Massachusetts Institute of Technology

JTL Transit Lab
URBAN MOBILITY LAB AT MIT

Trustworthy Spatial Intelligence

Learning, Calibrating, and Reasoning Toward World Models of Cities

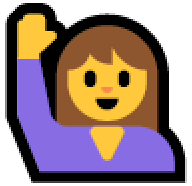
Dingyi Zhuang
Ph.D. Candidate @ MIT
Advisor: Prof. Jinhua Zhao

eMERGE seminar @ UC Berkeley
August 27, 2025

What's Up? 👋

Not just the greeting… but also benchmark datasets!

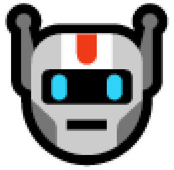
What's "Up"?



A dog

on

a table



A dog

right of

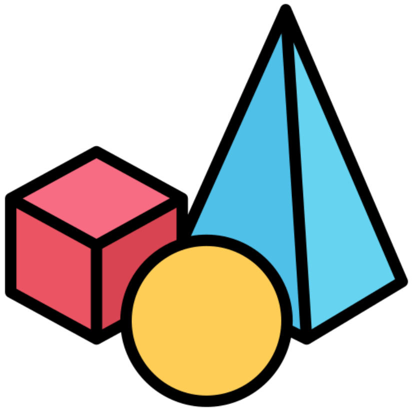
a table

Figures and contents adopted from paper "What's 'Up' with vision-language models? Investigating their struggle with spatial reasoning."

What does it really mean for a model to “understand space”?

What is Spatial Intelligence?

What things are

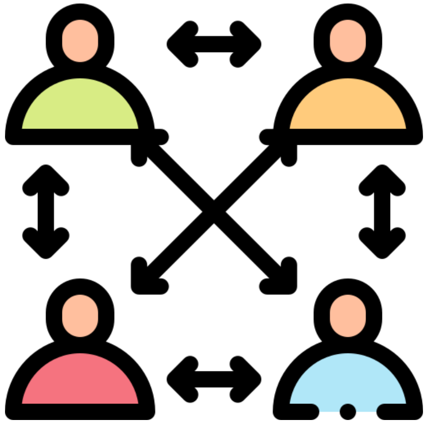


Object recognition

Where they are

Spatial localization

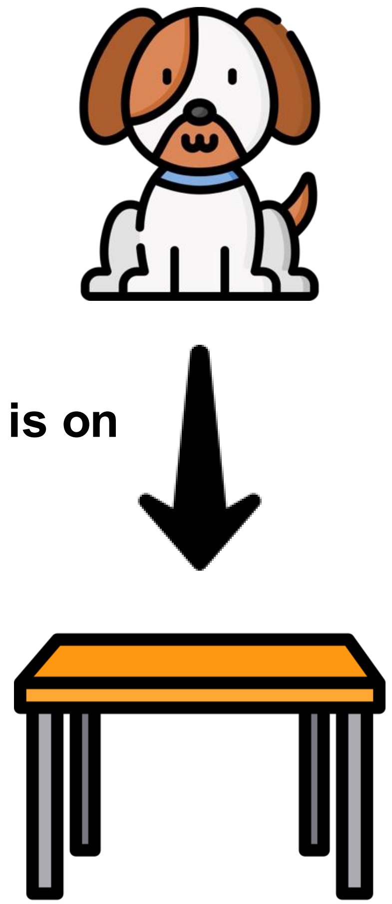
How they relate



Relationships & interactions

Adapted from "Frames of Mind: The Theory of Multiple Intelligence" and "How to Grow a Mind: Statistics, Structure, and Abstraction"

Relationships Are the Foundations of Spatial Intelligence



Objects in isolation = limited meanings

Locations gain meaning through connections

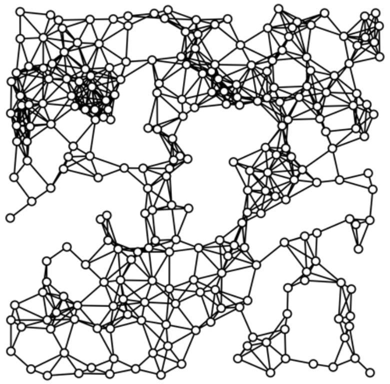
Relationship captures how entities influence or depend on one another

At its core, spatial intelligence is really about how well we can learn and represent relationships

Relationships in Actions

Macro & Micro Perspectives

Geographic Information System



Urban value emerges from proximity and connectivity

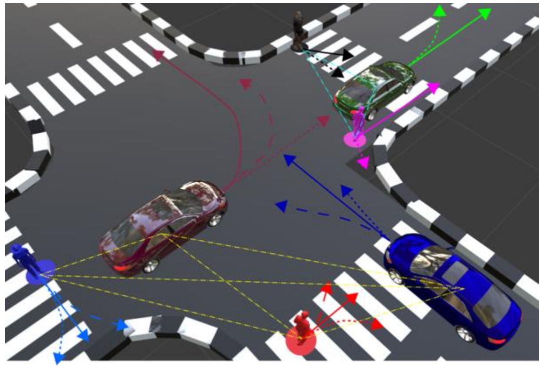
(AAAI 2020)[1]

Transportation Policies

Policy relies on relationships between central and peripheral zones.

(On going work)

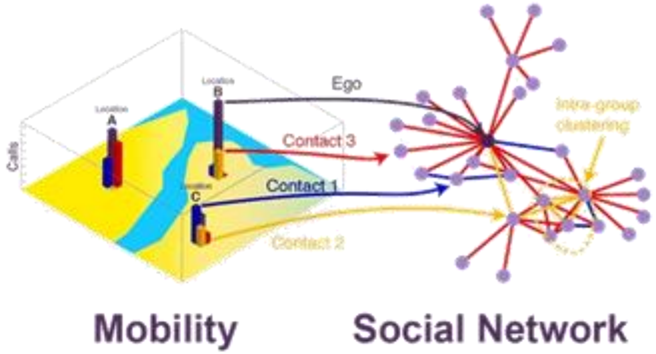
Autonomous Driving



Driving depends on relationships between vehicles, lanes, and pedestrians.

(On going work)

Social & Mobility Networks



Mobility emerges from spatial relationships linking people, places, and opportunities

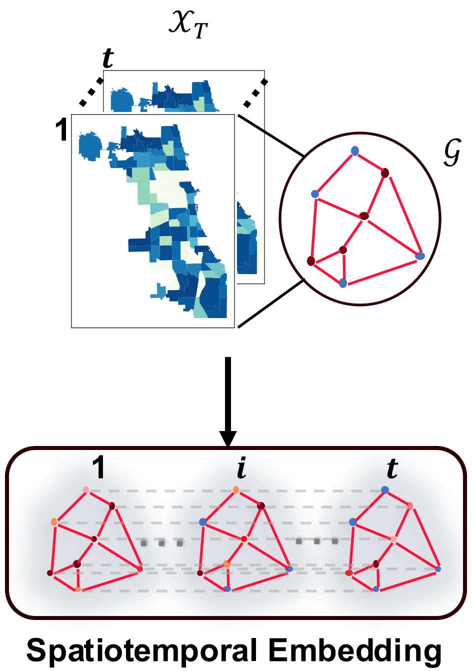
(On going work)

Figures adapted from "ScePT: Scene-consistent, Policy-based Trajectory Predictions for Planning" and "Social Circles"

Research Agenda

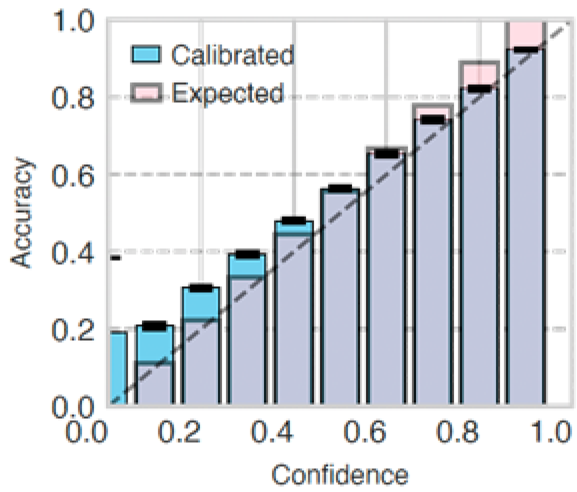
Learning, Calibrating, Reasoning about Relationships

Learning



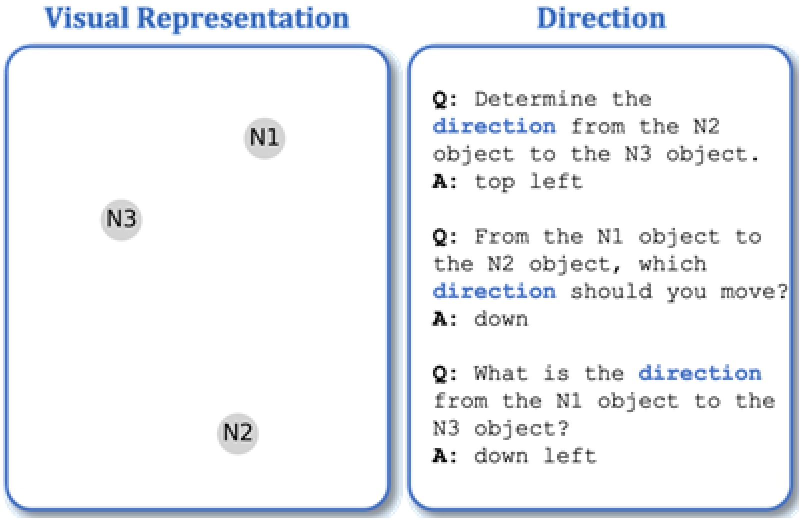
Leaning patterns in spatiotemporal data

Calibrating



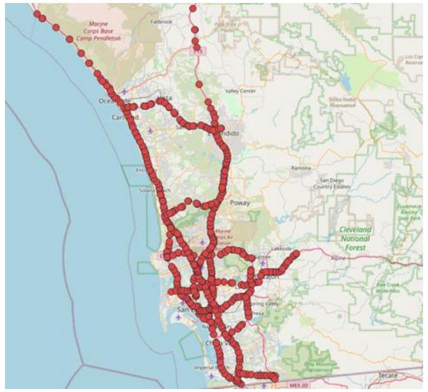
Calibrating model confidences and uncertainty

Reasoning



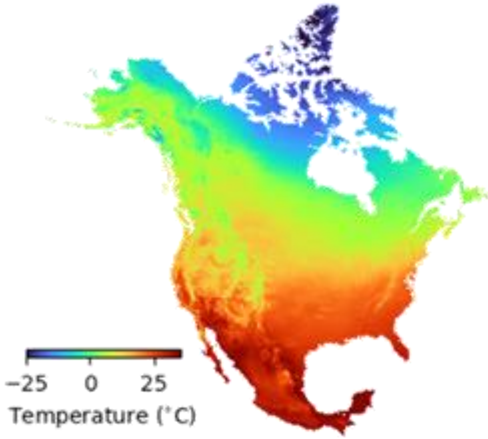
From pattern learning to reasoning with relationships

The Central Question: Learning Hidden Relationships in Cities

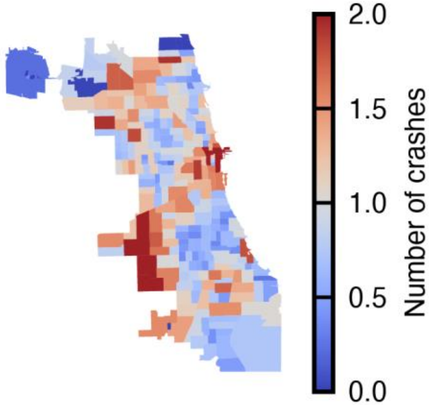


Highway Speed (San Diego)
(TRB 2025)[2]

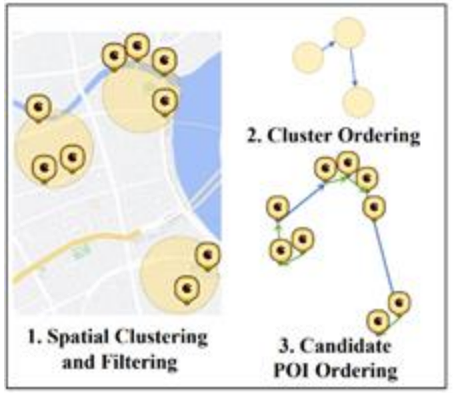
Transit Demand (Singapore)
(TR-C)[3]



Temperature (NA)



Traffic Crash (Chicago)
(SIGSPATIAL 2024)[4]



Point of Interests (Shanghai)
(EMNLP 2024)[5]

Challenges:

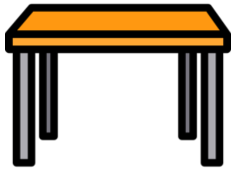
- Sparsity (high-resolution and missing information)
- High-dimensionality (city scale)
- Multi-modality (various data structures)

Figures adapted from Xinyu Chen (Job Talk, 2024)

Graphs as Lenses for Learning Hidden Patterns

A graph $\mathcal{G}$ is the combination of

Nodes $\mathcal{V}$
(entities)



is on

Edges $\mathcal{E}$
(relationships)

Adjacency Matrix $\boldsymbol{A}$ $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$

