MASKED AUTOENCODER-BASED SPATIO-TEMPORAL GRAPH NEURAL NETWORK FOR ROAD TRAFFIC FLOW RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Traffic sensor networks frequently suffer from missing data due to equipment failures and communication disruptions, degrading downstream intelligent transportation system (ITS) applications that depend on complete observations. We propose MAST-GNN, a Masked Autoencoder-based Spatio-Temporal Graph Neural Network that reconstructs missing traffic measurements through co-optimized spatial and temporal components tailored to the reconstruction objective. A two-layer graph convolutional network propagates information from observed to masked nodes along the road topology. A multi-scale temporal encoder captures short-term fluctuations, daily periodicity, and long-term trends through three parallel branches fused via a learnable concatenation-projection. Seven structured masking strategies, four spatial and three temporal, simulate realistic sensor outage patterns. We evaluate MAST-GNN on production traffic data from 95 signalized intersections in Beijing, at two granularities that share a single navigation-derived source, an intersection-level flow view and a direction-level turn-movement view, together with two public highway benchmarks. At a 30% mask rate, MAST-GNN attains $R^2 = 0.954$ at the intersection level and $R^2 = 0.912$ at the direction level, exceeding the strongest baseline at both granularities. Three paired ablations localize the gains: masked self-supervision (R^2 0.061→0.682), multi-scale temporal encoding (0.682→0.912), and graph propagation under cold start (0.886 versus -0.024 for non-parametric spatial interpolation). A cold-start study on twenty intersections with no training history exposes the complementary roles of the two information sources: temporal periodicity dominates under normal operation, whereas spatial topology becomes the sole recovery channel when temporal information fails. A mask-rate sweep quantifies this conditional value: the gain over the time-only baseline remains negligible under normal operation (-0.001 at 30% masking) and grows with temporal deprivation, reaching $+0.224$ at 90% masking; under cold start, graph propagation restores $R^2 = 0.886$ versus -0.024 for non-parametric spatial interpolation ($+0.910$). On public benchmarks, MAST-GNN achieves $R^2 = 0.731$ on METR-LA (207 highway sensors) and $R^2 = 0.886$ on PEMS-BAY (325 Bay Area sensors), demonstrating robust transfer to real-world highway networks. Cross-dataset analysis reveals that model capacity, graph connectivity, and feature dimensionality interact multiplicatively. The hidden dimension must exceed a graph-complexity-dependent threshold before auxiliary features can contribute substantially to reconstruction accuracy.

1 INTRODUCTION

Traffic flow data is the foundation of modern intelligent transportation systems (ITS), supporting applications from real-time congestion monitoring and route planning to long-term infrastructure investment decisions (Wu et al., 2021; Yu et al., 2018). Loop detectors, floating car systems, and connected-vehicle sensors now generate unprecedented volumes of spatio-temporal traffic observations. But no large-scale sensor deployment achieves complete data. Equipment malfunctions,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

communication link failures, power outages, and scheduled maintenance all produce substantial and often structured gaps in the observed traffic record.

The consequences of missing traffic data are severe. Downstream applications such as travel-time estimation, incident detection, and adaptive signal control all assume a complete or near-complete observation matrix. When missing values are filled with naive estimates, such as historical averages or simple interpolation, the resulting bias propagates through the entire analytics pipeline, degrading prediction accuracy and potentially leading to suboptimal operational decisions.

1.1 Limitations of Existing Approaches

Traditional statistical imputation methods, including historical mean substitution and linear interpolation, are computationally inexpensive but fundamentally limited: they treat each sensor independently and cannot exploit the rich spatial correlations that arise from the physical connectivity of the road network. Machine learning approaches such as K-nearest neighbors (KNN) and matrix factorization capture some non-linear structure but remain constrained by their inability to model the non-Euclidean topology of road graphs.

Deep learning methods, particularly those based on graph neural networks (GNNs), have shown considerable promise for traffic prediction tasks (Yu et al., 2018; Li et al., 2018; Wu et al., 2019; Jiang & Luo, 2022). Spatio-temporal graph convolutional networks (STGCN) (Yu et al., 2018), attention-enhanced variants (ASTGCN) (Guo et al., 2019), and diffusion convolutional recurrent neural networks (DCRNN) (Li et al., 2018) combine graph convolutions with temporal modules to jointly model spatial and temporal dependencies. Graph WaveNet (Wu et al., 2019) and T-GCN (Zhao et al., 2020) further explore adaptive graph learning and efficient temporal integration. However, these approaches are predominantly designed for supervised prediction tasks and assume that the input observations are complete. When applied to the imputation problem, they typically require ad-hoc preprocessing to fill missing values before model inference, creating a circular dependency.

More recently, masked self-supervised learning (SSL) has become a standard tool in natural language processing (BERT (Devlin et al., 2019)) and computer vision (MAE (He et al., 2022)). Its objective, masking parts of the input and training the model to reconstruct them, aligns directly with data imputation. Recent work on masked autoencoding for spatiotemporal data includes STD-MAE (Gao et al., 2024) and MagiNet (Zhou et al., 2025), which apply masking strategies to traffic imputation on METR-LA and PEMS-BAY. Yet existing applications of masked SSL to traffic data remain limited in three respects: (1) they typically employ only random masking, which fails to simulate the structured missingness patterns (e.g., regional sensor outages, peak-hour data loss) encountered in practice; (2) they underutilize multi-scale temporal dynamics, treating all time scales uniformly; and (3) they apply uniform reconstruction loss, ignoring the fact that errors at congested intersections during peak hours carry disproportionately higher operational costs.

1.2 Contributions

To address these limitations, we propose MAST-GNN, a self-supervised framework for traffic flow reconstruction on road networks. Rather than treating spatial and temporal components as isolated improvements, our design is guided by the principle that each component should be tailored to the *reconstruction* objective, not simply adapted from supervised prediction models. The main contributions are:

1. **A masked-autoreconstruction framework for production traffic data.** MAST-GNN reconstructs missing traffic measurements on real production data through co-optimized spatial and temporal components. On navigation-derived data from 95 signalized Beijing intersections, MAST-GNN attains $R^2 = 0.954$ at the intersection level and $R^2 = 0.912$ at the direction level at a 30% mask rate, exceeding the strongest baseline at both granularities.
2. **A reconstruction-oriented multi-scale temporal encoder.** Three parallel branches, short-term fully-connected, periodic sinusoidal, and long-term global pooling, capture complementary temporal granularities and are fused through a learnable projection. Paired ablation on the direction-level view isolates a $+0.23$ R^2 gain from temporal encoding ($0.682 \rightarrow 0.912$), which widens to $+0.37$ at the intersection level, showing that multi-scale time is most decisive for aggregate flow.

3. **The conditional value of graph structure.** A mask-rate sweep quantifies when spatial topology matters: the gain over the time-only baseline is negligible under normal operation (-0.001 at 30% masking), turns positive at 70% masking ($+0.012$), and grows to $+0.224$ at 90%. Under cold start, when no history exists, graph propagation restores $R^2 = 0.886$ versus -0.024 for non-parametric spatial interpolation ($+0.910$). Under normal operation temporal periodicity, compiled from signal timing, renders the graph redundant; under temporal deprivation it becomes the sole recovery channel. This conditional gain curve, invisible to time-only baselines, identifies the operational regime in which graph-based reconstruction is indispensable.
4. **Dual-granularity validation on a single real data source.** Decomposing each intersection into twelve turn movements shows that reconstruction difficulty is governed by semantic granularity rather than raw scale: the same architecture and data source yield $R^2 = 0.954$ on aggregate flow but $R^2 = 0.912$ on turn movements, where left-turn movements are the hardest (signal-modulated) and right-turn movements the most sparse.
5. **A systematic contrast between research benchmarks and production data.** Against public highway benchmarks (METR-LA $R^2 = 0.731$, PEMS-BAY $R^2 = 0.886$), production traffic data introduces data pathologies, structural sparsity, isolated nodes, multi-source integration, absent from clean research benchmarks, and graph connectivity interacts with feature dimensionality through a capacity threshold.

The remainder of this paper is organized as follows. Section 2 presents the methodology, Section 3 reports experimental results, and Section 4 concludes the paper; extended related work, discussion, and additional experimental detail are provided in Appendices A and B.

2 METHODOLOGY

2.1 PROBLEM DEFINITION

We model the urban road network as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$, where $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$ denotes the set of N intersections, $\mathcal{E}$ represents the set of road segments connecting intersections, and $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the adjacency matrix encoding road connectivity. Entry $A_{ij} = 1$ if a road segment directly connects intersection v_i to v_j, and $A_{ij} = 0$ otherwise.

