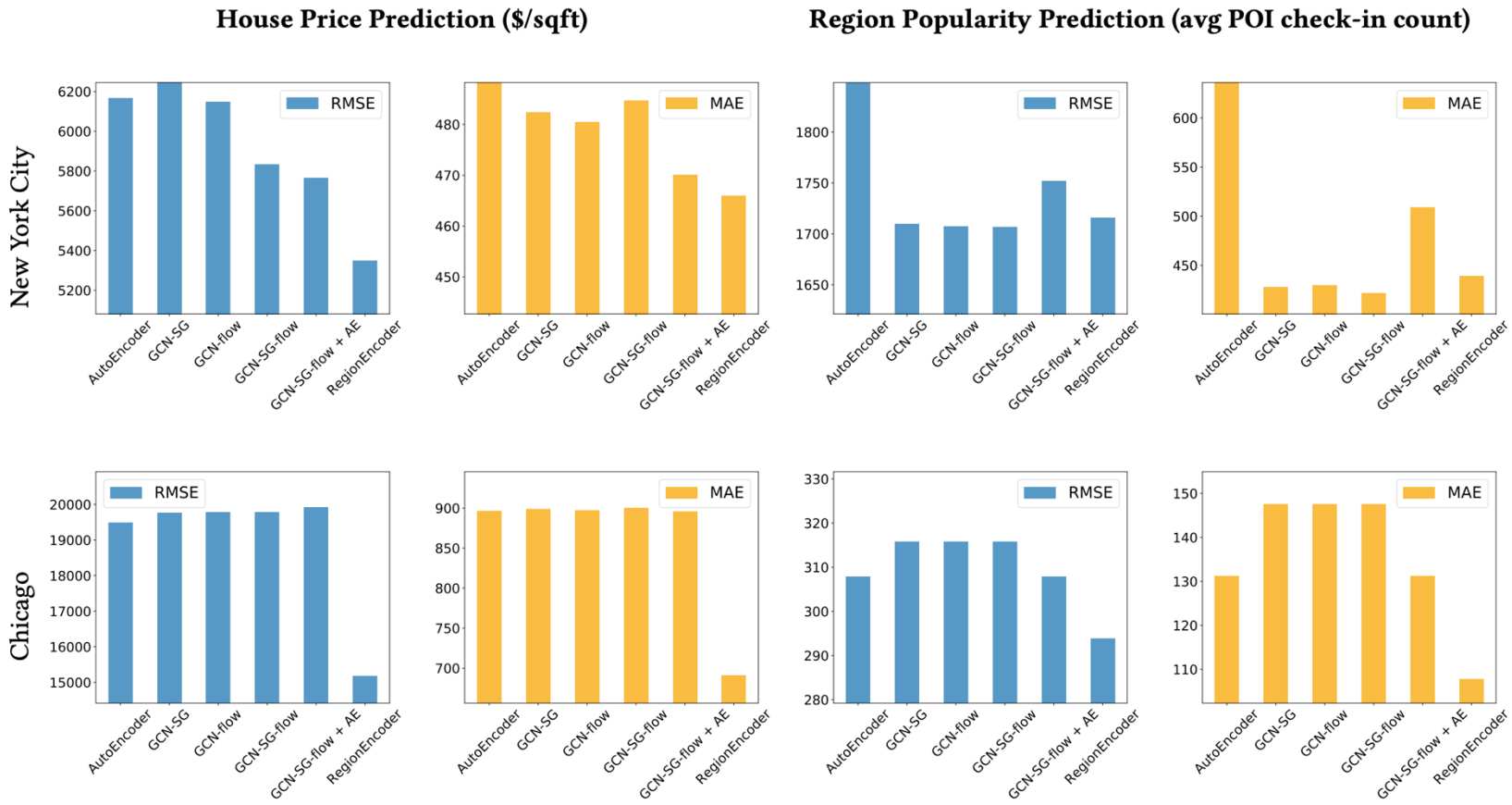
(d) NYC Regions by POI Popularity

1) Probe tasks

2) Price / Popularity

Multimodal helps (but not uniformly)

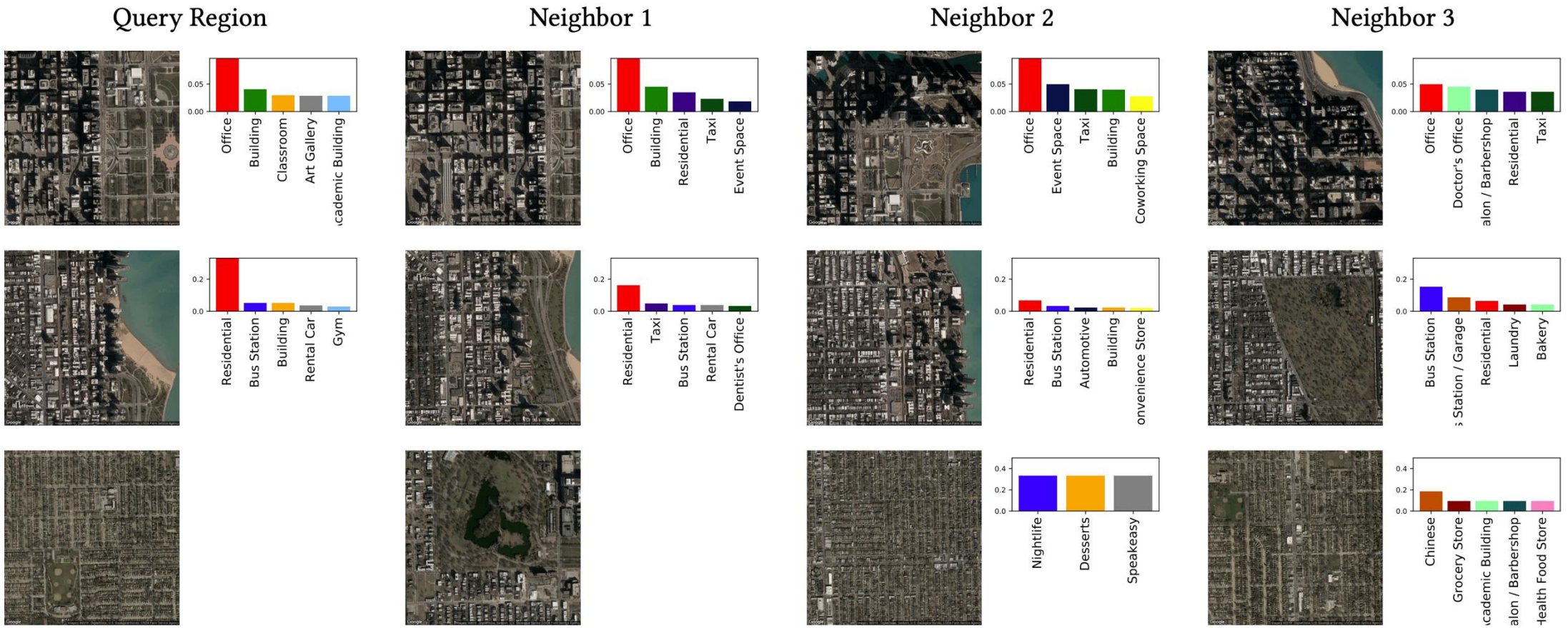
Joint > single-view…



… Context Matters

Limitations

Static Regions, ***NOT*** sensor placement



Agenda

1. Supervised Spatiotemporal Representation Learning (Inter-Sensor)

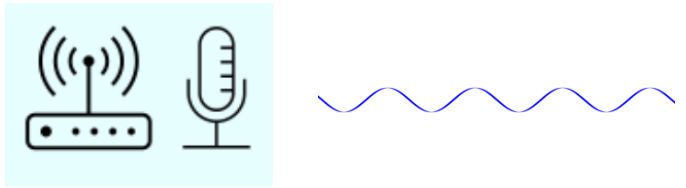
2. Self-supervised/Unsupervised Learning for Spatial Representation (City-Scale)

3. Self-supervised Distributed Sensing (Placement-Aware & Network Efficiency)

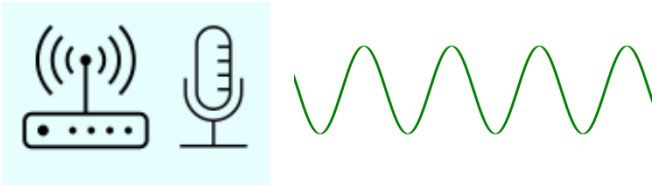
4. Zero-shot Reasoning on Representation

Distributed Sensing

How many cars are on the road?
What type of cars?