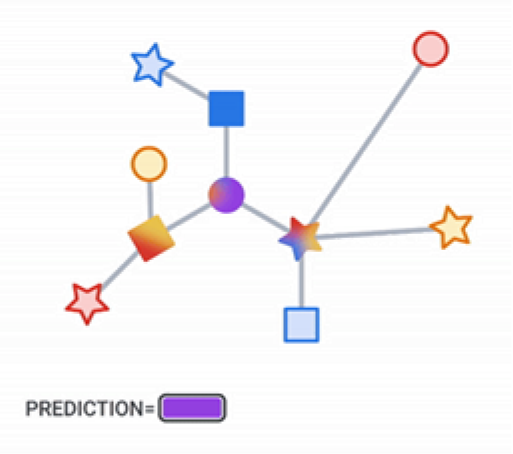
Road network

Scene graph

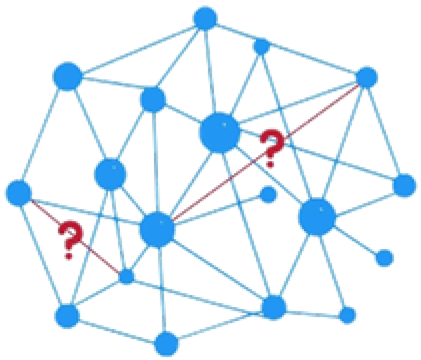
Transit network

- *Origin-destination (OD) networks*
- *Social networks*
- *Land-use interaction graphs*
- *Accessibility graphs*
- *Urban knowledge graphs*
- *Distance-based graphs*
- *…*

Graph Neural Networks (GNNs): Learning from Structure



- A GNN learns from graphs via **message passing**
- Each node **aggregates** neighbor info and **updates** its state
- Intuition: *“I adjust my choice based on what neighbors tell me”*



- Link prediction: the simplest way GNNs model relationships — decide if an edge should exist between two nodes.
- How: Based on updated node information
  - **Edge exists**: If two nodes become similar after exchanging neighbor info
  - **No edge:** If they remain very different

Figures adapted from the internet

Using GNNs to Learn Travel Demand from Relationships

Uncertainty Quantification of Sparse Travel Demand Prediction with Spatial-temporal GNNs

Dingyi Zhuang

Shenhao Wang

Haris N. Koutsopoulos

Jinhua Zhao

ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2022

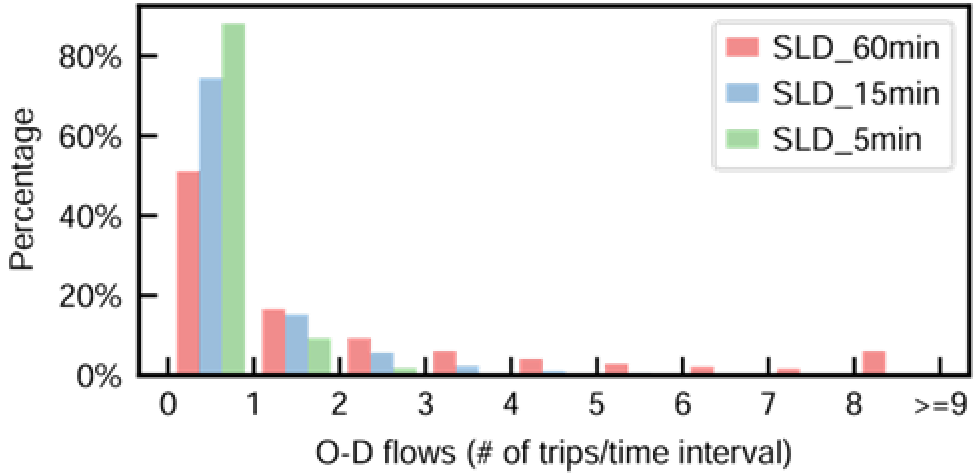
Oral Presentation, <10%

Sparsity in Spatiotemporal Transportation Data

For-Hire Vehicles in NYC

High-resolution OD demand are highly sparse, with many zero entries

Distribution of travel demand

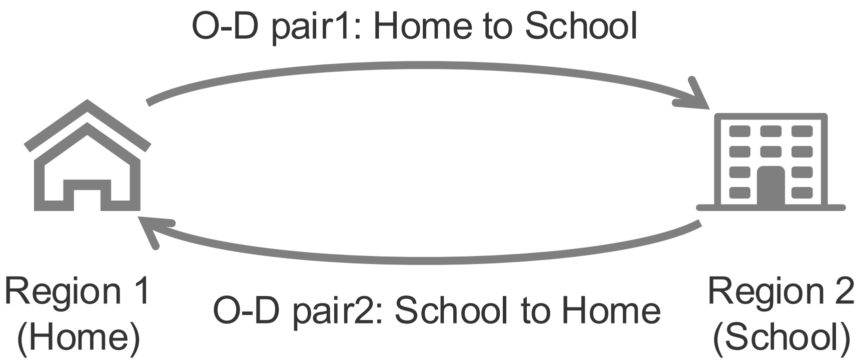


67 pick-up/drop-off zones, 67 × 67 OD pairs in total

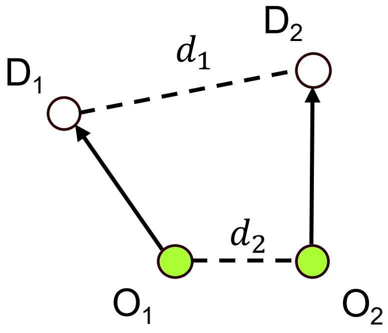
Sparsity is ubiquitous if scaling up spatial/temporal resolutions

Graph Representation of OD Demand

Nodes $\mathcal{V}$



Edges $\mathcal{E}$
(geographic proximity)



Adjacency
(distance-based)

$$\boldsymbol{A}_{1,2} = \sqrt{\frac{1}{2}(d_1^2 + d_2^2)}$$

OD pairs with nearby origins or destinations tend to have similar demand

Modeling Sparse OD Demand with Graph-Based Relationships

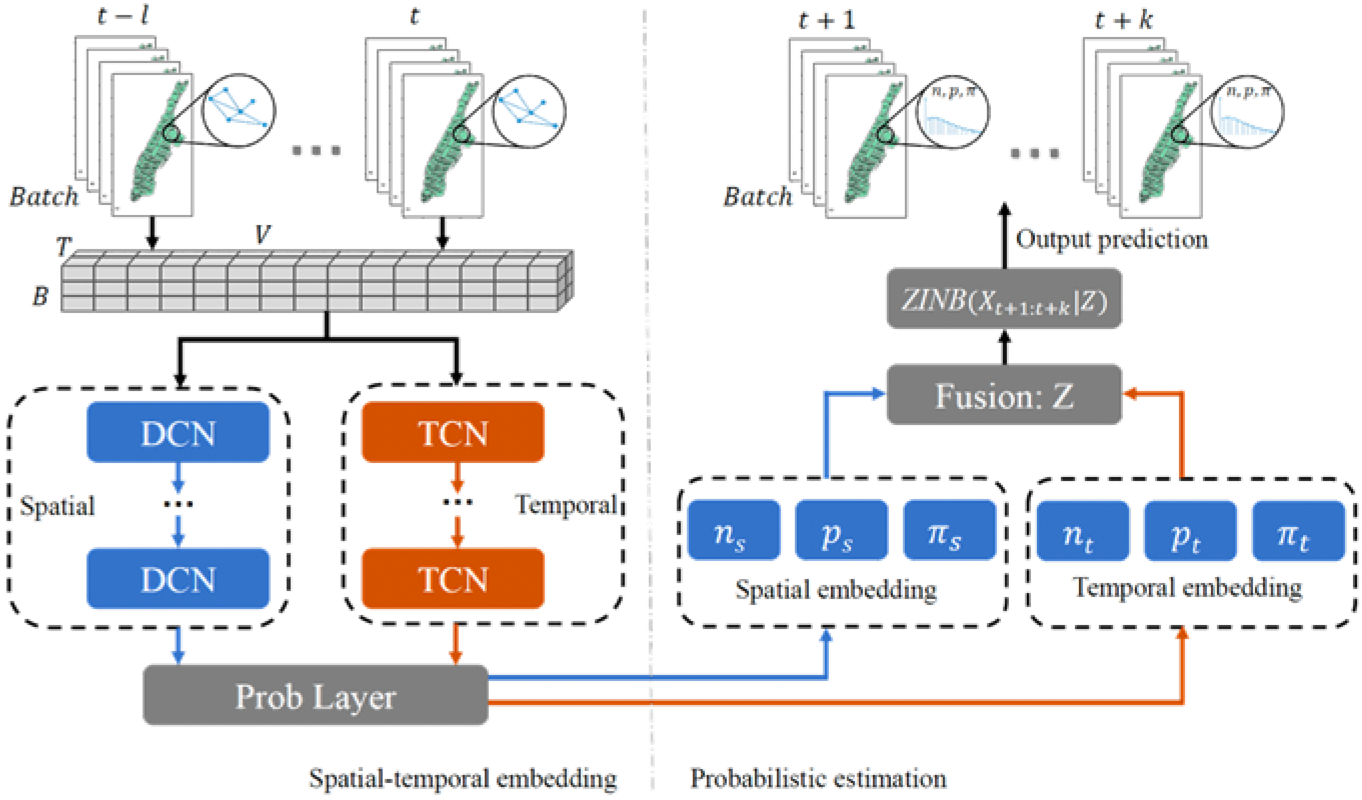
$$X \sim \pi\delta_0 + (1-\pi)NB(n,p)$$

Inputs:

OD graphs + demand time series

Outputs:

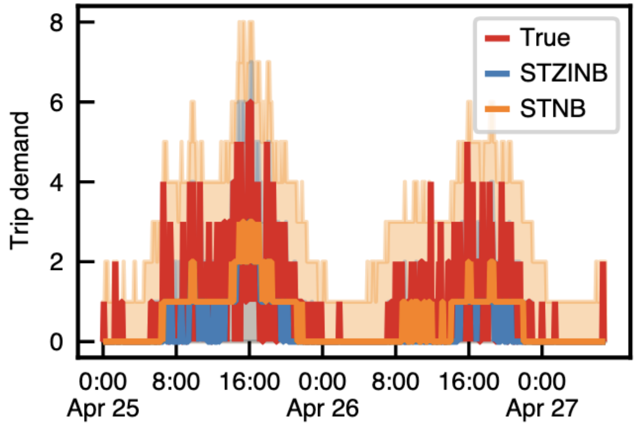
Probabilistic distribution of future demand



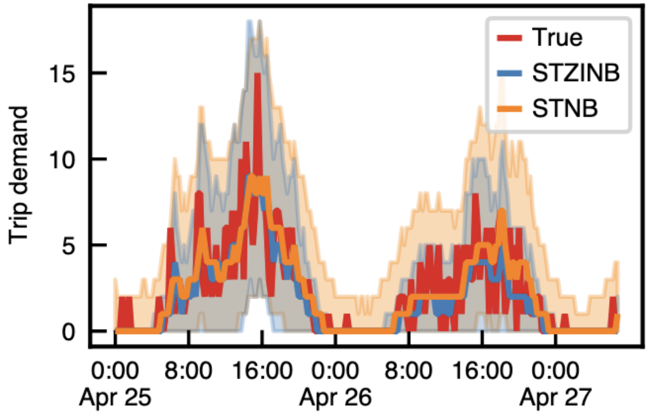
- Transform sparse OD demand into a probabilistic distribution
- Zero-inflated modeling handles excess zeros
- GNN capture spatial relationships among OD pairs

δ_0: Dirac delta distribution at zero (i.e. point mass at 0)
NB: Negative binomial distribution

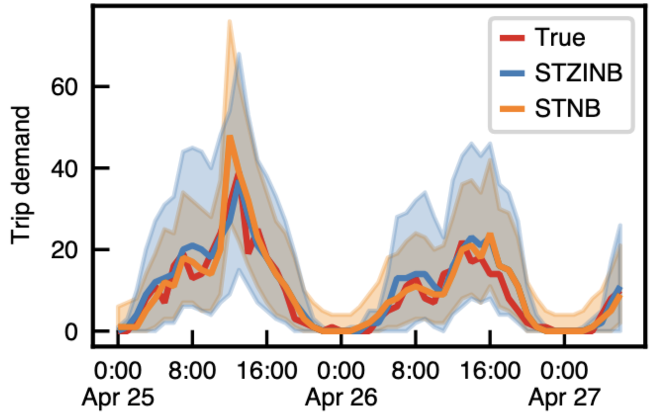
Results: Performance under Sparse Data



(a) SLD_5min case



(b) SLD_15min case



(c) SLD_60min case

- Overall **6%** accuracy gains compared to baselines
- Handling extreme sparse cases (**90%** data entries being zeroes)
- Efficient prediction intervals (≥ **55%** narrower than non–zero-inflated models)

https://github.com/ZhuangDingyi/STZINB

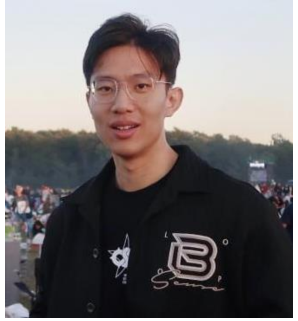
Improper Relationships in GNNs Can Propagate Untrustworthy and Inequitable Results

Mitigating Spatial Disparity in Urban Prediction Using Residual-Aware Spatiotemporal Graph Neural Networks: A Chicago Case Study


Dingyi Zhuang


Hanyong Xu


Yunhan Zheng


Xiaotong Guo


Shenhao Wang


Jinhua Zhao

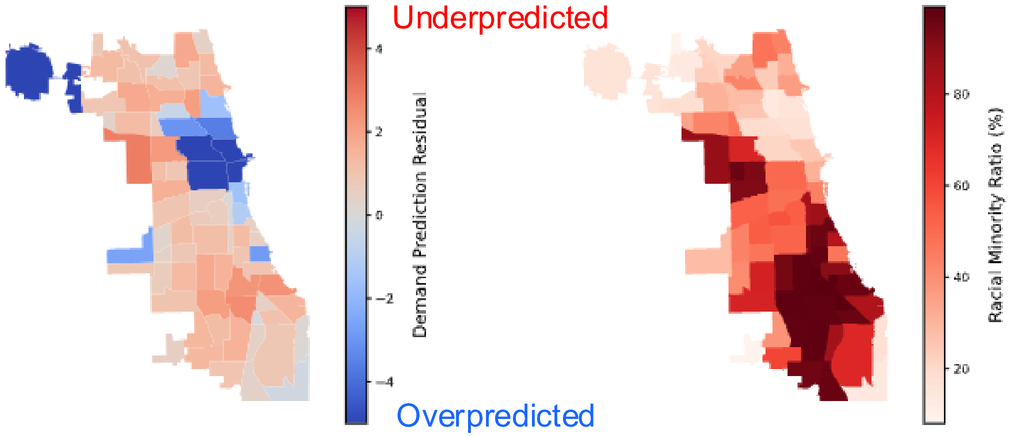
International World Wide Web Conference 2025

