PRELIMINARY PREPRINT VERSION – DO NOT CITE
This is not the version of record. The final version will appear in the Proceedings of IJCAI-ECAI 2026.

TrafficPDE: A Guidebook to Deployable AI-Driven Transportation Systems Through the Perception-Decision-Explanation Triangle

Ziyue Li
Department of Operations & Technology, Technical University of Munich
Heilbronn Data Science Center
Munich Data Science Institute
ziyue.li@tum.de

Abstract

Deploying AI in real-world intelligent transportation systems (ITS) remains hard: traffic data is noisy and incomplete, learned policies rarely transfer to new cities, and black-box models cannot earn the trust of operators and public officials. Drawing on years of joint experience in ITS industry and academia (including traffic signal control systems deployed across multiple cities and AI for one of the world's busiest metro networks), this paper presents **TrafficPDE**: a practitioner's guidebook organized around the **Perception-Decision-Explanation (PDE) triangle**. **Perception** tackles how to model traffic data reliably despite noise, missing sensors, and unobserved locations. **Decision** develops generalizable reinforcement learning agents that transfer across cities without costly re-calibration. **Explanation** builds causal DAG frameworks that make AI decisions interpretable to transportation practitioners. We show the three vertices reinforce each other, and close with three golden rules for any deployable ITS model: handle long-tail cases safely, operate from day one, and explain itself to people with a transportation background. TrafficPDE is not a survey of what has been done. It is a guidebook for what must be built.

1 Introduction

While many research foundations exist for intelligent transportation systems (ITS), several key challenges still block the realization of deployable ITS, and most of them point to *data*. The data from ITS is often characterized as spatiotemporal, high-dimensional, multi-modal, and noisy. An even more serious question is: "how much can we trust our data?" Take traffic flow data as an example. A common approach is via pre-installed loop detectors under the road: they generate magnetic fields and count a vehicle each time one passes. However, heavy snow or rain weakens the magnetic field, yielding inaccurate counts. A particularly counter-intuitive case: during severe congestion, the road is packed with stationary vehicles, so no car moves *through* the loop detector, and the count is nearly zero - the opposite of reality. Camera-based solutions also malfunction in limited visibility conditions such as foggy nights.

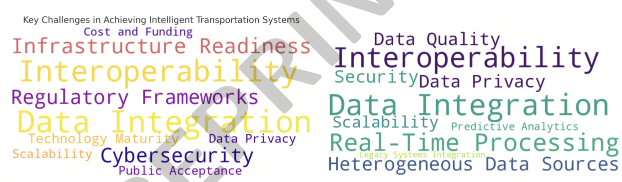

Figure 1: Key Challenges in Achieving Deployable ITS

We examine AI-driven Smart Transportation through the lens of the **PDE triangle**, not Partial Differential Equations, but **P**erception, **D**ecision, and **E**xplanation (Figure 2), as a framework to achieve deployable ITS:

- **Perception**: how to stably perceive and model the traffic environment, including its past, present, and future;
- **Decision**: how to make actionable and generalizable decisions based on the spatiotemporal, noisy, multi-modal perception;
- **Explanation**: how to explain the decisions, which is essential for AI algorithms to be trusted by operators and public officials, and therefore deployed in practice.

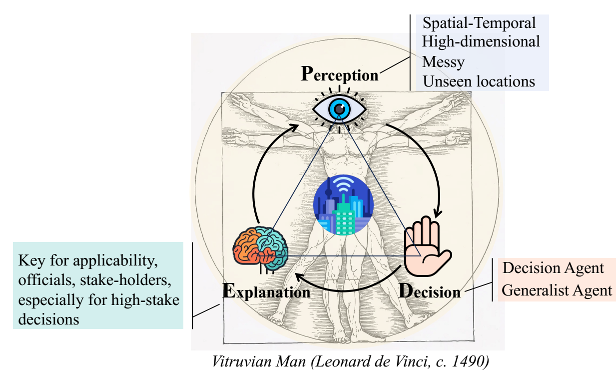

Figure 2: Traffic PDE Framework

2 Perception

A prerequisite for any downstream decision or explanation is reliable perception: understanding what is happening in the traffic network, what happened in the past, and what will happen in the future. We focus on three practical challenges: noisy data, scarce or absent data, and leveraging Large Language Models to strengthen perception.