Environment



Signals are not enough

Spatial Position

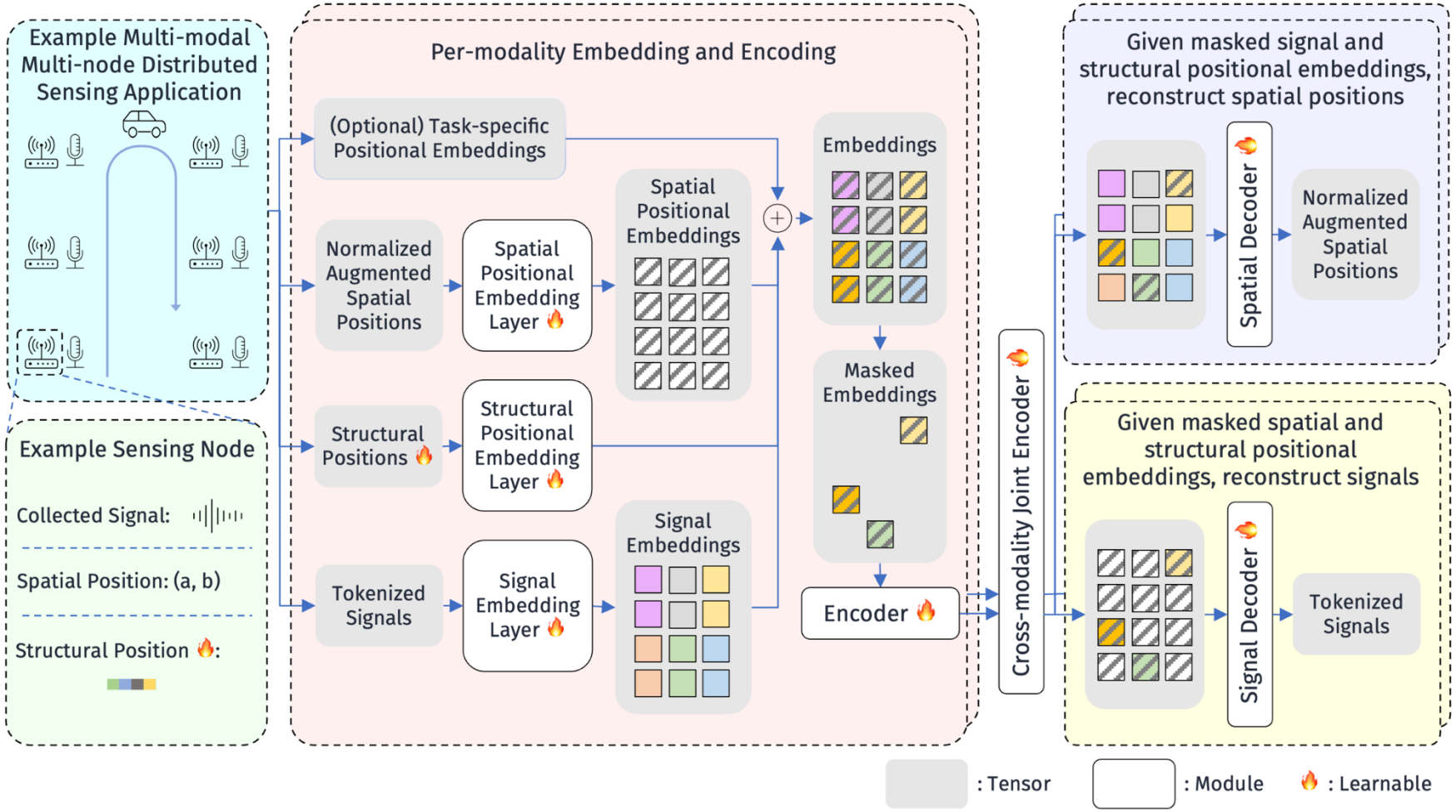
- Learn how geometry contributes to the signal.
- Layout may change across deployments.
- For example:

  *A **near** weak signal and a **far** strong signal can have a same representation.*

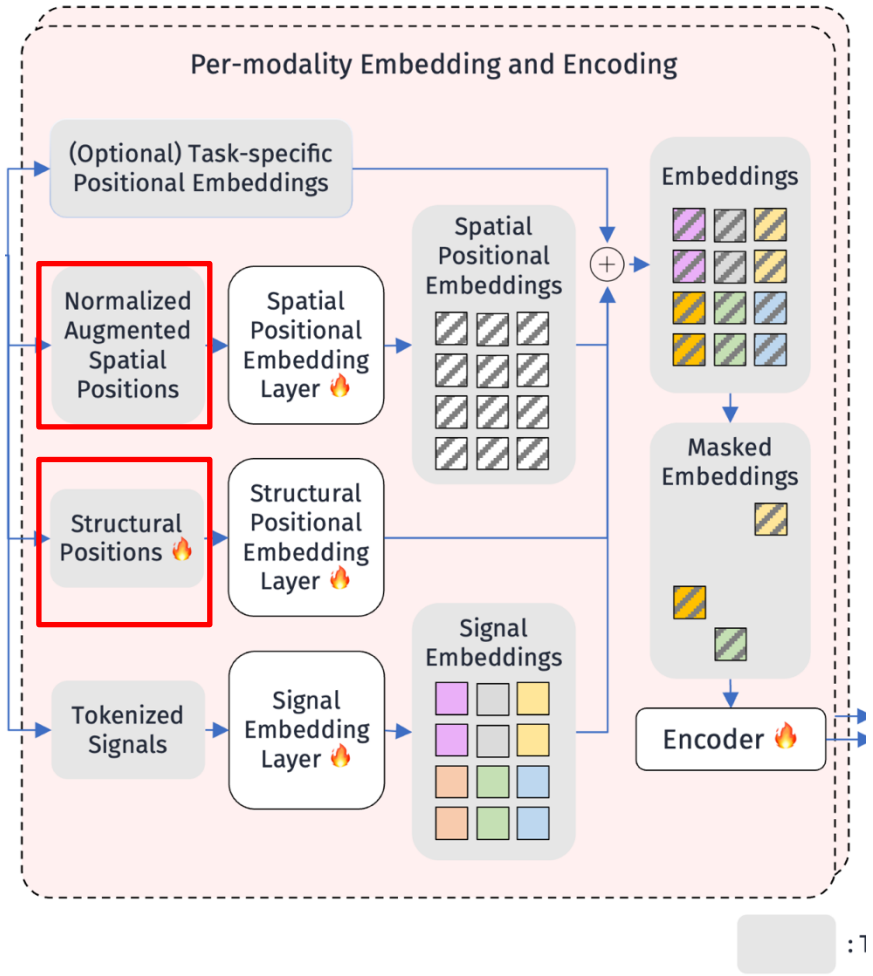
Structural Position

- Sensor-specific static bias
  - Sensitivity
  - Mounting orientation
  - Noise characteristics
- Manually labeling for all nodes is often costly and non-scalable.

SPAR Design


Example Multi-modal Multi-node Distributed Sensing Application
Example Sensing Node
Collected Signal:
Spatial Position: (a, b)
Structural Position :
Per-modality Embedding and Encoding
(Optional) Task-specific Positional Embeddings
Normalized Augmented Spatial Positions
Spatial Positional Embedding Layer
Spatial Positional Embeddings
Structural Positions
Structural Positional Embedding Layer
Tokenized Signals
Signal Embedding Layer
Signal Embeddings
Embeddings
Masked Embeddings
Encoder
Cross-modality Joint Encoder
Given masked signal and structural positional embeddings, reconstruct spatial positions
Spatial Decoder
Normalized Augmented Spatial Positions
Given masked spatial and structural positional embeddings, reconstruct signals
Signal Decoder
Tokenized Signals
: Tensor
: Module
: Learnable

Encoder



Spatial Position

GPS Node Location

Structural Position

A learnable vector specific to the device

Dual reconstruction objectives

Spatial Decoder

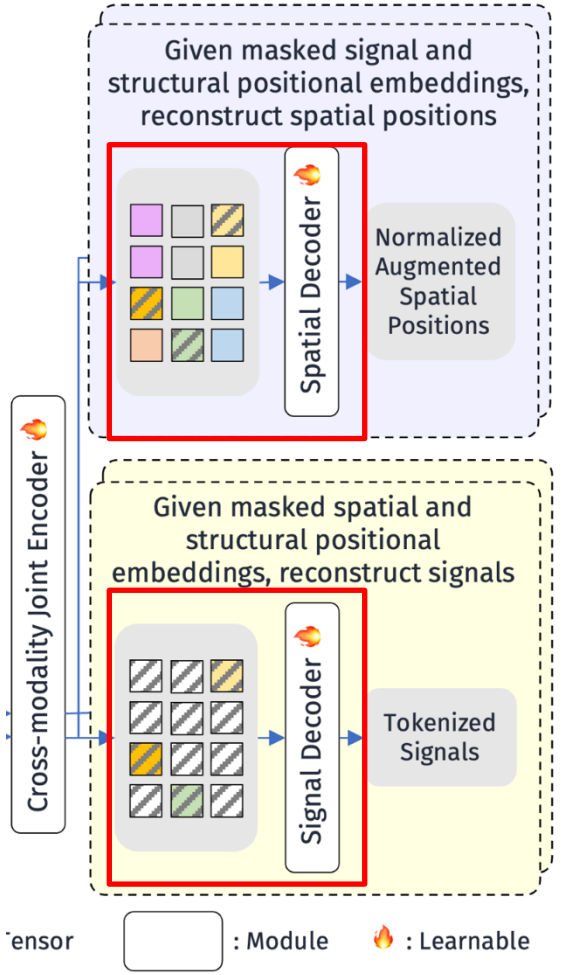
Inputs:

- Latent representation
- Signal embeddings (Unmasked)
- Structural embeddings (Unmasked)

Signal Decoder

Inputs:

- Latent representation
- Spatial embeddings (Unmasked)
- Structural embeddings (Unmasked)



Information Theory

Classical MAE has the following bound

$$-\mathbb{E}[L'] + C' \leq \sum_{k=1}^{K} I(\mathsf{X}^{(k)}; \widetilde{\mathsf{Z}}^{(k)})$$