At each intersection v_i and time step t, traffic demand is measured at two granularities derived from a single navigation source. The *intersection-level* view aggregates the twelve turn movements into a scalar flow $\mathbf{x}_i^{(t)} \in \mathbb{R}^1$. The *direction-level* view decomposes the same demand into its twelve turn movements, the Cartesian product of four approach directions (east, south, west, north) and three movement types (straight-through, left-turn, right-turn), yielding $\mathbf{x}_i^{(t)} \in \mathbb{R}^{12}$. The complete observation tensor is denoted $\mathbf{X} \in \mathbb{R}^{T \times N \times F}$, with $F = 1$ for the flow view and $F = 12$ for the direction view. In our production dataset, $N = 95$ signalized intersections and $T = 30 \times 288 = 8{,}640$ spans 30 days of five-minute slots.

The traffic flow reconstruction problem is formulated as masked self-supervised learning. Given the observed tensor $\mathbf{X}$, we generate a binary mask $\mathbf{M} \in \{0, 1\}^{T \times N}$ and produce the masked input:

$$\tilde{\mathbf{X}} = \mathbf{X} \odot (1 - \mathbf{M}) \tag{1}$$

where $\odot$ denotes element-wise multiplication broadcast along the feature dimension. The model f_θ is trained to reconstruct the original observations from the masked input:

$$\hat{\mathbf{X}} = f_\theta(\tilde{\mathbf{X}}, \mathbf{A}) \tag{2}$$

The training objective minimizes the reconstruction error on the masked positions, enabling the model to learn rich spatio-temporal representations that capture both spatial dependencies through the graph structure and temporal dynamics across time steps.

2.2 MODEL ARCHITECTURE

The proposed MAST-GNN consists of four major components: (1) a spatial GCN encoder, (2) a multi-scale temporal encoder, (3) a temporal convolution module, and (4) a fully-connected decoder. The overall architecture is illustrated in Fig. 1.

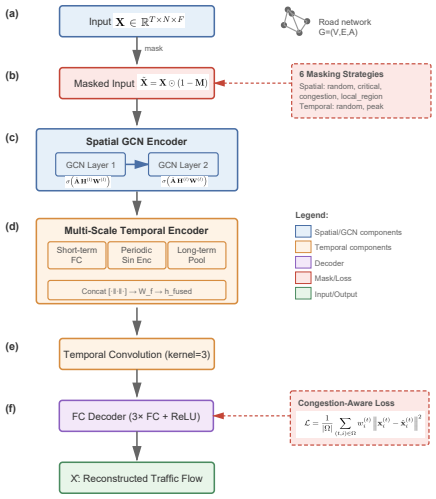

Figure 1: Architecture of the proposed MAST-GNN. **(a)** Input traffic observation tensor with the road network graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$. **(b)** Masked autoencoding: seven structured masking strategies simulate realistic sensor outage patterns (spatial: random, critical-node, congestion, local-region; temporal: random-temporal, peak-hour, missing-segment). **(c)** Two-layer GCN spatial encoder captures road-topology dependencies via normalized adjacency message passing. **(d)** Multi-scale temporal encoder with three parallel branches, short-term FC, periodic sinusoidal, and long-term global pooling, fused via concatenation and linear projection. **(e)** 1D temporal convolution (kernel=7) for local sequential smoothing. **(f)** Fully-connected decoder reconstructs the original traffic observations.

2.2.1 Spatial GCN Encoder

We employ a two-layer graph convolutional network to capture spatial dependencies through the road network topology. The adjacency matrix is first augmented with self-loops:

$$\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}_N \tag{3}$$

and symmetrically normalized:

$$\hat{\mathbf{A}} = \tilde{\mathbf{D}}^{-1/2} \tilde{\mathbf{A}} \, \tilde{\mathbf{D}}^{-1/2} \tag{4}$$

where $\tilde{\mathbf{D}}$ is the diagonal degree matrix with $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$.

Each GCN layer performs message passing:

$$\mathbf{H}^{(l+1)} = \sigma\Big(\hat{\mathbf{A}} \, \mathbf{H}^{(l)} \mathbf{W}^{(l)}\Big) \tag{5}$$

where $\mathbf{H}^{(0)} = \tilde{\mathbf{X}}$ is the masked input features, $\mathbf{W}^{(l)}$ is the learnable weight matrix of layer l, and $\sigma(\cdot)$ is the ReLU activation function. The two-layer configuration is chosen based on empirical evidence: deeper GCN architectures (3–4 layers) exhibit over-smoothing, where node representations become indistinguishable after excessive message-passing steps (Wu et al., 2021).

2.2.2 Multi-Scale Temporal Encoder

Traffic flow exhibits temporal patterns at multiple scales: short-term fluctuations (minute-to-minute variations), periodic patterns (daily and weekly cycles), and long-term trends (gradual changes in traffic demand). We design a multi-scale temporal encoder with three parallel branches to capture these complementary dynamics:

Short-term branch: A two-layer fully-connected network processes the GCN output at each time step independently:

$$\mathbf{h}_{\text{short}}^{(t)} = \text{ReLU}\Big(\mathbf{W}_{s2} \cdot \text{ReLU}\Big(\mathbf{W}_{s1} \, \mathbf{h}^{(t)} + \mathbf{b}_{s1}\Big) + \mathbf{b}_{s2}\Big) \tag{6}$$

This branch captures immediate temporal dynamics and local non-linear transformations.

Periodic branch: Sinusoidal time encoding injects explicit periodic information:

$$\mathbf{e}^{(t)} = \left[\sin\left(\frac{2\pi t}{T}\right), \cos\left(\frac{2\pi t}{T}\right)\right]^\top \tag{7}$$

which is concatenated with the node features and projected through a fully-connected layer. This encoding provides the model with hour-of-day information that is critical for distinguishing peak and off-peak patterns.

Long-term branch: Global mean pooling across all time steps followed by a fully-connected layer extracts the overall traffic trend:

$$\mathbf{h}_{\text{long}} = \mathbf{W}_l \cdot \left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{h}^{(t)}\right) + \mathbf{b}_l \tag{8}$$

The pooled representation is broadcast back to all time steps, providing a global context.

Fusion: The three branches are concatenated and projected to the hidden dimension:

$$\mathbf{h}_{\text{fused}}^{(t)} = \mathbf{W}_f \left[\mathbf{h}_{\text{short}}^{(t)} \| \mathbf{e}_{\text{proj}}^{(t)} \| \mathbf{h}_{\text{long}}\right] + \mathbf{b}_f \tag{9}$$

where $\|$ denotes concatenation and $\mathbf{W}_f \in \mathbb{R}^{d \times 3d}$ projects from $3d$ to d dimensions.

2.2.3 Temporal Convolution

A 1D temporal convolution with kernel size 7 (a ± 3-step window, i.e., up to ± 15 minutes at five-minute resolution) is applied to the fused features to capture local temporal dependencies:

$$\mathbf{z}^{(t)} = \text{ReLU}\left(\sum_{k=-3}^{3} \mathbf{W}_c^{(k)}\, \mathbf{h}_{\text{fused}}^{(t+k)} + \mathbf{b}_c\right) \tag{10}$$

This module smooths the temporal features and captures short-range sequential patterns that complement the multi-scale encoding.

2.2.4 Decoder

The decoder consists of three fully-connected layers that map the encoded representations back to the original feature space:

$$\begin{aligned}\hat{\mathbf{x}}^{(t)} = \mathbf{W}_{d3} \cdot \text{ReLU}\Big(\mathbf{W}_{d2} \cdot \text{ReLU}\big(\mathbf{W}_{d1}\, \mathbf{z}^{(t)} \\ + \mathbf{b}_{d1}\big) + \mathbf{b}_{d2}\Big) + \mathbf{b}_{d3}\end{aligned} \tag{11}$$

2.3 Masking Strategies

Unlike prior work that relies solely on random masking, we design seven structured masking strategies (four spatial and three temporal) that simulate realistic missing-data patterns in traffic sensor networks: random, critical-node, congestion-based, and local-region *spatial masks*, and random-temporal, peak-hour, and missing-segment *temporal masks*. Their definitions are provided in Appendix B.

In practice, spatial and temporal masks can be combined to create more complex missing-data patterns. The default configuration applies both spatial and temporal masking at a rate of 30%.

