TOWARDS EFFICIENT, SCALABLE, AND TRANSFERABLE SPATIO-TEMPORAL FORECASTING: A PERIODIC PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Transferable spatio-temporal forecasting seeks to reuse knowledge from data-rich regions when target observations are scarce. Recurring patterns offer transferable information, yet their signal levels, fluctuation amplitudes, and short-term evolution vary across regions. We propose PACT, a graph-free forecasting framework that uses recent observations and future daily-cycle positions to guide the application of shared temporal regularities. PACT predicts local changes relative to each node's latest observation in a normalized space, then refines these predictions using information aggregated across nodes. Its core mechanism conditions a compact global summary on both the future cycle position and the current regional state, enabling node-specific retrieval of context for each forecast step. Shared parameters accommodate different node sets without adjacency matrices or cross-region node correspondence. The compact aggregation and retrieval operations avoid pairwise node interactions, yielding computation that scales linearly with node count at fixed model dimensions. Experiments on real-world traffic datasets demonstrate strong forecasting performance across regions and sensing scales, together with efficient target adaptation and inference. Code is available at https://anonymous.4open.science/r/PACT-F67B/.

1 INTRODUCTION

Spatio-temporal forecasting supports planning and management in urban systems by anticipating changes across locations, with traffic prediction providing an important example (Zheng et al., 2014; Wang et al., 2020). The task is to predict future observations at multiple locations from their historical measurements, accounting for both temporal evolution and dependencies across locations. Accurate prediction typically requires substantial historical data, which may be unavailable in regions with newly established monitoring systems or recently deployed sensors. Learning reliable forecasting models from these limited observations alone is therefore difficult. Transferring knowledge from data-rich regions offers a way to support prediction in such settings (Wang et al., 2019; Yao et al., 2019; Lu et al., 2022), raising the question of which regularities can be reused across regions and how they should be applied to the target observations.

Periodic patterns offer a basis for this sharing: recurring activities, such as daily travel, can produce similar temporal rhythms across regions. Existing cross-city forecasting studies learn reusable traffic patterns and exploit frequency similarities to transfer predictive knowledge (Liu et al., 2023b; 2024b). However, the usefulness of these patterns depends on how they unfold in the target region. Similar daily rhythms can exhibit different signal levels and fluctuation amplitudes, while similar recent trajectories can have different continuations at different positions in the daily cycle. These differences motivate examining transfer from a periodic perspective, considering both the target region's recent observations and the cycle position of the time being predicted.

Figure 1 illustrates these differences using traffic observations. In Figure 1(a), Chengdu and Shenzhen exhibit similar daily shapes after standardization, despite differences in their original traffic levels and fluctuation amplitudes. This suggests that shared temporal shapes should be used together with the target region's observed scale. Figure 1(b) shows two recent trajectories that closely resemble each other but subsequently move in opposite directions at different positions in the daily

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.

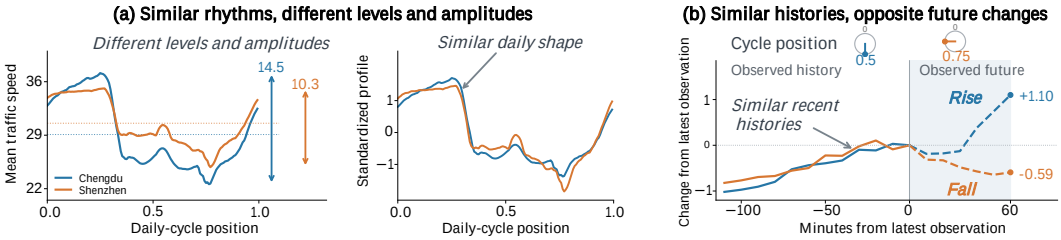

Figure 1: Challenges in transferring periodic patterns. (a) Chengdu and Shenzhen have similar daily shapes but different levels and amplitudes. Dotted lines indicate means, arrows show maximum–minimum ranges, and standardized profiles reveal shape similarity. (b) Two selected Shenzhen windows have similar histories but opposite net changes after 60 minutes. Curves are centered at the latest observation; solid and dashed lines denote observed histories and futures. Circular indicators show the relative daily-cycle positions of the first future samples.

cycle. This case illustrates why recent observations alone may leave the future evolution ambiguous and motivates considering the cycle position of the time being predicted.

