PAST: A PRIMARY-AUXILIARY SPATIO-TEMPORAL NETWORK FOR TRAFFIC TIME SERIES IMPUTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Traffic time series imputation is essential for the reliability of intelligent transportation systems, yet real-world missingness ranges from scattered random gaps to prolonged fiber gaps at single sensors and block gaps spanning several connected sensors. Existing imputation models infer missing values from relationships among observed measurements, and the available evidence diminishes as gaps become longer and spatially wider. In contrast, timestamps and sensor identities remain available for every entry and encode the recurring temporal and location-specific structure of traffic. Based on this observation, we distinguish *primary patterns*, which are learned from observed measurements and therefore depend on the missingness mask, from *auxiliary patterns*, which are learned from mask-invariant context. We propose the **P**rimary-**A**uxiliary **S**patio-**T**emporal network (PAST), which consists of two interacting modules. The Graph-Integrated Module propagates observations over mask-dependent temporal graphs and the road network, and is regularized by an interval-aware edge dropout that encourages reconstruction beyond the nearest observations. The Cross-Gated Module encodes temporal and spatial context through bidirectional gating, injects it into each graph layer, and learns a residual correction of the primary estimate. Experiments on METR-LA, PeMS-Bay, and LargeST-SD under 27 missing configurations show that PAST achieves the lowest RMSE in 53 of 54 settings against seven established baselines and in 17 of 18 settings against CSDI, ImputeFormer, and OPCR. Under fiber and block missingness, PAST reduces RMSE by 16.3% on average and by up to 36.9% relative to the strongest baseline.

1 INTRODUCTION

Traffic sensors support congestion monitoring, signal control, and incident detection in intelligent transportation systems (Hu et al., 2022). Their records are frequently incomplete owing to sensor failures, transmission errors, and environmental disturbances (Cottam et al., 2025; Li et al., 2014), and such gaps can render adaptive signal control unresponsive or delay incident detection (Mei et al., 2023; Tak et al., 2016). Accurate imputation is therefore required for isolated missing readings as well as for the prolonged and spatially concentrated outages that occur in practice.

Missing traffic data are commonly categorized into three types (Chan et al., 2023), as illustrated in Figure 1. *Random* missingness affects scattered entries and typically results from sporadic transmission errors. *Fiber* missingness removes a continuous interval at a single sensor, for example during a detector outage. *Block* missingness removes a continuous interval at several connected sensors, as in a regional failure (Xing et al., 2022). At the same missing rate, a randomly missing value is usually surrounded by observed neighbors in time and space, whereas entries inside a fiber or block gap have few or none.

Existing imputation models differ mainly in how they exploit relationships among measurements, including similarity and regression (Jonsson & Wohlin, 2004; Azur et al., 2011), recurrence, adversarial training, and message passing (Cao et al., 2018; Yoon et al., 2018; Cini et al., 2022), diffusion and transformer architectures (Tashiro et al., 2021; Du et al., 2023; Nie et al., 2024), and attention over the road topology (Xu et al., 2022; Wei et al., 2024; Liang et al., 2022). In all of these models,

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.

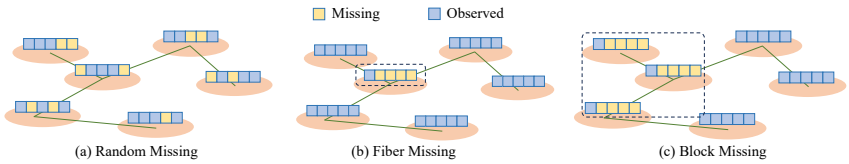


Figure 1: **Three types of missing data in traffic networks.** Random missingness affects isolated entries, fiber missingness a continuous interval at one sensor, and block missingness a continuous interval at a group of connected sensors.

the estimate of a missing entry is inferred from other observed values. Such evidence is abundant for random gaps but becomes scarce as gaps grow longer or wider, since temporal predecessors and spatial neighbors are removed together. Extending the input window to earlier days increases cost and fails when those days are also incomplete.

Some information, however, is unaffected by missingness. The timestamp and the sensor identity are known for every entry, and traffic is strongly organized by both: peak hours recur with daily and weekly schedules, and each location exhibits persistent behavior determined by its road geometry and surroundings. We therefore distinguish the available information according to its dependence on the missingness mask. **Primary patterns** are learned from observed measurements and describe the current traffic state accurately, but degrade as more data are missing. **Auxiliary patterns** are learned from temporal and spatial context and remain fully available inside any gap. The two sources are complementary, and their relative importance depends on the missingness structure, with long or wide gaps relying more heavily on recurring context.

Based on this distinction, we propose the **P**rimary-**A**uxiliary **S**patio-**T**emporal network (PAST). The Graph-Integrated Module (GIM) extracts primary patterns. It represents each incomplete sequence as a mask-dependent temporal graph in which observed entries send messages to missing ones, followed by propagation over the road network, and an interval-aware dropout removes short observed-to-missing edges with higher probability. The Cross-Gated Module (CGM) extracts auxiliary patterns from temporal and spatial embeddings through bidirectional gating, injects them into each graph layer of GIM, and is trained to predict the residual of GIM in the manner of gradient boosting (Ke et al., 2017).

We evaluate PAST on METR-LA, PeMS-Bay, and LargeST-SD under 27 missing configurations against ten baselines. PAST achieves the lowest RMSE in 53 of 54 settings against the seven established baselines, and its average RMSE reduction increases from 8.3% under random missingness to 16.3% under fiber and block missingness. Reorganizing the same network into a temporal and a spatial branch removes a large part of this improvement, and interval-aware dropout reduces the growth of error with gap length.

Our contributions are summarized as follows:

- We organize imputation by information availability, distinguishing mask-dependent primary patterns from mask-invariant auxiliary patterns, and show that this organization is the main source of robustness to structured missingness.
- We design PAST, which couples a Graph-Integrated Module with interval-aware dropout and a Cross-Gated Module through context injection and residual learning.
- We evaluate PAST on three datasets with 27 missing configurations and ten baselines, together with analyses of gap length, dropout rules, alternative decompositions, components, variability, and cost.

2 Related Work

Time series imputation. Classical imputation methods include interpolation, KNN (Jonsson & Wohlin, 2004), and MICE (Azur et al., 2011). Deep imputation models are trained on incomplete data with self-supervision, using bidirectional recurrence in BRITS (Cao et al., 2018), adversarial training in GAIN (Yoon et al., 2018), disentangled temporal representations in TIDER (LIU et al.,

2023), self-attention in SAITS (Du et al., 2023), and conditional diffusion in CSDI, PriSTI, and CoFILL (Tashiro et al., 2021; Liu et al., 2023; He et al., 2025). For spatio-temporal data, GRIN combines message passing with recurrence (Cini et al., 2022), SPIN applies attention over sparse observations (Marisca et al., 2022), ImputeFormer exploits low-rank structure (Nie et al., 2024), and OPCR propagates information from observed to unobserved entries in highly sparse data (Deng et al., 2024). Traffic-specific models introduce cyclic training with self-attention (Xu et al., 2022), self-attention graph convolution (Wei et al., 2024), memory-augmented dynamic graphs with separate temporal and spatial modeling (Liang et al., 2022), semantic graph transformers (Zhang et al., 2024), dynamic graph recurrence (Kong et al., 2023), and mixtures of spatio-temporal attention experts (Wang et al., 2025), and several of them address fiber or block gaps. Most of these models decompose the problem along the temporal and spatial dimensions, and all of them rely on observed measurements. PAST instead decomposes the problem according to information availability and dedicates a module to the context that remains available under any mask.

Context and exogenous information. Calendar information is widely used in forecasting. Informer, Autoformer, and FEDformer incorporate timestamp encodings (Zhou et al., 2021; Wu et al., 2021; Zhou et al., 2022), STID adds spatial and temporal identity embeddings (Shao et al., 2022), and TimeXer embeds exogenous variables (Wang et al., 2024). Imputation models also employ positional or temporal encodings (Du et al., 2023; Tashiro et al., 2021), typically as additional input features of a single network. In PAST, context is processed by a dedicated module that conditions every graph layer and corrects the measurement-driven estimate where observations are absent.

Graph operators under missingness. Spatial layers in traffic models range from fixed diffusion or polynomial filters (Kipf & Welling, 2017; Li et al., 2018; Yu et al., 2018; Wu et al., 2019) to learned aggregators such as GraphSAGE (Hamilton et al., 2017) and attention mechanisms such as GATv2 (Brody et al., 2022). Attention weights are computed from neighbor features, which inside a block gap are themselves estimated. PAST adapts to missingness through the graph structure, whose connectivity is determined by the mask and the road geometry.