🏆 ***Best Paper Award** at WebST Workshop*

How Spatial Disparity Emerges

Transportation Network Companies (TNCs)

Pick-up demand from TNCs



STGCN **prediction residual** distribution in Chicago

Minority rate distribution in Chicago

Nodes $\mathcal{V}$: Region

Edges $\mathcal{E}$: Geographic proximity

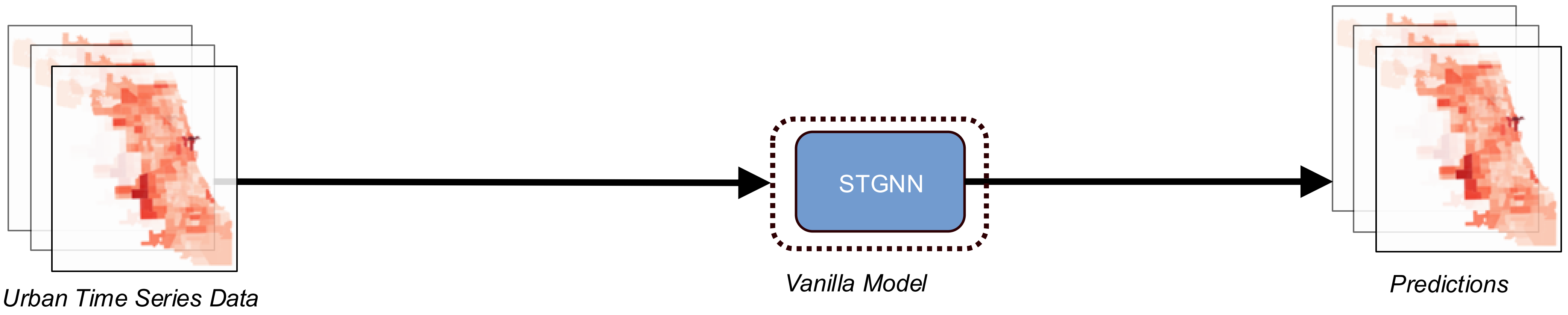
- Adjacency matrix propagates information from neighbors.

What do you observe?

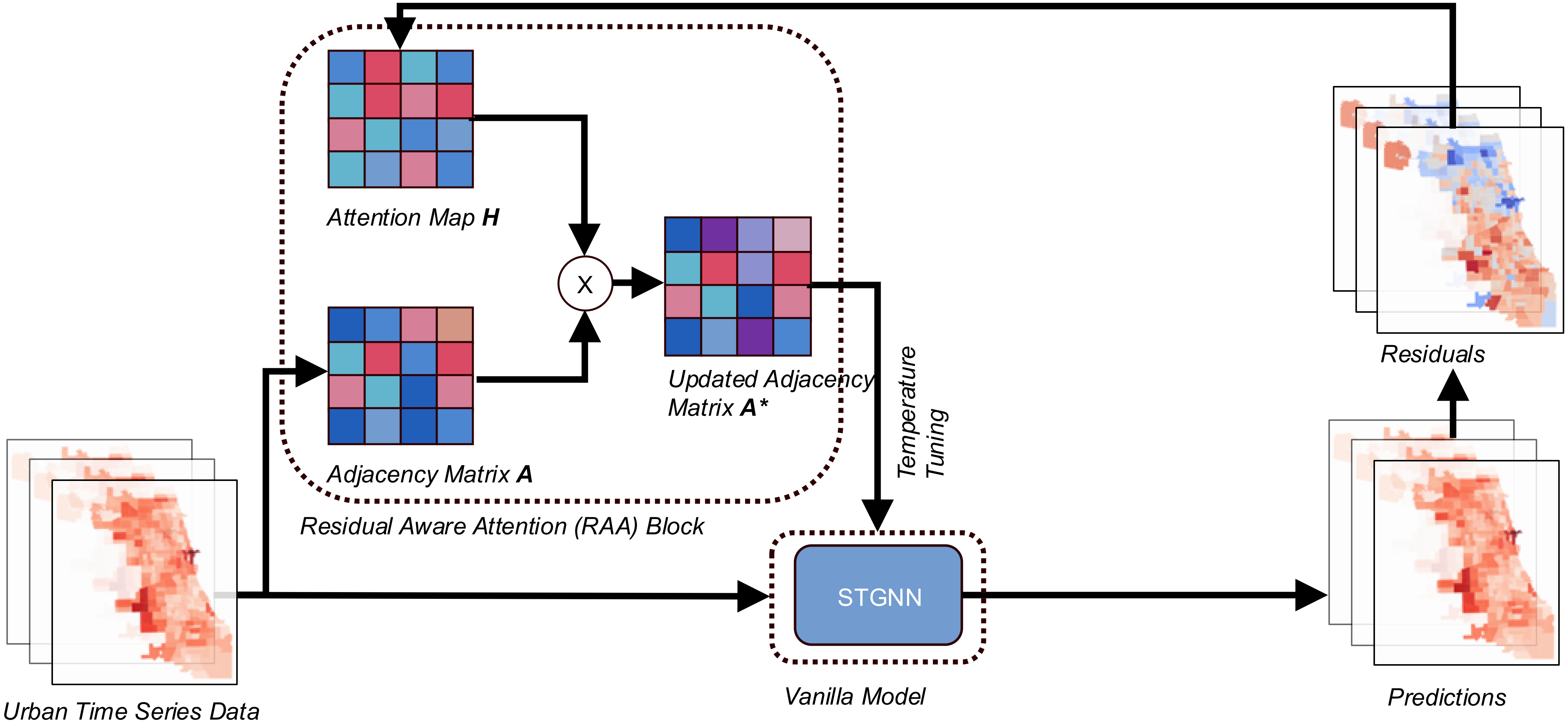
- If your neighbors are underpredicted, so are you. Leading to underserved demand.
- In transportation, this means entire regions can be systematically underpredicted.
- ***Accurate predictions do not necessarily lead to equitable outcomes***

Residual-Aware Attention: Rethinking Adjacency

Residual-aware block: adjusts adjacency weights using residual signs



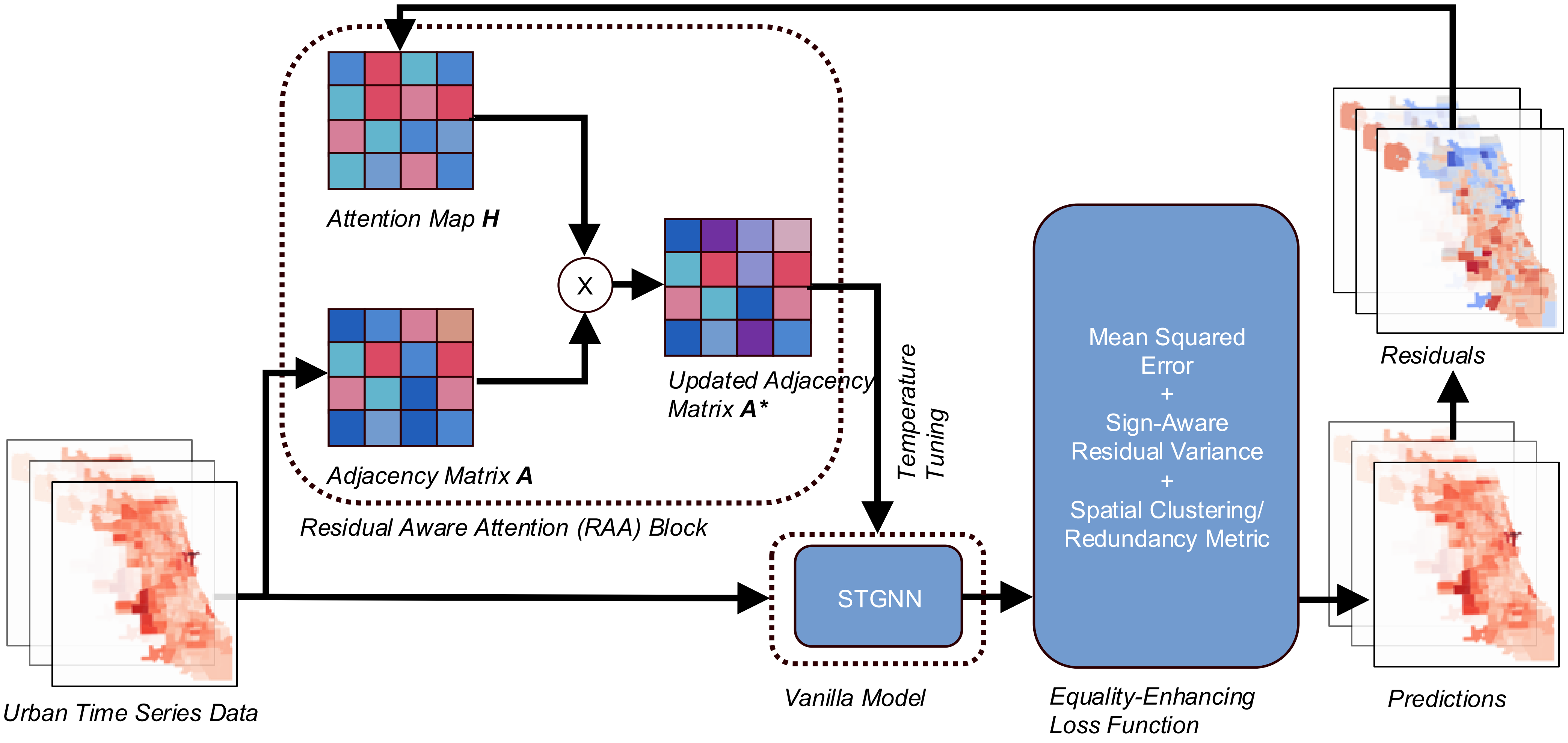
Residual-Aware Attention: Training



$$A_{adapted} = \boldsymbol{A} \odot \boldsymbol{H}$$

$$\boldsymbol{H} = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

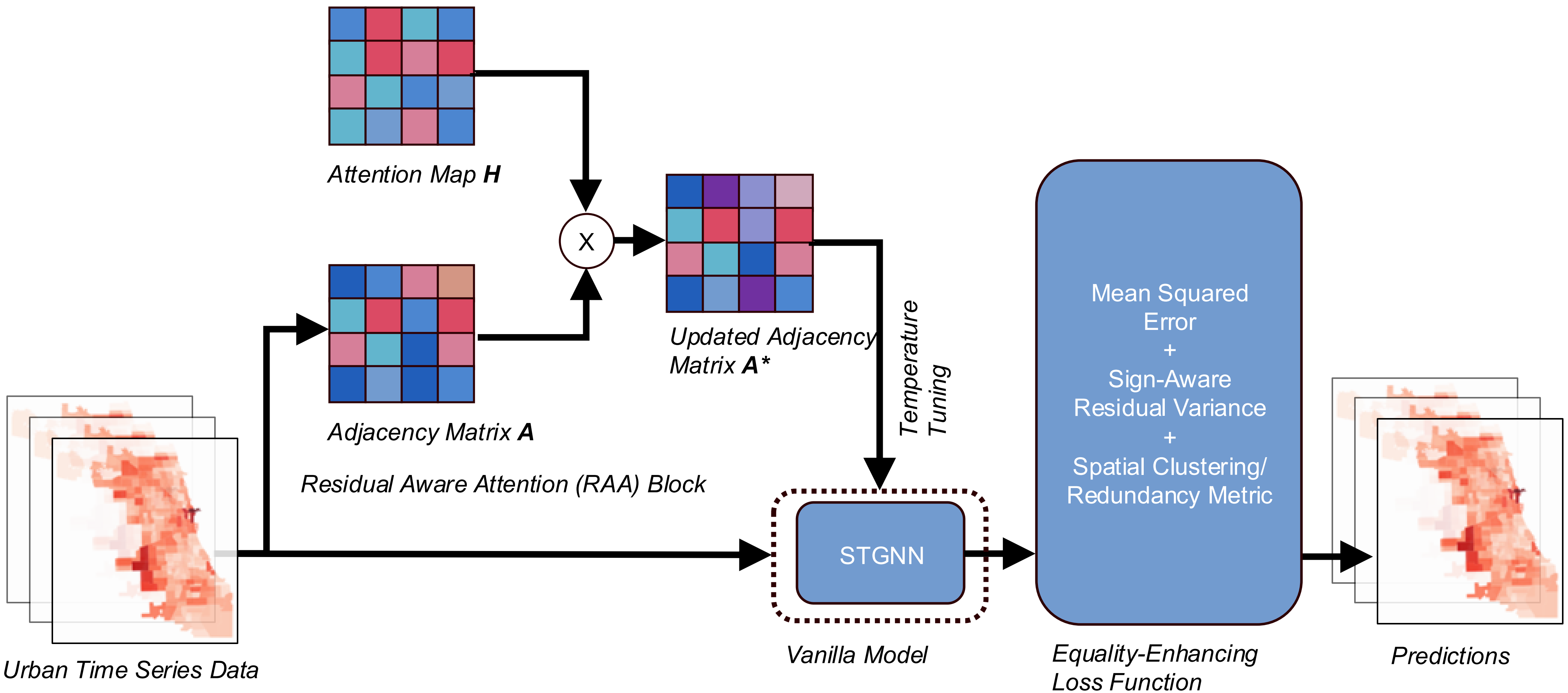
Residual-Aware Attention: Training



Include the spatial disparity D_s and fairness metrics D_f like Moran's I into loss function