2.4 Training Procedure

The model is trained using the Adam optimizer with a cosine annealing learning rate schedule:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})\left(1 + \cos\left(\frac{t \cdot \pi}{T_{\text{epochs}}}\right)\right) \tag{12}$$

where $\eta_{\max} = 0.0015$ and $\eta_{\min} = 10^{-6}$. Gradient clipping is applied with a maximum norm of 1.2, and L_2 regularization with coefficient 1.5×10^{-5} is added to prevent overfitting. The hidden dimension is set to 128 and the model is trained for 600 epochs.

The training objective is a congestion-aware weighted mean-squared-error on the masked positions. Let $\Omega \subseteq \{(t, i)\}$ denote the masked (time, node) slots and let $\hat{\mathbf{x}}_i^{(t)}$ denote the reconstruction:

$$\mathcal{L}(\theta) = \frac{1}{|\Omega|\, F} \sum_{(t,i)\in\Omega} w_{t,i} \left\| \mathbf{x}_i^{(t)} - \hat{\mathbf{x}}_i^{(t)} \right\|^2, \tag{13}$$

where the per-slot weight combines a peak-hour factor and a congestion factor,

$$w_{t,i} = p_t\, q_i, \qquad p_t = 1 + \beta \mathbb{1}[\, t \in \text{AM/PM peaks}\,], \qquad c_i = \frac{\mathrm{Var}_i}{\bar{\mathrm{Var}}}, \tag{14}$$

$$q_i = \frac{1 + \alpha\, c_i}{\frac{1}{N}\sum_j (1 + \alpha\, c_j)}, \qquad \bar{\mathrm{Var}} = \frac{1}{N}\sum_j \mathrm{Var}_j, \tag{15}$$

with Var_i the empirical variance of node i's primary feature over the training window and $\mathbb{1}[\cdot]$ the indicator of the morning (07:00–09:00) and evening (17:00–19:00) peak slots. This weighting assigns larger loss to reconstruction errors at high-variance, congested intersections during peak hours. The peak and congestion emphasis is strongest early in training and is interpolated toward uniform weights as training progresses, so that the model first learns to prioritize operationally critical slots and then refines the full mask distribution. The congestion strength α and peak weight β are set per experiment family (production data: $\alpha = 1.2$, $\beta = 2.0$; public benchmarks: $\alpha = 1.5$, $\beta = 1.0$).

3 Experiments and Results

3.1 Dataset

We evaluate MAST-GNN on production traffic data from 95 signalized intersections in Beijing, spanning 30 days at five-minute resolution. The data integrates two municipal sources: (1) a navigation-derived turning-flow feed, which records traffic demand per turn movement at five-minute granularity, and (2) the signal-timing database, which associates each approach lane with a phase, turn direction, and green split. From this single source we derive two strictly nested views. The *intersection-level flow view* aggregates the twelve turn movements into a scalar demand per intersection, yielding $\mathbf{X}_{\text{flow}} \in \mathbb{R}^{8640\times 95\times 1}$. The *direction-level view* retains all twelve movements, yielding $\mathbf{X}_{\text{dir}} \in \mathbb{R}^{8640\times 95\times 12}$.

A detailed characterization of production-data topology and structural sparsity is given in Appendix B.

3.2 Evaluation Metrics

We report three standard regression metrics:

$$\text{MSE} = \frac{1}{|\Omega|} \sum_{(t,i)\in\Omega} \left\| \mathbf{x}_i^{(t)} - \hat{\mathbf{x}}_i^{(t)} \right\|^2 \tag{16}$$

$$\text{MAE} = \frac{1}{|\Omega|} \sum_{(t,i)\in\Omega} \left\| \mathbf{x}_i^{(t)} - \hat{\mathbf{x}}_i^{(t)} \right\|_1 \tag{17}$$

$$R^2 = 1 - \frac{\sum_{(t,i)\in\Omega} \left\| \mathbf{x}_i^{(t)} - \hat{\mathbf{x}}_i^{(t)} \right\|^2}{\sum_{(t,i)\in\Omega} \left\| \mathbf{x}_i^{(t)} - \bar{\mathbf{x}} \right\|^2} \tag{18}$$

where $\bar{\mathbf{x}}$ is the mean of the ground-truth values over masked positions. R^2 is the primary metric, with values closer to 1 indicating better reconstruction.

All experiments are run with a fixed random seed (42). The Beijing production experiments follow a fixed time-based split (24 training days, 6 held-out evaluation days); the public-benchmark experiments train on multiple one-week sliding windows and evaluate on a held-out window. We report point estimates of the metrics. Because R^2 is computed over the masked positions only, its denominator varies with the mask rate; cross-mask-rate comparisons therefore reflect both reconstruction quality and the composition of the evaluation set. Beijing models train for 600 epochs; public-benchmark models for 500 epochs. Tables report single representative runs; independent runs under an identical configuration differ by roughly ± 0.002 in R^2 (the direction-level 30% configuration, for instance, measures 0.910–0.913 across runs).

3.3 Baselines

We compare MAST-GNN against non-parametric, graph-free temporal, and graph-temporal baselines, all evaluated under an identical random-mask protocol at 30% masking; their definitions and the evaluation protocol are detailed in Appendix B.

3.4 Intersection-Level Reconstruction

Intersection-level reconstruction across mask rates is reported in Table 10; the corresponding analysis is given in Appendix B.

3.4.1 Comparison with Baselines

Table 4 compares MAST-GNN against baselines on the intersection-level view at the 30% mask rate. The granularity analysis behind these comparisons is provided in Appendix B.

3.5 Experiments on Public Benchmark Datasets

To validate the generalizability of MAST-GNN beyond a single-city deployment, we additionally evaluate on two widely-used public traffic benchmarks: METR-LA (Li et al., 2018) and PEMS-BAY (Li et al., 2018).

3.5.1 Adaptation and Results

We adapt MAST-GNN to these datasets by adjusting the temporal resolution ($T = 288$ time steps per day), sensor count ($N = 207$ / $N = 325$), and hidden dimension ($d = 128$, matching Beijing). The same masking strategies are applied without dataset-specific hyperparameter tuning; the benchmarks train on multiple one-week sliding windows (batch size 64) for 500 epochs. At $d = 128$, the model contains approximately 0.42M parameters, comparable to the Beijing configuration.

For PEMS-BAY we use the released adjacency of the DCRNN benchmark (2,694 directed edges including self-loops). For METR-LA we construct adjacency from inter-sensor distances with a 5,000 m threshold, which yields an undirected graph without self-loops (407 edges, average degree 3.9). This self-built graph is sparser than the directed graph commonly used in prior traffic-forecasting studies on METR-LA, so the absolute R^2 values reported for METR-LA are not directly comparable with published numbers from those studies; the METR-LA/PEMS-BAY contrast in this paper is meaningful only internally, under our uniform protocol.

Table 1 reports reconstruction performance at 30% mask rate across five of the seven masking strategies at the full hidden dimension after 500 training epochs. The results reveal three findings that critically refine our understanding of the method's transfer behavior.

The capacity, connectivity, and convergence analysis behind these results is provided in Appendix B.

3.6 Direction-Level Demand Imputation

We train MAST-GNN on the direction-level view, raising the feature dimension to $F{=}12$ so that each turn movement constitutes an independent reconstruction channel, while keeping the architecture, mask protocol, and training schedule identical to the intersection-level experiments (24 training days, 6 held-out days).

Table 1: Reconstruction on Public Benchmarks (30% Mask Rate, 500 Epochs, Hidden Dim=128)

	METR-LA		PEMS-BAY	
Masking Strategy	R^2	**MAE**	R^2	**MAE**
`local_region`	0.7129	0.172	0.8613	0.114
`random`	0.7305	0.178	0.8629	0.117
`congestion`	0.6818	0.172	0.8859	0.099
`peak`	0.6464	0.255	0.6538	0.208
`missing_segment`	0.6257	0.161	0.7631	0.107

The mask-rate sweep for the direction-level view is reported in Table 8 (Appendix B).

3.6.1 Comparison with Baselines and Ablations

Table 5 compares MAST-GNN against baselines at the 30% mask rate. The paired-ablation and baseline analysis corresponding to Table 5 is provided in Appendix B.

Per-turn-movement reconstruction statistics are reported in Table 9 (Appendix B).

3.6.2 Graph Resilience under Cold Start

The complementary roles of temporal periodicity and spatial topology are sharpest when temporal information is unavailable altogether. We emulate a *cold start*, the commissioning of a new intersection with no historical data, by selecting the twenty highest-degree intersections and masking them through *all* training and test slots, forcing their reconstruction from spatial neighbors alone. Table 6 reports the result.

The cold-start analysis behind Table 6, including the quantification of the graph-only recovery gain, is provided in Appendix B.