To address these challenges, we propose PACT, a graph-free framework for transferable spatio-temporal forecasting. The model normalizes observations using each region's training data and encodes recent changes relative to each node's latest observation. This separates the current reference level from the changes to be predicted while retaining fluctuation magnitudes in normalized units. Observations from other nodes provide additional context about recent changes across the region. PACT combines this encoded context with the future daily-cycle position to condition the aggregation of node representations into a compact global summary. Each node retrieves a correction from the summary to refine its local change forecast, which is added to the latest observation and restored to the original scale. Shared node-wise mappings accommodate different node sets without adjacency matrices or node correspondence, while the compact aggregation keeps computation linear in node count at fixed model dimensions. Our contributions are summarized as follows:

- We propose a framework for transferable spatio-temporal forecasting that combines relative-change prediction with periodic context, grounding the use of shared temporal regularities in target observations.
- We develop conditional global aggregation, where future daily-cycle positions and recent encoded changes jointly determine how node information is summarized. Node-specific retrieval refines local forecasts with parameters independent of node count and computation that scales linearly at fixed model dimensions.
- We evaluate PACT on cross-city traffic forecasting with limited target observations and large-scale regional transfer tasks. Ablation, source-composition, and efficiency studies further examine its components, transfer behavior, and computational cost.

2 Related Work

Spatio-Temporal Forecasting. Spatio-temporal forecasting models temporal evolution and dependencies among spatial locations to predict future observations. Graph-based methods represent these dependencies through information propagation: DCRNN models diffusion on a directed graph, while Graph WaveNet learns an adaptive adjacency matrix (Li et al., 2018; Wu et al., 2019). Periodic structure provides additional temporal information. ASTGCN combines recent, daily, and weekly observations for traffic forecasting (Guo et al., 2019); general time-series methods such as Autoformer and CoST capture seasonal patterns through decomposition and representation learning (Wu et al., 2021; Woo et al., 2022). GMAN uses temporal embeddings to relate historical representations to future prediction steps (Zheng et al., 2020). These studies motivate using both recurring patterns and prediction-time information. Our work investigates how such information can guide cross-node aggregation when a forecasting model is transferred to a new spatial region.

Transferable Spatio-Temporal Forecasting. Transferable spatio-temporal forecasting uses knowledge from source regions to support prediction where observations are limited. Early studies investigate cross-city transfer and meta-learning to reuse knowledge across urban prediction

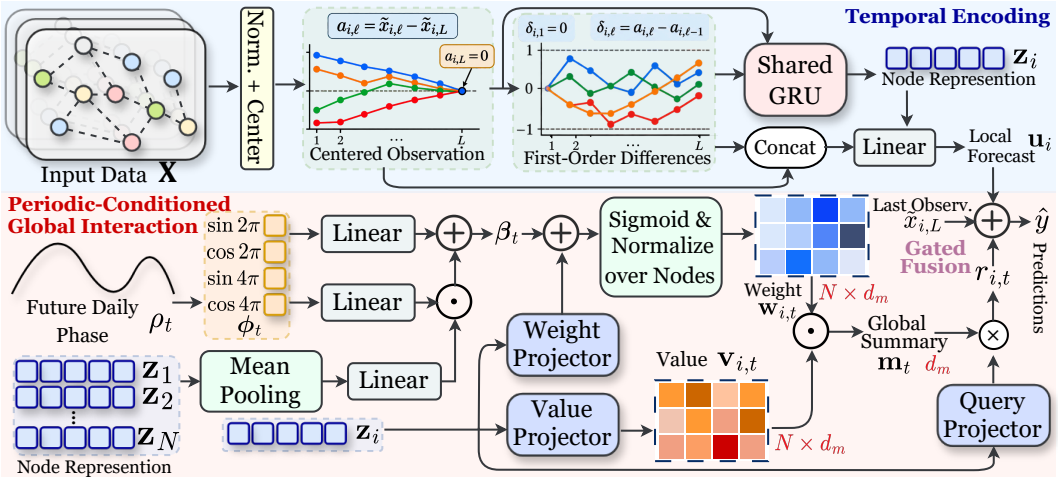

Figure 2: Overview of PACT. Relative-change temporal encoding produces node representations and local forecasts, while future daily-cycle positions and the current city state condition global aggregation. Each node retrieves a correction from the resulting global summary to refine its final prediction.