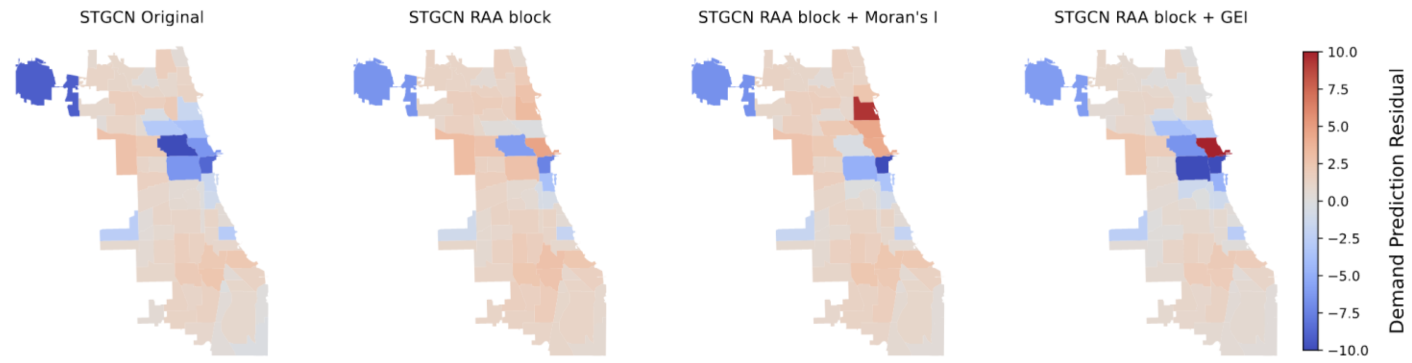
$$\mathcal{L}_{joint} = \mathcal{L}_{prediction} + \lambda_s D_s + \lambda_d D_f$$

Residual-Aware Attention: Inference

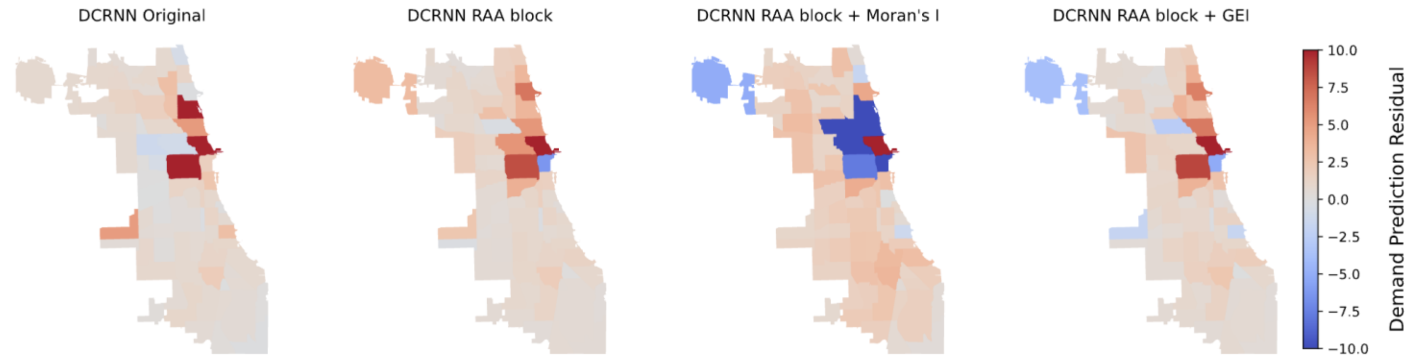


The attention map is fixed during model inference

Results: More Equitable with Minimal Trade-off



(a) STGCN Residual Distribution



(b) DCRNN Residual Distribution

Small accuracy loss

MAE **7%**

SMAPE **12%**

Fairness metrics improvements

GEI **18%**

Moran's I **80%**

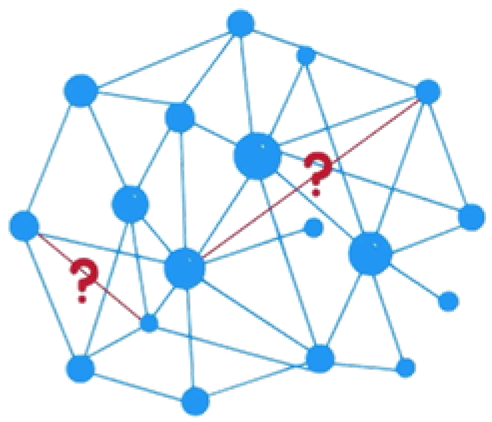
SDI **47%**

MAE: Mean Absolute Error
SMAPE: Symmetric Mean Absolute Percentage Error
GEI: Generalized Entropy Index, measures spatial disparity
Moran's I: measures spatial autocorrelation
SDI: Scaled Disparity Index, measures demographic disparity

From Learning to Trust

Can We Really Rely on These Models?

Graphs as a lens



Trust?

- Graphs are powerful representations.
- They learn hidden patterns.
- They could make accurate predictions.
- But… can we trust their predictions?

What is Calibration?

Calibration = aligning model confidence with reality

50% chance a pedestrian is on the crosswalk across 100 similar simulations/training

80% of the labels show the pedestrian is on the crosswalk, you are **underconfident**!

Calibration bridges the gaps

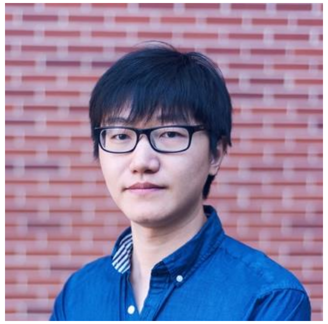
Calibrating Graph Link Predictions for Trustworthy Topology in Autonomous Driving

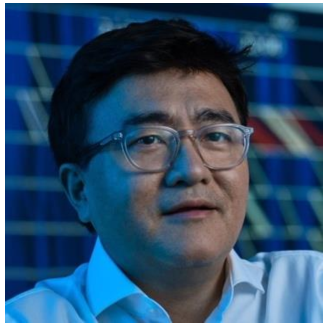
GNN-Based Topology Refinement for Map Generation


Dingyi Zhuang


Xiaoqi Wang


David Paz


Wenbin He

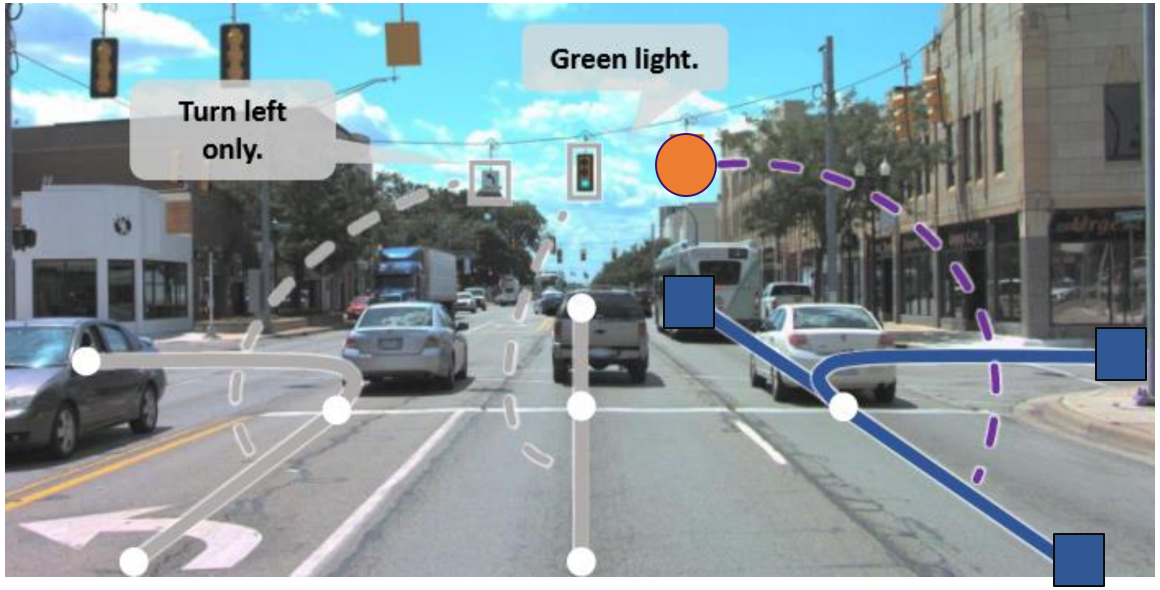

Liu Ren

Internship work at Bosch Center for Artificial Intelligence during Summer 2025
In submission to ICLR 2026

BOSCH

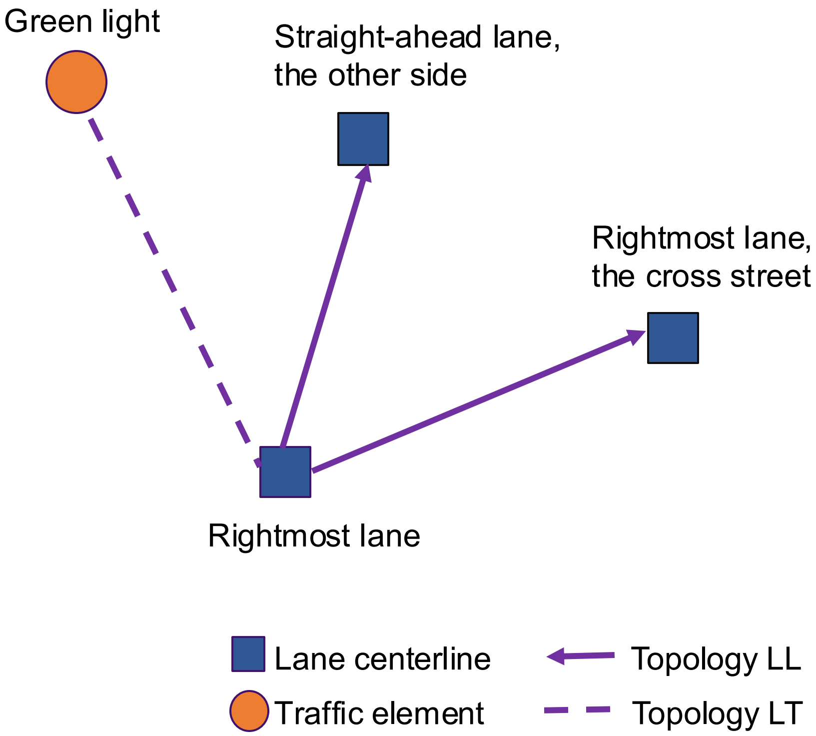
Driving Scene Graph

An autonomous vehicle navigating towards an intersection, this is what front-view camera sees.



Which lane to drive?

Which traffic signal to follow?



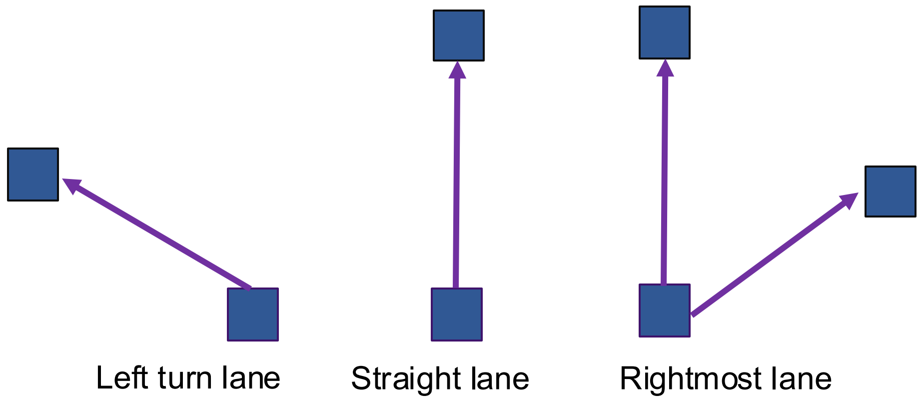
Figures adopted from: "Graph-based topology reasoning for driving scenes."

Driving Scene Topology Graph

How to ensure topology is reliable?