$\boldsymbol{X}$: Signal Embeddings

$\boldsymbol{Z}$: Latent Representation

Information Theory

$$-\mathbb{E}[L] + C \leq \sum_{k=1}^{K} I(\mathsf{X}^{(k)}; \widetilde{\mathsf{Z}}^{(k)} | \mathsf{S}^{(k)}, \mathsf{R}^{(k)}) + I(\mathsf{S}^{(k)}; \widetilde{\mathsf{Z}}^{(k)} | \mathsf{X}^{(k)}, \mathsf{R}^{(k)}).$$

$\boldsymbol{X}$: Signal Embeddings
$\boldsymbol{Z}$: Latent Representation
$\boldsymbol{S}$: Spatial Embeddings
$\boldsymbol{R}$: Structural Embeddings

Information Theory

How much information about the signal X is stored in Z, assuming we already know placement S and structure R.

$$-\mathbb{E}[L] + C \leq \sum_{k=1}^{K} I(\mathsf{X}^{(k)}; \widetilde{\mathsf{Z}}^{(k)} | \mathsf{S}^{(k)}, \mathsf{R}^{(k)}) + I(\mathsf{S}^{(k)}; \widetilde{\mathsf{Z}}^{(k)} | \mathsf{X}^{(k)}, \mathsf{R}^{(k)}).$$

$\boldsymbol{X}$: Signal Embeddings
$\boldsymbol{Z}$: Latent Representation
$\boldsymbol{S}$: Spatial Embeddings
$\boldsymbol{R}$: Structural Embeddings

How much information about placement S is stored in Z, assuming we already know the signal X.

Minimizing the loss directly means to maximize the signal and placement information

Evaluation

Datasets

M3N-VC dataset: moving vehicles

Ridgecrest Seismicity Dataset: Earthquake events

RealWorldHAR dataset: Human activity recognition

Baselines

Classical MAE, Contrastive Learning, other strong general-purpose methods

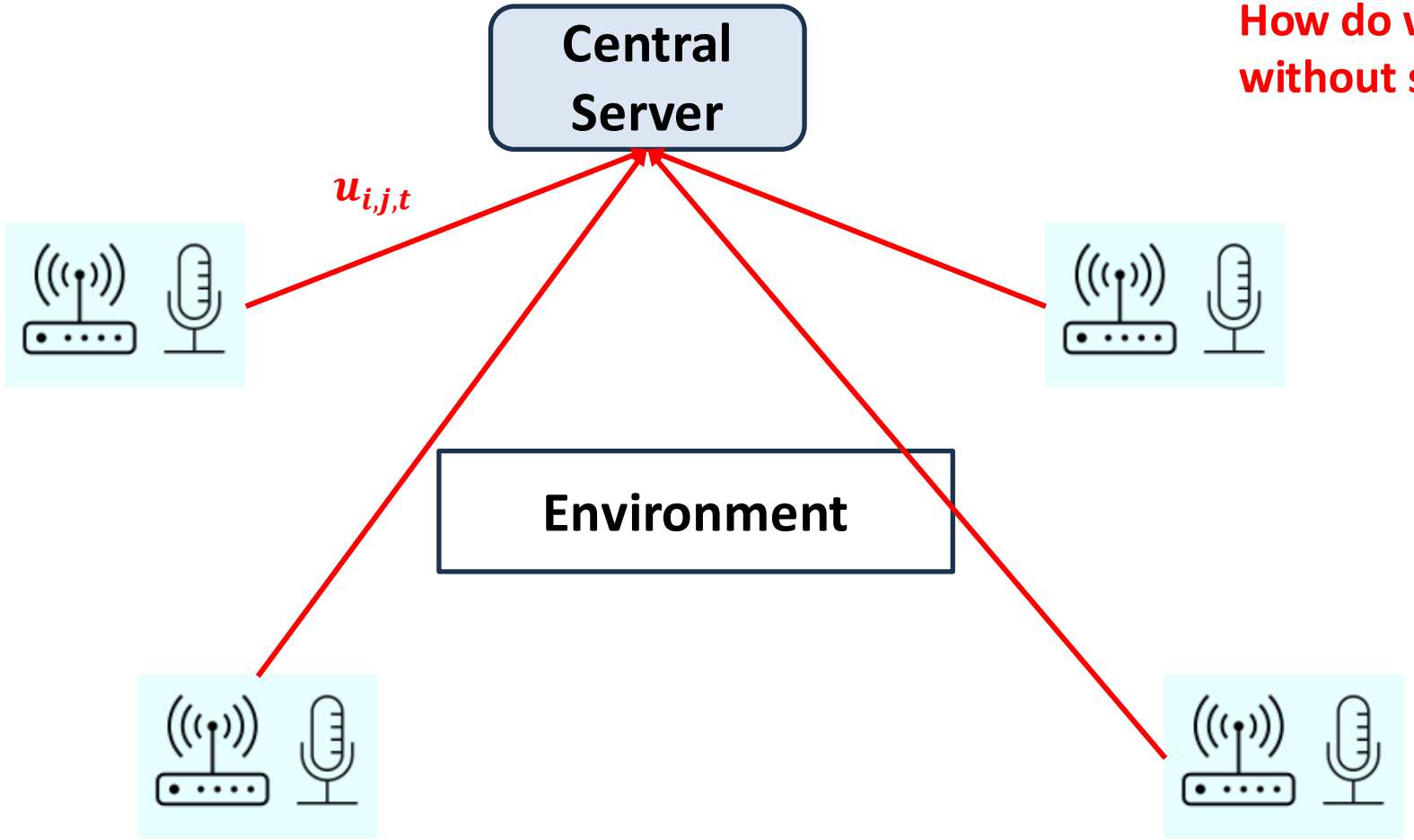
Performance

Method	M3N-VC Classification		Ridgecrest Earthquake Localization		RealWorld-HAR Recognition	
	Accuracy (%) (↑)	F1 (%) (↑)	MSE (km^2) (↓)	Dist. Err. (km) (↓)	Accuracy (%) (↑)	F1 (%) (↑)
CMC	89.53 ± 7.62	89.33 ± 7.78	94.25 ± 6.67	10.38 ± 0.63	74.97 ± 1.23	74.82 ± 2.18
Cosmo	94.21 ± 0.50	94.04 ± 0.54	98.24 ± 13.77	10.44 ± 0.83	84.37 ± 0.33	85.30 ± 0.43
SimCLR	95.53 ± 0.73	95.41 ± 0.74	99.87 ± 11.31	10.29 ± 0.52	84.36 ± 0.47	85.49 ± 0.36
AudioMAE	99.06 ± 0.23	99.03 ± 0.24	33.65 ± 3.51	5.65 ± 0.29	89.18 ± 0.32	90.11 ± 0.53
CAV-MAE	98.97 ± 0.04	98.94 ± 0.04	31.58 ± 3.57	5.48 ± 0.37	88.12 ± 0.24	89.05 ± 0.35
FOCAL	93.62 ± 0.75	93.46 ± 0.76	131.50 ± 1.48	12.53 ± 0.09	84.98 ± 0.73	86.24 ± 0.77
FreqMAE	92.72 ± 0.75	92.55 ± 0.79	54.08 ± 5.44	7.14 ± 0.25	83.43 ± 0.56	84.07 ± 0.50
PhyMask	83.38 ± 2.33	82.68 ± 2.27	56.39 ± 3.27	7.67 ± 0.39	84.79 ± 3.23	82.15 ± 9.13
SPAR	**99.27** ± **0.07**	**99.26** ± **0.07**	**23.46** ± **2.77**	**5.37** ± **0.24**	**89.63** ± **0.57**	**90.45** ± **0.63**

Performance

Method	M3N-VC Classification		Ridgecrest Earthquake Localization		RealWorld-HAR Recognition	
	Accuracy (%) (↑)	F1 (%) (↑)	MSE (km^2) (↓)	Dist. Err. (km) (↓)	Accuracy (%) (↑)	F1 (%) (↑)
CMC	89.53 ± 7.62	89.33 ± 7.78	94.25 ± 6.67	10.38 ± 0.63	74.97 ± 1.23	74.82 ± 2.18
Cosmo	94.21 ± 0.50	94.04 ± 0.54	98.24 ± 13.77	10.44 ± 0.83	84.37 ± 0.33	85.30 ± 0.43
SimCLR	95.53 ± 0.73	95.41 ± 0.74	99.87 ± 11.31	10.29 ± 0.52	84.36 ± 0.47	85.49 ± 0.36
AudioMAE	99.06 ± 0.23	99.03 ± 0.24	33.65 ± 3.51	5.65 ± 0.29	89.18 ± 0.32	90.11 ± 0.53
CAV-MAE	98.97 ± 0.04	98.94 ± 0.04	31.58 ± 3.57	5.48 ± 0.37	88.12 ± 0.24	89.05 ± 0.35
FOCAL	93.62 ± 0.75	93.46 ± 0.76	131.50 ± 1.48	12.53 ± 0.09	84.98 ± 0.73	86.24 ± 0.77
FreqMAE	92.72 ± 0.75	92.55 ± 0.79	54.08 ± 5.44	7.14 ± 0.25	83.43 ± 0.56	84.07 ± 0.50
PhyMask	83.38 ± 2.33	82.68 ± 2.27	56.39 ± 3.27	7.67 ± 0.39	84.79 ± 3.23	82.15 ± 9.13
SPAR	**99.27** ± **0.07**	**99.26** ± **0.07**	**23.46** ± **2.77**	**5.37** ± **0.24**	**89.63** ± **0.57**	**90.45** ± **0.63**