2.1 Handle the Noise

Real-world traffic data suffers from a critical training-testing discrepancy: models are trained on clean data, yet deployed on messy, corrupted sensor readings. Many solutions focus on improving sensor quality, but this is costly.

Industry Insights: *Upgrading testing data quality means higher deployment cost for the ITS industry.*

A more practical solution is to close the gap from the other side: deliberately inject noise, corruption, and missing values into training data so the model learns to be robust against them. This is the central idea of *contrastive self-supervised learning* [Chen *et al.*, 2020]: create two corrupted views of the same input as a positive pair, and train representations to be invariant to such perturbations. The contrastive objective takes the InfoNCE form:

$$\mathcal{L}_{CL} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\big(\mathrm{sim}(\mathbf{z}_i,\,\mathbf{z}_i^{+})/\tau\big)}{\sum_{j=1}^{N}\exp(\mathrm{sim}(\mathbf{z}_i,\,\mathbf{z}_j)/\tau)} \tag{1}$$

where $\mathbf{z}_i$ and $\mathbf{z}_i^{+}$ are representations of a positive pair, $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity, and τ is a temperature hyperparameter. We apply this framework across three traffic data modalities.

Traffic Demand Time Series. Useful augmentations include jittering, temporal warping, and masking, applied across temporal and contextual dimensions to capture both local patterns and global structure [Wen *et al.*, 2021; Eldele *et al.*, 2021; Yue *et al.*, 2022; Deldari *et al.*, 2021; Wang *et al.*, 2023]. The key insight is that a model trained to recognize the same sequence under different degradation conditions will be naturally robust to the sensor noise and missing data encountered in real deployment.

Road Network. Physically connected or neighboring road segments form natural positive pairs, under the assumption that nearby roads share similar traffic patterns [You *et al.*, 2020; Mao *et al.*, 2022]. The contrastive loss is defined over a contextual neighboring graph $\mathcal{C}(s_i)$: $\mathcal{L}_{RN} = -\frac{1}{|\mathcal{S}|}\sum_{s_i\in\mathcal{S}}\left[\frac{1}{|\mathcal{C}(s_i)|}\sum_{s_j\in\mathcal{C}(s_i)}\mathcal{I}(\mathbf{h}_{s_i},\mathbf{h}_{s_j})\right]$, where $\mathcal{I}(\cdot,\cdot)$ is a mutual information estimator and $\mathcal{C}(s_i)$ is the set of contextually neighboring nodes of s_i.

Trajectory. Augmentation strategies such as point dropping, detour insertion, and temporal reordering construct two views of the same trip [Li *et al.*, 2018; Mao *et al.*, 2022]. The contrastive loss encourages representations of two augmented views of the same trajectory τ_i to be closer than those of different trajectories: $\mathcal{L}_{Traj} = -\frac{1}{|\mathcal{T}|}\sum_{\tau_i\in\mathcal{T}}\mathcal{I}(\mathbf{h}_{\tau_i'},\mathbf{h}_{\tau_i})$. This joint learning across all three modalities produces representations that capture static road geometry, dynamic sensor patterns, and movement behavior simultaneously, benefiting downstream tasks such as travel time estimation, route recommendation, and anomaly detection.

2.2 Small Data or even No Data

In real ITS deployments, many locations have never had sensors installed, meaning there is literally no data at all.

Industry Insights: *In one of our pilot cities with over 30 million residents, we could only deploy sensors at 120 intersections. Most locations have no data at all.*

The task of estimating values at unobserved locations from observed sensors is called *spatial kriging*. We approach this from two complementary directions.

Doing Subtraction - Mask and Recover. We randomly mask sensor nodes during training and recover their readings from the remaining observed graph via message passing [Wu *et al.*, 2021], which teaches the model to generalize to unseen locations at test time: $\mathcal{L}_{mask} = \frac{1}{|\mathcal{M}|}\sum_{v\in\mathcal{M}}\|\hat{\mathbf{x}}_v - \mathbf{x}_v\|^2$. However, physical neighbors are not always the most informative reference: two adjacent roads can belong to entirely different traffic regimes, while distant roads can share nearly identical patterns. We address this by selecting informative reference nodes regardless of physical proximity, using contrastive-prototypical learning [Li *et al.*, 2024] and heterogeneous spatial relation encoding that jointly models proximity, functional similarity, and transition probability [Zheng *et al.*, 2023].