3 Problem Formulation

Data and mask. Consider a traffic network with N sensors observed over a window of L time steps, represented by $\boldsymbol{X} \in \mathbb{R}^{L \times N}$, where $x_{t,i}$ denotes the reading of sensor i at step t. The mask $\boldsymbol{M} \in \{0,1\}^{L \times N}$ satisfies $m_{t,i} = 1$ if $x_{t,i}$ is observed and $m_{t,i} = 0$ otherwise. The road network is a weighted graph $G_{\mathcal{S}} = \langle V_{\mathcal{S}}, E_{\mathcal{S}} \rangle$ with adjacency matrix $\boldsymbol{A}_{\mathcal{S}} \in \mathbb{R}^{N \times N}$. Each entry is further associated with a context $c_{t,i} = (\tau_t, i)$, where τ_t contains the week, hour, and minute of step t and i is the sensor identity. The goal is to estimate $x_{t,i}$ for all entries with $m_{t,i} = 0$ given $\boldsymbol{M} \odot \boldsymbol{X}$, $\boldsymbol{M}$, $\boldsymbol{A}_{\mathcal{S}}$, and the contexts.

Missingness types. Given a missing rate r, the three types differ in the shape of the missing set. Random missingness removes each entry independently. Fiber missingness removes continuous intervals at individual sensors, with lengths drawn uniformly up to l steps. Block missingness removes intervals of up to l steps simultaneously at s connected sensors. The three types are nested, since a fiber of length one is a random gap and a block spanning one sensor is a fiber. We therefore treat them as progressively more challenging evaluation settings for the same imputation model.

Primary and auxiliary information. The inputs are divided into two groups according to their relation with the mask. *Primary* information, $(\boldsymbol{M} \odot \boldsymbol{X}, \boldsymbol{M}, \boldsymbol{A}_{\mathcal{S}})$, comprises the measurements and the structure through which they interact, and the amount of information it provides about a missing entry depends on $\boldsymbol{M}$. *Auxiliary* information, $\{c_{t,i}\}$, is identical under every mask. An entry deep inside a block has little primary support in its vicinity, whereas its auxiliary context is as complete as that of any observed entry. PAST extracts *primary patterns* from the first group and *auxiliary patterns* from the second, and learns how to weight them for each entry. This division serves as an inductive bias on where evidence remains under missingness. The two groups are complementary, since auxiliary patterns cannot reflect irregular events such as incidents, and primary patterns provide little support inside long gaps.

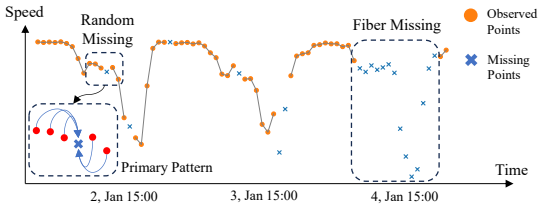


Figure 2: **Traffic speed at a PeMS-Bay sensor** (Li et al., 2018). A random gap (left) can be recovered from adjacent observations through primary patterns. A fiber gap (right) removes local support, and its values follow the recurring daily profile of the sensor.

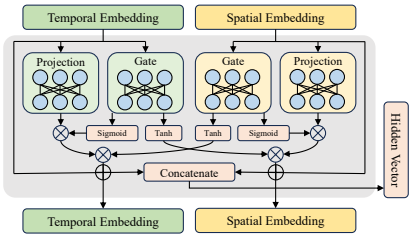


Figure 3: **One cross-gated layer.** Context vectors are projected, gated within their stream by a sigmoid, and modulated across streams by a $\tanh$ gate.

4 METHOD

4.1 OVERVIEW

Figure 2 illustrates the roles of the two groups of information. A randomly missing reading is largely determined by the surrounding observations. A fiber gap of several hours contains no such observations, but the sensor typically behaves similarly during the same hours on previous days, and PAST learns this recurring structure from the timestamp and sensor identity across all training windows.

The overall architecture is shown in Figure 4. GIM embeds the incomplete sequence together with its mask and processes it with n graph-integrated layers, while CGM embeds the sensor identities and timestamps and processes them with n cross-gated layers. The ℓ-th layer of CGM passes a hidden vector to the ℓ-th layer of GIM, so that context conditions the propagation of measurements at every depth. Linear decoders map the final states to $\boldsymbol{Y}_{\mathrm{GIM}}, \boldsymbol{Y}_{\mathrm{CGM}} \in \mathbb{R}^{L\times N}$, and the completed sequence retains the observations and sums the two predictions at missing entries:

$$\boldsymbol{Y} = \boldsymbol{M} \odot \boldsymbol{X} + (\mathbf{1} - \boldsymbol{M}) \odot (\boldsymbol{Y}_{\mathrm{GIM}} + \boldsymbol{Y}_{\mathrm{CGM}}). \tag{1}$$

Following gradient boosting (Ke et al., 2017), GIM is fitted to the observed values and CGM is fitted to the residual of GIM:

$$\mathcal{L}_{\mathrm{GIM}} = \mathrm{MSE}\left(\boldsymbol{M}\odot\boldsymbol{Y}_{\mathrm{GIM}},\ \boldsymbol{M}\odot\boldsymbol{X}\right), \qquad \mathcal{L}_{\mathrm{CGM}} = \mathrm{MSE}\left(\boldsymbol{M}\odot\boldsymbol{Y}_{\mathrm{CGM}},\ \boldsymbol{M}\odot(\boldsymbol{X}-\boldsymbol{Y}_{\mathrm{GIM}})\right). \tag{2}$$

With this target, the sum in Equation (1) approximates $\boldsymbol{X}$, and CGM models the component of the signal that is not explained by the observations in the window. This component is largest where observations are scarce.

4.2 GRAPH-INTEGRATED MODULE

Each graph-integrated layer (Figure 5) first propagates information along time within each sensor and then across sensors at each time step. The temporal connectivity is determined by the mask, and the spatial connectivity by the road network.

Temporal graph. For sensor i, the temporal graph $G_{\mathcal{T}}^{(i)}$ contains one vertex per time step and a context vertex, indexed by 0, which carries the hidden vector from CGM. A vertex v sends a message to a vertex u if v is observed or v is the context vertex:

$$[\boldsymbol{A}_{\mathcal{T}}^{(i)}]_{uv} = \begin{cases} 1, & v = 0 \text{ or } m_{v,i} = 1, \\ 0, & \text{otherwise}, \end{cases} \qquad u, v \in \{0, 1, \ldots, L\}.$$

Consequently, observed steps form a fully connected bidirectional subgraph, each missing step receives messages from all observed steps, and missing steps do not act as sources. The temporal layer aggregates and projects the vertex features $\boldsymbol{H} \in \mathbb{R}^{(L+1)\times d}$ as

$$\boldsymbol{H}' = \boldsymbol{W}_{\mathcal{T}}^{\top}\left(\boldsymbol{D}_{\mathcal{T}}^{-1}\boldsymbol{A}_{\mathcal{T}}^{(i)}\boldsymbol{H}\right) + \boldsymbol{b}_{\mathcal{T}}, \tag{3}$$

where $\boldsymbol{D}_{\mathcal{T}}$ is the in-degree matrix. Unlike recurrent and convolutional encoders, which require absent inputs to be filled, the temporal graph adapts its support to the mask and connects each missing step directly to all observations in the window.

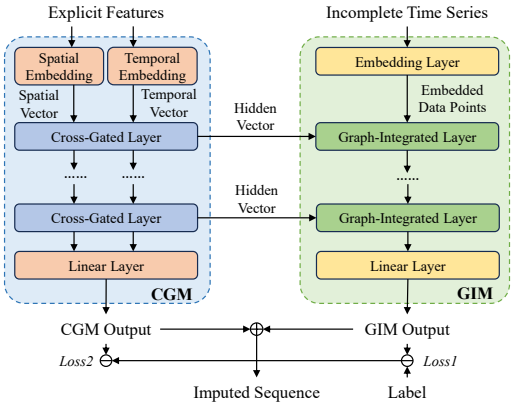


Figure 4: **Framework of PAST.** GIM (right) processes the incomplete series, and CGM (left) processes sensor and time context and sends a hidden vector to each graph-integrated layer. GIM is fitted to the label (*Loss1*) and CGM to the residual of GIM (*Loss2*). The two outputs are summed at missing entries.

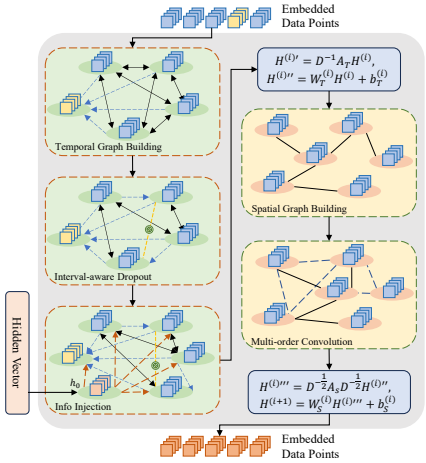


Figure 5: **One graph-integrated layer.** Temporal graph building, interval-aware dropout, and context injection (left) are followed by spatial propagation over the road network (right).