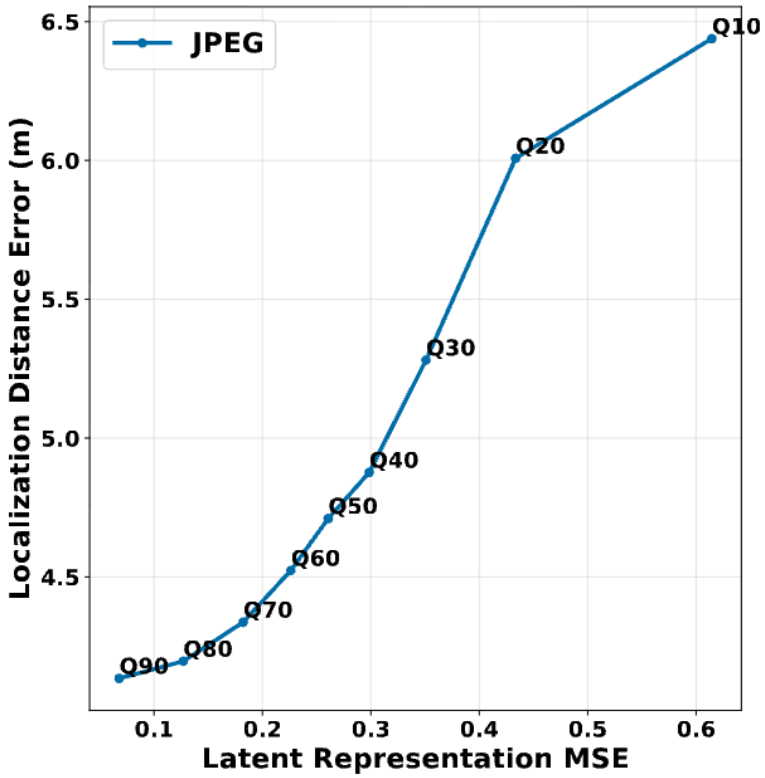
Distributed Sensing – Data Transmission

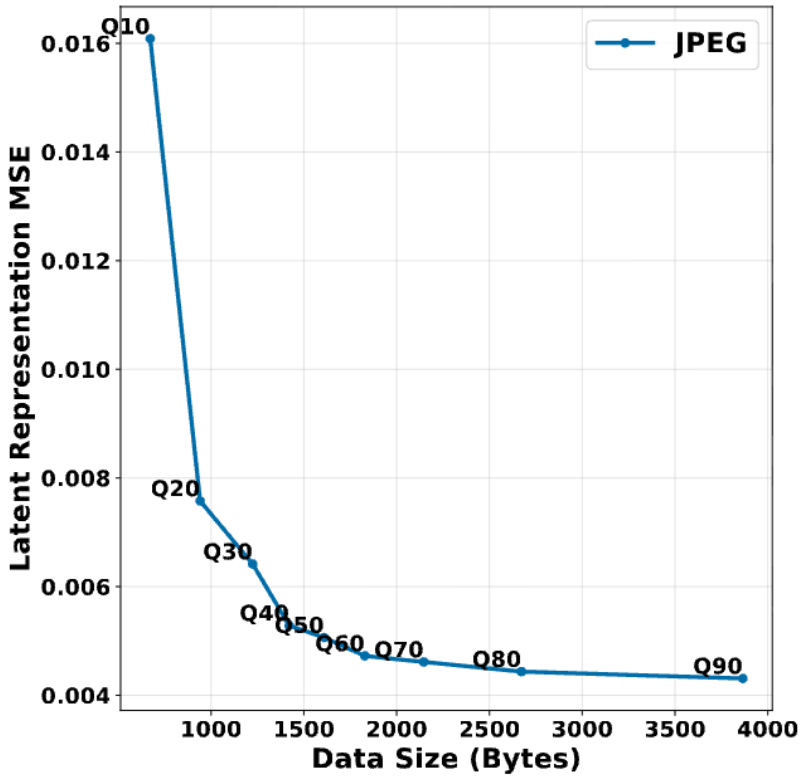


How do we minimize the total bandwidth without sacrificing too much accuracy?

Key Insights

Compression as Latent Representation Shift





ZipFM Design

Objective: Minimize total bitrate while the distortion $d_t(B_t)$ is less than a constraint c

$$\min_{\mathcal{B}_t} \quad \sum_{i,j} b_{i,j,t} \quad \text{s.t.} \quad d_t(\mathcal{B}_t) \leq c.$$

which can be converted into: (by KKT conditions)

$$\frac{\partial d_t}{\partial b^*_{i,j,t}} = -\frac{1}{\mu} \quad \text{and} \quad d_t(\mathcal{B}^*_t) = c,$$

The marginal decrease in distortion per additional bit should be equal across all data sources!

ZipFM Design (Cont.)

Compression set: $u_{i,j,t} \in \{1, \ldots, U\}$

Feedback Control

$$u_{i,j,t+1} = \begin{cases} u_{i,j,t} - 1 & \text{if } \frac{\partial d_t}{\partial b_{i,j,t}} + \frac{1}{\mu} \leq -h, \quad u_{i,j,t} > 1 \\ u_{i,j,t} + 1 & \text{if } \frac{\partial d_t}{\partial b_{i,j,t}} + \frac{1}{\mu} \geq h, \quad u_{i,j,t} < U \\ u_{i,j,t} & \text{otherwise} \end{cases}$$

ZipFM Design (Cont.)

Feedback Control

If the slope is smaller than the target slop, it means one extra bit reduces distortion more than target. Valuable stream!

$$u_{i,j,t+1} = \begin{cases} u_{i,j,t} - 1 & \text{if } \frac{\partial d_t}{\partial b_{i,j,t}} + \frac{1}{\mu} \leq -h, \quad u_{i,j,t} > 1 \\ u_{i,j,t} + 1 & \text{if } \frac{\partial d_t}{\partial b_{i,j,t}} + \frac{1}{\mu} \geq h, \quad u_{i,j,t} < U \\ u_{i,j,t} & \text{otherwise} \end{cases}$$

ZipFM Design (Cont.)

Feedback Control

$$u_{i,j,t+1} = \begin{cases} u_{i,j,t} - 1 & \text{if } \frac{\partial d_t}{\partial b_{i,j,t}} + \frac{1}{\mu} \leq -h, \quad u_{i,j,t} > 1 \\ u_{i,j,t} + 1 & \text{if } \frac{\partial d_t}{\partial b_{i,j,t}} + \frac{1}{\mu} \geq h, \quad u_{i,j,t} < U \\ u_{i,j,t} & \text{otherwise} \end{cases}$$

One extra bit reduces distortion less than target. This stream is not very useful

ZipFM Design (Cont.)

Feedback Control

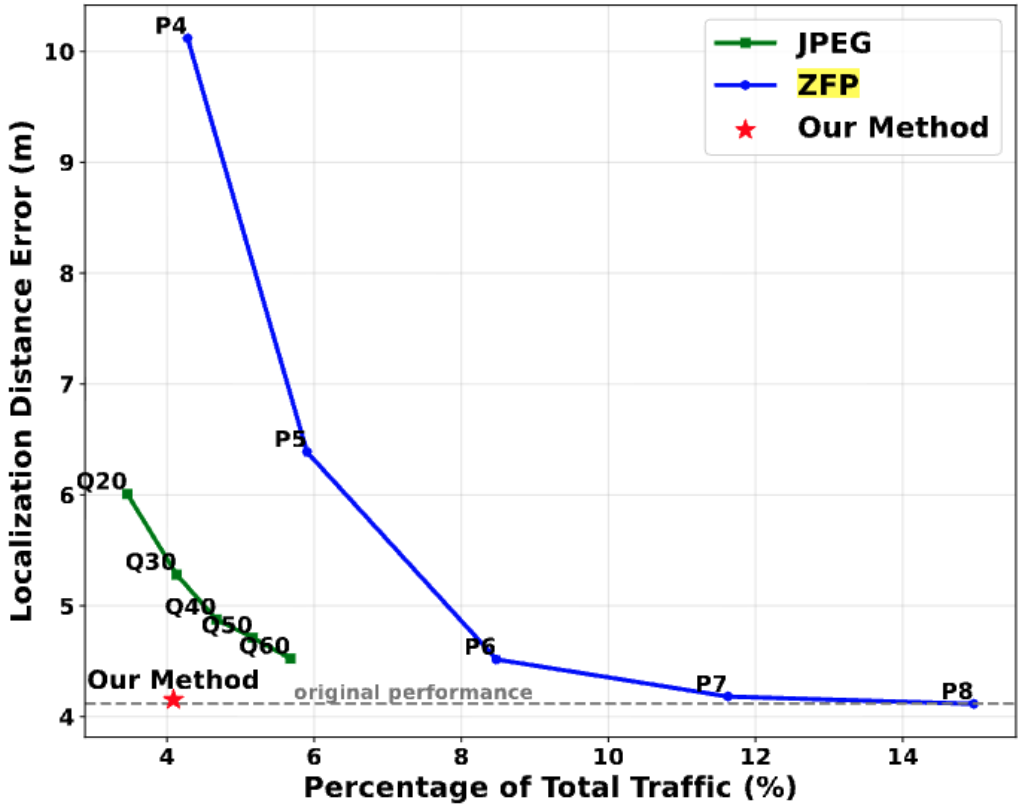
$$u_{i,j,t+1} = \begin{cases} u_{i,j,t} - 1 & \text{if } \frac{\partial d_t}{\partial b_{i,j,t}} + \frac{1}{\mu} \leq -h, \quad u_{i,j,t} > 1 \\ u_{i,j,t} + 1 & \text{if } \frac{\partial d_t}{\partial b_{i,j,t}} + \frac{1}{\mu} \geq h, \quad u_{i,j,t} < U \\ u_{i,j,t} & \text{otherwise} \end{cases}$$