tasks (Wang et al., 2019; Yao et al., 2019). More recent approaches differ in the knowledge they transfer: ST-GFSL learns node-level meta knowledge for adaptation and graph reconstruction, while TransGTR learns transferable graph structures (Lu et al., 2022; Jin et al., 2023). TPB stores reusable traffic patterns in a learned bank, and FEPCross exploits frequency similarities through frequency-enhanced pretraining (Liu et al., 2023b; 2024b). FlashST uses prompt tuning to adapt pretrained forecasting models (Li et al., 2024). Periodic information is thus already part of cross-region transfer research. PACT focuses on how shared regularities are used during prediction: recent encoded changes and the future daily-cycle position jointly condition a compact global summary, from which each node retrieves a correction to its local forecast.

3 PROBLEM DEFINITION

Problem 1 (Spatio-temporal forecasting). Given L historical steps from N nodes, let $\mathbf{X} \in \mathbb{R}^{L\times N\times C}$ denote the observations and $\mathbf{Y} \in \mathbb{R}^{T\times N\times C}$ the values over the next T steps, where C is the feature dimension. The task is to learn $\hat{\mathbf{Y}} = f_\Theta(\mathbf{X})$, where Θ denotes the model parameters.

Problem 2 (Transferable spatio-temporal forecasting). Let $\mathcal{D}_s$ denote a data-rich source-domain dataset and $\mathcal{D}_{\text{tar}}^{\text{sup}}$ a limited target support set, where each domain represents a region observing the same variables. The goal is to use these data to forecast subsequent target windows with labels disjoint from the support set. Source and target domains may differ in node count, signal scale, and temporal dynamics; no cross-domain node correspondence is assumed.

4 METHODOLOGY

4.1 OVERVIEW

As illustrated in Figure 2, PACT consists of relative-change temporal encoding, conditional global aggregation, and node-specific retrieval. Given normalized historical observations, a shared temporal encoder captures each node's recent changes and produces a local change forecast. For each forecast step, the future daily-cycle position and the mean encoded node state jointly condition the aggregation of node representations into a compact global summary. Each node then retrieves a correction from the corresponding summary according to its own representation, which is combined with the local forecast and latest observation to restore the prediction to the original scale. All modules share parameters across nodes, enabling transfer across cities with different node sets without requiring an adjacency matrix.

4.2 RELATIVE-CHANGE TEMPORAL ENCODING

Encoding Relative Changes. Regions may exhibit similar periodic patterns while differing in signal levels and fluctuation amplitudes. To reduce dependence on absolute signal values, PACT encodes recent changes in a normalized space and anchors forecasts to the latest observation. For brevity, we describe forecasting a single observed variable and omit the feature dimension from the notation. For each domain, observations are normalized as $\widetilde{x}_{i,\ell} = (x_{i,\ell} - \mu)/\sigma_x$, where μ and $\sigma_x > 0$ are the mean and standard deviation estimated from its available training observations, shared across nodes, and fixed during inference. For node i, we center the normalized history at its latest observation and compute first-order differences:

$$a_{i,\ell} = \widetilde{x}_{i,\ell} - \widetilde{x}_{i,L}, \qquad \delta_{i,\ell} = \begin{cases} 0, & \ell = 1, \\ a_{i,\ell} - a_{i,\ell-1}, & \ell > 1. \end{cases} \tag{1}$$

The centered history retains fluctuations relative to the latest observation, while the difference channel explicitly represents successive changes in normalized units.

A shared gated recurrent unit (GRU) processes the two-channel sequence $\mathbf{e}_{i,\ell} = [a_{i,\ell}, \delta_{i,\ell}]^\top$. Its final hidden state provides the node representation:

$$\mathbf{z}_{i,\ell} = \mathrm{GRU}_\theta(\mathbf{e}_{i,\ell}, \mathbf{z}_{i,\ell-1}), \qquad \mathbf{z}_i = \mathbf{z}_{i,L} \in \mathbb{R}^{d_z}, \tag{2}$$

where $\mathbf{z}_{i,0} = \mathbf{0}$ and the encoder parameters θ are shared across nodes and cities. The representation $\mathbf{z}_i$ supports both local prediction and the conditional global aggregation introduced below, where global refers to information aggregated across the observed nodes within the current city.