Interval-aware dropout. A model trained on randomly masked windows tends to rely on the closest observations, which are the first to be removed by a fiber or block gap. During training, we therefore drop each observed-to-missing edge with a probability that decays with the temporal distance Δt between its endpoints:

$$p(\Delta t) = e^{-\alpha \Delta t + \beta}, \qquad \beta = \log \frac{\bar{p}\, L^2}{\sum_{u=1}^{L} \sum_{v=1}^{L} e^{-\alpha |u-v|}}. \tag{4}$$

The offset β keeps the mean of p over all pairs of steps at a nominal rate $\bar{p}$, so that α controls the distribution of dropped edges while their overall proportion stays fixed. For $\alpha = 0$, the rule reduces to uniform edge dropout (Srivastava et al., 2014). As α increases, short edges are removed more frequently, and the model is trained on examples in which the nearest evidence is unavailable. Edges within the observed subgraph and edges from the context vertex are not dropped.

Spatial propagation. Road adjacency weights are $a_{ij} = \exp(-d_{ij}^2/\sigma^2)$ for connected sensors and 0 otherwise, where d_{ij} is the distance between sensors and σ is the standard deviation of the distances. Using the normalized operator $\boldsymbol{S} = \boldsymbol{D}_{\mathcal{S}}^{-1/2} \boldsymbol{A}_{\mathcal{S}} \boldsymbol{D}_{\mathcal{S}}^{-1/2}$, the spatial layer applies K rounds of one-hop propagation to the temporal output and fuses all orders:

$$\boldsymbol{Z}_0 = \boldsymbol{H}', \qquad \boldsymbol{Z}_k = \boldsymbol{S}\boldsymbol{Z}_{k-1} \;\; (k = 1, \ldots, K), \qquad \boldsymbol{H}^{\text{next}} = \phi\big([\boldsymbol{Z}_0 \| \boldsymbol{Z}_1 \| \cdots \| \boldsymbol{Z}_K]\, \boldsymbol{W}_{\mathcal{S}} + \boldsymbol{b}_{\mathcal{S}}\big), \tag{5}$$

where $\boldsymbol{S}$ acts along the sensor dimension at every time step and ϕ is a nonlinearity. $\boldsymbol{Z}_k$ aggregates information over k-hop neighborhoods through repeated sparse products, so the dense power $\boldsymbol{A}_{\mathcal{S}}^k$ is never formed. Concatenating orders 0 to K allows the layer to reach sensors outside a block while preserving local features. We adopt a distance-based operator instead of attention (Brody et al., 2022), since attention weights would be computed from neighbor features that are themselves reconstructed inside a block, whereas distances are exact for every sensor.

4.3 Cross-Gated Module

CGM encodes the context of each entry. A sensor embedding table provides the spatial vector $\boldsymbol{v}_s \in \mathbb{R}^d$ of sensor i, and embeddings of week, hour, and minute are concatenated and projected into the temporal vector $\boldsymbol{v}_t \in \mathbb{R}^d$ of step t. A cross-gated layer (Figure 3) refines the pair $(\boldsymbol{v}_s, \boldsymbol{v}_t)$ with four $d \times d$ projections, $\boldsymbol{v}_{sp} = \boldsymbol{W}_{sp}\boldsymbol{v}_s$, $\boldsymbol{v}_{sg} = \boldsymbol{W}_{sg}\boldsymbol{v}_s$, $\boldsymbol{v}_{tp} = \boldsymbol{W}_{tp}\boldsymbol{v}_t$, and $\boldsymbol{v}_{tg} = \boldsymbol{W}_{tg}\boldsymbol{v}_t$, and updates

$$\boldsymbol{v}_s' = \boldsymbol{v}_s + \boldsymbol{v}_{sp} \odot \mathrm{sigmoid}(\boldsymbol{v}_{sg}) \odot \tanh(\boldsymbol{v}_{tg}), \qquad \boldsymbol{v}_t' = \boldsymbol{v}_t + \boldsymbol{v}_{tp} \odot \mathrm{sigmoid}(\boldsymbol{v}_{tg}) \odot \tanh(\boldsymbol{v}_{sg}). \tag{6}$$

The sigmoid gate selects features within a stream, as in a gated linear unit (Dauphin et al., 2017). The $\tanh$ gate of the other stream modulates the update with a signed factor, which allows the time of day to amplify or suppress features of a sensor and vice versa. This mechanism models the interaction between temporal and spatial context, such as a sensor that is congested only during the evening peak, which is difficult to express by summing or concatenating the two embeddings. The four matrices contain $4d^2$ parameters, the same number as a single $2d \times 2d$ linear layer. The refined vectors are passed to the next cross-gated layer, and their concatenation, projected to width d, forms the hidden vector injected into the corresponding graph-integrated layer.

4.4 Training and Complexity

Both losses in Equation (2) are minimized during training, and interval-aware dropout is applied only during training. For n layers, width d, spatial order K, and $|E_{\mathcal{S}}|$ road edges, temporal aggregation in Equation (3) costs $O(nNL^2d)$ in the worst case of a fully observed window, spatial propagation in Equation (5) costs $O(nKL|E_{\mathcal{S}}|d)$, and projections and gating cost $O(n(K+1)NLd^2)$. The temporal term is quadratic in the window length because each missing step is connected to every observation, whereas the spatial term is linear in the number of road edges.

5 Experiments

5.1 Setup

Datasets and missingness. We use METR-LA (207 sensors), PeMS-Bay (325 sensors) (Li et al., 2018), and the San Diego subset of LargeST (716 sensors and 1,739 road edges) (Liu et al., 2023), aggregated to 15-minute intervals (Appendix B). Each dataset is evaluated under nine configurations, namely random missingness with $r \in \{0.2, 0.4, 0.6\}$, fiber missingness with maximum length $l \in \{24, 48, 96\}$ steps (6, 12, and 24 hours), and block missingness with $(l, s) \in \{(48, 5), (48, 10), (96, 5)\}$. Fiber and block configurations use a missing rate of 0.4. The first 80% of each series forms the training period and the remaining 20% the held-out period. Following Cini et al. (2022), we report *in-sample* imputation, which reconstructs missing entries within the training period, and *out-of-sample* imputation, in which a trained model completes windows from the held-out period. Both settings share the masking procedure, so the out-of-sample setting evaluates robustness on unseen data. MAE results are given in Appendix D.

Baselines and implementation. We compare PAST with seven established methods over all 27 configurations, namely KNN (Jonsson & Wohlin, 2004), linear interpolation, MICE (Azur et al., 2011), TIDER (LIU et al., 2023), BRITS (Cao et al., 2018), GRIN (Cini et al., 2022), and the traffic-specific STCPA (Xu et al., 2022). We further compare with three recent methods based on diffusion, low-rank transformers, and sparse-data propagation, namely CSDI (Tashiro et al., 2021), ImputeFormer (Nie et al., 2024), and OPCR (Deng et al., 2024), under a representative configuration of each type. The first four established methods operate on a given incomplete matrix and are evaluated in-sample (Appendix C). PAST uses $n = 3$ layers, width $d = 64$, window length $L = 96$, spatial order $K = 2$, and dropout parameters $\alpha = 0.1$ and $\bar{p} = 0.1$ for all datasets and configurations. It is trained with Adam at a learning rate of 10^{-4} on a single NVIDIA A6000 GPU, and baselines use the hyperparameters recommended by their authors.

5.2 Main Results

Table 1 compares PAST with all ten baselines, and Figure 6a summarizes the improvement over all 27 configurations. Against the seven established baselines, PAST achieves the lowest RMSE in 53 of the 54 dataset–configuration–split settings (Table 6), and in 51 of 54 settings in terms of MAE (Table 7). Against the three recent methods, PAST performs best in 17 of 18 settings. OPCR is the strongest of these methods in-sample and slightly outperforms PAST on random METR-LA (0.2018 versus 0.2082), but its out-of-sample error increases considerably under structured gaps, for example from 0.3790 to 0.5510 on block METR-LA, while that of PAST increases from 0.3090 to 0.3442.

Random missingness. PAST achieves the lowest RMSE in eight of the nine in-sample settings and in all nine out-of-sample settings. The only exception is METR-LA at $r = 0.6$, where GRIN

Table 1: **Imputation RMSE under a representative configuration of each missingness type.** All configurations use missing rate 0.4. Non-parametric and transductive baselines are evaluated in-sample only. Bold marks the best result, and the gray row gives the relative change of PAST with respect to the strongest baseline. Results for all 27 configurations are reported in Appendix D.