3.6.3 The Conditional Value of Graph Structure

The cold-start result suggests a sharper reading of the graph-free versus graph-based contrast: spatial structure is not uniformly redundant, but conditionally valuable, its contribution grows as temporal information is withdrawn. We test this directly by sweeping the mask rate and, at each rate, training both the full MAST-GNN and the T-CNN graph-free temporal baseline under an identical protocol. Table 2 reports the R^2 gap between the two.

The sweep results in Table 2 and their operational implications are analyzed in Appendix B.

Table 2: Graph Gain versus Temporal Deprivation (Direction-Level). At each mask rate both models are re-trained from scratch under an otherwise identical protocol; the gain therefore measures adaptation to a given level of temporal deprivation. This adaptation setting differs from the mask-rate robustness sweep of Table 8, where a single model trained at 30% masking is evaluated under increasing test-time mask rates.

Temporal deprivation	**Graph-free**	**MAST-GNN**	**Graph gain**
30% masking (normal)	0.911	0.910	−0.001
50% masking	0.904	0.902	−0.002
70% masking	0.878	0.890	+0.012
90% masking (severe)	0.634	0.857	+0.224
Cold start (no history)	−0.024	0.886	+0.910

Extended analysis and discussion (the complementary mechanisms of masked reconstruction, granularity semantics, the benchmark–production contrast, computational cost, capacity-driven cross-dataset transfer, operational implications, limitations, and data availability) are provided in Appendix A, together with additional experimental detail (production-data characteristics, benchmark

dataset statistics, per-strategy and per-movement results, mask-rate sweeps, and masking-strategy definitions) in Appendix B.

4 CONCLUSION

This paper presented MAST-GNN, a masked autoencoder-based spatio-temporal graph neural network for reconstructing missing traffic measurements, evaluated on real production data at two complementary granularities. The model combines a two-layer GCN spatial encoder, a multi-scale temporal encoder with three parallel branches, and a masked self-supervised training objective.

On navigation-derived data from 95 Beijing signalized intersections, MAST-GNN attains $R^2 = 0.954$ on aggregate flow and $R^2 = 0.912$ on direction-level turn movements at a 30% mask rate, exceeding the strongest temporal baseline and graph-temporal competitors. Three paired ablations localize the gains: masked self-supervision ($+0.62$), multi-scale temporal encoding ($+0.23$ at the direction level, $+0.37$ at the intersection level), and graph propagation under cold start (from -0.024 to 0.886). The cold-start study exposes the central insight of the work: the two information sources are complementary rather than redundant, temporal periodicity dominates under normal operation, whereas spatial topology becomes the sole recovery channel when history is stripped away. A mask-rate sweep corroborates this boundary condition quantitatively: the graph gain over a graph-free temporal baseline is negligible under normal operation and grows with temporal deprivation, reaching $+0.910$ under cold start. The turn movement, as the atomic unit of signal control, makes direction-level reconstruction an act of recovering control semantics rather than merely filling missing values. On public benchmarks, MAST-GNN transfers cleanly ($R^2 = 0.731$ on METR-LA, 0.886 on PEMS-BAY), and the capacity-threshold analysis shows that graph connectivity and feature dimensionality interact multiplicatively.

Future work will explore dynamic graph construction, cross-city transfer, and the coupling of direction-level imputation with downstream signal-timing optimization, so that reconstruction uncertainty propagates explicitly into green-split decisions, closing the loop between data completion and control.

AI USE STATEMENT

(This section is required and does not count toward the page limit.)

In this work, we used generative AI tools to improve the clarity of the abstract and to assist with language editing and copy-editing tasks during manuscript preparation. We have not used generative AI tools for the core methodological design, experiments, or data analysis. Additionally, we used generative AI tools for assistance with LaTeX formatting and reference management. We have reviewed all AI-assisted work: AI-suggested text edits were manually verified against the underlying experimental results, and AI-formatted references were checked against the original sources. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All experiments in this paper are controlled by a fixed random seed and the complete training protocol (Adam optimizer, cosine-annealing schedule with $\eta_{\max} = 0.0015$ and $\eta_{\min} = 10^{-6}$, gradient clipping at norm 1.2, L_2 coefficient 1.5×10^{-5}, hidden dimension 128, 600 epochs) is reported in Section 3. The two public benchmarks (METR-LA, PEMS-BAY) are fully public; dataset statistics, adjacency construction, and the seven masking protocols are described in Section 3.3, with definitions in Appendix B. The Beijing production data cannot be released in raw form (see Data Availability, Appendix A), but we provide a synthetic data generator with statistically equivalent spatio-temporal correlations, the complete preprocessing pipeline, and the model implementation, released in the camera-ready version. Model and optimization details are described in Section 3.4 (Training Procedure).

Ethics statement

This paper analyzes traffic-flow data from signalized intersections. The public benchmarks used (METR-LA, PEMS-BAY) are widely used, openly released datasets. The Beijing production data contains no personally identifiable information; it consists solely of aggregated vehicle-turn-movement counts per five-minute interval. The data was collected and used under an agreement with the Beijing Municipal Commission of Transportation for research purposes. No human subjects, privacy-sensitive records, or harmful applications are involved. The authors see no conflicts of interest.

References

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *Proc. ICML*, pp. 1597–1607, 2020.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*, pp. 4171–4186, 2019.

W. Du, D. Cote, and Y. Liu. Saits: Self-attention-based imputation for time series. *Expert Syst. Appl.*, 219:119871, 2023.

H. Gao, R. Jiang, Z. Dong, J. Deng, Y. Ma, and X. Song. Spatial-temporal-decoupled masked pre-training for spatiotemporal forecasting. In *Proc. IJCAI*, 2024.

J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchler, C. Doersch, B. Avila Pires, D. Guo, M. Gheshlaghi Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Proc. NeurIPS*, 2020.

M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu. Pct: Point cloud transformer. *Comput. Vis. Media*, 7:187–199, 2021.

S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proc. AAAI*, pp. 922–929, 2019.

W. L. Hamilton, R. Ying, and J. Leskovec. Inductive representation learning on large graphs. In *Proc. NeurIPS*, pp. 1024–1034, 2017.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proc. IEEE/CVF CVPR*, pp. 16000–16009, 2022.

Z. Hou, X. Liu, Y. Cen, Y. Dong, H. Yang, C. Wang, and J. Tang. Graphmae: Self-supervised masked graph autoencoders. In *Proc. KDD*, pp. 1766–1776, 2022.

W. Jiang and J. Luo. Graph neural network for traffic forecasting: A survey. *Expert Syst. Appl.*, 207, 2022.

T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *Proc. ICLR*, 2017.

Y. Li, R. Yu, C. Shahabi, and Y. Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *Proc. ICLR*, 2018.

I. Marisca, A. Cini, and C. Alippi. Learning to reconstruct missing data from spatiotemporal graphs with sparse observations. In *Proc. NeurIPS*, 2022.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proc. ICML*, pp. 8748–8763, 2021.

C. Song, Y. Lin, S. Guo, and H. Wan. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In *Proc. AAAI*, pp. 914–921, 2020.

P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks. In *Proc. ICLR*, 2018.

X. Wei, Y. Zhang, S. Wang, X. Zhao, Y. Hu, and B. Yin. Self-attention graph convolution imputation network for spatio-temporal traffic data. *IEEE Trans. Intell. Transp. Syst.*, 25, 2024.

Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang. Graph wavenet for deep spatial-temporal graph modeling. In *Proc. IJCAI*, pp. 1907–1913, 2019.

Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proc. KDD*, pp. 753–763, 2020.

Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.*, 32(1):4–24, 2021.

K. Xu, W. Hu, J. Leskovec, and S. Jegelka. How powerful are graph neural networks? In *Proc. ICLR*, 2019.

X. Yi, Y. Zheng, J. Zhang, and T. Li. St-mvl: Filling missing values in geo-sensory time series data. In *Proc. IJCAI*, pp. 2704–2710, 2016.

B. Yu, H. Yin, and Z. Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proc. IJCAI*, pp. 3634–3640, 2018.

L. Zhao, Y. Song, C. Zhang, Y. Liu, P. Wang, T. Lin, M. Deng, and H. Li. T-gcn: A temporal graph convolutional network for traffic prediction. *IEEE Trans. Intell. Transp. Syst.*, 21(9):3848–3858, 2020.

H. Zhou, S. Zhang, J. Peng, C. Zhang, J. Li, H. Xiong, and W. Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proc. AAAI*, pp. 11106–11115, 2021.