Lane topology graph

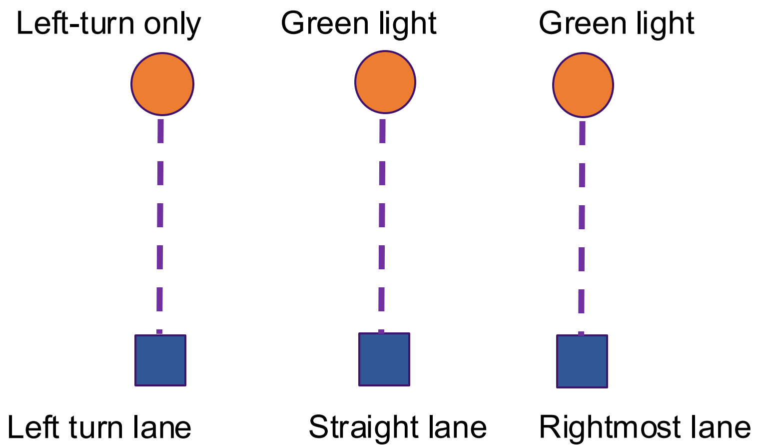
Consists of centerlines as well as their connectivity



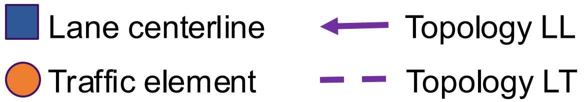
Full lane topology graph of the example

Lane-traffic element topology graph

Lane-to-traffic element assignment
(e.g., lights, signs, markers)



Full lane-traffic element topology graph of the example



HD Maps – Current Solutions

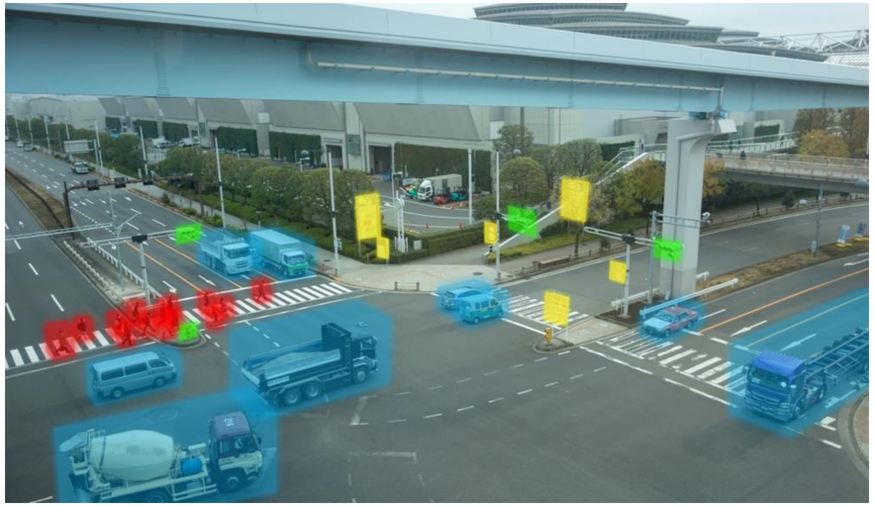
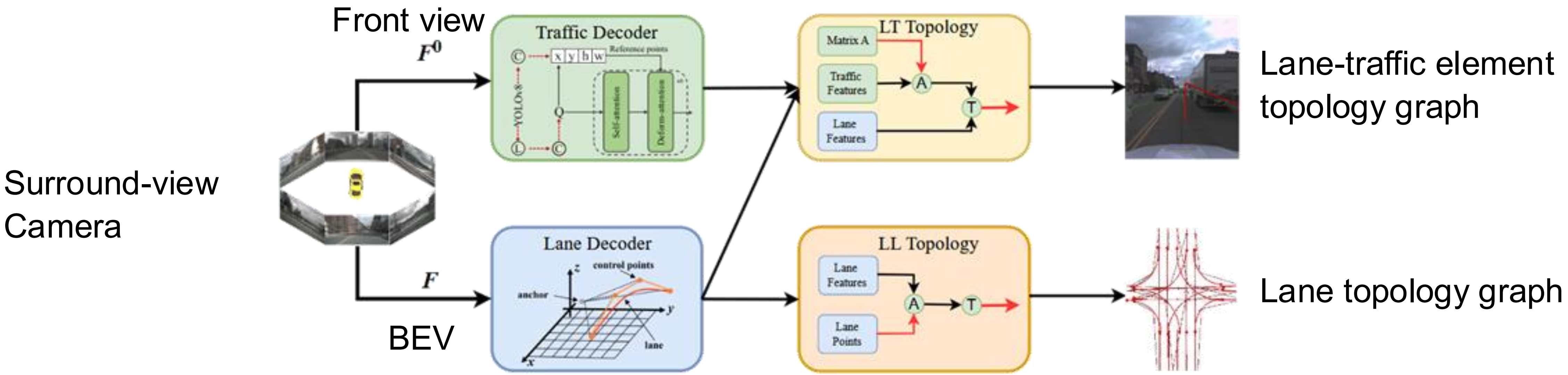


Fig: Vehicle navigation with High Definition (HD) mapping

- Precise geometry of each lane
- Capture how lanes connect (e.g., left turn)
- How traffic elements controls lanes
- Driving rules: speed limits, priority rules at intersections, and so on

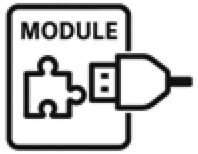
Topology Reasoning – Concept

Real-time understanding of objects and predicting topology graphs in driving scenes



Figures adopted from: "TopoMLP: A Simple yet Strong Pipeline for Driving Topology Reasoning."

Graph Self-Supervised Learning (GSSL) based Refinement/Calibration

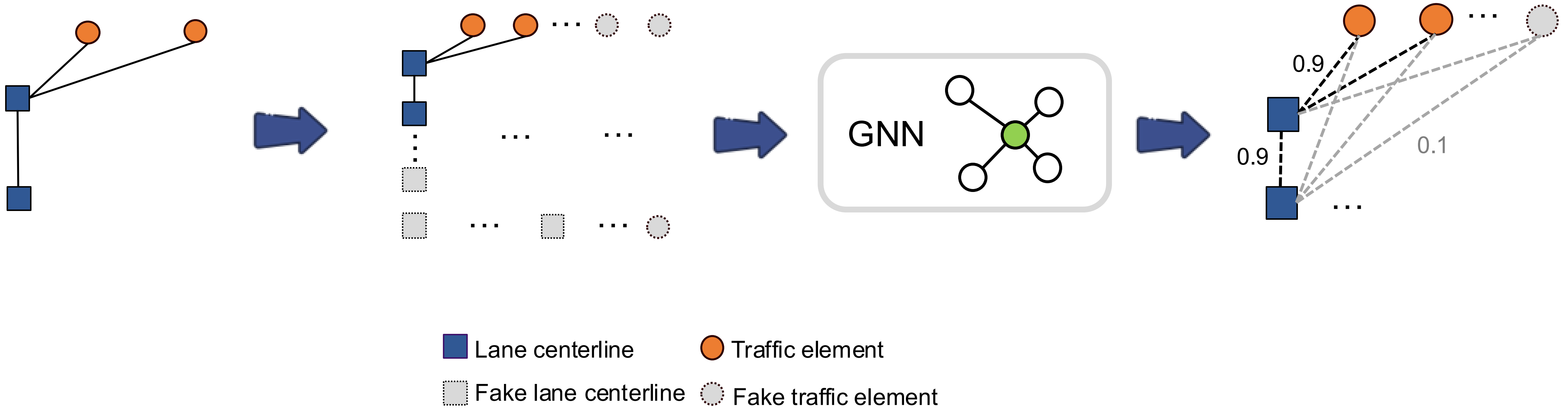


A **universal refinement module, plug-and-play** with existing topology reasoning models to calibrate topology graph

- Link prediction
- Differentiate **true** v.s. **fake** connections
- **Calibrate** confidence of the relationship

Ground-truth topology from annotation

Graph augmentation



SSL: The Engine Behind Modern AI

- Learn from raw data without human labels
- Pretext tasks: predict missing words (text), masked pixels (vision), missing edges (graphs)
- Rich representations
- Foundation of most frontier AI systems today

ChatGPT

LLaMA

OpenLaneV2 Dataset and Evaluations

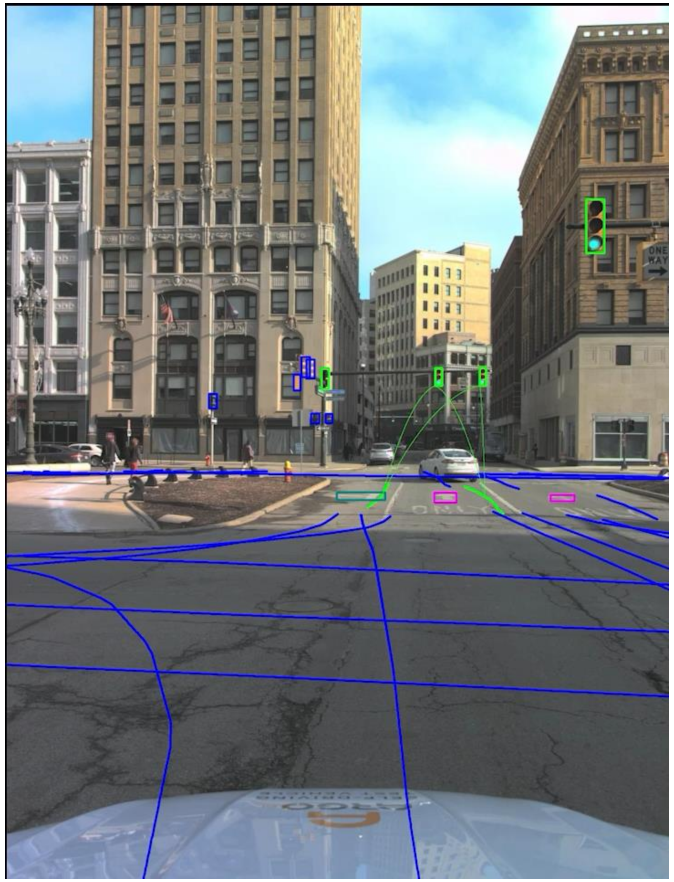
- Combine **Argoverse 2** and **nuScence** datasets.
- ~2,000 annotated road scenes with 72K frames @ 2Hz
- Annotate lanes polylines, traffic elements bounding boxes, and the driving scene topology.

Top_{ll} & Top_{lt} mAP of predicted topology

Uncalibrated Results

- Lane-traffic elements topology not well detected
- Lanes-lane topology inconsistent

TopoNet

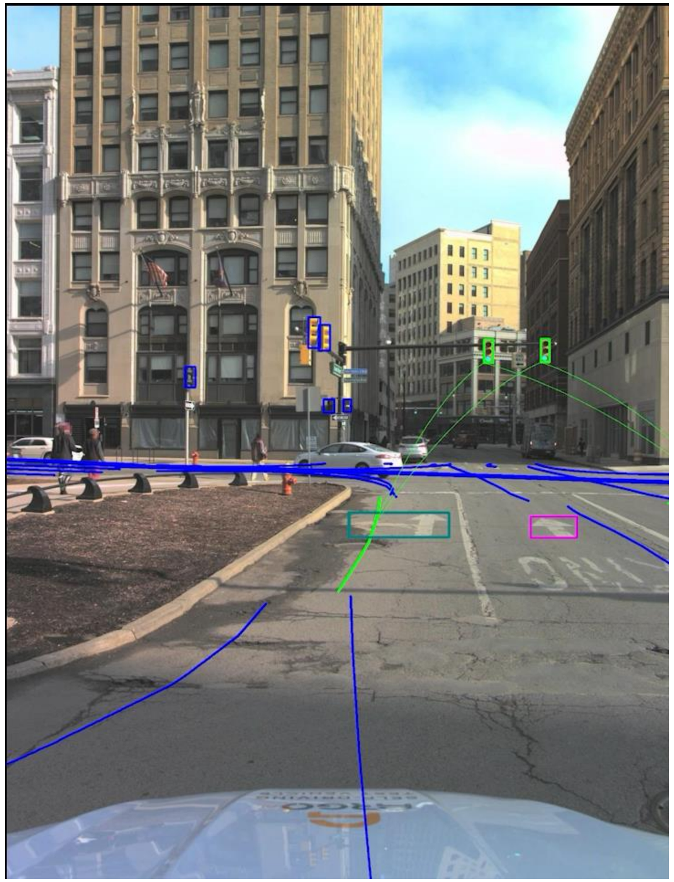


Ground-truth

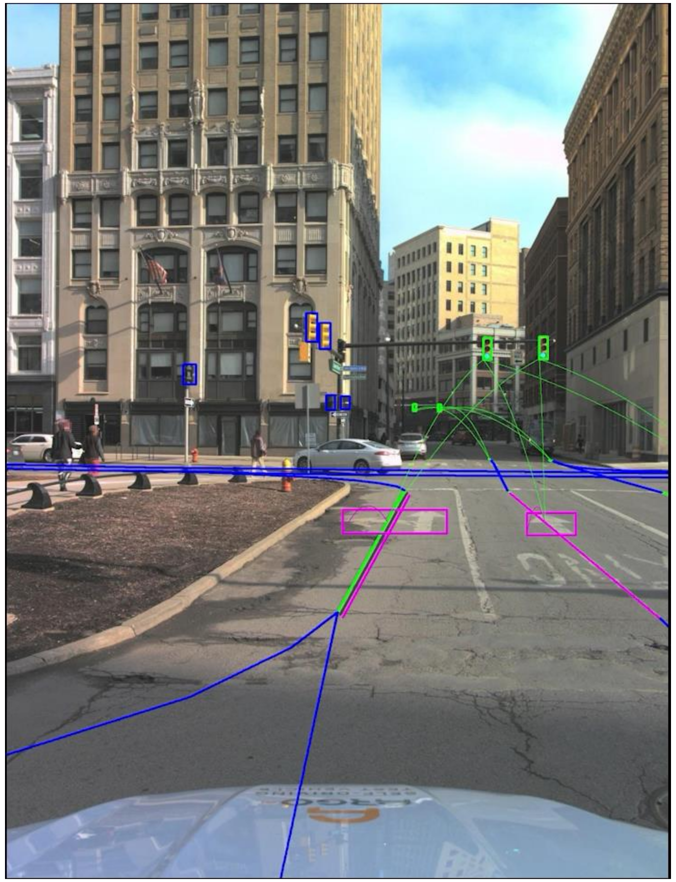
Uncalibrated Results

- Lane-traffic elements topology not well detected
- Lanes-lane topology inconsistent