Local Prediction. The local branch predicts changes over the next T steps by combining the encoded representation with a direct linear projection of the temporal features:

$$\mathbf{u}_i = \mathbf{W}_{\mathrm{loc}}\mathbf{z}_i + \mathbf{W}_{\mathrm{dir}}\mathbf{f}_i + \mathbf{b}_{\mathrm{loc}} \in \mathbb{R}^T, \qquad \mathbf{f}_i = [a_{i,1}, \ldots, a_{i,L}, \delta_{i,2}, \ldots, \delta_{i,L}]^\top. \tag{3}$$

Here, $\mathbf{W}_{\mathrm{loc}} \in \mathbb{R}^{T \times d_z}$ and $\mathbf{W}_{\mathrm{dir}} \in \mathbb{R}^{T \times (2L-1)}$ are learned projections, and $\mathbf{b}_{\mathrm{loc}} \in \mathbb{R}^T$ is a learned bias. The recurrent pathway captures temporal dependencies, while the direct pathway preserves a linear contribution from the observed changes. The resulting $\mathbf{u}_i$ is a local change forecast. It is subsequently refined using global information, added to the latest observation, and transformed back to the city's original scale.

4.3 CONDITIONAL GLOBAL AGGREGATION

Phase and State Conditioning. Similar recent trajectories can have different continuations at different positions in the daily cycle. PACT therefore uses the future cycle position together with recent observations to condition the construction of a global summary for each forecast step. Let $\rho_t \in [0, 1)$ denote the daily-cycle position of the future time associated with forecast step $t \in \{1, \ldots, T\}$. We represent this position using two Fourier harmonics:

$$\boldsymbol{\phi}_t = [\sin(2\pi\rho_t), \cos(2\pi\rho_t), \sin(4\pi\rho_t), \cos(4\pi\rho_t)]^\top \in \mathbb{R}^4. \tag{4}$$

This encoding preserves the circular structure of the daily cycle and requires no future observations. We summarize recent temporal changes across the domain by averaging the encoded node histories, $\mathbf{s} = N^{-1}\sum_{i=1}^{N} \mathbf{z}_i \in \mathbb{R}^{d_z}$. This state and the future phase jointly determine a condition vector:

$$\boldsymbol{\beta}_t = \mathbf{W}_\phi \boldsymbol{\phi}_t + (\mathbf{W}_s \mathbf{s}) \odot (\mathbf{W}_{s\phi} \boldsymbol{\phi}_t) \in \mathbb{R}^{d_m}, \tag{5}$$

where $\odot$ denotes element-wise multiplication, $\mathbf{W}_\phi, \mathbf{W}_{s\phi} \in \mathbb{R}^{d_m \times 4}$ and $\mathbf{W}_s \in \mathbb{R}^{d_m \times d_z}$ are learned projections, and d_m is the summary dimension. The phase term distinguishes future positions in the daily cycle, while the state-dependent term allows the condition at the same cycle position to vary with the observed recent changes.

Global Aggregation. For forecast step t, each node i contributes a value vector and a positive weight vector:

$$\mathbf{v}_{i,t} = \mathbf{W}_t^V \mathbf{z}_i, \qquad \boldsymbol{\kappa}_{i,t} = \sigma(\mathbf{W}_K \mathbf{z}_i + \mathbf{b}_K + \boldsymbol{\beta}_t)\,. \tag{6}$$

Both vectors lie in $\mathbb{R}^{d_m}$. The learned matrices $\mathbf{W}_t^V, \mathbf{W}_K \in \mathbb{R}^{d_m \times d_z}$ and bias $\mathbf{b}_K \in \mathbb{R}^{d_m}$ are shared across nodes and cities; the value projection is specific to step t. The condition $\boldsymbol{\beta}_t$ incorporates the future phase and current city state, and σ applies sigmoid element-wise. We normalize the weights across nodes and aggregate their values:

$$\mathbf{w}_{i,t} = \boldsymbol{\kappa}_{i,t} \oslash \left(\sum_{j=1}^{N} \boldsymbol{\kappa}_{j,t} \right), \qquad \mathbf{m}_t = \sum_{i=1}^{N} \mathbf{w}_{i,t} \odot \mathbf{v}_{i,t}, \tag{7}$$

where $\oslash$ denotes element-wise division. Each entry of the global summary $\mathbf{m}_t$ is therefore a weighted average across nodes. Applying the shared condition before sigmoid allows it to adjust the nodes' relative weights, making the summary depend on the prediction conditions.