Doing Addition - Insert and Infer. Rather than masking existing nodes, we insert virtual nodes at target unobserved locations and generate pseudo labels from the model's own predictions, then retrain iteratively [Xu *et al.*, 2025].

Subtraction and addition offer complementary tools: subtraction trains on existing sensor configurations; addition explicitly expands coverage to locations where no sensor has ever existed.

2.3 Large Language Model

LLMs have demonstrated remarkable generalization ability, including for time series tasks in a zero-shot setting [Gruver *et al.*, 2023; Zhou *et al.*, 2023], motivating their use for traffic perception. Traffic data shares structural properties with language sequences - periodicity, trends, and contextual dependencies - and LLMs carry rich world knowledge that may help contextualize urban dynamics. We investigate three key design questions.

Spatiotemporal Tokenization. The first challenge is how to feed multivariate spatiotemporal traffic data into an LLM. We project each sensor's time series into the LLM's embedding space, treating sensor variables as token sequences and enabling multi-step traffic forecasting [Liu *et al.*, 2024]. Incorporating the road graph structure via spatiotemporal GNN encoders further captures inter-sensor spatial dependencies that pure sequence models miss [Liu *et al.*, 2025a], allowing the model to reason about how congestion on one road affects connected roads.

Cross-Modality Alignment. A more principled approach is to align a lightweight time series encoder with the LLM's representation space, rather than directly tokenizing raw numbers [Liu *et al.*, 2025c]. This cross-modality alignment transfers the LLM's language priors to the time series backbone without the instability of direct numerical tokenization.

Efficient Deployment via Distillation. LLMs are too large for real-time ITS deployment. We address this via privileged knowledge distillation [Liu *et al.*, 2025b]: a large LLM teacher, which has access to privileged future context during training, distills its knowledge into a compact student model ten times smaller.

Industry Insights: *To our knowledge, ITS industry is not yet ready for LLM-based real-time prediction, due to latency, computational cost, and interpretability concerns. However, an LLM assistant for mobility analysis and urban planning, leveraging its reasoning ability rather than raw prediction speed, is gaining real traction among practitioners.*

3 Decision

How to make a generalizable decision? We take traffic signal control (TSC) as a concrete use case, driven by a fundamental question that follows from good traffic prediction: *what can AI actually do for the society with it?*

The Industry Baseline. SCATS (Sydney Coordinated Adaptive Traffic System) remains dominant in the TSC industry worldwide. It works by measuring vehicle arrivals at road detectors to estimate local demand in real time, then adjusting cycle lengths, phase splits, and green-wave offsets accordingly, all within a fixed cyclic phase structure that operators can inspect and understand. Critically, SCATS is training-free: when deployed to a new city, a simple reparameterization of its parameters is sufficient and the system works from day one. This is the bar that any learned agent must clear to be practically adopted.

RL Formulation. In TSC, a phase is a combination of non-conflicting traffic movements; with standard intersection geometry, this yields up to 8 possible phases. The action at each timestep is to select the next phase and allocate a pre-fixed green duration, a control scheme known as *acyclic control*. This makes TSC a sequential decision problem naturally modeled as a Markov Decision Process (MDP) $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \gamma, \pi \rangle$, where the state $s \in \mathcal{S}$ encodes traffic volume and current phase, the action $a \in \mathcal{A}$ selects the next phase, and the agent learns a policy $\pi(a|s)$ to maximize the expected discounted return: $G^t := \sum_{l=0}^{\infty} \gamma^l r^{t+l}$.

Generalist Agent: We explore three tracks toward making such an agent generalizable to new, unseen cities and scenarios.

Diversity: Train Across Many Scenarios. Inspired by the success of generalist agents in language modeling [Reed *et al.*, 2022], one direct approach is to expose the agent to as many diverse traffic scenarios as possible during training. We proposed GESA [Jiang *et al.*, 2024a], a generalist TSC agent co-trained across seven different city scenarios without any manual annotation. GESA achieves **9.39%** higher rewards when transferred directly to a new unseen scenario, demonstrating that diversity of training experience is a practical path toward generalization.