	Random ($r = 0.4$)			Fiber ($l = 24$)			Block ($l = 48$, $s = 10$)		
Method	METR	PeMS	LargeST	METR	PeMS	LargeST	METR	PeMS	LargeST
In-sample imputation (training period)									
KNN	0.7342	0.9212	0.7858	0.7317	0.9184	0.7851	0.7346	0.9240	0.8052
Linear	0.3167	0.2551	0.1173	0.5910	0.6739	0.3123	0.6883	0.8418	0.5286
MICE	0.3459	0.3329	0.1104	0.5832	0.6683	0.1746	0.6662	0.8096	0.2473
TIDER	1.1140	1.1350	0.8792	1.0400	1.0140	0.8820	1.0920	0.8674	0.9447
BRITS	0.3281	0.3204	0.2301	0.3546	0.3449	0.1981	0.3825	0.3682	0.2106
GRIN	0.2397	0.2224	0.0952	0.3282	0.3504	0.1299	0.6804	0.3700	0.1455
STCPA	0.3030	0.3254	0.1520	0.4283	0.4347	0.1855	0.4707	0.4813	0.1952
CSDI	0.2785	0.2472	0.1090	0.4952	0.5631	0.1957	0.5185	0.6621	0.3263
ImputeFormer	0.3391	0.3338	0.1593	0.4258	0.4463	0.1865	0.5116	0.4975	0.2302
OPCR	**0.2018**	0.2144	0.0954	0.2951	0.3253	0.1210	0.3790	0.3222	0.1323
PAST (ours)	0.2082	**0.1996**	**0.0914**	**0.2867**	**0.3098**	**0.1196**	**0.3090**	**0.3220**	**0.1306**
Relative change	+3.2%	−6.9%	−4.0%	−2.8%	−4.8%	−1.2%	−18.5%	−0.1%	−1.3%
Out-of-sample imputation (held-out period)									
BRITS	0.4475	0.3978	0.2558	0.5449	0.4304	0.2395	0.5459	0.4614	0.2393
GRIN	0.2419	0.2476	0.0931	0.3739	0.3935	0.1391	0.6725	0.4502	0.1590
STCPA	0.3965	0.4064	0.2528	0.5457	0.5447	0.2275	0.6591	0.5588	0.2488
CSDI	0.2711	0.2634	0.1079	0.4774	0.5878	0.1910	0.5499	0.6729	0.3320
ImputeFormer	0.3524	0.3492	0.1633	0.4399	0.4669	0.1926	0.5417	0.5153	0.2337
OPCR	0.2874	0.2553	0.0962	0.5303	0.4548	0.1444	0.5510	0.4749	0.1609
PAST (ours)	**0.2108**	**0.2104**	**0.0916**	**0.3083**	**0.3461**	**0.1282**	**0.3442**	**0.3634**	**0.1404**
Relative change	−12.9%	−15.0%	−1.6%	−17.5%	−12.0%	−7.8%	−36.5%	−19.3%	−11.7%

obtains 0.2754 and PAST 0.2820. The average reductions relative to the strongest baseline are 8.2% in-sample and 8.5% out-of-sample. Since most missing entries have observed neighbors in this regime, local interpolation remains effective and the benefit of auxiliary context is smaller than under structured gaps.

Fiber missingness. PAST achieves the lowest RMSE in all fiber settings, with average reductions of 12.8% in-sample and 18.2% out-of-sample. Methods that rely on local continuity degrade rapidly with the gap length. As l grows from 24 to 96 steps on METR-LA, the in-sample RMSE of linear interpolation increases from 0.5910 to 0.8584 and the out-of-sample RMSE of GRIN from 0.3739 to 0.5224, whereas that of PAST increases from 0.3083 to 0.3870.

Block missingness. Block missingness removes temporal and spatial support simultaneously. PAST achieves the lowest RMSE in every block setting, with average reductions of 14.4% in-sample and 19.9% out-of-sample. Increasing the spatial span from $s = 5$ to $s = 10$ at $l = 48$ leaves the in-sample RMSE of PAST on METR-LA nearly unchanged (0.3114 and 0.3090), whereas that of GRIN rises from 0.4015 to 0.6804. In the out-of-sample setting on METR-LA with $(l, s) = (48, 10)$, PAST reduces RMSE from 0.5459 for the strongest baseline to 0.3442, a reduction of 36.9%.

Out-of-sample robustness and variability. From the training period to the held-out period, the RMSE of PAST increases by 7.7% on average over the 27 configurations, compared with 8.7% for GRIN, 27.6% for STCPA, and 29.1% for BRITS. Since the baselines degrade most under structured gaps, the margin of PAST is larger out-of-sample for every type (Figure 6a). Auxiliary patterns are learned over the entire training period independently of the observations within a window, which contributes to their stability. Across repeated runs, the standard deviation of PAST is at most 0.0063 RMSE, with a median of 0.0016 (Table 11), which is an order of magnitude smaller than its margins under structured missingness.

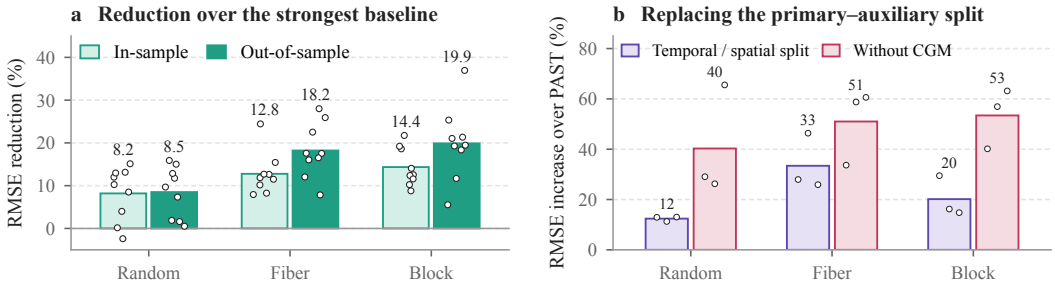


Figure 6: **Analysis of the improvements of PAST.** (a) RMSE reduction of PAST relative to the strongest of the seven established baselines. Bars show means and dots the nine dataset–configuration settings of each type. (b) In-sample RMSE increase when the same network is reorganized into a temporal and a spatial branch, or when CGM is removed. Dots correspond to the three datasets.

Table 2: **Organization and component ablations** (in-sample RMSE; random $r = 0.4$, fiber $l = 48$, block $l = 48$, $s = 10$). T/S split: temporal and spatial branches instead of primary and auxiliary; CG→LL: linear layers instead of cross-gating; G→RNN: recurrent temporal layer. Mean: average increase over PAST. Out-of-sample results are given in Appendix E.

	Random			Fiber			Block			
Variant	METR	PeMS	LargeST	METR	PeMS	LargeST	METR	PeMS	LargeST	Mean
PAST (ours)	**0.2082**	**0.1996**	**0.0914**	**0.3009**	**0.3179**	**0.1193**	**0.3090**	**0.3220**	**0.1306**	–
T/S split	0.2351	0.2221	0.1033	0.3850	0.4653	0.1502	0.4001	0.3742	0.1499	+22.0%
w/o CGM	0.2687	0.2520	0.1513	0.4022	0.5047	0.1916	0.4331	0.5053	0.2131	+48.2%
w/o GIM	0.8131	0.4885	0.2185	0.8523	0.5024	0.2195	0.8494	0.4937	0.2137	+132.4%
CG→LL	0.2700	0.2129	0.1094	0.3849	0.3428	0.1353	0.4475	0.3355	0.1608	+19.7%
G→RNN	0.2387	0.2280	0.1012	0.3376	0.3501	0.1440	0.3420	0.3605	0.1348	+12.1%

5.3 Effect of the Primary–Auxiliary Organization

To separate the effect of the organization from that of having two branches, we modify only the inputs of each branch. In the *temporal/spatial split*, the spatial layers are moved from GIM to CGM and the timestamps from CGM to GIM, with all other components unchanged. As shown in Figure 6b and Table 2, this variant performs worse than PAST in all nine settings, with average RMSE increases of 12.4% under random, 33.4% under fiber, and 20.2% under block missingness. The degradation is largest for the structured types, where the auxiliary branch of PAST retains information that is unavailable to a temporal or spatial branch.

Removing CGM leads to the same conclusion. RMSE increases by 40.3% under random missingness and by 51.0% and 53.4% under fiber and block missingness in-sample, and by 29.8%, 35.8%, and 36.3% out-of-sample. Context therefore improves accuracy in every setting, with the largest benefit when local support is lost. Removing GIM causes a much larger degradation (132.4% on average in Table 2), since context alone cannot represent the current traffic state.

The design of each module also matters. Replacing the cross-gating of Equation (6) with linear layers of the same size (CG→LL) increases RMSE by 19.7% on average and by up to 44.8%, which indicates the importance of modeling the interaction between temporal and spatial context. Replacing the graph temporal layer with a recurrent–graph layer (G→RNN) increases RMSE by 12.1%, as recurrence accumulates errors across long gaps.