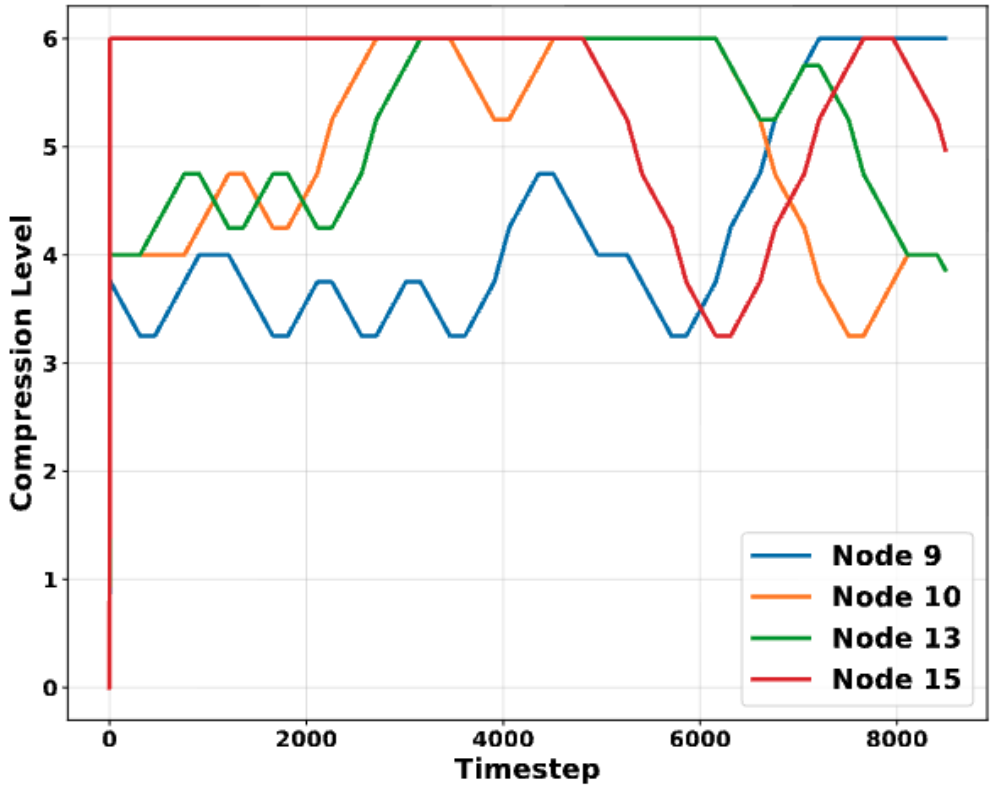
Estimating the Derivative in Practice

Periodically send two versions of the data using compression levels $u_{i,j,t}$ and $u_{i,j,t} + 1$.

Evaluation

Compare with uniform compression algorithms: JPEG, WebP, and ZFP





Limitations

- Assumes Centralized Backbone.
- Assume a shared total communication budget.
  - Per-node bitrate caps? (WiFi v.s. LTE, Power Consumption)
- The Convexity Assumption Is Not Guaranteed.
- Discrete Feedback May Oscillate.

Agenda

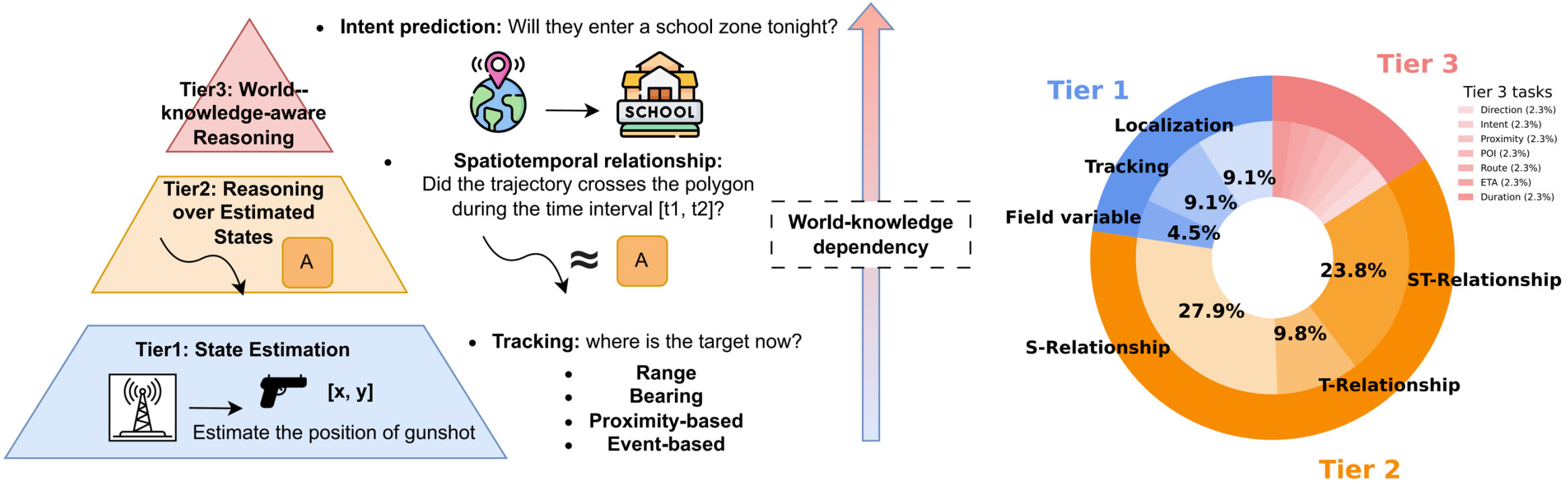
1. Supervised Spatiotemporal Representation Learning (Inter-Sensor)

2. Self-supervised/Unsupervised Learning for Spatial Representation (City-Scale)

3. Self-supervised Distributed Sensing (Placement-Aware & Network Efficiency)

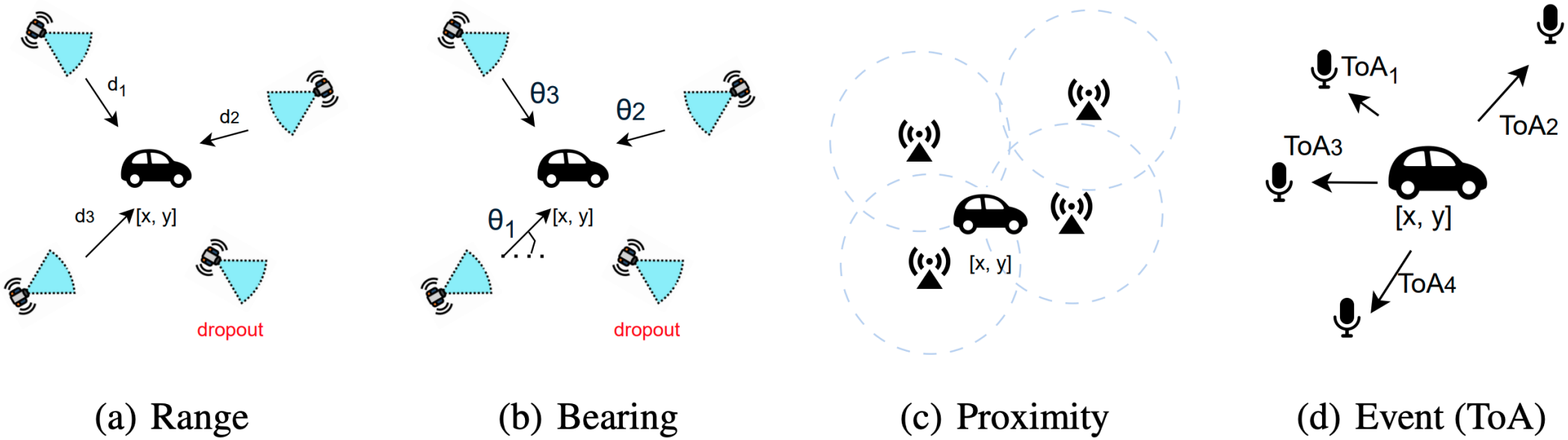
4. Zero-shot Reasoning on Representation

Can LLMs reason about spatio-temporal data?