J. Zhou, B. Lü, Z. Liu, S. Pan, X. Feng, H. Wei, G. Zheng, X. Wang, and C. Zhou. MagiNet: Mask-aware graph imputation network for incomplete traffic data. *ACM Trans. Knowl. Discov. Data*, 19, 2025.

A Supplementary Discussion and Analysis

A.1 The Complementary Mechanisms of Masked Reconstruction

Three paired ablations on the direction-level view isolate the contributions of MAST-GNN's core mechanisms. *Masked self-supervision* is the foundation: a GCN autoencoder trained without masking attains $R^2 = 0.061$ but reaches 0.682 once trained to reconstruct masked inputs, a $+0.62$ gain showing that the model must do more than copy its input to succeed. *Multi-scale temporal encoding* contributes the second-largest gain ($+0.23$, from 0.682 to 0.912), and its contribution inverts across granularities: at the intersection level it supplies $+0.37$ (from 0.589 to 0.954), because aggregate flow carries a stronger daily periodicity that no spatial mechanism can substitute. *Graph propagation* contributes the third and most operationally distinct mechanism, whose value is invisible in normal operation but becomes decisive under cold start (Section 3.6.2).

Together these ablations reveal a complementary division of labor: masked self-supervision teaches reconstruction under incompleteness, multi-scale time exploits periodicity, and the road graph provides a resilience channel that activates when temporal information fails. The two-layer GCN depth is deliberate: deeper stacks (3–4 layers) over-smooth node representations, and the 2-layer configuration offers the best receptive-field/distinctiveness trade-off.

A.2 Granularity Determines Task Semantics

Evaluating the same architecture on two strictly nested views of one data source quantifies how reconstruction difficulty tracks semantic granularity rather than raw scale. On aggregate flow MAST-GNN attains $R^2 = 0.954$; on the twelve turn movements it attains $R^2 = 0.912$. The gap is not

a loss of capability but a shift of task: aggregate flow is a smoothed, temporally dominant quantity, whereas each turn movement is sparse, intermittently active, and directly modulated by signal phasing. The per-movement breakdown makes this concrete, left-turn movements are consistently the hardest (mean $R^2 = 0.81$) because their demand is determined by green splits, and straight-through movements the easiest (0.92). Recognizing that the turn movement is the atomic unit of signal control, direction-level reconstruction is best understood as the recovery of control semantics, not merely the filling of missing values.

A.3 Research Benchmarks versus Production Data

The public benchmarks (METR-LA, PEMS-BAY) evaluate reconstruction on clean, single-source, uniformly-missing highway data, where the dominant challenge is sensor-network capacity. Production turn-movement data introduces pathologies that these benchmarks do not capture: structural sparsity (right turns vanish off-peak), isolated nodes (unreachable by message passing), and multi-source integration (navigation demand aligned with signal timing). The cold-start experiment (Section 3.6.2) is possible only on production data, where a newly commissioned intersection has no history by construction. We therefore argue that graph-based imputation evaluation should extend beyond clean benchmarks to production data whose missingness is structured and whose graph is incomplete, if the goal is deployment readiness rather than benchmark ranking.

A.4 Computational Complexity

Table 3 summarizes the computational cost of MAST-GNN and comparable architectures. With 0.42M parameters, MAST-GNN is markedly lighter than attention-based alternatives, and its fully-connected and pooling-based design (rather than multi-head attention) keeps inference fast for real-time deployment. On a single NVIDIA RTX 4090 GPU, MAST-GNN trains on a full month of direction-level signalized-intersection data (95 intersections $\times$ 12 movements $\times$ 6,912 five-minute slots) in roughly 40 seconds for 600 epochs, and reconstructs the held-out six days in a single forward pass.

Table 3: Computational Complexity Comparison

Model	**Params**
GCN-AE	0.18M
ASTGCN	1.80M
GMAN	2.40M
DCRNN	1.20M
MAST-GNN	**0.42M**

A.5 Cross-Dataset Performance Analysis: Capacity-Driven Transfer

The scaling from $d = 32$ to $d = 128$ transforms the cross-dataset performance landscape. At $d = 128$, PEMS-BAY reaches $R^2 = 0.886$ and METR-LA $R^2 = 0.731$, converging to the same performance band as the production turn-movement data ($R^2 = 0.912$). This subsection identifies the mechanisms underlying this non-linear scaling behavior and demonstrates that model capacity, graph connectivity, and feature dimensionality interact in non-trivial ways across datasets.

A.5.1 The Capacity Threshold Effect

The most consequential finding is the existence of a *capacity threshold* below which learning fails entirely on complex graphs. At $d = 32$, PEMS-BAY barely learns ($R^2_{\text{random}} = 0.311$) while METR-LA achieves meaningful reconstruction ($R^2 = 0.622$). This asymmetry is explained by the graph Laplacian spectrum: PEMS-BAY's larger eigengap (due to its denser, more heterogeneous adjacency) requires more eigenmodes to represent spatial dependencies, placing a higher lower bound on the hidden dimension needed to capture graph structure.

At $d = 128$, both thresholds are crossed. PEMS-BAY's R^2 jumps to 0.886 (a $+185\%$ relative improvement on the best strategy), while METR-LA improves to 0.731 ($+17\%$). The differential

response, PEMS-BAY gaining $11\times$ more relative improvement from the same capacity increase, indicates that PEMS-BAY was far below its capacity threshold at $d = 32$, while METR-LA was closer to saturation. This capacity-asymmetric scaling has practical consequences: when deploying the same architecture to a new sensor network, the minimum viable hidden dimension is dictated by graph complexity, not node count alone.

A.5.2 Graph Connectivity as a Multiplier

The inversion of the METR-LA/PEMS-BAY performance ranking at $d = 128$, PEMS-BAY now outperforms METR-LA on all five masking strategies, reveals graph connectivity as an *amplifier* of model capacity. PEMS-BAY's denser adjacency (2694 edges, average degree 8.3) enables 2-layer GCN message-passing to reach a larger fraction of the graph per propagation step. At low hidden dimensions, this denser connectivity contributes noise rather than signal (insufficient capacity to represent the richer message space). At sufficient capacity, the same connectivity becomes a powerful information conduit.

METR-LA's sparser road network (407 edges, average degree 3.9) imposes a structural ceiling that additional capacity cannot overcome. The GCN receptive field at depth 2 covers only approximately 25% of nodes, leaving isolated sensor clusters–particularly in peripheral highway segments–without adequate spatial context. This structural limitation, not algorithmic adequacy, explains the residual R^2 gap between METR-LA and PEMS-BAY at $d = 128$.

A.5.3 Feature Dimensionality: The Residual Bottleneck

With model capacity and graph connectivity accounted for, the remaining gap between the single-feature public benchmark and the twelve-turn-movement production data is attributable to feature dimensionality. PEMS-BAY provides a single real measurement (speed), from which one synthetic occupancy feature is derived. The production data instead provides twelve correlated turn-movement channels that jointly constrain the reconstruction target, and this richer multi-channel manifold, together with dense connectivity, explains why production turn-movement data ($R^2 = 0.912$) and PEMS-BAY ($R^2 = 0.886$) converge to the same performance band.

This feature gap has a specific structural interpretation: the public benchmark setting corresponds to reconstructing a d-dimensional manifold from a 1-dimensional projection, while the production turn-movement setting corresponds to reconstruction from a k-channel embedding where k contains multiple correlated views of the target. The practical upshot is clear: when deploying MAST-GNN to new sensor networks, equipping each sensor with correlated multi-channel measurements would close the majority of the remaining accuracy gap, independent of further architectural refinement.

A.5.4 Convergence Dynamics and Generalization

At $d = 128$, the training and evaluation R^2 values converge to within 1 percentage point on PEMS-BAY ($R^2_{\text{train}} = 0.886$, $R^2_{\text{eval}} = 0.886$ for congestion) and within 12 points on METR-LA ($R^2_{\text{train}} = 0.855$, $R^2_{\text{eval}} = 0.731$). This near-zero generalization gap at sufficient capacity confirms that the architecture does not overfit to training windows. The residual METR-LA gap is attributable to the structural ceiling imposed by sparse connectivity, not to generalization failure.

The convergence timeline is instructive: PEMS-BAY requires approximately 400 epochs to cross $R^2_{\text{train}} = 0.80$ and continues improving through epoch 497. METR-LA converges more gradually, reaching best performance at epoch 472. Both timelines are substantially longer than the 163-epoch convergence observed at $d = 32$, consistent with the larger parameter space requiring more optimization steps. The consistent monotonic improvement across 500 epochs (with episodic fluctuations from strategy rotation) suggests that even longer training could yield marginal additional gains.