Knowledge: Local and Global Priors. Diversity alone is not enough if the agent cannot disentangle what knowledge is specific to a city and what is universal. DuaLight [Lu *et al.*, 2024] addresses this by explicitly separating local knowledge (scenario-specific patterns) from global knowledge (shared decision principles across cities), combining both to improve both in-distribution performance and out-of-distribution transfer.

Architecture: Transformer as a Generalizable Decision Backbone. A key architectural insight is that the MDP trajectory, i.e., a sequence of states, actions, and rewards $s_t, a_t, r_t, s_{t+1}, \ldots$, has the same sequential structure as a language sentence. Works such as Decision Transformer [Chen *et al.*, 2021] and Trajectory Transformer [Janner *et al.*, 2021] exploit this by treating decision-making as next-token prediction with a Transformer backbone. Transformers have also been found to exhibit strong meta-learning ability, allowing a single model to adapt quickly to new tasks [Melo, 2022].

Building on this, we designed X-Light [Jiang *et al.*, 2024b] with a two-level Transformer architecture: the Lower Transformer aggregates states, actions, and rewards among the target intersection and its neighbors *within* a city, while the Upper Transformer learns general decision trajectories *across* cities. When directly transferred to unseen scenarios, X-Light surpasses all baselines by **+7.91%** on average and up to **+16.3%** in some cases.

Industry Insights: *No city has yet adopted RL agents for long-term TSC deployment. Three barriers remain: (1) Operators need to understand what the traffic lights are doing - a "green wave" policy is interpretable, but an RL policy is not. (2) The sim-to-real gap is significant: RL TSC typically uses acyclic control, which most cities do not support; real deployments prefer SCATS-style cyclic control. (3) Looking further ahead, do we even need a centralized controller when vehicles can cooperate as a connected fleet?*

To close the gap between RL performance and industrial deployability, we proposed GuideLight [Jiang *et al.*, 2024c]. Rather than replacing SCATS, GuideLight learns *alongside* it: the agent outputs cyclic phases (compatible with real-world infrastructure) and is guided by behavior cloning from a SCATS teacher, encouraging the RL policy to stay close to interpretable, rule-based behavior. A curriculum learning schedule gradually transitions the agent from easy, teacher-guided scenarios to harder, self-directed ones, producing a model that satisfies industrial requirements while still outperforming purely rule-based baselines.

4 Explanation

Making decisions deployable in the public sector requires more than accuracy - it requires *explainability*. A broad family of post-hoc interpretability methods exists to attribute model outputs to input features [Ribeiro *et al.*, 2016; Lundberg and Lee, 2017], but they explain *what* the model uses, not *why* the traffic system behaves as it does. Here we focus on *causal structure learning*: discovering the underlying cause-and-effect relationships that drive traffic congestion, which is what city planners need to act on.

Causal structure is typically represented as a Directed Acyclic Graph (DAG), where each node is a variable and each directed edge encodes a causal dependency. Learning this structure is inherently hard: the search space over all possible DAGs grows super-exponentially, and the problem is NP-Complete [Chickering, 1996]. NoTears [Zheng *et al.*, 2018] was the first to break this barrier by reformulating the combinatorial DAG search as a *continuous optimization* problem. The key insight is an algebraic characterization of acyclicity:

$$h(\mathbf{W}) = \mathrm{tr}\big(e^{\mathbf{W}\circ\mathbf{W}}\big) - d = 0 \iff \mathbf{W} \text{ is a DAG} \quad (2)$$

where $\mathbf{W} \in \mathbb{R}^{d\times d}$ is the weighted adjacency matrix and $\circ$ denotes the Hadamard product. This turns DAG learning into a smooth constrained optimization:

$$\min_{\mathbf{W}} \frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda\|\mathbf{W}\|_1 \quad \text{s.t.} \quad h(\mathbf{W}) = 0 \quad (3)$$

$\mathbf{X} \in \mathbb{R}^{n\times d}$ is the observation matrix and λ controls sparsity.

We extend NoTears to model traffic congestion, working with the rich multi-modal data available at a traffic intersection: scalar variables including OD demand, turning probability, and traffic light cycle time; vector variables including long and short cycle lengths, irrational guidance lane, irrational phase sequence , and congestion state; and functional variables including occupancy and mean speed.