Perceive -> Relate -> Decide

Benchmark built from sensing modalities



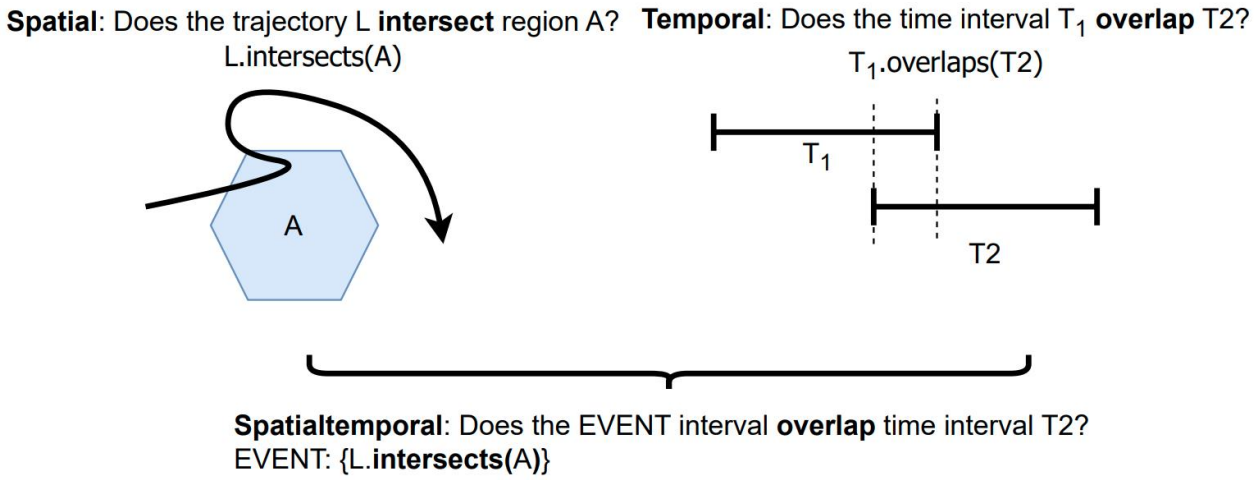
(a) Range (b) Bearing (c) Proximity (d) Event (ToA)

Figure 2: State estimation. Four types of sensor modalities for localization and tracking.

Sensors define constraints

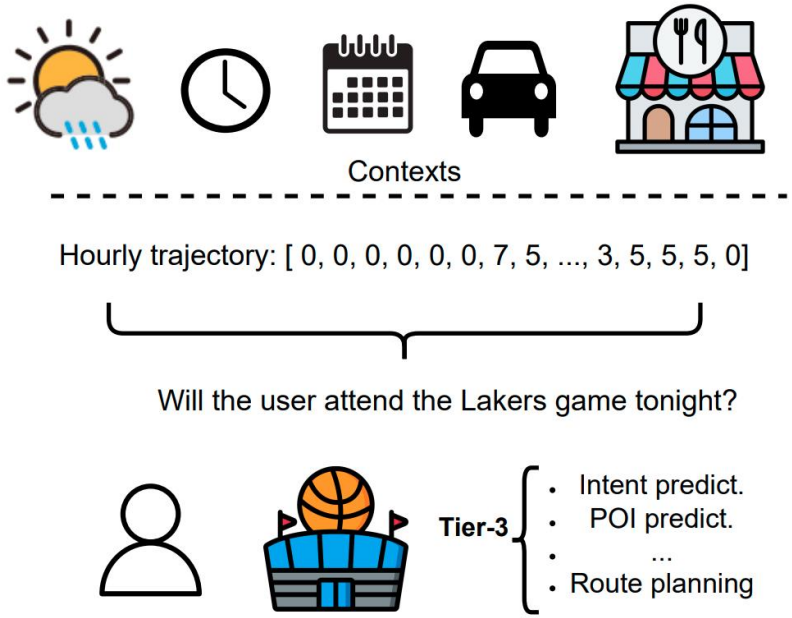
From geometry to logic to world knowledge