A.5.5 Implications for Practitioners

These findings have three concrete implications for deploying graph-based traffic reconstruction models. First, choosing hidden dimension requires estimating the target network's graph complexity, a dense urban grid (PEMS-BAY-like) needs $d \geq 64$, while a sparse highway network (METR-LA-like) operates effectively at $d \geq 32$. Second, graph connectivity is a first-order performance

driver that is often underappreciated; sensor placement should prioritize edge density in the resulting adjacency graph. Third, feature engineering yields the highest return on investment once the capacity threshold is crossed: adding a single correlated measurement (flow rate, at minimal hardware cost) would conservatively add 5–8 R^2 percentage points on speed-only deployments.

A.6 Practical Implications

The proposed MAST-GNN has direct practical implications for ITS deployments:

- At the typical 30% missing-data rate, the model reconstructs turn-movement demand with $R^2 = 0.912$ at the direction level and 0.954 at the intersection level, recovering signal-control-ready observations from partial sensor data.
- Under cold start, commissioning a new intersection with no history, graph propagation restores $R^2 = 0.886$ from spatial topology alone, enabling plug-and-play deployment without a warm-up period.
- The graceful degradation across mask rates (to $R^2 = 0.654$ at 70% masking) provides a predictable operating envelope for network operators planning sensor redundancy.

A.7 Limitations and Future Work

We discuss limitations across four themes.

Static graph structure: The adjacency matrix is constructed from fixed road topology. Effective traffic correlations may differ from physical connectivity due to detours, temporary road closures, or evolving travel patterns. Dynamic graph construction methods (Wu et al., 2019; 2020) could learn time-varying adjacency matrices that capture functional rather than purely structural connectivity.

Geographic and temporal generalization: While the evaluation covers a private Beijing dataset and two public benchmarks (METR-LA, PEMS-BAY), it does not span diverse geographic and regulatory contexts. Transfer learning experiments across cities with different road network morphologies and traffic management regimes would strengthen generalizability claims. The current implementation uses dense PyTorch tensors; migrating to PyTorch Geometric or DGL would enable efficient sparse operations on larger graphs.

Reconstruction vs. forecasting: The model focuses on reconstruction (imputation) rather than forecasting. Extending MAST-GNN to jointly perform imputation and multi-step prediction, potentially via the Informer architecture (Zhou et al., 2021) for long-sequence forecasting, is a promising direction. Integration with real-time adaptive traffic control systems (e.g., SCATS, SCOOT) requires further engineering for streaming data and concept drift.

External factors: Integrating exogenous data, weather conditions, special events, public transit schedules, could improve reconstruction accuracy during anomalous conditions. The Beijing dataset includes only traffic and spatial features; enriching the feature space with external sources remains for future work.

A.8 Data Availability

The Beijing traffic dataset used in this study contains proprietary sensor data provided under agreement with the Beijing Municipal Commission of Transportation. The data cannot be publicly released in raw form. To facilitate reproducibility, we provide: (1) a synthetic data generator that produces traffic patterns with statistically equivalent spatio-temporal correlations, spatial graph structure, and missing-data distributions; (2) the complete preprocessing pipeline and the model implementation, released in the camera-ready version; and (3) the METR-LA and PEMS-BAY experimental configurations, which use fully public datasets. Researchers seeking access to the original Beijing data should contact the corresponding author to discuss data use agreements.

B Additional Background and Experimental Details

B.1 Production Data Characteristics

The direction-level view decomposes every intersection into its twelve turn movements, the atomic unit on which signal timing and capacity analysis operate.

Topology. The road network consists of 95 signalized intersections connected through 177 undirected road segments (351 directed edges), yielding an average node degree of 3.73. Five intersections are isolated, linked to no adjacent segment, and are therefore unreachable by graph message passing.

Structural sparsity. Direction-level demand exhibits pronounced structural sparsity that is absent from the aggregate flow view. Right-turn movements vanish during off-peak and night hours (nonzero rate 20%–28%), whereas straight-through movements remain active throughout the day (nonzero rate 41%–47%, mean 15–17 vehicles per slot). Left-turn demand carries the highest variance because it is directly modulated by signal phasing and green splits. This heterogeneity renders naive spatial interpolation ineffective, as quantified in the direction-level experiments below.

B.2 Extended Related Work

B.3 Traffic Data Imputation Methods

Early approaches to traffic data imputation relied on statistical methods. Historical mean substitution replaces missing values with the average of observations at the same time of day and day of week, implicitly assuming strict periodicity. Linear interpolation estimates missing values from temporally adjacent observations, which works for short gaps but fails for extended outages. K-nearest neighbors (KNN) imputation identifies sensors with similar historical patterns and uses their current readings as proxies. While these methods are simple and interpretable, they cannot capture the complex non-linear spatio-temporal correlations inherent in traffic data.

Matrix factorization approaches decompose the traffic observation matrix into low-rank factors, capturing latent patterns in the data. Multi-task learning frameworks combine imputation with auxiliary prediction objectives; ST-MVL fuses multiple views of spatio-temporal correlations (Yi et al., 2016). However, these methods assume linear relationships and struggle with the highly non-linear dynamics of traffic flow, particularly during congestion transitions and incident-induced disruptions.

Deep learning methods have significantly advanced traffic data imputation. Autoencoder architectures learn compressed representations of traffic patterns and reconstruct missing values through the decoder pathway. Recurrent neural networks (RNNs), particularly bidirectional LSTMs, model temporal dependencies in both forward and backward directions. Attention mechanisms further enhance temporal modeling by allowing the model to selectively focus on relevant time steps. Recent graph attention imputation networks such as SA-GCIN (Wei et al., 2024) combine self-attention with graph convolutions for spatio-temporal traffic data recovery. Self-attention-based imputation methods such as SAITS (Du et al., 2023) demonstrate that transformer architectures with masked self-supervised learning achieve state-of-the-art performance on general multivariate time-series imputation benchmarks, though they do not exploit graph-structured spatial dependencies. Recent graph-based autoencoders combine spatial structure with temporal reconstruction for improved imputation under missing data. Comprehensive surveys by Jiang and Luo (Jiang & Luo, 2022) provide a thorough review of deep learning approaches for traffic data imputation.

B.4 Graph Neural Networks for Traffic Modeling

The road network is naturally represented as a graph, with intersections as nodes and road segments as edges. Graph neural networks (GNNs) have emerged as the dominant paradigm for modeling spatial dependencies in traffic data (Wu et al., 2021). Graph convolutional networks (GCNs) generalize the convolution operation to non-Euclidean domains, enabling feature aggregation along the road topology (Kipf & Welling, 2017). Graph attention networks (GATs) introduce learnable attention coefficients for weighted neighbor aggregation (Veličković et al., 2018), while GraphSAGE enables inductive learning through sampled neighborhood aggregation (Hamilton et al., 2017). The-

oretical analysis by Xu et al. (Xu et al., 2019) characterizes the expressive power of different GNN architectures.

STGCN (Yu et al., 2018) combines spectral graph convolutions with gated temporal convolutions for traffic forecasting. ASTGCN (Guo et al., 2019) further incorporates spatial and temporal attention mechanisms. DCRNN (Li et al., 2018) models traffic flow as a diffusion process on the road graph, capturing both spatial and temporal dependencies through a seq2seq architecture. Graph WaveNet (Wu et al., 2019) introduces an adaptive adjacency matrix to complement the predefined road graph, learning hidden spatial dependencies. T-GCN (Zhao et al., 2020) integrates temporal convolution with graph convolution for efficient traffic prediction. GMAN (Song et al., 2020) employs multi-attention mechanisms across spatial and temporal dimensions. Recent work on dynamic graph learning (Wu et al., 2019) further adapts graph structures to evolving traffic conditions.

Despite their success, most GNN-based traffic models are designed for supervised prediction tasks and assume complete input observations. Their application to the imputation problem, where the input itself is incomplete, remains underexplored. A few recent works have begun to address this gap. SPIN (Marisca et al., 2022) learns to reconstruct missing data from sparse spatiotemporal observations. However, such approaches typically employ only a single masking strategy and do not systematically evaluate diverse structured masking patterns.