5.4 Effect of Interval-Aware Dropout

Figure 7a reports the error as a function of the length of the missing interval containing each entry. With uniform dropout ($\alpha = 0$), RMSE increases from 0.1949 for isolated entries to 0.3466 for 12-hour intervals. With interval-aware dropout (Equation (4), $\alpha = 0.1$), it increases only to 0.3009, a

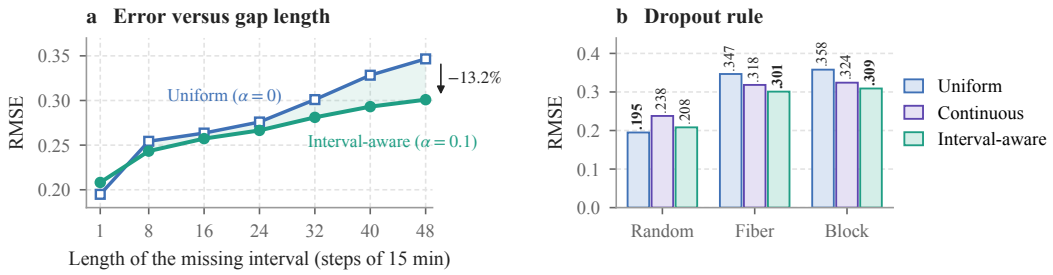


Figure 7: **Effect of interval-aware dropout** (METR-LA, in-sample). (a) RMSE versus the length of the missing interval for uniform and interval-aware dropout at the same nominal rate. (b) RMSE of three dropout rules under random ($r = 0.4$), fiber ($l = 48$), and block ($l = 48$, $s = 10$) missingness.

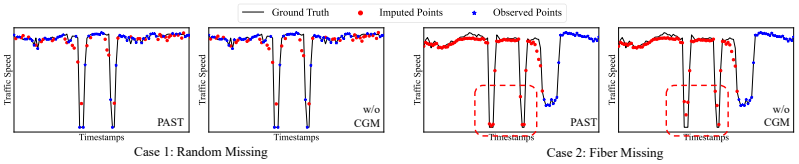


Figure 8: **Recovery examples on METR-LA.** Random gaps (left pair) and a fiber gap (right pair), imputed by PAST and by its variant without CGM. Blue points are observed and red points are imputed.

reduction of 13.2% at the longest length, and the error is lower at every length greater than one step. Only for isolated entries, whose nearest observations are always available, does RMSE increase slightly, by 0.013.

Figure 7b examines whether this benefit stems from the distance-dependent design or from targeting neighborhoods in general. A continuous rule that always removes the three nearest steps on each side of a missing entry also improves fiber and block recovery over uniform dropout, but it discards local evidence required for random gaps and yields the highest error under random missingness (0.2378). Interval-aware dropout removes nearby edges with high probability but not deterministically, performs best under both structured types, and remains within 0.013 of uniform dropout on random gaps, with α controlling the trade-off (Figure 9).

Sensitivity and cost. Spatial propagation is essential under structured missingness. Without it ($K = 0$), fiber RMSE nearly doubles from 0.3009 to 0.5690, and $K = 2$ performs best for fiber and block gaps. A larger width slightly improves structured gaps and degrades random gaps, and $d = 64$ balances the two (Figure 9). One training epoch takes 2.2, 5.7, and 35.3 minutes on the three datasets, with 22.5, 41.1, and 36.9 GB of GPU memory (Appendix H).

Qualitative analysis. Figure 8 shows two windows of one METR-LA sensor. Under random missingness, both PAST and the variant without CGM closely follow the ground truth. In the fiber example, the first 60 steps (about 15 hours) are missing. The variant without CGM deviates noticeably from the ground truth inside the gap, whereas PAST follows the variation within the gap more closely because this variation recurs in the daily profile of the sensor and is captured by its temporal and spatial context.

6 Conclusion

We presented PAST, which combines measurement-driven primary patterns and context-driven auxiliary patterns through a Graph-Integrated Module with interval-aware dropout and a Cross-Gated Module coupled by context injection and residual learning. Its improvements increase with the length and spatial extent of the gaps and are attributable to the separation of information by availability. The principle extends to other series driven by internal dynamics and calendar or entity context, and scaling PAST to state-wide networks is left for future work.

REFERENCES

Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49, 2011.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *International Conference on Learning Representations*, 2022.

Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: Bidirectional recurrent imputation for time series. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31, Montréal,Canada, 2018. Curran Associates, Inc.

Robin Kuok Cheong Chan, Joanne Mun-Yee Lim, and Rajendran Parthiban. Missing traffic data imputation for artificial intelligence in intelligent transportation systems: Review of methods, limitations, and challenges. *IEEE Access*, 11:34080–34093, 2023. doi: 10.1109/ACCESS.2023.3264216.

Andrea Cini, Ivan Marisca, and Cesare Alippi. Filling the g_ap_s: Multivariate time series imputation by graph neural networks. In *International Conference on Learning Representations*, pp. 1–20, Vienna, Austria, 2022. OpenReview.net.

Adrian Cottam, Xiaofeng Li, and Yao-Jan Wu. Vemlan: Imputing missing data for failed freeway traffic sensors. *Journal of Intelligent Transportation Systems*, 0(0):1–15, 2025. doi: 10.1080/15472450.2025.2477785.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 933–941, Sydney, Australia, 06-11 Aug 2017. PMLR.

Leyan Deng, Defu Lian, Chenwang Wu, and Enhong Chen. Learning from highly sparse spatio-temporal data. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Wenjie Du, David Côté, and Yan Liu. Saits: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219:119619, 2023. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2023.119619.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Wenying He, Jieling Huang, Junhua Gu, Ji Zhang, and Yude Bai. Filling the missings: Spatiotemporal data imputation by conditional diffusion. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 2874–2882, 2025.

Hanwen Hu, Yirong Chen, Jingchao Ouyang, Shiyou Qian, Jian Cao, Han Han, Jie Wang, and Michael Lepech. @me: A fine-grained route recommendation system to grab impatient passengers. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, Padua, Italy, 2022. IEEE. doi: 10.1109/IJCNN55064.2022.9892084.

P. Jonsson and C. Wohlin. An evaluation of k-nearest neighbour imputation using likert data. In *10th International Symposium on Software Metrics, 2004. Proceedings.*, pp. 108–118, Washington DC, United States, 2004. IEEE Computer Society. doi: 10.1109/METRIC.2004.1357895.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, Long Beach, CA, USA, 2017. Curran Associates, Inc.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Xiangjie Kong, Wenfeng Zhou, Guojiang Shen, Wenyi Zhang, Nali Liu, and Yao Yang. Dynamic graph convolutional recurrent imputation network for spatiotemporal traffic missing data. *Knowledge-Based Systems*, 261:110188, 2023. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2022.110188.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting, 2018. URL https://arxiv.org/abs/1707.01926.

Yuebiao Li, Zhiheng Li, and Li Li. Missing traffic data: comparison of imputation methods. *IET Intelligent Transport Systems*, 8(1):51–57, 2014. doi: https://doi.org/10.1049/iet-its.2013.0052.

Yuebing Liang, Zhan Zhao, and Lijun Sun. Memory-augmented dynamic graph convolution networks for traffic data imputation with diverse missing patterns. *Transportation Research Part C: Emerging Technologies*, 143:103826, 2022. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2022.103826.

Mingzhe Liu, Han Huang, Hao Feng, Leilei Sun, Bowen Du, and Yanjie Fu. PriSTI: A conditional diffusion framework for spatiotemporal imputation. In *IEEE International Conference on Data Engineering*, pp. 1927–1939, 2023.

SHUAI LIU, Xiucheng Li, Gao Cong, Yile Chen, and YUE JIANG. Multivariate time-series imputation with disentangled temporal representations. In *The Eleventh International Conference on Learning Representations*, pp. 1–19, Online, 2023. OpenReview.net.

Xu Liu, Yutong Xia, Yuxuan Liang, Junfeng Hu, Yiwei Wang, LEI BAI, Chao Huang, Zhenguang Liu, Bryan Hooi, and Roger Zimmermann. Largest: A benchmark dataset for large-scale traffic forecasting. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 75354–75371, New Orleans, USA, 2023. Curran Associates, Inc.

Ivan Marisca, Andrea Cini, and Cesare Alippi. Learning to reconstruct missing data from spatiotemporal graphs with sparse observations. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Hao Mei, Junxian Li, Bin Shi, and Hua Wei. Reinforcement learning approaches for traffic signal control under missing data. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23, Macao, P.R.China, 2023. ACM. ISBN 978-1-956792-03-4. doi: 10.24963/ijcai.2023/251.

Tong Nie, Guoyang Qin, Wei Ma, Yuewen Mei, and Jian Sun. ImputeFormer: Low rankness-induced transformers for generalizable spatiotemporal imputation. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024. doi: 10.1145/3637528.3671751.

Zezhi Shao, Zhao Zhang, Fei Wang, Wei Wei, and Yongjun Xu. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, pp. 4454–4458, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392365. doi: 10.1145/3511808.3557702.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1): 1929–1958, January 2014. ISSN 1532-4435.

Sehyun Tak, Soomin Woo, and Hwasoo Yeo. Data-driven imputation method for traffic data in sectional units of road links. *IEEE Transactions on Intelligent Transportation Systems*, 17(6): 1762–1771, 2016. doi: 10.1109/TITS.2016.2530312.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 24804–24816, Online, 2021. Curran Associates, Inc.