Tier2: *logic*



(a) Reasoning over states

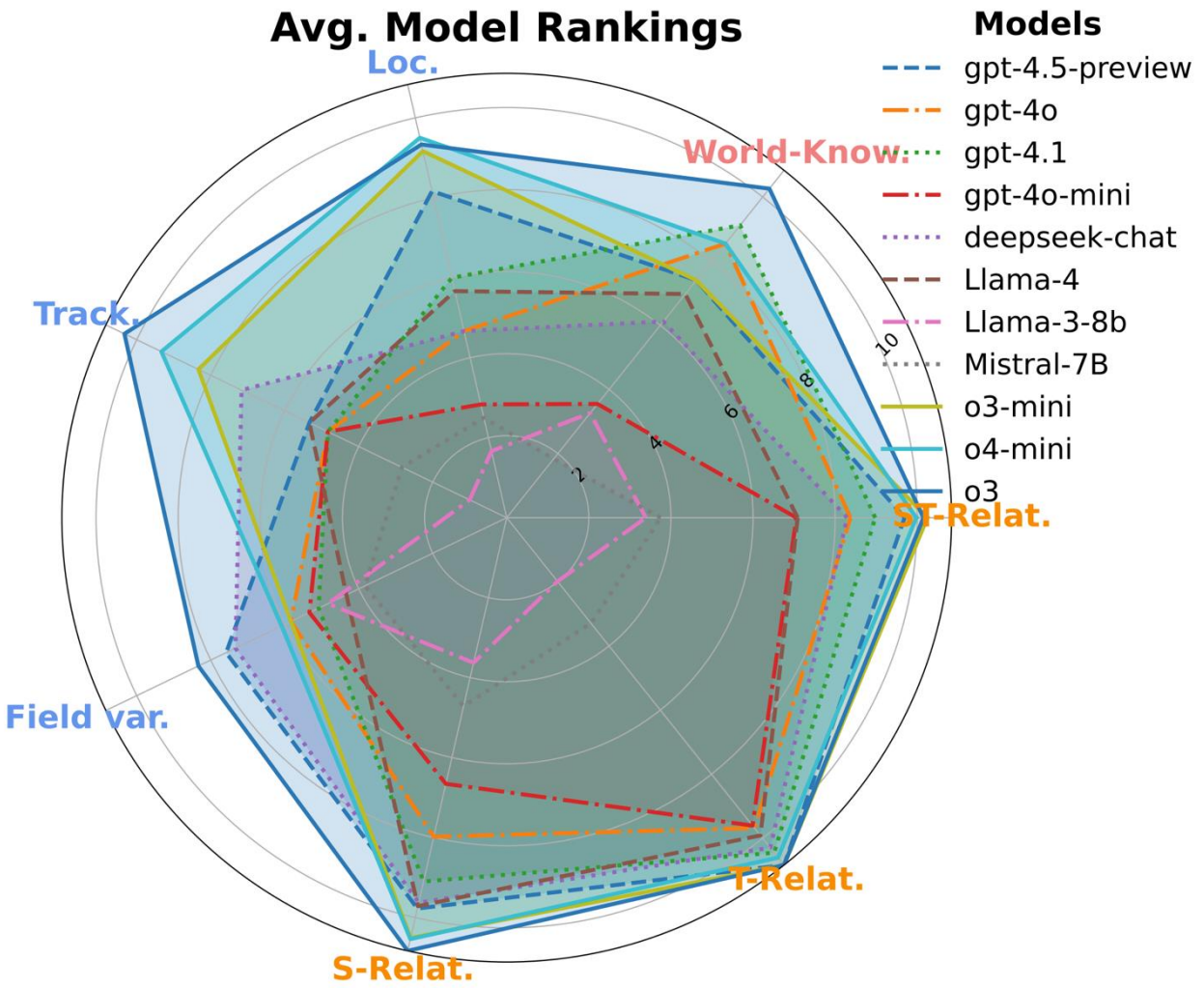
Tier3: *context*



(b) Knowledge-aware reasoning

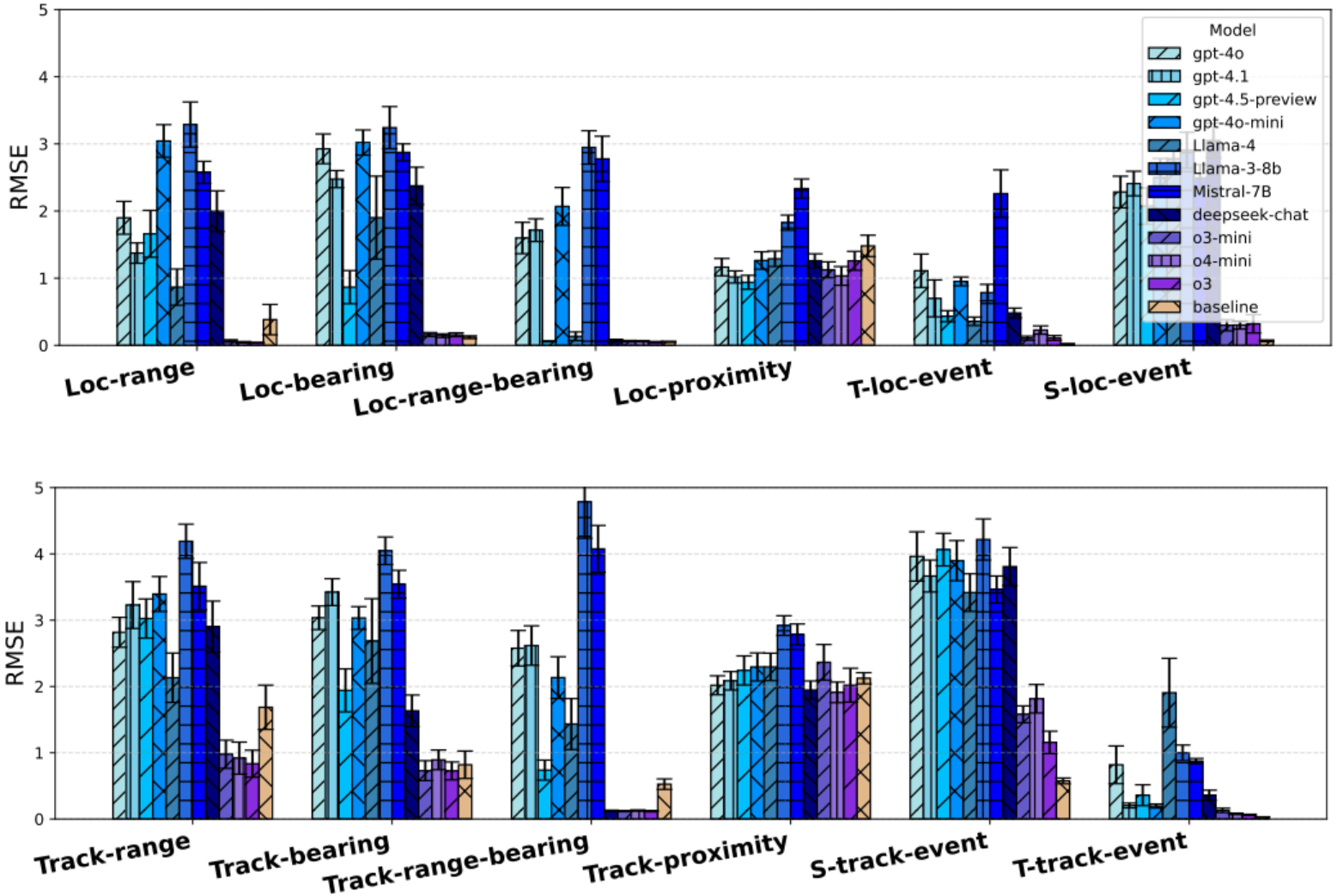
LRMs beat LLMs on geometry-heavy tiers

Key Idea***: Geometry is the wall…***



Compare to first-principles baselines

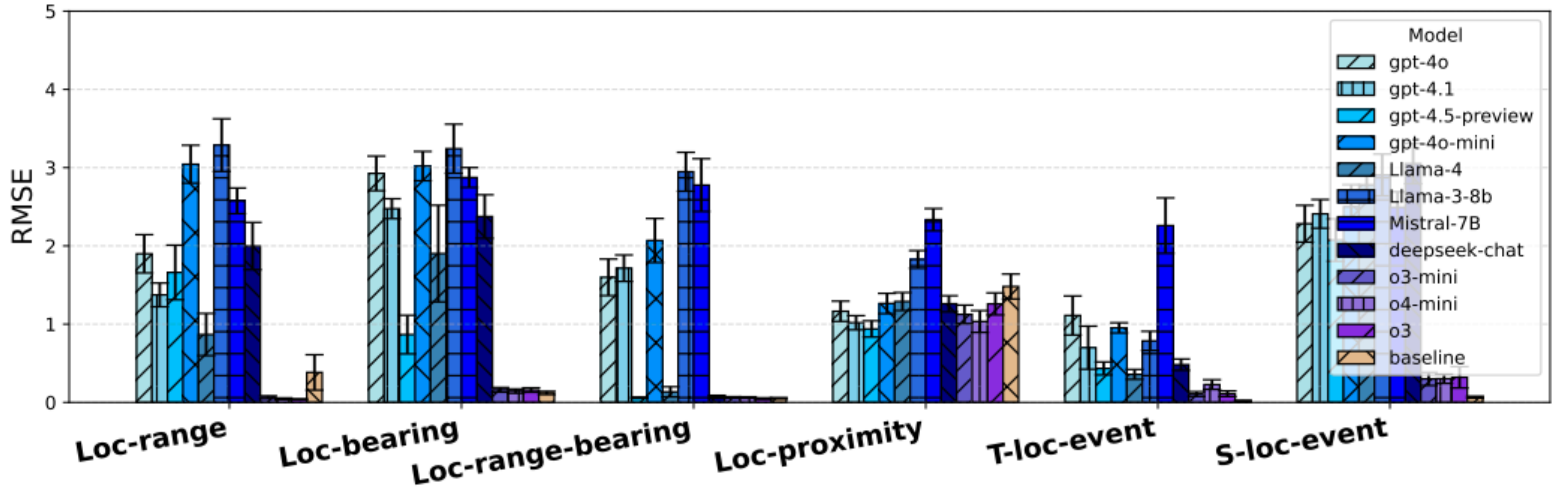
Key Idea*:* Baselines still win. Why?

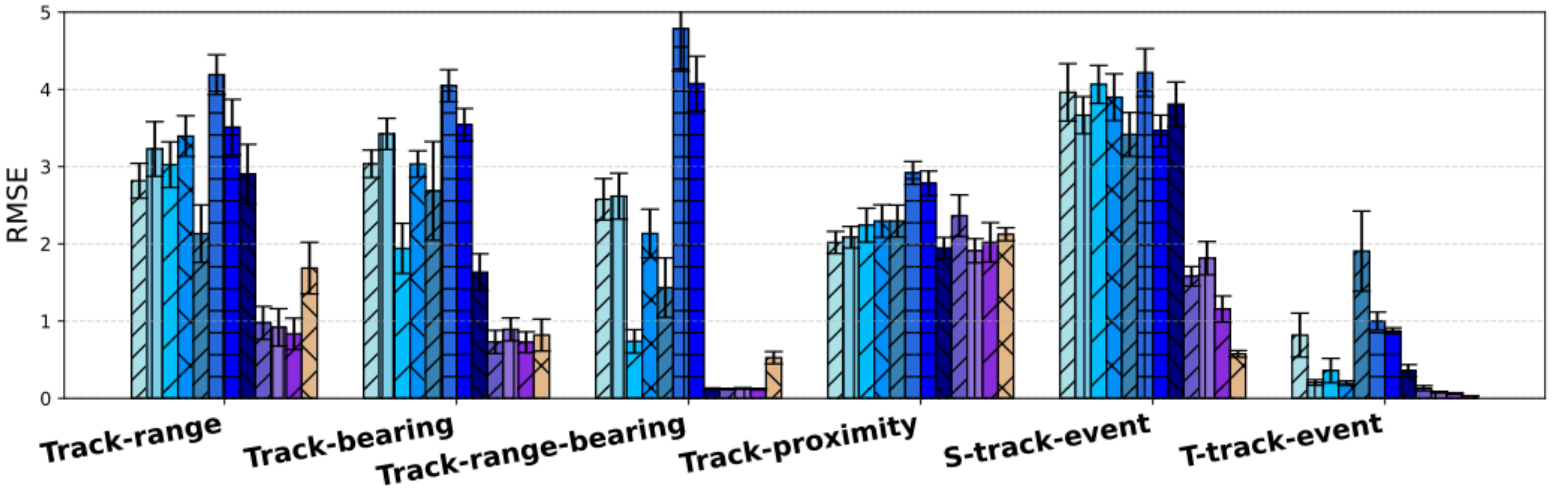


Compare to first-principles baselines

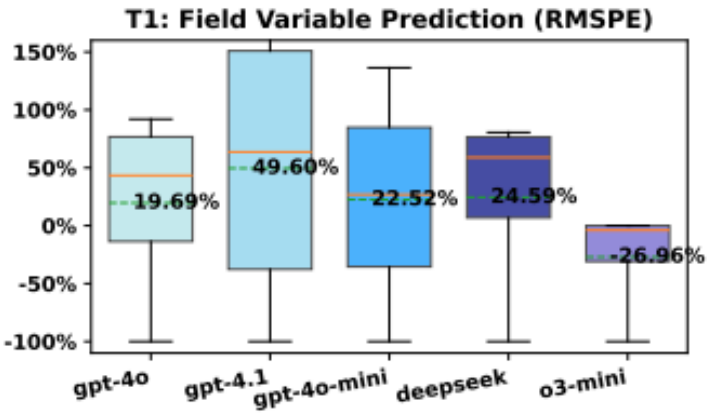
Key Idea*:* Baselines still win. Why?