Standard DAG methods assume scalar variables, but traffic sensor readings are inherently time-varying functions. MultiFun-DAG [Lan *et al.*, 2024] extends the NoTears framework to functional data by modeling the causal relationship between two functional variables as a function-on-function regression, where the causal coefficient is itself a kernel function rather than a scalar.

A road network consists of multiple intersections, each with its own set of variables, a signalized intersection has traffic-light variables that an unsignalized one does not. MM-DAG [Lan *et al.*, 2023] jointly learns DAGs across multiple locations by (1) supporting *overlapping and distinct* node sets per location, and (2) enforcing global *consensus* across locations while allowing local uniqueness. The key contribution is the Causality Difference (CD) measure, a differentiable penalty on structural disagreement between DAGs with different node sets.

More Applications. Following the same causal DAG framework, we have applied this family to EV charging station selection [Junker *et al.*, 2025], e-scooter mobility hub placement [Jin *et al.*, 2026], energy crisis understanding [Peter *et al.*, 2024], and traffic incident root cause analysis [Gou *et al.*, 2026], showing that causal structure learning generalizes well beyond traffic signal control.

Industry Insights: *The key challenge is that there is no ground truth for the learned DAG. The best we can do is offer the discovered causal structure to public sector planners and city officials, and leave the final decision to them. Our role is a decision assistant, not a decision maker.*

5 Interrelations

The three vertices of the PDE triangle are not independent, but they reinforce each other in non-obvious ways.

Better Perception Helps Decision: Better perception of the environment directly improves the quality of decisions. In X-Light [Jiang *et al.*, 2024b], we add a predictive head that forecasts the next traffic state, giving the RL agent a forward-looking representation rather than relying solely on the current observation. In CoSLight [Ruan *et al.*, 2024], a lightweight MLP-based collaborator selection module helps each intersection identify which neighboring intersections are most relevant to coordinate with, a form of structured perceptual attention that makes the multi-agent decision process more efficient and scalable.

Explanation Helps Perception: Learning a causal structure as an auxiliary task can improve predictive performance, even when the final goal is perception rather than explanation. In DCGCN [Lin *et al.*, 2023], jointly learning two causal graphs, i.e., a within-view causal graph capturing each node's internal temporal dependencies, and a cross-view causal graph capturing spatial causal relations across nodes, acts as a structured inductive bias that improves traffic prediction accuracy.

Explanation Helps Decision: Causal structure can also directly guide the decision process. In our ongoing work, we learn a DAG over neighboring traffic intersections from data, and use the discovered causal dependencies to structure the collaboration graph for multi-agent RL. The DAG provides a principled, interpretable prior on which intersections should coordinate, leading to better policies and, importantly, a decision process that operators can inspect and understand.

6 Conclusion

We have presented a vision of AI-driven Smart Transportation organized around the PDE triangle: Perception, Decision, and Explanation. Each vertex addresses a distinct barrier to real-world deployment — noisy and incomplete data, poor generalization across cities, and the opacity of learned policies — and the three are deeply intertwined rather than independent research tracks. Looking back at the full arc of this work, three questions have guided every design choice we made, and we believe they should guide the field:

Final Golden Rules: *Three golden checklist for a deployable ML model in ITS:* ***1) Can the model handle long-tail cases safely? 2) Can the model run from day one of deployment? 3) Can the model explain itself to people with a transportation background?***

Ethical Statement

There are no ethical issues.

References

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[Chen *et al.*, 2021] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

[Chickering, 1996] David Maxwell Chickering. Learning Bayesian networks is NP-complete. In *Learning from Data: Artificial Intelligence and Statistics V*, volume 112 of *Lecture Notes in Statistics*, pages 121–130. Springer, 1996.

[Deldari *et al.*, 2021] Shohreh Deldari, Daniel V Smith, Hao Xue, and Flora D Salim. Time series change point detection with self-supervised contrastive predictive coding. In *Proceedings of the web conference 2021*, pages 3124–3135, 2021.

[Eldele *et al.*, 2021] Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2352–2359. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.