Yiming Wang, Hao Peng, Senzhang Wang, Haohua Du, Chunyang Liu, Jia Wu, and Guanlin Wu. STAMImputer: Spatio-temporal attention MoE for traffic data imputation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 3435–3443, 2025.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 469–498, Vancouver, Canada, 2024. Curran Associates, Inc.

Xiulan Wei, Yong Zhang, Shaofan Wang, Xia Zhao, Yongli Hu, and Baocai Yin. Self-attention graph convolution imputation network for spatio-temporal traffic data. *IEEE Transactions on Intelligent Transportation Systems*, 25(12):19549–19562, 2024. doi: 10.1109/TITS.2024.3461735.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22419–22430, Online, 2021. Curran Associates, Inc.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph WaveNet for deep spatial-temporal graph modeling. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 1907–1913, 2019.

Jiping Xing, Wei Wu, Qixiu Cheng, and Ronghui Liu. Traffic state estimation of urban road networks by multi-source data fusion: Review and new insights. *Physica A: Statistical Mechanics and its Applications*, 595:127079, 2022. ISSN 0378-4371. doi: https://doi.org/10.1016/j.physa.2022.127079.

Qianxiong Xu, Sijie Ruan, Cheng Long, Liang Yu, and Chen Zhang. Traffic speed imputation with spatio-temporal attentions and cycle-perceptual training. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, pp. 2280–2289, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392365. doi: 10.1145/3511808.3557480.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GAIN: Missing data imputation using generative adversarial nets. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5689–5698, Stockholm, Sweden, 10–15 Jul 2018. PMLR.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 3634–3640, Stockholm, Sweden, 7 2018. International Joint Conferences on Artificial Intelligence Organization. doi: 10.24963/ijcai.2018/505. URL https://doi.org/10.24963/ijcai.2018/505.

Kunpeng Zhang, Feng Zhou, Lan Wu, Na Xie, and Zhengbing He. Semantic understanding and prompt engineering for large-scale traffic data imputation. *Information Fusion*, 102:102038, 2024. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.102038.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):11106–11115, May 2021.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27268–27286, Baltimore MD,USA, 2022. PMLR.

A NOTATION

Table 3: **Notation.**

Symbol	Meaning	Symbol	Meaning
L, N	window length, number of sensors	$G_{\mathcal{S}}$, $\boldsymbol{A}_{\mathcal{S}}$	road graph and its adjacency
$\boldsymbol{X}$, $\boldsymbol{M}$	readings and observation mask	$G_{\mathcal{T}}^{(i)}$, $\boldsymbol{A}_{\mathcal{T}}^{(i)}$	temporal graph of sensor i
$c_{t,i} = (\tau_t, i)$	timestamp and sensor identity	$\boldsymbol{S}$	normalized road operator
$\boldsymbol{Y}_{\mathrm{GIM}}$, $\boldsymbol{Y}_{\mathrm{CGM}}$	outputs of the two modules	K	spatial order
$\boldsymbol{Y}$	completed series	α, $\bar{p}$	dropout decay and nominal rate
r	missing rate	n, d	number of layers, width
l, s	maximum gap length, block span	$\boldsymbol{v}_s$, $\boldsymbol{v}_t$	spatial and temporal context vectors

B DATASETS

- **METR-LA** contains traffic speeds from 207 loop detectors on Los Angeles highways, recorded every 5 minutes from March 1 to June 30, 2012 (34,272 steps).
- **PeMS-Bay** contains traffic speeds from 325 sensors in the San Francisco Bay Area, recorded every 5 minutes from January 1 to May 31, 2017 (52,116 steps).
- **LargeST-SD** is the San Diego subset of LargeST (Liu et al., 2023), a benchmark of traffic flow from 8,600 sensors on California highways over 2017–2021. It contains 716 sensors and 1,739 road edges, making it the largest network in our study.

All series are averaged over non-overlapping windows to 15-minute intervals. Road adjacency applies a Gaussian kernel to the distances between connected sensors, as described in Section 4.

C BASELINES AND IMPLEMENTATION

Established baselines. **KNN** (Jonsson & Wohlin, 2004) averages the k most similar observed instances. **Linear** interpolation connects adjacent observations of each sensor. **MICE** (Azur et al., 2011) iterates regressions of each variable on the others. **TIDER** (LIU et al., 2023) learns disentangled trend, seasonality, and bias representations of a multivariate series. **BRITS** (Cao et al., 2018) imputes with bidirectional recurrence. **GRIN** (Cini et al., 2022) combines graph message passing with bidirectional recurrence. **STCPA** (Xu et al., 2022) is a traffic-specific model with cyclic training and self-attention.

Recent baselines. **CSDI** (Tashiro et al., 2021) is a conditional score-based diffusion imputer. **ImputeFormer** (Nie et al., 2024) is a transformer with low-rank inductive biases for spatio-temporal imputation; it serves as the representative attention-based method, as it outperforms earlier self-attention imputers such as SAITS (Du et al., 2023). **OPCR** (Deng et al., 2024) propagates information from observed to unobserved entries for highly sparse spatio-temporal data. These three models are evaluated under the representative configuration of each type in Table 1. MDGCN (Liang et al., 2022) and GT-TDI (Zhang et al., 2024) also address several missingness types and are not included, as no official implementations are available.

Protocol. All methods are evaluated on the same masks and splits. Baselines use the hyperparameters recommended in their papers or official code. PAST uses a single configuration for all datasets and missingness types (Table 4); the only dataset-dependent setting is the batch size, which is reduced on LargeST-SD to fit GPU memory.

D FULL RESULTS

Tables 6 and 7 report RMSE and MAE for all 27 dataset–configuration pairs in both settings. The in-sample columns include all established baselines, and the out-of-sample columns include the learned baselines that can be applied to new windows.

Table 4: **Hyperparameters of PAST.**

Parameter	Symbol	Value	Parameter	Symbol	Value
Layers	n	3	Window length	L	96
Width	d	64	Spatial order	K	2
Dropout decay	α	0.1	Nominal dropout rate	$\bar{p}$	0.1
Optimizer	–	Adam	Learning rate	–	10^{-4}
Batch size (METR, PeMS)	–	32	Batch size (LargeST)	–	8

Table 5: **Missingness types handled and information used by related imputers.**

Model	Missingness types targeted	Information used
KNN, MICE, BRITS, CSDI, STCPA	Random	Measurements
GRIN, GT-TDI	Random, fiber	Measurements
MDGCN	Random, fiber, block	Measurements
PAST	Random, fiber, block	Measurements + context

Table 6: **RMSE for all 27 dataset–configuration pairs.** Random uses $r \in \{0.2, 0.4, 0.6\}$; fiber and block use missing rate 0.4. Bold marks the best result within each setting.