Effect of Conditional Aggregation. For the aggregation in equation 7 without the numerical denominator floor, fix the node representations, a forecast step t, and a summary coordinate k. Write $\xi_i = (\mathbf{W}_K \mathbf{z}_i + \mathbf{b}_K)_k$, $v_i = v_{i,t,k}$, and $\beta = \beta_{t,k}$. Let $p_i = \sigma(\xi_i + \beta)$, $w_i = p_i / \sum_{j=1}^{N} p_j$, and $m = \sum_{i=1}^{N} w_i v_i$.

Proposition 4.1 (Conditional concentration and response)**.** *Increasing β preserves the ordering of the node weights. For every $h \in \{1, \ldots, N\}$, the total weight assigned to the h highest-ranked nodes is non-increasing in β. The weights approach* $\mathrm{softmax}(\boldsymbol{\xi})$ *as $\beta \to -\infty$ and uniform weights as $\beta \to +\infty$. Moreover,*

$$\frac{\partial m}{\partial \beta} = -\operatorname{Cov}_w(v, p), \qquad \left| \frac{\partial m}{\partial \beta} \right| \le \frac{\max_i v_i - \min_i v_i}{4}. \tag{8}$$

Here Cov_w denotes covariance over nodes with probabilities w_i. The shared condition thus adjusts weight concentration without changing node rankings, while the summary response depends on the covariance between the aggregated values and sigmoid activations. The response bound has no explicit dependence on the number of nodes. The proof and the treatment of the numerical denominator floor are provided in Appendix I.

4.4 Node-Specific Retrieval

Each node retrieves information from the global summary according to its encoded history. For forecast step t, we construct a query vector and compute a scalar correction:

$$\mathbf{q}_{i,t} = \mathbf{W}_t^Q \mathbf{z}_i, \qquad r_{i,t} = \frac{\mathbf{q}_{i,t}^\top \mathbf{m}_t}{\sqrt{d_m}}, \tag{9}$$

where $\mathbf{W}_t^Q \in \mathbb{R}^{d_m \times d_z}$ is a learned projection specific to forecast step t and shared across nodes and domains. The query determines how the global summary contributes to node i.

We combine the correction with the local change forecast and latest observation, then restore the prediction to the original scale:

$$\hat{y}_{i,t} = \mu + \sigma_x \left(\widetilde{x}_{i,L} + u_{i,t} + \gamma r_{i,t} \right), \tag{10}$$

where $\hat{y}_{i,t}$ denotes the prediction in the original scale for node i at future step t, and γ is a learned scalar controlling the global correction. This formulation anchors the forecast to the node's current level and combines local temporal changes with phase-conditioned global information.

Scale and Shift Equivariance. Equation 10 can equivalently be written as $\hat{y}_{i,t} = x_{i,L} + \sigma_x(u_{i,t} + \gamma r_{i,t})$, so the normalization mean cancels exactly. Under a common transformation $\mathbf{X}' = a\mathbf{X} + b$ with $a > 0$ and $\sigma_x' = a\sigma_x$, the normalized relative-change inputs remain unchanged. With fixed model parameters and unchanged timestamps, both the local forecasts and global corrections therefore remain unchanged, yielding $\widehat{\mathbf{Y}}' = a\widehat{\mathbf{Y}} + b$. This property concerns common changes in signal level and scale rather than differences in temporal dynamics.