[Gou *et al.*, 2026] Xiaochuan Gou, Ziyue Li, Tian Lan, Junpeng Lin, Bingyu Zhao, Chen Zhang, Di Wang, Xiangliang Zhang, et al. Traffident: A dataset for understanding the interplay between traffic dynamics and incidents. *Advances in Neural Information Processing Systems*, 38, 2026.

[Gruver *et al.*, 2023] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. Large language models are zero-shot time series forecasters. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[Janner *et al.*, 2021] Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. *Advances in neural information processing systems*, 34:1273–1286, 2021.

[Jiang *et al.*, 2024a] Haoyuan Jiang, Ziyue Li, Zhishuai Li, Lei Bai, Hangyu Mao, Wolfgang Ketter, and Rui Zhao. A general scenario-agnostic reinforcement learning for traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, 25(9):11330–11344, 2024.

[Jiang *et al.*, 2024b] Haoyuan Jiang, Ziyue Li, Hua Wei, Xuantang Xiong, Jingqing Ruan, Jiaming Lu, Hangyu Mao, and Rui Zhao. X-light: Cross-city traffic signal control using transformer on transformer as meta multi-agent reinforcement learner. In Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 94–102. International Joint Conferences on Artificial Intelligence Organization, 8 2024. Main Track.

[Jiang *et al.*, 2024c] Haoyuan Jiang, Xuantang Xiong, Ziyue Li, Hangyu Mao, Guanghu Sui, Jingqing Ruan, Yuheng Cheng, Hua Wei, Wolfgang Ketter, and Rui Zhao. Guidelight:" industrial solution" guidance for more practical traffic signal control agents. *arXiv preprint arXiv:2407.10811*, 2024.

[Jin *et al.*, 2026] Meng Jin, Melanie Handrich, Simone Martinenz, Nicholas Hoeser, and Ziyue Li. From causal discovery to implementation: An agentic ai framework for e-scooter mobility hub planning across 29 german cities. *arXiv preprint arXiv:2606.25484*, 2026.

[Junker *et al.*, 2025] Julius Stephan Junker, Rong Hu, Ziyue Li, and Wolfgang Ketter. Data-driven optimization of ev charging station placement using causal discovery. In *2025 IEEE 21st International Conference on Automation Science and Engineering (CASE)*, pages 2722–2729. IEEE, 2025.

[Lan *et al.*, 2023] Tian Lan, Ziyue Li, Zhishuai Li, Lei Bai, Man Li, Fugee Tsung, Wolfgang Ketter, Rui Zhao, and Chen Zhang. Mm-dag: Multi-task dag learning for multi-modal data-with application for traffic congestion analysis. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1188–1199, 2023.

[Lan *et al.*, 2024] Tian Lan, Ziyue Li, Junpeng Lin, Zhishuai Li, Lei Bai, Man Li, Fugee Tsung, Rui Zhao, and Chen Zhang. Multifun-dag: Multivariate functional directed acyclic graph. *arXiv preprint arXiv:2404.13836*, 2024.

[Li *et al.*, 2018] Xiucheng Li, Kaiqi Zhao, Gao Cong, Christian S Jensen, and Wei Wei. Deep representation learning for trajectory similarity computation. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pages 617–628. IEEE, 2018.

[Li *et al.*, 2024] Zhishuai Li, Yunhao Nie, Ziyue Li, Lei Bai, Yisheng Lv, and Rui Zhao. Non-neighbors also matter to Kriging: A new contrastive-prototypical learning. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 46–54. PMLR, 02–04 May 2024.

[Lin *et al.*, 2023] Junpeng Lin, Ziyue Li, Zhishuai Li, Lei Bai, Rui Zhao, and Chen Zhang. Dynamic causal graph convolutional network for traffic prediction. In *2023 IEEE 19th International Conference on Automation Science and Engineering (CASE)*, pages 1–8. IEEE, 2023.

[Liu *et al.*, 2024] Chenxi Liu, Sun Yang, Qianxiong Xu, Zhishuai Li, Cheng Long, Ziyue Li, and Rui Zhao. Spatial-temporal large language model for traffic prediction. In *2024 25th IEEE international conference on mobile data management (MDM)*, pages 31–40. IEEE, 2024.