		In-sample								Out-of-sample			
Cfg.	Data	KNN	Linear	MICE	TIDER	BRITS	GRIN	STCPA	**PAST**	BRITS	GRIN	STCPA	**PAST**
Random missingness, r													
0.2	METR	0.7187	0.2860	0.2913	1.1390	0.3109	0.2218	0.2915	**0.1952**	0.4219	0.2191	0.3561	**0.1979**
	PeMS	0.9022	0.2139	0.2385	0.8363	0.2898	0.1900	0.3037	**0.1897**	0.3671	0.2069	0.3876	**0.2030**
	LargeST	0.7828	0.1075	0.0978	0.8812	0.1880	0.0906	0.1536	**0.0769**	0.2244	0.0869	0.1658	**0.0767**
0.4	METR	0.7342	0.3167	0.3459	1.1140	0.3281	0.2397	0.3030	**0.2082**	0.4475	0.2419	0.3965	**0.2108**
	PeMS	0.9212	0.2551	0.3329	1.1350	0.3204	0.2224	0.3254	**0.1996**	0.3978	0.2476	0.4064	**0.2104**
	LargeST	0.7858	0.1173	0.1104	0.8792	0.2301	0.0952	0.1520	**0.0914**	0.2558	0.0931	0.2528	**0.0916**
0.6	METR	0.7646	0.3665	0.4223	0.9362	0.3540	**0.2754**	0.3396	0.2820	0.5011	0.2903	0.4240	**0.2888**
	PeMS	0.9619	0.3259	0.4505	1.0890	0.3697	0.2946	0.3569	**0.2564**	0.4375	0.3177	0.4429	**0.2672**
	LargeST	0.7949	0.1378	0.1316	0.8726	0.2200	0.1069	0.1543	**0.0978**	0.2549	0.1064	0.1862	**0.0986**
Fiber missingness, l													
24	METR	0.7317	0.5910	0.5832	1.0400	0.3546	0.3282	0.4283	**0.2867**	0.5449	0.3739	0.5457	**0.3083**
	PeMS	0.9184	0.6739	0.6683	1.0140	0.3449	0.3504	0.4347	**0.3098**	0.4304	0.3935	0.5447	**0.3461**
	LargeST	0.7851	0.3123	0.1746	0.8820	0.1981	0.1299	0.1855	**0.1196**	0.2395	0.1391	0.2275	**0.1282**
48	METR	0.7320	0.7201	0.6866	1.2060	0.3558	0.3919	0.4641	**0.3009**	0.6054	0.4440	0.6023	**0.3440**
	PeMS	0.9180	0.9008	0.8305	0.8205	0.3607	0.3702	0.4731	**0.3179**	0.4386	0.4355	0.5719	**0.3589**
	LargeST	0.7867	0.5509	0.2630	0.8639	0.2284	0.1579	0.1925	**0.1193**	0.2688	0.1782	0.2437	**0.1283**
96	METR	0.7371	0.8584	0.8220	1.0590	0.3832	0.5006	0.4514	**0.3391**	0.6129	0.5224	0.5457	**0.3870**
	PeMS	0.9233	1.0981	0.9730	0.8674	0.3816	0.4013	0.4583	**0.3501**	0.4625	0.4621	0.5887	**0.3880**
	LargeST	0.7844	0.9371	0.5696	0.8638	0.1878	0.1638	0.1831	**0.1430**	0.2316	0.1974	0.2347	**0.1648**
Block missingness, (l, s)													
(48, 5)	METR	0.7363	0.6969	0.6775	1.0820	0.3826	0.4015	0.4357	**0.3114**	0.5856	0.4391	0.6020	**0.3454**
	PeMS	0.9257	0.8759	0.8384	1.0240	0.3778	0.3808	0.4421	**0.3247**	0.4500	0.4464	0.5332	**0.3595**
	LargeST	0.8070	0.5445	0.2497	0.9059	0.2244	0.1478	0.1875	**0.1348**	0.2510	0.1517	0.2462	**0.1433**
(48, 10)	METR	0.7346	0.6883	0.6662	1.0920	0.3825	0.6804	0.4707	**0.3090**	0.5459	0.6725	0.6591	**0.3442**
	PeMS	0.9240	0.8418	0.8096	0.8674	0.3682	0.3700	0.4813	**0.3220**	0.4614	0.4502	0.5588	**0.3634**
	LargeST	0.8052	0.5286	0.2473	0.9447	0.2106	0.1455	0.1952	**0.1306**	0.2393	0.1590	0.2488	**0.1404**
(96, 5)	METR	0.7348	0.8343	0.7898	1.0210	0.3896	0.4394	0.4475	**0.3414**	0.5444	0.4808	0.6395	**0.3590**
	PeMS	0.9211	1.1007	0.9337	0.9770	0.3859	0.4058	0.4630	**0.3412**	0.4645	0.4702	0.5531	**0.3792**
	LargeST	0.8074	0.9240	0.4380	0.9004	0.2811	0.1766	0.1851	**0.1382**	0.3282	0.1891	0.2412	**0.1493**

Table 7: **MAE for all 27 dataset–configuration pairs.** Settings as in Table 6.

		In-sample								Out-of-sample			
Cfg.	Data	KNN	Linear	MICE	TIDER	BRITS	GRIN	STCPA	**PAST**	BRITS	GRIN	STCPA	**PAST**
Random missingness, r													
0.2	METR	0.4520	0.1284	0.1575	0.8104	0.1316	0.1331	0.1390	**0.1007**	0.2172	0.1314	0.1766	**0.1038**
	PeMS	0.5274	0.1025	0.1297	0.5253	0.1354	**0.0977**	0.1328	0.0981	0.1827	0.1015	0.1762	**0.1008**
	LargeST	0.5729	0.0697	0.0594	0.7271	0.1205	0.0590	0.0790	**0.0488**	0.1365	0.0557	0.1013	**0.0479**
0.4	METR	0.4612	0.1423	0.1945	0.7981	0.1379	0.1280	0.1358	**0.1018**	0.2317	0.1305	0.1805	**0.1036**
	PeMS	0.5376	0.1180	0.1752	0.7159	0.1485	0.1108	0.1436	**0.1002**	0.1961	0.1166	0.1810	**0.1035**
	LargeST	0.5748	0.0750	0.0731	0.7210	0.1481	0.0656	0.0845	**0.0591**	0.1612	0.0633	0.1257	**0.0583**
0.6	METR	0.4768	0.1653	0.2435	0.6769	0.1495	**0.1425**	0.1503	0.1475	0.2570	**0.1517**	0.1890	0.1534
	PeMS	0.5571	0.1446	0.2353	0.6728	0.1701	0.1554	0.1596	**0.1261**	0.2120	0.1599	0.2011	**0.1299**
	LargeST	0.5812	0.0857	0.0831	0.7172	0.1451	0.0726	0.0838	**0.0631**	0.1608	0.0711	0.1432	**0.0627**
Fiber missingness, l													
24	METR	0.4585	0.2926	0.3117	0.7460	0.1472	0.1604	0.1703	**0.1408**	0.2486	0.1819	0.3202	**0.1506**
	PeMS	0.5365	0.3073	0.3218	0.6232	0.1595	0.1692	0.1882	**0.1579**	0.2083	0.1844	0.2278	**0.1696**
	LargeST	0.5743	0.1935	0.1060	0.7262	0.1263	0.0867	0.1131	**0.0753**	0.1464	0.0896	0.1426	**0.0806**
48	METR	0.4598	0.3850	0.3660	0.8525	0.1507	0.2011	0.1677	**0.1466**	0.2942	0.2295	0.3169	**0.1635**
	PeMS	0.5350	0.4338	0.3908	0.5214	0.1755	0.1837	0.1817	**0.1640**	0.2146	0.2085	0.2346	**0.1774**
	LargeST	0.5754	0.3593	0.1498	0.7063	0.1465	0.1098	0.1133	**0.0745**	0.1666	0.1191	0.1411	**0.0915**
96	METR	0.4626	0.4830	0.4415	0.7711	0.1711	0.2924	0.1770	**0.1607**	0.3116	0.3104	0.3272	**0.1890**
	PeMS	0.5383	0.5623	0.4700	0.5470	0.1734	0.2107	0.1907	**0.1606**	0.2247	0.2376	0.2491	**0.1942**
	LargeST	0.5737	0.6541	0.3105	0.7085	0.1208	0.1044	0.1152	**0.0859**	0.1411	0.1184	0.1588	**0.0951**
Block missingness, (l, s)													
(48, 5)	METR	0.4602	0.3671	0.3716	0.7723	0.1609	0.2160	0.1792	**0.1479**	0.2825	0.2378	0.3120	**0.1575**
	PeMS	0.5394	0.4182	0.3987	0.6356	0.1735	0.1927	0.1955	**0.1632**	0.2212	0.2193	0.2529	**0.1759**
	LargeST	0.5910	0.3554	0.1495	0.7338	0.1407	0.0944	0.1152	**0.0843**	0.1549	0.1008	0.1512	**0.0888**
(48, 10)	METR	0.4617	0.3590	0.3617	0.7727	0.1589	0.3403	0.1711	**0.1482**	0.2555	0.3422	0.3272	**0.1647**
	PeMS	0.5380	0.3967	0.3830	0.5825	0.1723	0.1808	0.1839	**0.1623**	0.2248	0.2091	0.2448	**0.1761**
	LargeST	0.5905	0.3408	0.1469	0.7770	0.1391	0.1048	0.1145	**0.0850**	0.1546	0.1112	0.1440	**0.0893**
(96, 5)	METR	0.4603	0.4621	0.4331	0.9567	0.1878	0.2461	0.1806	**0.1689**	0.2607	0.2726	0.3505	**0.1790**
	PeMS	0.5371	0.5614	0.4559	0.6106	0.1780	0.2084	0.1932	**0.1731**	0.2274	0.2373	0.2557	**0.1877**
	LargeST	0.5923	0.6438	0.2425	0.7358	0.1431	0.1188	0.1164	**0.0905**	0.1567	0.1248	0.1654	**0.0961**

E Ablations

Variants. *Temporal/spatial split* moves the spatial layers of GIM into CGM and moves the timestamp input from CGM to GIM, leaving all other components unchanged; the branches then encode temporal and spatial structure of the measurements instead of primary and auxiliary information. *w/o CGM* removes the Cross-Gated Module, including context injection and the residual output. *w/o GIM* predicts from context only. *CG→LL* replaces each cross-gated layer by linear layers with ReLU of the same size. *G→RNN* replaces the graph temporal layer of GIM by a hybrid recurrent–graph layer. Tables 8 and 9 report both settings; removing either module increases the error in every setting, and the relative contribution of CGM is largest under fiber and block missingness in both splits.

F Sensitivity

Dropout decay α. Uniform dropout ($\alpha = 0$) performs best for random gaps, and every positive α improves fiber and block recovery, with the best values at 0.1 and 0.2, respectively. The default $\alpha = 0.1$ provides a balanced setting for all three types.