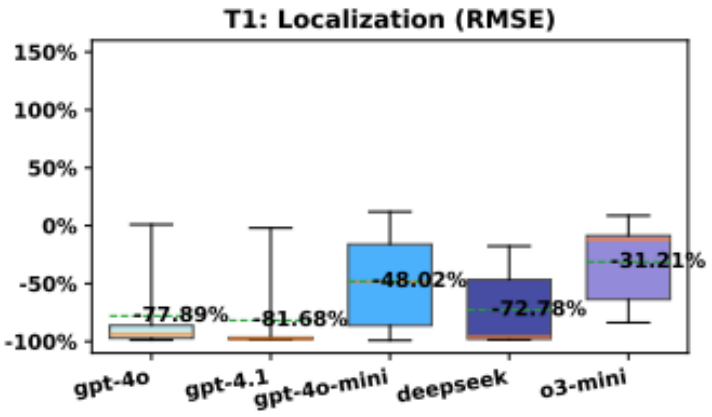
Reasoning-optimization

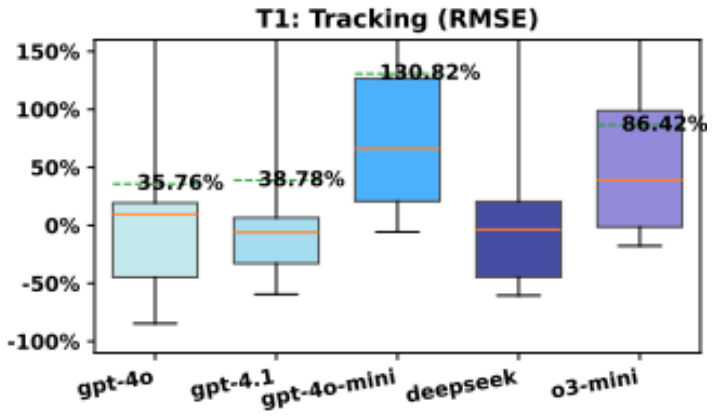




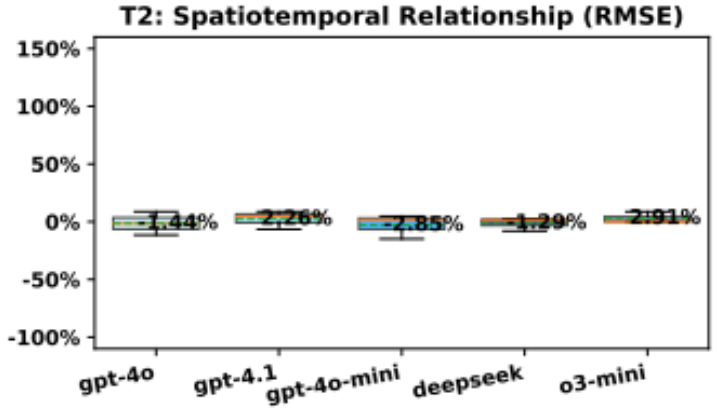
Tool use helps… and can hurt

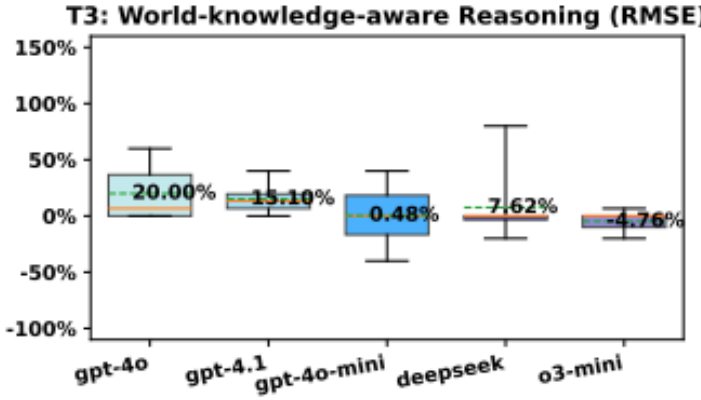




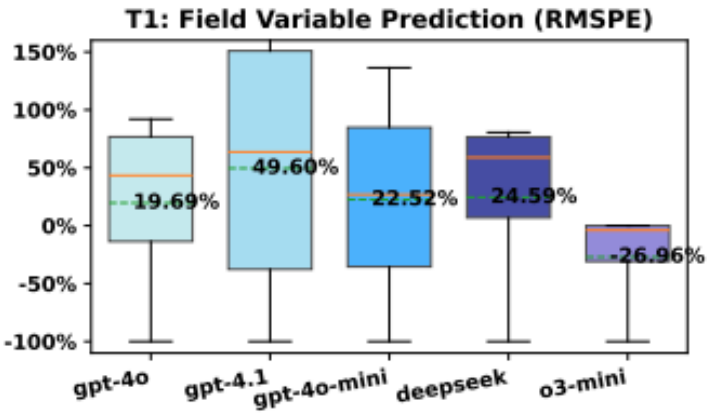


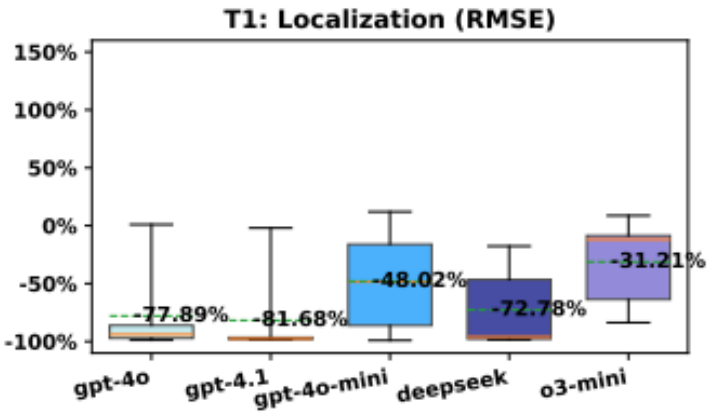
Tools ≠ solved

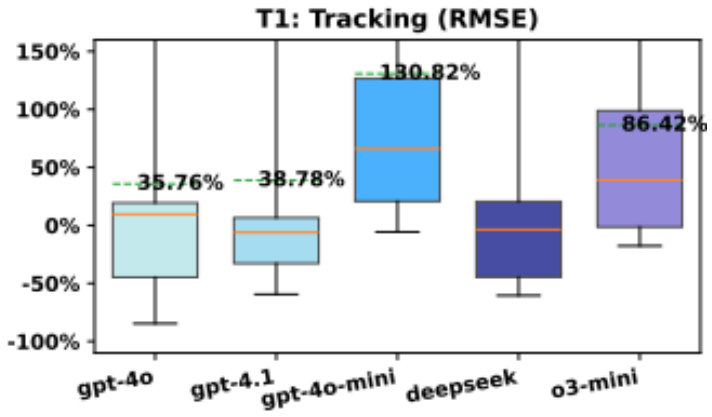




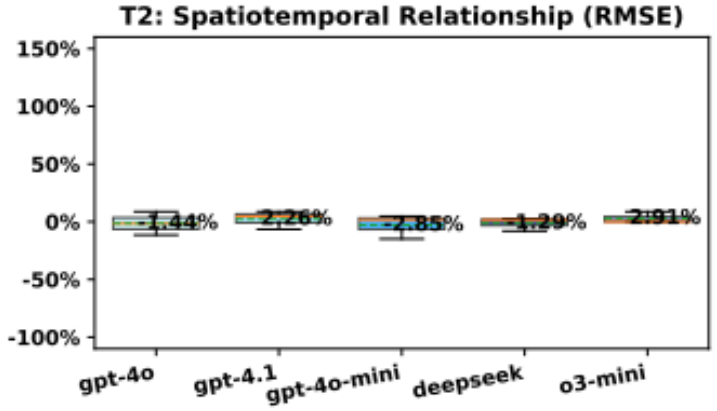
Tool use helps… and can hurt

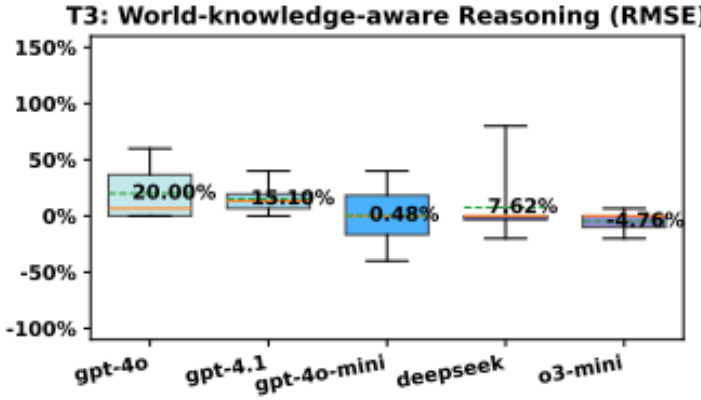






Needs Guardrails





References

[1] Yinghui Zhang, Hu An, Yaxuan Xing, Yang Liu, and Tiankui Zhang. "Learning temporal and spatial features jointly: A unified framework for space-time data prediction in industrial IoT networks." IEEE Sensors Journal 23, no. 16 (2023): 18752-18764.

[2] Liu, Jing, et al. "Distributional and spatial-temporal robust representation learning for transportation activity recognition." Pattern Recognition (2023).

[3] Porter Jenkins, Ahmad Farag, Suhang Wang, and Zhenhui Li. "Unsupervised representation learning of spatial data via multimodal embedding." In Proceedings of the 28th ACM international conference on information and knowledge management, pp. 1993-2002. 2019.

[4] Yizhuo Chen, Tianchen Wang, You Lyu, Yanlan Hu, Jinyang Li, Tomoyoshi Kimura, Hongjue Zhao, Yigong Hu, Denizhan Kara, and Tarek Abdelzaher. "Spar: Self-supervised placement-aware representation learning for multi-node iot systems." arXiv e-prints (2025): arXiv-2505.

[5] Wang, Tianchen, Yizhuo Chen, Hongjue Zhao, You Lyu, Jinyang Li, Tomoyoshi Kimura, Yigong Hu et al. "On Network-Efficient Multimodal Multi-Vantage Foundation Models for Distributed Sensing." In 2025 IEEE 22nd International Conference on Mobile Ad-Hoc and Smart Systems (MASS), pp. 19-27. IEEE, 2025.

[6] Pengrui Quan, Brian Wang, Kang Yang, Liying Han, and Mani Srivastava. "Benchmarking Spatiotemporal Reasoning in LLMs and Reasoning Models: Capabilities and Challenges." arXiv preprint arXiv:2505.11618 (2025).

Thanks for Listening

Any Questions?