B.5 Masked Self-Supervised Learning

Masked self-supervised learning has revolutionized representation learning across modalities. BERT (Devlin et al., 2019) demonstrated that masking random tokens in text and training a transformer to predict them yields powerful language representations. MAE (He et al., 2022) extended this principle to images, showing that masking up to 75% of image patches and reconstructing them produces strong visual features. PCT (Guo et al., 2021) applied masked point modeling to 3D point clouds. Contrastive learning frameworks such as SimCLR (Chen et al., 2020) and BYOL (Grill et al., 2020) provide complementary self-supervised paradigms, while CLIP (Radford et al., 2021) demonstrates the power of large-scale vision-language pre-training. GraphMAE (Hou et al., 2022) extends masked autoencoding to graph-structured data. In the traffic domain, STD-MAE (Gao et al., 2024) applies spatial-temporal-decoupled masked pre-training for forecasting, while MagiNet (Zhou et al., 2025) introduces a mask-aware graph imputation network specifically for incomplete traffic data on METR-LA and PEMS-BAY.

In the traffic domain, the masked SSL paradigm naturally maps to the data imputation problem: masking observed traffic values and training the model to reconstruct them is equivalent to learning an imputation function. The core idea of masked self-supervised learning, introduced by Devlin et al. (Devlin et al., 2019) for text and extended to vision by He et al. (He et al., 2022), randomly masking portions of the input and training the model to reconstruct them. The potential of structured masking strategies, designed to simulate realistic sensor outage patterns such as regional failures and peak-hour data loss, has not been fully explored. Our work bridges this gap by introducing seven domain-specific masking strategies and demonstrating their complementary effects on reconstruction performance.

B.6 Baselines

We compare MAST-GNN against three families of baselines, all evaluated under an identical random-mask protocol at 30% masking:

Non-parametric: (i) zero-fill, which replaces masked values with zero, equivalent to a constant predictor; (ii) node-wise temporal mean, which fills each masked entry with the per-node historical mean; and (iii) spatial k-nearest-neighbor (KNN), which imputes a masked node from the mean of its graph neighbors along the same turn movement.

Graph-free temporal models: (i) SAITS (Du et al., 2023), a masked self-attention imputation network; and (ii) T-CNN, a temporal-only convolutional baseline that processes each of the 95×12 node–direction sequences independently without any inter-node interaction, thereby isolating the marginal value of spatial structure. Both treat the input as a multivariate time series without an explicit road graph.

Graph-temporal models: (i) GCN-AE, a vanilla two-layer GCN autoencoder trained *without* masking, which isolates the contribution of masked self-supervision; and (ii) MTGNN (Wu et al., 2020), which learns a graph-structure embedding and combines it with temporal convolution.

Methods designed for supervised traffic *prediction* (e.g., DCRNN (Li et al., 2018), STGCN (Yu et al., 2018)) are not directly comparable, as they assume complete inputs and optimize a forecasting rather than a masked-reconstruction objective.

B.7 Auxiliary Signal-Control Features

Beyond the reconstruction target, each intersection is annotated with a rich auxiliary feature vector assembled from four families of signal-control covariates. *Per-direction green-split ratios* (12 dimensions) record the fraction of each signal cycle allotted to each turn movement. *Per-direction green-light durations* (12 dimensions) record the corresponding absolute green times. A *direction-level supply–demand matrix* (24 dimensions) pairs twelve demand channels with twelve volume-to-capacity (v/c) ratios obtained from signal timing. Seven *30-day dynamic statistics* of turn-movement demand, daily mean, peak, standard deviation, coefficient of variation, trend slope, kurtosis, and skewness, complete a 55-dimensional feature ensemble ($12 + 12 + 24 + 7 = 55$). This feature ensemble, absent from the single-channel public highway benchmarks, encodes the very signal-timing covariates that anchor turn-movement periodicity, and it underpins our diagnosis that signal phasing, rather than an arbitrary periodicity prior, is the dominant generator of demand structure.

B.8 Intersection-Level Mask-Rate Analysis

We first evaluate MAST-GNN on the intersection-level flow view ($F = 1$), where each intersection carries its aggregate demand, the row-wise sum of its twelve turn movements. Table 10 reports reconstruction across mask rates. At the default 30% mask rate the model attains $R^2 = 0.954$, and degrades gracefully as the mask rate grows; the collapse to $R^2 = 0.524$ at 70% masking marks the point at which fewer than a third of the observations remain to anchor reconstruction.

B.9 Intersection-Level Baseline Analysis

Two observations stand out. First, temporal information dominates: the node-wise temporal mean already attains $R^2 = 0.451$, and the graph-free temporal model (SAITS 0.946) is the strongest baseline, confirming that aggregate flow is governed by a strong daily periodicity. Second, spatial structure alone is weak at this granularity: the graph-only encoder reaches only $R^2 = 0.589$, and adding the multi-scale temporal encoder contributes a $+0.37$ gain ($0.589 \rightarrow 0.954$), larger than the direction-level counterpart, showing that multi-scale time is most decisive for aggregate flow. MAST-GNN exceeds the strongest baseline (SAITS 0.946 vs. MAST-GNN 0.954) while retaining an explicit, interpretable road topology as its spatial prior.

Table 4: Intersection-Level Reconstruction vs. Baselines (30% Mask Rate). The MAST-GNN row is a second evaluation pass of the same configuration reported in the mask-rate sweep (Table 10, RMSE 10.31); the two passes round to $R^2 = 0.954$ and differ by 0.05 in RMSE, within the run-to-run spread noted in the experimental-setup paragraph of Section 3.

Method	R^2	**RMSE (veh/5 min)**
Zero-fill	0.000	48.05
MLP (graph-free)	0.000	47.93
Temporal mean	0.451	35.60
GCN-AE (graph-only)	0.589	30.83
MTGNN (Wu et al., 2020)	0.930	12.65
SAITS (Du et al., 2023)	0.946	11.07
MAST-GNN (Proposed)	**0.954**	**10.36**

Taken together, the intersection-level results establish a reference point: aggregate flow is dominated by temporal periodicity, and the multi-scale temporal encoder supplies the largest single gain. We

next test whether the same architecture and protocol transfer to a strictly nested but structurally sparser reconstruction target, the direction-level view.

B.10 Direction-Level Ablation Analysis

Three paired contrasts localize the sources of accuracy.

Masked self-supervision. A GCN autoencoder trained *without* masking attains only $R^2 = 0.061$: exposed to complete inputs during training, it learns a trivial identity mapping and fails when evaluated on masked data. Masked training alone raises this to $R^2 = 0.682$ (the graph-only encoder), a $+0.62$ gain that isolates masked self-supervision as the mechanism forcing the model to reconstruct from partial evidence.

Multi-scale temporal encoding. Adding the multi-scale temporal encoder to the graph-only model raises R^2 from 0.682 to 0.912 ($+0.23$), inverting the intersection-level ordering where temporal encoding contributed the larger $+0.37$. The inversion is informative: at the direction level, where periodicity is weaker and each turn movement is sparse, spatial message passing contributes a larger share.

Graph propagation versus temporal dominance. The graph-free temporal models (T-CNN 0.910, SAITS 0.904) nearly match the full model, because a strong daily periodicity already anchors reconstruction. Yet the non-parametric spatial baseline (KNN 0.068) and the graph-free MLP (0.069) both collapse, while the graph-only encoder (0.682) is the first model to exceed the temporal-mean baseline (0.572) by a wide margin. MAST-GNN attains $R^2 = 0.912$, exceeding the strongest baseline while adding an explicit, interpretable, and cold-start-robust spatial prior (Section 3.6.2).

Table 5: Direction-Level Reconstruction vs. Baselines (30% Mask Rate)

Method	R^2	**RMSE (veh/5 min)**
Zero-fill	-0.109	10.16
Spatial KNN	0.068	9.31
MLP (graph-free)	0.069	9.27
GCN-AE (no masking)	0.061	9.30
Temporal mean	0.572	6.31
GCN-only (masked)	0.682	5.43
MTGNN (Wu et al., 2020)	0.754	4.77
T-CNN (graph-free, time-only)	0.910	2.89
SAITS (Du et al., 2023)	0.904	2.98
MAST-GNN (Proposed)	**0.912**	**2.85**

B.11 Cold-Start Analysis

Non-parametric spatial interpolation fails entirely ($R^2 = -0.024$): borrowing the same turn movement from a physically adjacent intersection is uninformative, because a left turn at one intersection feeds a different movement at its downstream neighbor. Learned graph propagation inverts this, MAST-GNN restores $R^2 = 0.886$ for the cold-start intersections, within 0.026 of its normal-operation $R^2 = 0.912$, so a new intersection can be brought online from spatial topology alone. This is the resilience property: under normal operation temporal periodicity dominates and spatial topology is nearly redundant, but when history is stripped away, the road graph becomes the sole recovery channel.