Spatial order K. Without spatial propagation, the error under structured gaps nearly doubles. $K = 1$ recovers most of the accuracy, $K = 2$ performs best for fiber and block gaps, and $K = 3$ slightly improves random gaps while degrading fiber gaps.

Table 8: **Component ablations, RMSE.** CG→LL replaces cross-gating by linear layers with ReLU; G→RNN replaces the graph temporal layer of GIM by a hybrid recurrent–graph layer. Random uses $r = 0.4$, fiber $l = 48$, block $(l, s) = (48, 10)$.

		In-sample					Out-of-sample				
Setting	Data	PAST	-CGM	-GIM	CG→LL	G→RNN	PAST	-CGM	-GIM	CG→LL	G→RNN
Random	METR	**0.2082**	0.2687	0.8131	0.2700	0.2387	**0.2108**	0.2622	1.0633	0.2609	0.2628
	PeMS	**0.1996**	0.2520	0.4885	0.2129	0.2280	**0.2104**	0.2544	0.5382	0.2216	0.2451
	LargeST	**0.0914**	0.1513	0.2185	0.1094	0.1012	**0.0916**	0.1319	0.3265	0.0996	0.1154
Fiber	METR	**0.3009**	0.4022	0.8523	0.3849	0.3376	**0.3440**	0.3854	1.1021	0.3644	0.4459
	PeMS	**0.3179**	0.5047	0.5024	0.3428	0.3501	**0.3589**	0.5219	0.5385	0.3762	0.3927
	LargeST	**0.1193**	0.1916	0.2195	0.1353	0.1440	**0.1283**	0.1922	0.3218	0.1434	0.1625
Block	METR	**0.3090**	0.4331	0.8494	0.4475	0.3420	**0.3442**	0.4043	1.1031	0.4072	0.4351
	PeMS	**0.3220**	0.5053	0.4937	0.3355	0.3605	**0.3634**	0.5192	0.5266	0.3693	0.4045
	LargeST	**0.1306**	0.2131	0.2137	0.1608	0.1348	**0.1404**	0.2085	0.3079	0.1687	0.1543

Table 9: **Component ablations, MAE.** Variants as in Table 8.

		In-sample					Out-of-sample				
Setting	Data	PAST	-CGM	-GIM	CG→LL	G→RNN	PAST	-CGM	-GIM	CG→LL	G→RNN
Random	METR	**0.1018**	0.1325	0.6611	0.1386	0.1252	**0.1036**	0.1183	1.1333	0.1222	0.1548
	PeMS	**0.1002**	0.1246	0.2386	0.1079	0.1197	**0.1035**	0.1262	0.2897	0.1109	0.1244
	LargeST	**0.0591**	0.0956	0.1467	0.0710	0.0728	**0.0583**	0.0964	0.2066	0.0748	0.0702
Fiber	METR	**0.1466**	0.2291	0.7265	0.2353	0.1629	**0.1635**	0.2052	1.2147	0.1874	0.2281
	PeMS	**0.1640**	0.2927	0.2523	0.1969	0.1695	**0.1774**	0.2799	0.5422	0.1979	0.1948
	LargeST	**0.0745**	0.1321	0.1482	0.0875	0.0919	**0.0915**	0.1506	0.2036	0.1045	0.1025
Block	METR	**0.1482**	0.2305	0.7221	0.2593	0.1761	**0.1647**	0.2244	1.2169	0.2471	0.2181
	PeMS	**0.1623**	0.2700	0.2438	0.1762	0.1839	**0.1761**	0.2714	0.2782	0.1877	0.2019
	LargeST	**0.0850**	0.1540	0.1457	0.1125	0.0857	**0.0893**	0.1485	0.1948	0.1090	0.0959

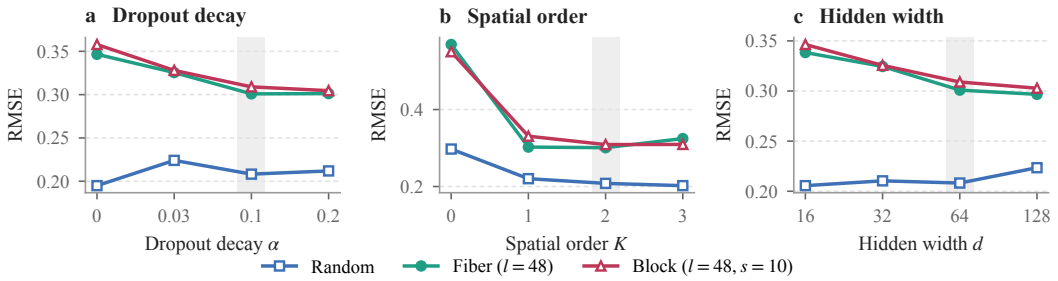


Figure 9: **Sensitivity of PAST to one hyperparameter at a time** (METR-LA, in-sample RMSE). The shaded column marks the default.

Width d. Smaller widths are slightly better for random gaps and larger widths for structured gaps, with modest differences overall. The default $d = 64$ balances the two at moderate cost.

G Run-to-Run Variability

Table 11 reports the standard deviation of RMSE across repeated training runs. The standard deviation of PAST is below 0.0065 in every setting, with a median of 0.0016, whereas those of BRITS, GRIN, and STCPA reach 0.0135, 0.0118, and 0.0190. All are far smaller than the margins of PAST under structured missingness in Table 1.

Table 10: **Sensitivity of PAST** on METR-LA (in-sample; random $r = 0.4$, fiber $l = 48$, block $l = 48$, $s = 10$). One parameter is varied at a time; the default is in bold in the header. Bold in the body marks the best value in each row.

	RMSE				MAE			
Dropout decay α								
Mask	0	0.03	**0.1**	0.2	0	0.03	**0.1**	0.2
Random	**0.1949**	0.2241	0.2082	0.2119	**0.0969**	0.1024	0.1018	0.1001
Fiber	0.3466	0.3255	**0.3009**	0.3012	0.1621	0.1533	0.1466	**0.1445**
Block	0.3579	0.3280	0.3090	**0.3046**	0.1791	0.1600	**0.1482**	0.1517
Spatial order K								
Mask	0	1	**2**	3	0	1	**2**	3
Random	0.2974	0.2204	0.2082	**0.2023**	0.1329	0.1169	0.1018	**0.0987**
Fiber	0.5690	0.3025	**0.3009**	0.3246	0.2974	0.1489	**0.1466**	0.1599
Block	0.5502	0.3308	**0.3090**	0.3094	0.2903	0.1562	0.1482	**0.1459**
Hidden width d								
Mask	16	32	**64**	128	16	32	**64**	128
Random	**0.2056**	0.2104	0.2082	0.2235	**0.1009**	0.1017	0.1018	0.1222
Fiber	0.3384	0.3245	0.3009	**0.2967**	0.1710	0.1592	**0.1466**	0.1493
Block	0.3464	0.3256	0.3090	**0.3028**	0.1788	0.1619	0.1482	**0.1450**

Table 11: **Standard deviation of RMSE across repeated runs** in the representative configuration of each missingness type. KNN and linear interpolation are deterministic.

	Random			Fiber			Block		
Method	METR	PeMS	LargeST	METR	PeMS	LargeST	METR	PeMS	LargeST
KNN	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Linear	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
MICE	0.0002	0.0001	0.0002	0.0001	0.0000	0.0002	0.0002	0.0004	0.0001
TIDER	0.0003	0.0001	0.0006	0.0012	0.0009	0.0006	0.0009	0.0024	0.0002
BRITS	0.0135	0.0101	0.0035	0.0078	0.0066	0.0085	0.0087	0.0045	0.0091
GRIN	0.0012	0.0029	0.0017	0.0102	0.0118	0.0114	0.0093	0.0067	0.0096
STCPA	0.0039	0.0078	0.0043	0.0149	0.0190	0.0154	0.0040	0.0019	0.0039
PAST	0.0023	0.0005	0.0063	0.0044	0.0026	0.0006	0.0009	0.0013	0.0016

Table 12: **Training cost of PAST** on one NVIDIA A6000 GPU (48 GB).

Dataset	Sensors	Batch size	Time per epoch	Peak GPU memory
METR-LA	207	32	2 min 11 s	22.5 GB
PeMS-Bay	325	32	5 min 43 s	41.1 GB
LargeST-SD	716	8	35 min 20 s	36.9 GB

H Computational Cost

Table 12 reports the training cost. Following Section 4, the cost per layer is $O(NL^2d)$ for temporal aggregation, $O(KL|E_S|d)$ for spatial propagation, and $O((K+1)NLd^2)$ for projections and gating. Spatial propagation uses the sparse one-hop operator repeatedly and never forms dense powers of the adjacency, so its cost scales with the number of road edges instead of N^2. The temporal term is quadratic in the window length and linear in the number of sensors. For the larger network of LargeST-SD, the batch size is reduced to fit in GPU memory. Restricting observed-to-missing edges to a temporal neighborhood or sampling them would make this term linear in L, which provides a direct way to scale PAST to state-wide networks.