[Liu *et al.*, 2025a] Chenxi Liu, Kethmi Hirushini Hettige, Qianxiong Xu, Cheng Long, Shili Xiang, Gao Cong, Ziyue Li, and Rui Zhao. St-llm+: Graph enhanced spatio-temporal large language models for traffic prediction. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[Liu *et al.*, 2025b] Chenxi Liu, Hao Miao, Qianxiong Xu, Shaowen Zhou, Cheng Long, Yan Zhao, Ziyue Li, and Rui Zhao. Efficient multivariate time series forecasting via calibrated language models with privileged knowledge distillation. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, pages 3165–3178. IEEE, 2025.

[Liu *et al.*, 2025c] Chenxi Liu, Qianxiong Xu, Hao Miao, Sun Yang, Lingzheng Zhang, Cheng Long, Ziyue Li, and

Rui Zhao. Timecma: Towards llm-empowered multivariate time series forecasting via cross-modality alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18780–18788, 2025.

[Lu *et al.*, 2024] Jiaming Lu, Jingqing Ruan, Haoyuan Jiang, Ziyue Li, Hangyu Mao, and Rui Zhao. Dualight: Enhancing traffic signal control by leveraging scenario-specific and scenario-shared knowledge. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '24, page 1283–1291, Richland, SC, 2024. International Foundation for Autonomous Agents and Multiagent Systems.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[Mao *et al.*, 2022] Zhenyu Mao, Ziyue Li, Dedong Li, Lei Bai, and Rui Zhao. Jointly contrastive representation learning on road network and trajectory. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1501–1510, 2022.

[Melo, 2022] Luckeciano C Melo. Transformers are meta-reinforcement learners. In *international conference on machine learning*, pages 15340–15359. PMLR, 2022.

[Peter *et al.*, 2024] Philipp K Peter, Yulin Li, Ziyue Li, and Wolfgang Ketter. Deep causal learning to explain and quantify the geo-tension's impact on natural gas demand. In *2024 IEEE PES Innovative Smart Grid Technologies Europe (ISGT EUROPE)*, pages 1–5. IEEE, 2024.

[Reed *et al.*, 2022] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *Transactions on Machine Learning Research*, 2022.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[Ruan *et al.*, 2024] Jingqing Ruan, Ziyue Li, Hua Wei, Haoyuan Jiang, Jiaming Lu, Xuantang Xiong, Hangyu Mao, and Rui Zhao. Coslight: Co-optimizing collaborator selection and decision-making to enhance traffic signal control. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2500–2511, 2024.

[Wang *et al.*, 2023] Luxuan Wang, Lei Bai, Ziyue Li, Rui Zhao, and Fugee Tsung. Correlated time series self-supervised representation learning via spatiotemporal bootstrapping. In *2023 IEEE 19th international conference on automation science and engineering (CASE)*, pages 1–7. IEEE, 2023.

[Wen *et al.*, 2021] Qingsong Wen, Liang Sun, Fan Yang, Xiaomin Song, Jingkun Gao, Xue Wang, and Huan Xu. Time series data augmentation for deep learning: A survey. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4653–4660. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Survey Track.

[Wu *et al.*, 2021] Yuankai Wu, Dingyi Zhuang, Aurelie Labbe, and Lijun Sun. Inductive graph neural networks for spatiotemporal kriging. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 4478–4485, 2021.

[Xu *et al.*, 2025] Qianxiong Xu, Cheng Long, Ziyue Li, Sijie Ruan, Rui Zhao, and Zhishuai Li. Kits: Inductive spatio-temporal kriging with increment training strategy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 12945–12953, 2025.

[You *et al.*, 2020] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In *Advances in Neural Information Processing Systems*, volume 33, pages 5812–5823, 2020.

[Yue *et al.*, 2022] Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 8980–8987, 2022.

[Zheng *et al.*, 2018] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

[Zheng *et al.*, 2023] Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, Jianzhong Qi, Chaochao Chen, and Longbiao Chen. Increase: Inductive graph representation learning for spatio-temporal kriging. In *Proceedings of the ACM Web Conference 2023*, pages 673–683, 2023.

[Zhou *et al.*, 2023] Tian Zhou, Peisong Niu, Liang Sun, Rong Jin, et al. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems*, 36:43322–43355, 2023.