TopoNet



Ground-truth

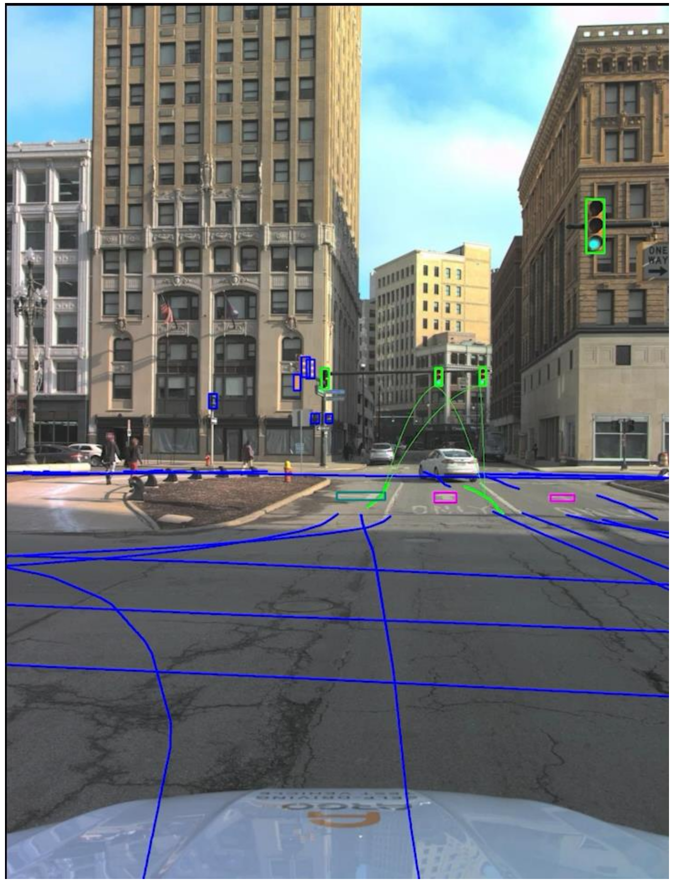


Calibrated Results

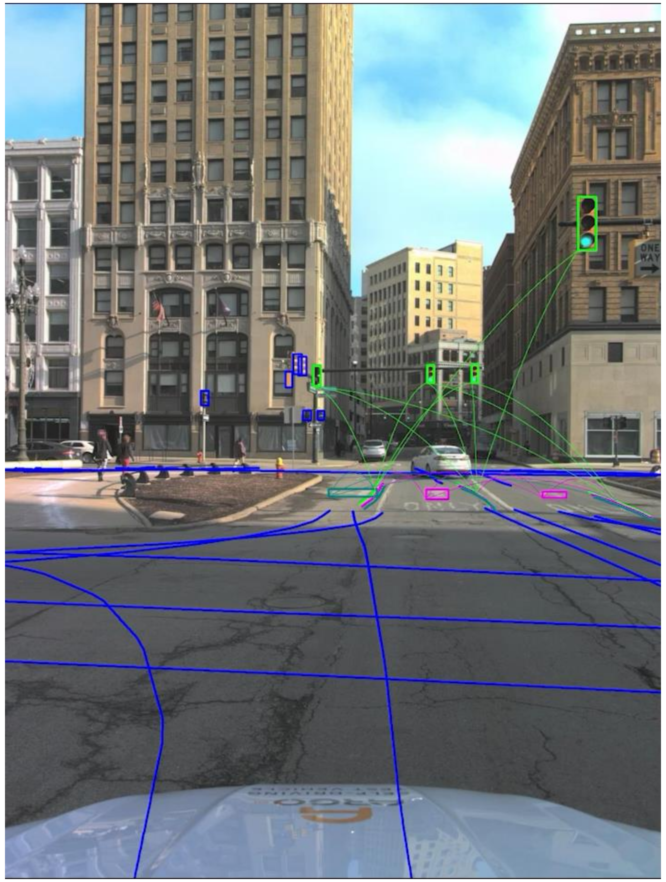
Top_{ll} **100%**

Top_{lt} **20%**

TopoNet



After calibration



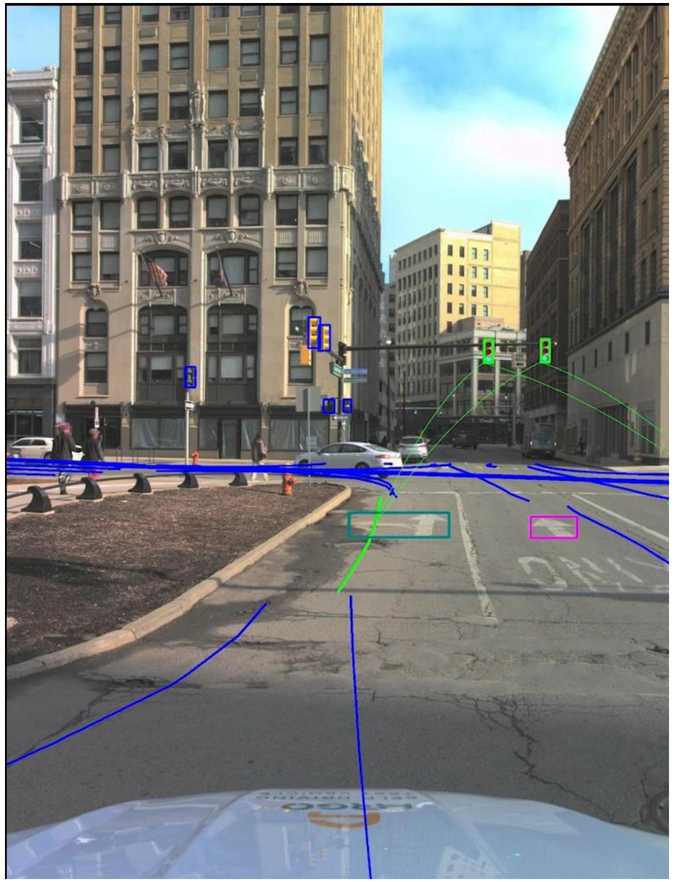
Ground-truth

Calibrated Results

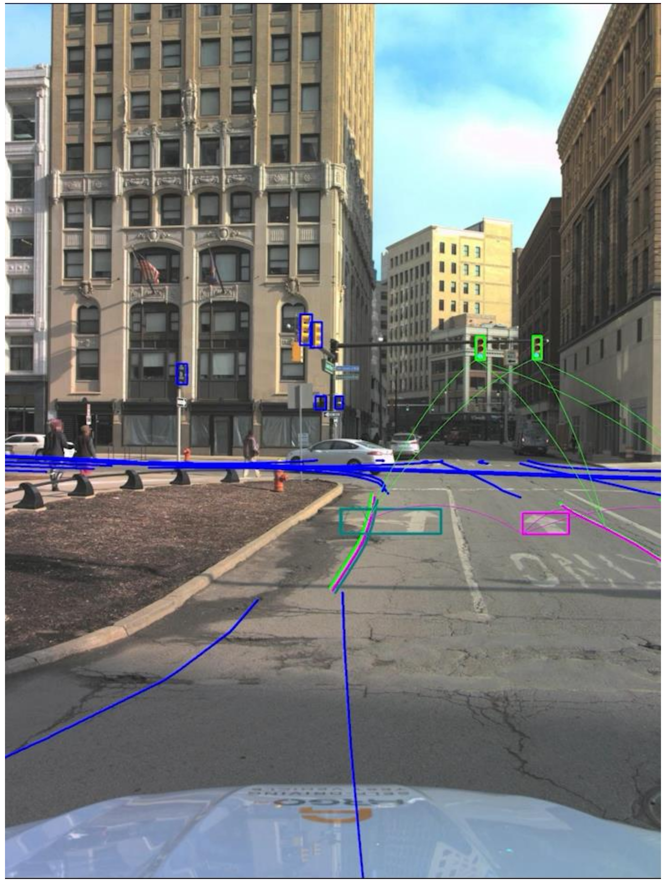
Top_{ll} **100%**

Top_{lt} **20%**

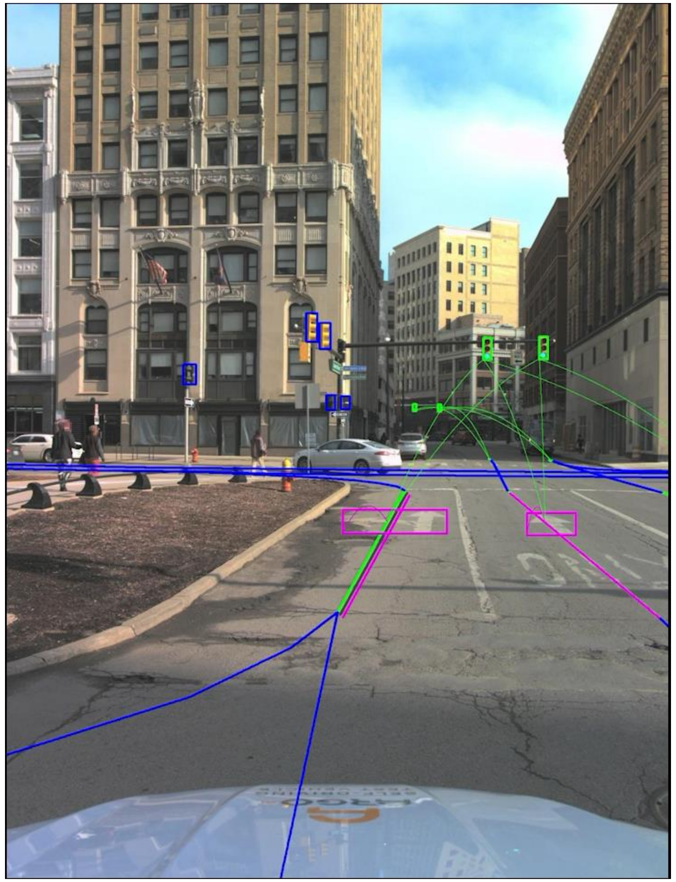
TopoNet



After calibration

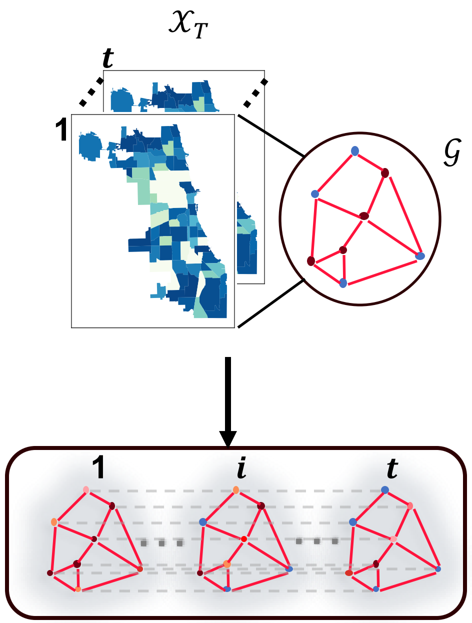


Ground-truth



Learning + Calibrating = Trustworthy Patterns

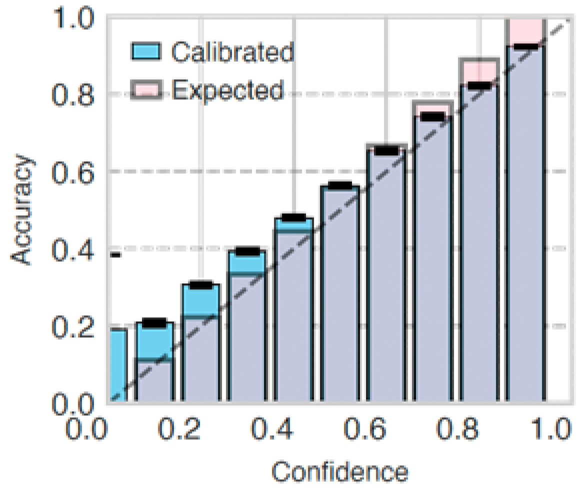
Learning



Spatiotemporal Embedding

Leaning patterns in spatiotemporal data

Calibrating



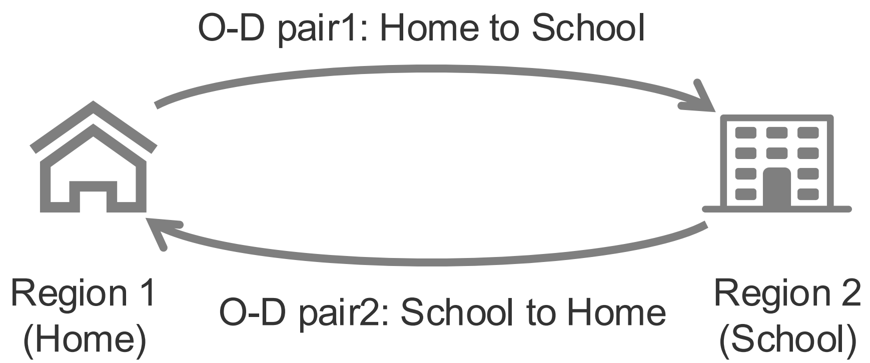
Calibrating model confidences and uncertainty

Trustworthy Patterns: Predictions we can rely on

Patterns ≠ Intelligence

Patterns Tell Us What; Reasoning Explains Why

Patterns
(what tends to happen)



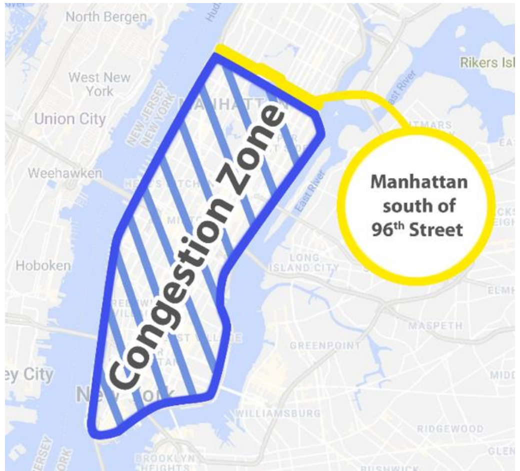
OD pairs with nearby origins or destinations tend to have similar demand

Reasoning
(Why it happens, and when the rule applies)

Nearby origins link to the same job centers, and nearby destinations share accessibility

Why Reasoning Matters

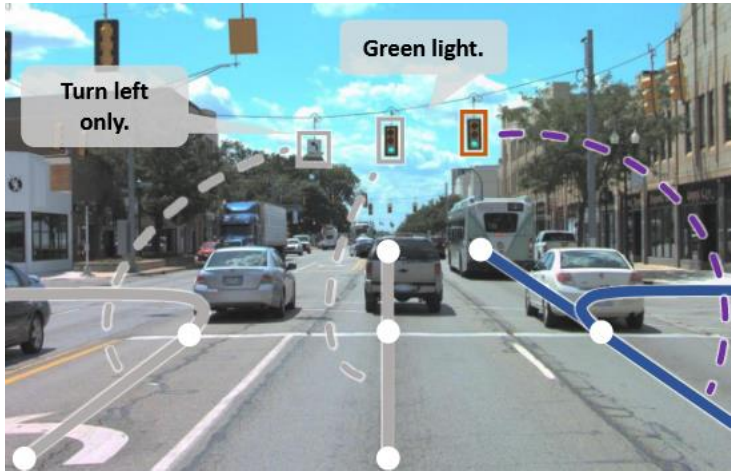
Policy-critical



Urban Planning

requires causal and social reasoning about accessibility and equity.