Table 6: Cold-Start Reconstruction on 20 Intersections with No History

Method	**Cold-start** R^2	**Normal** R^2
Spatial KNN	-0.024	—
MAST-GNN (graph propagation)	**0.886**	0.913

B.12 Conditional Value of Graph Structure

Under normal operation (30% masking) the graph contributes nothing: T-CNN attains $R^2 = 0.911$ and MAST-GNN 0.910, a marginally negative gap of -0.001. The gap remains flat through 50% masking (-0.002). As masking crosses 70%, the gap turns positive ($+0.012$), and at 90% masking, where fewer than a tenth of the observations remain, the graph-free model collapses to $R^2 = 0.634$ while MAST-GNN holds 0.857, a $+0.224$ gain. The cold-start row closes the spectrum: with temporal statistics stripped entirely, non-parametric spatial interpolation scores -0.024 and graph propagation restores 0.886, a $+0.910$ gain.

The increase of the graph gain with temporal deprivation is the central structural finding of our analysis. It recasts the earlier complementarity in quantitative form: signal timing compiles spatial coordination into a daily periodicity, so the graph is redundant whenever that periodicity is observable, and becomes the sole recovery channel precisely when it is not. This is not a weakness of graph modeling but a boundary condition on its value, and it predicts, operationally, that graph-based reconstruction is most indispensable for low-sampling-rate sensors, intermittently-connected deployments, and newly commissioned intersections.

B.13 Masking Strategy Definitions

B.13.1 Spatial Masking

Spatial masks operate on the node (intersection) dimension, zeroing out all time steps for selected intersections.

Random spatial mask: Each intersection is independently masked with probability p_s. This is the simplest strategy and serves as a baseline.

Critical-node mask: Intersections with higher degree centrality (more connecting roads) are preferentially masked. This simulates failures at major junctions, which are both more impactful and more likely to have redundant sensing.

Congestion-based mask: Intersections currently experiencing congestion (congestion index above a threshold) are masked. This trains the model to reconstruct data at the most operationally critical locations.

Local-region mask: A seed intersection is selected and its k-hop neighborhood in the road graph is masked simultaneously. This simulates regional sensor outages caused by localized infrastructure failures (e.g., power outage affecting a cluster of detectors).

B.13.2 Temporal Masking

Temporal masks operate on the time-step dimension, zeroing out all intersections at selected time steps.

Random temporal mask: Each time step is independently masked with probability p_t.

Peak-hour mask: Morning peak (7:00–9:00) or evening peak (17:00–19:00) periods are masked as contiguous blocks. This simulates scenarios where data collection systems are overloaded during high-traffic periods.

Missing-segment mask: A contiguous segment of L consecutive time steps is masked simultaneously across all intersections, simulating a sustained communication failure or power outage. Our five-minute-resolution datasets, both the production turn-movement data and the public benchmarks, provide the 288 daily time steps required to define meaningful contiguous temporal gaps.

B.14 Benchmark Dataset Description

B.14.1 Dataset Description

METR-LA collects traffic speed data from 207 loop detectors on Los Angeles highways over 4 months (March–June 2012) at 5-minute resolution. **PEMS-BAY** records traffic flow, occupancy,

and speed from 325 sensors in the San Francisco Bay Area over 6 months (January–June 2017) at 5-minute resolution. Both datasets construct sensor adjacency from physical road distances.

Table 7: Dataset Statistics

Dataset	Nodes	Interval	Features
METR-LA	207	5 min	Speed, Occupancy
PEMS-BAY	325	5 min	Speed
Beijing (Ours)	95	5 min	Turn-movement demand (12) + 55-dim signal covariates

B.15 Public Benchmark Capacity and Connectivity Analysis

Finding 1: Model capacity is the dominant bottleneck, not features. At $d = 32$ (CPU-only experiments in preliminary work), PEMS-BAY barely learned ($R^2_{\text{random}} = 0.311$). Scaling to $d = 128$ increases performance to $R^2 = 0.863$, a $+177\%$ relative improvement that exceeds the combined gains from all architectural improvements at fixed capacity. METR-LA shows a more modest $+17\%$ gain ($0.622 \rightarrow 0.731$), indicating that its sparser adjacency graph (407 edges vs. PEMS-BAY's 2,369 non-self edges) imposes a more fundamental information bottleneck.

Finding 2: PEMS-BAY outperforms METR-LA at sufficient capacity. This inverts the preliminary expectation, PEMS-BAY has 57% more sensors ($N = 325$) and the Bay Area's heterogeneous traffic patterns, yet achieves higher reconstruction accuracy on every strategy. The explanation emerges from graph structure: PEMS-BAY's substantially denser adjacency provides more spatial message-passing pathways, enabling the GCN to propagate information more effectively from observed to masked nodes. This demonstrates that graph connectivity, not node count, is the primary spatial bottleneck.

Finding 3: Training and evaluation R^2 values converge at sufficient capacity. At $d = 128$, the gap between best training R^2 and evaluation R^2 collapses to near zero for both datasets (PEMS-BAY: $R^2_{\text{train}} = 0.886$, $R^2_{\text{eval}} = 0.886$ on congestion; METR-LA: 0.855, 0.731). This rules out overfitting as a confounding factor at low capacity and confirms that the 32→128 scaling primarily addresses underfitting rather than introducing memorization.

The congestion masking strategy achieves the highest R^2 on PEMS-BAY (0.886) but is mid-ranked on METR-LA (0.682). This asymmetry is consistent with the graph structure hypothesis: congestion-based masking removes the highest-degree nodes, and PEMS-BAY's denser graph provides more alternative paths for information flow. Peak-hour masking remains the most challenging strategy across both datasets, as expected given the higher traffic variability during rush hours.

These results were obtained at $d = 128$ with 500 training epochs, 8–16 sliding windows, and no dataset-specific hyperparameter tuning. The clear monotonic learning curves and near-zero train/eval gap at convergence suggest that longer training could yield additional marginal gains, though the primary benefit of extended epochs is convergence assurance rather than accuracy. The residual gap between METR-LA and PEMS-BAY is attributable to METR-LA's sparse road network representation, its 407 edges (average degree 3.9) provide substantially less spatial context per node than PEMS-BAY's 2,694 edges (average degree 8.3). This structural limitation is dataset-inherent and cannot be overcome by increasing model capacity alone. Future work should explore learned adjacency mechanisms (e.g., adaptive graph learning) to mitigate sparse road network representations.

B.16 Direction-Level Mask-Rate Sweep

B.16.1 Effect of Mask Rate

Table 8 reports reconstruction across mask rates. At the 30% mask rate the model attains $R^2 = 0.912$ with an RMSE of 2.85 vehicles per 5-minute slot, lower than the intersection-level $R^2 = 0.954$ because turn movements are sparser, temporally intermittent, and directly modulated by signal phasing. The degradation to $R^2 = 0.654$ at 70% masking, a rate far beyond the 30% used in training, indicates that reconstruction relies on genuine spatio-temporal structure rather than on artifact fitting.

Table 8: Effect of Mask Rate on Direction-Level Reconstruction (Robustness Sweep). A single model trained at 30% masking is evaluated under higher test-time mask rates; the low R^2 at 70% masking therefore reflects out-of-distribution masking rather than the per-rate re-training setting of Table 2.

Mask Rate	R^2	**RMSE (veh/5 min)**
20%	0.913	2.84
30%	0.912	2.85
50%	0.872	3.44
70%	0.654	5.66

B.17 Per-Turn-Movement Reconstruction

B.17.1 Per-Direction Reconstruction and Isolated Nodes

Table 9 reports R^2 for each of the twelve turn movements, which ranges from 0.764 to 0.923. Straight-through movements reconstruct most reliably (mean $R^2 = 0.92$), whereas left-turn movements are consistently the hardest (mean $R^2 = 0.81$), consistent with their direct dependence on signal phasing and green splits. Right-turn movements occupy a middle band with the widest spread across approaches, reflecting heterogeneous sparsity. Excluding the five isolated intersections, which receive no message-passing information, raises the overall R^2 from 0.912 to 0.915.

Table 9: Per-Turn-Movement Reconstruction R^2 (30% Mask Rate)

Approach	**Straight**	**Left**	**Right**
East	0.911	0.764	0.780
South	0.923	0.830	0.875
West	0.918	0.781	0.904
North	0.923	0.848	0.828

B.18 Intersection-Level Mask-Rate Sweep

Table 10: Intersection-Level Reconstruction across Mask Rates ($F = 1$)

Mask Rate	R^2	**RMSE (veh/5 min)**
20%	0.952	10.51
30%	0.954	10.31
50%	0.878	16.79
70%	0.524	33.14