Safety-critical



Autonomous Driving

requires reasoning about lanes, pedestrians, traffic rules, and social norm – not just patterns

General AI systems for transportation and urban planning require spatial reasoning to effectively navigate environments and support real-world interactions

Spatial Reasoning: Fine-tuning VLMs

Sparkle: Mastering basic Spatial Capabilities in Vision Language Models Elicits Generalization to Composite Spatial Reasoning

Yihong Tang* Ao Qu* Zhaokai Wang* Dingyi Zhuang*

Zhaofeng Wu, Wei Ma, Shenhao Wang, Yunhan Zheng, Zhan Zhao, Jinhua Zhao

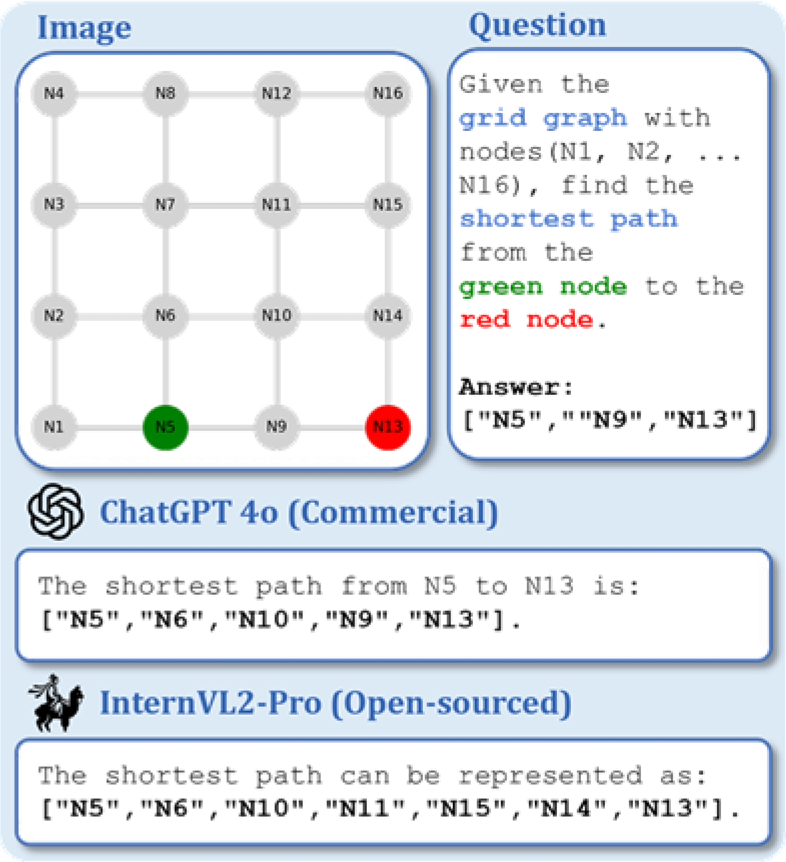
Accepted at EMNLP Findings 2025

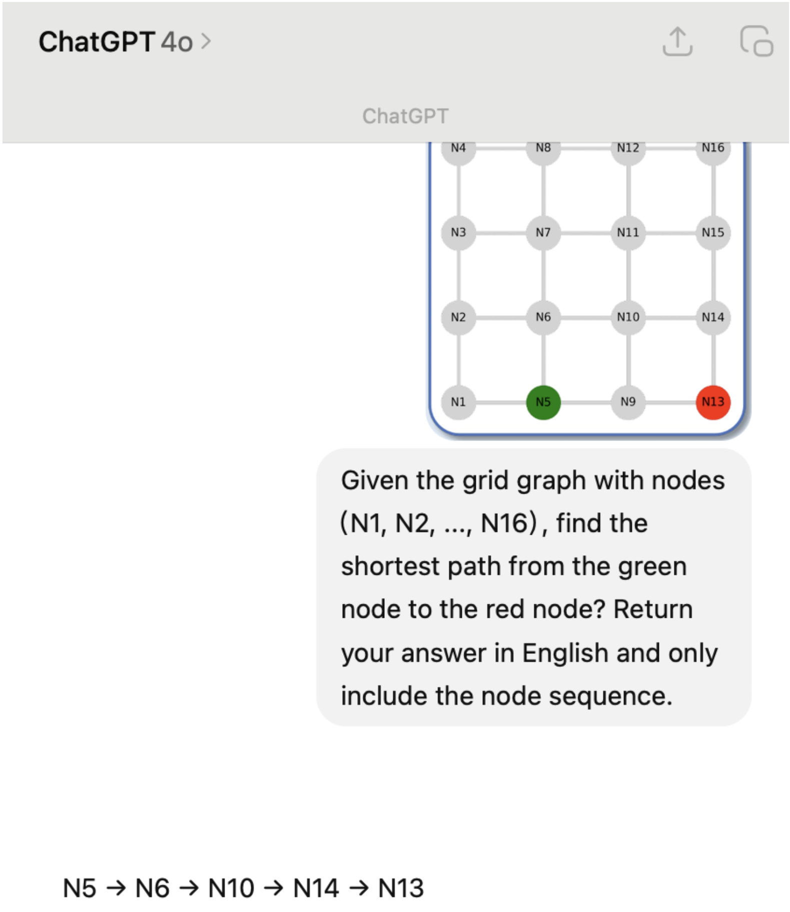
🏆 ***Best Paper Award**, MKLM Workshop @ ICJAI 2025*

* Equal contribution

Spatial Reasoning Gap in VLMs

State-of-the-art Vision Language Models (VLMs) fail to solve the pathfinding problem, a simple 2D spatial reasoning task





ChatGPT 4o still makes mistakes…

Screenshot on Aug 26, 2025

Slides adapted from Yihong Tang

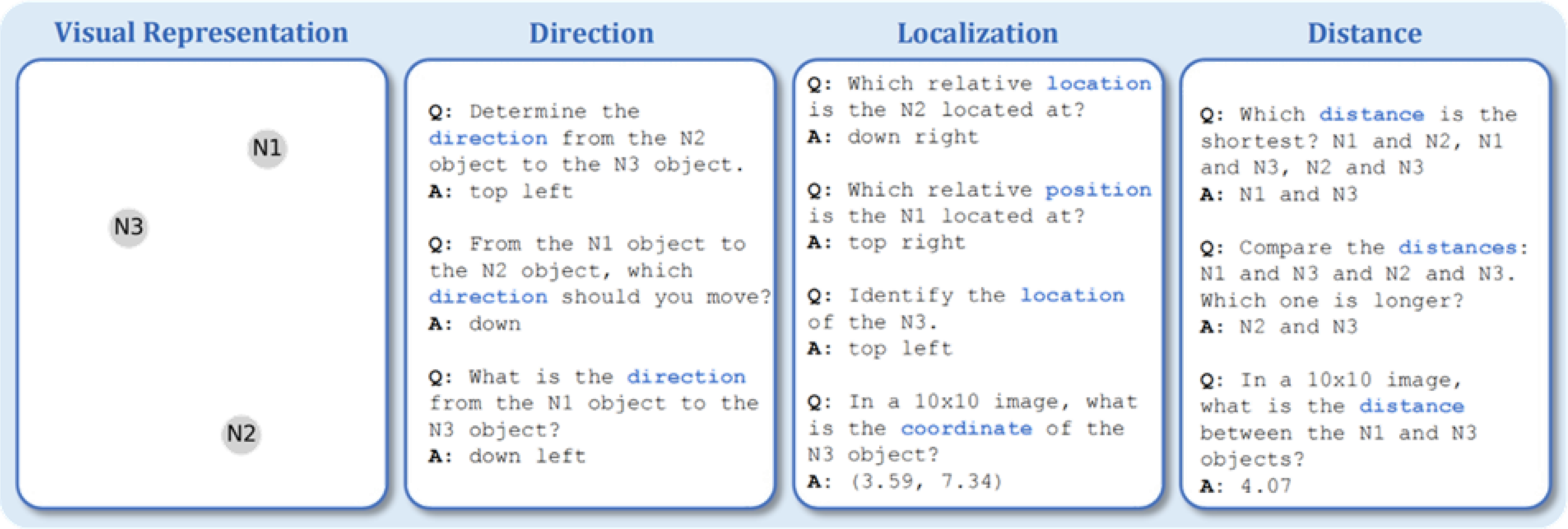
Key Research Questions

How well do existing models perform on 2D spatial reasoning tasks?

What are the fundamental capabilities that underpin spatial reasoning?

Can mastering basic capabilities lead to better performance on more complex, composite tasks?

Disentangling Spatial Reasoning

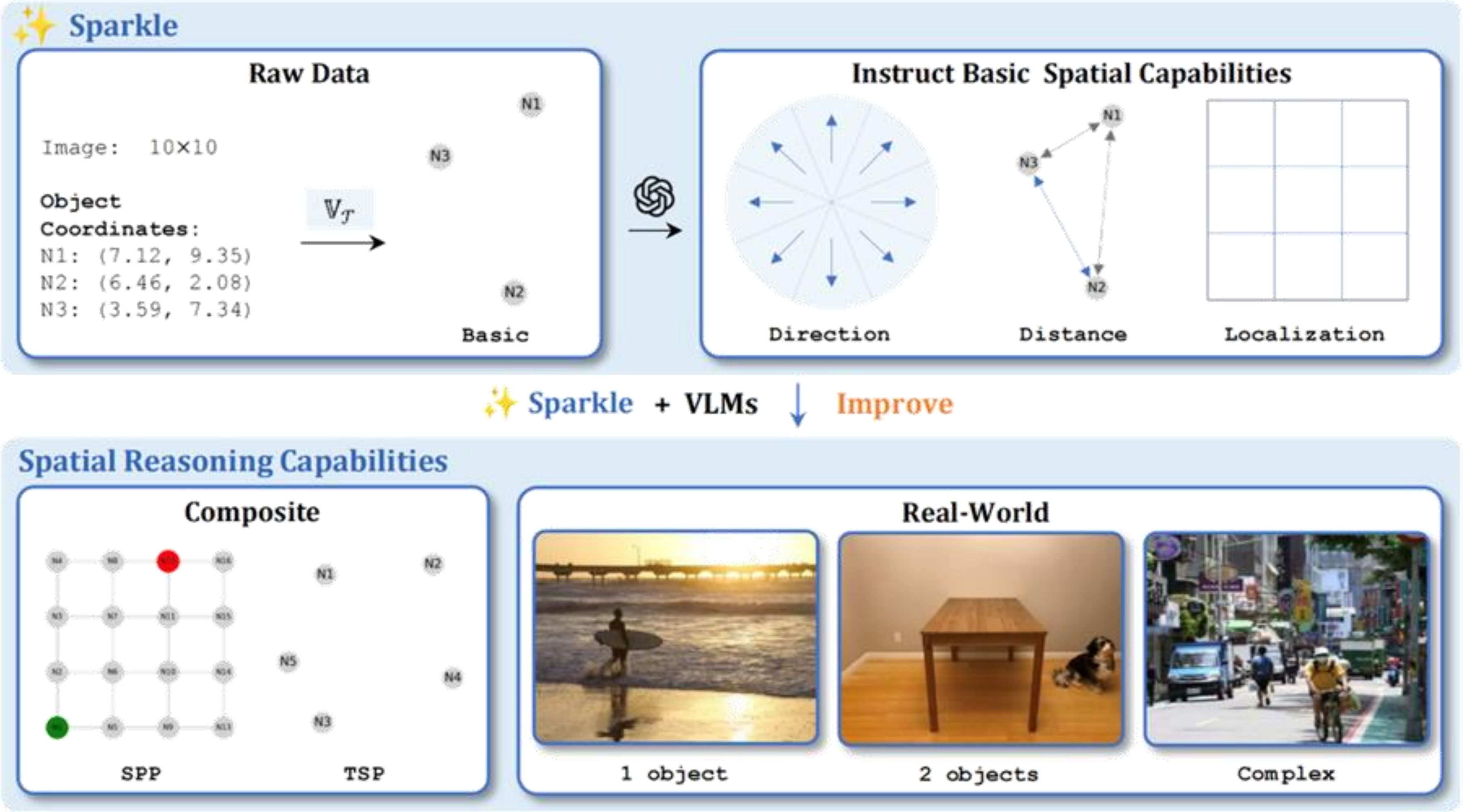


Direction: Understanding the relative orientation between objects

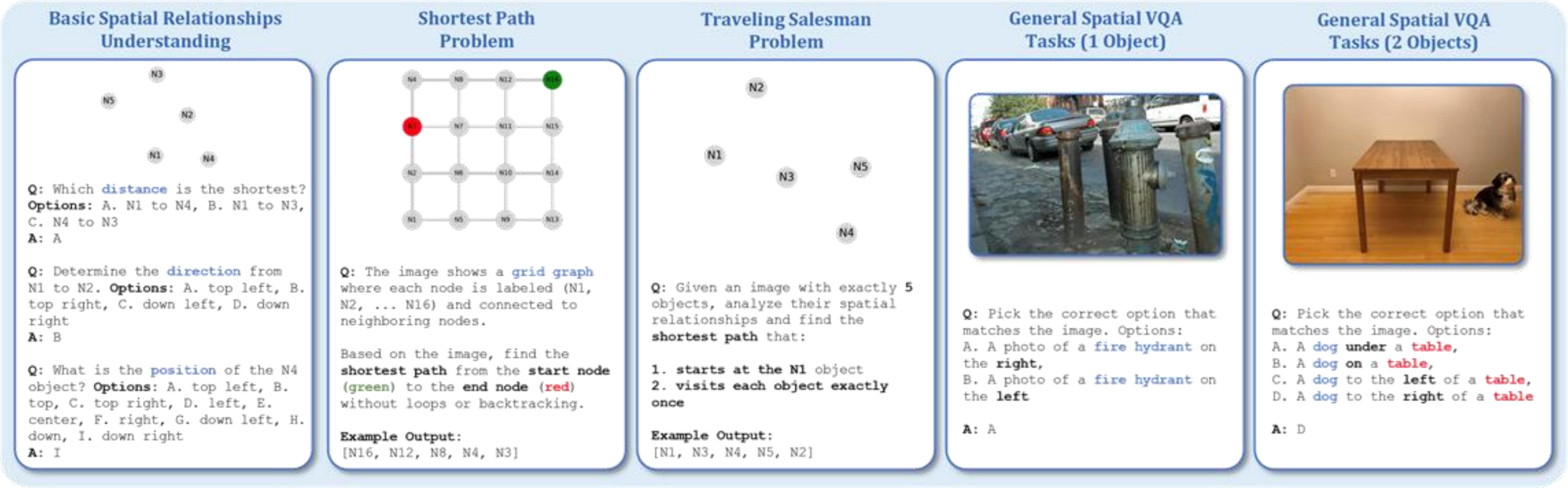
Localization: Determining an object's precise position in space

Distance: Measuring the spatial displacement between objects

The Sparkle Framework



A Multi-level Approach



Compared to the baseline VLM model

Basic tasks

20% - 165%

Composite tasks

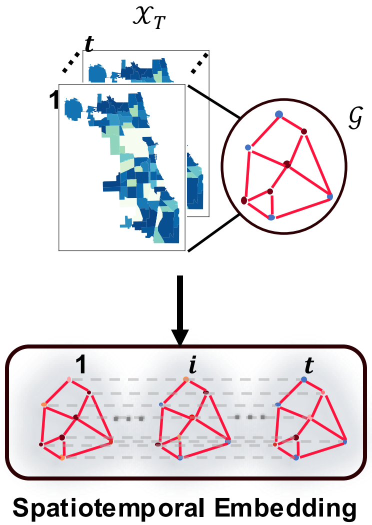
20% - 283%

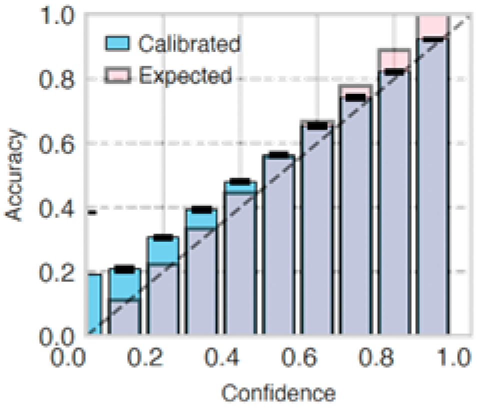
General tasks

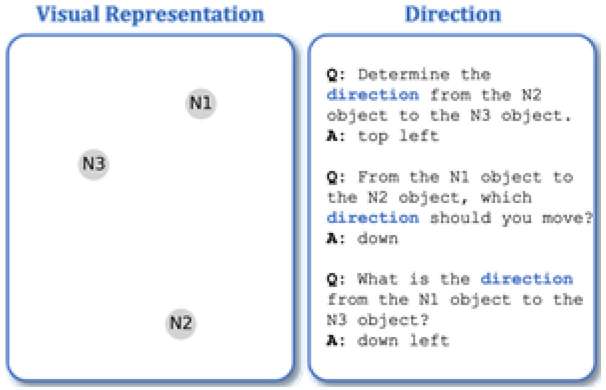
5% - 25%

Wrap-up

Learning	Calibrating	Reasoning







Find Patterns

Build Trust

Explain why

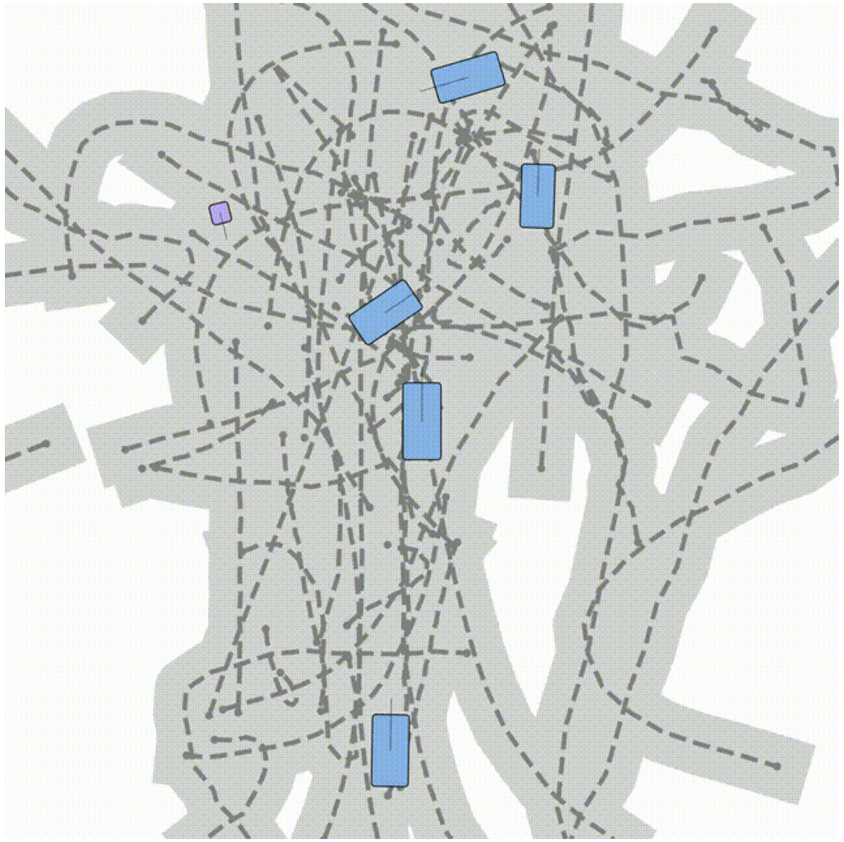
Trustworthy Patterns

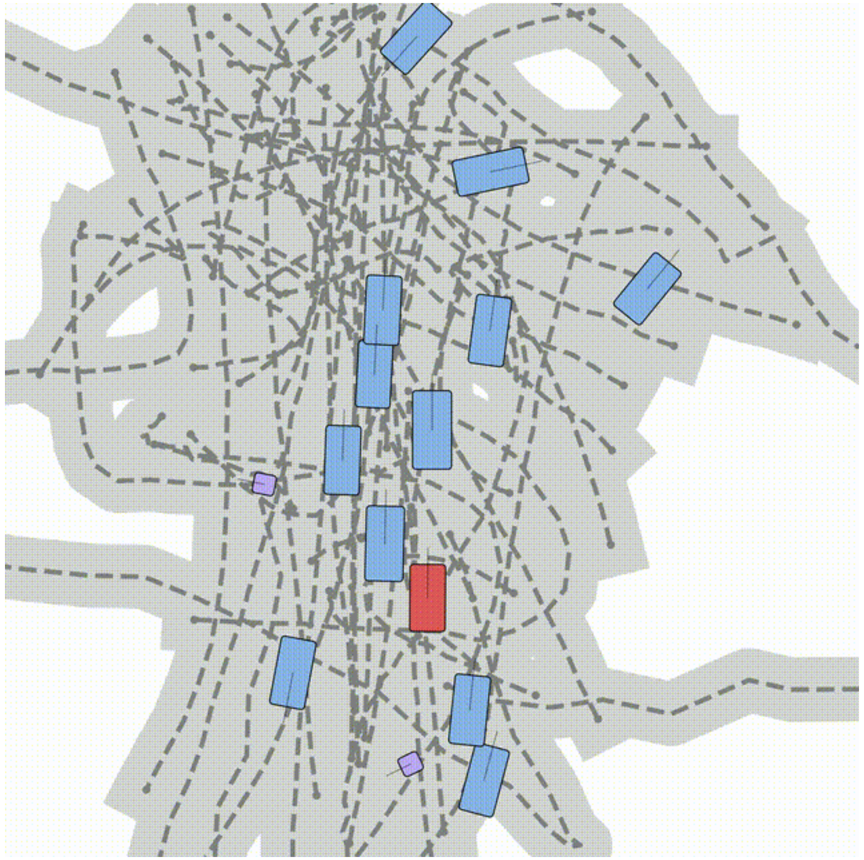
≈ Trustworthy Spatial Intelligence

A Bigger Question

Can we build models for transportation and urban systems that don't just recognize patterns, or even just reason about them — but that actually internalize them, so they can simulate, plan, and ask 'what-if' questions?

Scenario Dreamer

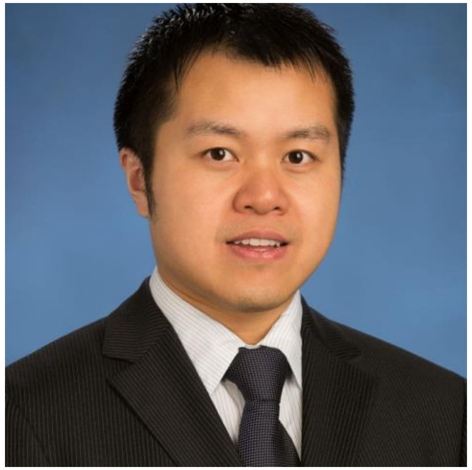




Figures and GIFs adopted from: "Scenario Dreamer: Vectorized Latent Diffusion for Generating Driving Simulation Environments."

What is “World Model”

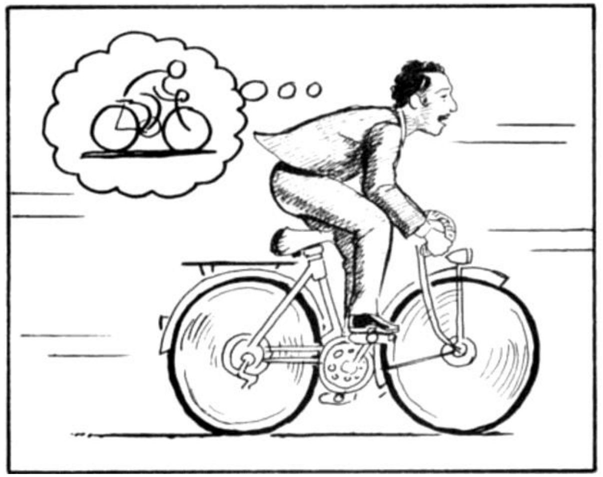
David Ha



Jürgen Schmidhuber

Yann LeCun

*Origin in **cognitive science & robotics**: internal model of the world for prediction & planning.*



World models as the foundations of autonomous intelligence; self-supervised learning to internalize physics & dynamics

Figures adopted from “World Models”

How Does My Work Connect

Indirectly, but address core ingredients

Contribution

Learning

Calibrating

Reasoning

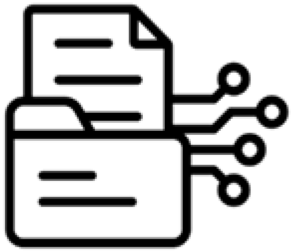
Still Missing

Richer dynamics:
integrated environment simulator

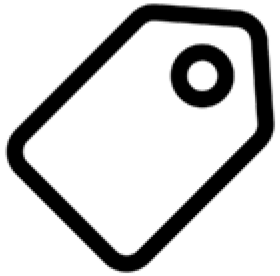
Calibration under decision-making:
Online, real-time planning with feedbacks

Higer-level reasoning:
Counterfactual simulation, embodied reasoning in urban system

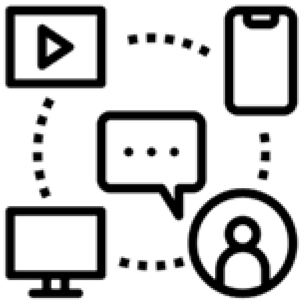
The Road Ahead: World Models of Cities



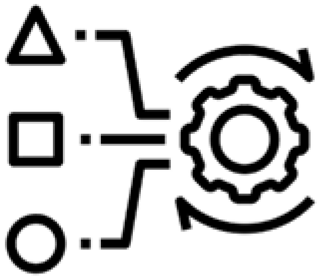
Transportation data synthesis & augmentation



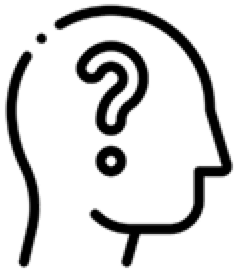
Auto-labeling



Multi-modality



Dynamic simulation



Embodied Reasoning

Thank you!

Questions?

Publication Reference

1. *Wu, Y., Zhuang, D., Labbe, A., & Sun, L. (2021, May). Inductive graph neural networks for spatiotemporal kriging. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 5, pp. 4478-4485).*
2. *Cao, X., Zhuang, D., Zhao, J., & Wang, S. (2025). Virtual Nodes Improve Long-term Traffic Prediction. arXiv preprint arXiv:2501.10048.*
3. *Zhuang, D., Hao, S., Lee, D. H., & Jin, J. G. (2020). From compound word to metropolitan station: Semantic similarity analysis using smart card data. Transportation Research Part C: Emerging Technologies, 114, 322-337.*
4. *Zhuang, D., Bu, Y., Wang, G., Wang, S., & Zhao, J. (2024, October). Sauc: Sparsity-aware uncertainty calibration for spatiotemporal prediction with graph neural networks. In Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems (pp. 160-172).*
5. *Tang, Y., Wang, Z., Qu, A., Yan, Y., Wu, Z., Zhuang, D., ... & Ma, W. (2024). ItiNera: Integrating spatial optimization with large language models for open-domain urban itinerary planning. arXiv preprint arXiv:2402.07204.*