4.5 Training and Complexity

Source Pretraining and Target Adaptation. PACT is pretrained on source-city observations and subsequently fine-tuned on the available target-city data. For a dataset $\mathcal{D}$ of historical-input and future-target pairs, we define the normalized forecasting error as

$$\mathcal{L}_{\mathrm{MAE}}(\Theta;\mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{(\mathbf{X},\mathbf{Y})\in\mathcal{D}}\frac{1}{NT}\sum_{t=1}^{T}\left\|\frac{\hat{\mathbf{y}}_t-\mathbf{y}_t}{\sigma_x}\right\|_1, \tag{11}$$

where $\hat{\mathbf{y}}_t, \mathbf{y}_t \in \mathbb{R}^N$ contain the predicted and observed values at forecast step t, and σ_x is the corresponding city's normalization standard deviation. Target adaptation initializes the model from the pretrained parameters and updates all trainable parameters. Each city's normalization statistics are estimated from its training observations and kept fixed during evaluation. Shared node-wise parameters allow adaptation to a different target node set without requiring node correspondence. Training configurations are provided in the experimental setup.

Computational Complexity. For a single input window, temporal encoding costs $O(NLd_z^2)$, and the value and query projections cost $O(NTd_zd_m)$. Weight normalization, aggregation, and retrieval cost $O(NTd_m)$, while the direct local prediction costs $O(NLT)$. Computation therefore scales linearly with N when the history length, forecast horizon, and representation dimensions are fixed. The global summaries occupy $O(Td_m)$ space; node-wise representations and projected values require additional memory linear in N. The model constructs no $N \times N$ adjacency or attention matrix, and its parameter count is independent of the number of nodes.

5 Experiments

5.1 Experimental setup

Datasets. We use four real-world traffic speed datasets: METR-LA, PEMS-BAY, Chengdu, and Shenzhen. They contain 207, 325, 524, and 627 nodes, respectively, with sampling intervals of 5 minutes for METR-LA and PEMS-BAY and 10 minutes for Chengdu and Shenzhen.

Evaluation Protocol. We evaluate four transfer directions: METR-LA $\leftrightarrow$ PEMS-BAY and Chengdu $\leftrightarrow$ Shenzhen. For each direction, PACT is pretrained on source-city data and fine-tuned on the first two days of labeled target observations, with no additional target validation set. Target checkpoints minimize adaptation-set MAE averaged across forecast steps; remaining observations are used for testing. We report original-scale MAE and MAPE at 5/15/30 minutes for METR-LA/PEMS-BAY and 10/30/60 minutes for Chengdu/Shenzhen. MAPE (%) excludes zero ground-truth speeds.

Baselines. We consider three groups of baselines: (1) the training-free historical average method **HA**; (2) conventional forecasting models, including **DCRNN** (Li et al., 2018), **GWNet** (Wu et al., 2019), **STFGNN** (Li & Zhu, 2021), **DLinear** (Zeng et al., 2023), **PatchTST** (Nie et al., 2023) and **iTransformer** (Liu et al., 2024a); and (3) methods for few-shot adaptation or transferable traffic forecasting, including **ST-GFSL** (Lu et al., 2022), **STEP** (Shao et al., 2022), **TransGTR** (Jin et al., 2023), **FlashST** (Li et al., 2024), **TPB** (Liu et al., 2023b), and **FEPCross** (Liu et al., 2024b). The conventional forecasting models use Reptile (Nichol et al., 2018) to obtain source-learned initializations before target adaptation. Baseline hyperparameters follow the configurations recommended in their original papers or official implementations.

5.2 Transferable Forecasting Performance

Table 1 compares forecasting performance on four transfer tasks with two days of target observations. PACT achieves the lowest mean MAE and MAPE at every reported horizon across all four directions, reducing MAE by approximately 5.23% on average relative to the strongest baseline in each setting. The gains are particularly pronounced at the 5-minute horizon: on METR-LA $\rightarrow$ PEMS-BAY, MAE and MAPE decrease by 12.00% and 12.57%, respectively; on PEMS-BAY $\rightarrow$ METR-LA, the corresponding reductions are 8.97% and 9.67%. These results demonstrate the effectiveness of PACT under limited target supervision.

5.3 Ablation Studies

We examine the roles of source pretraining, global aggregation, and node-specific retrieval on Shenzhen $\rightarrow$ Chengdu. Table 2 compares six variants. Target Only trains the full model from random

Table 1: Overall performance comparison on cross-city few-shot forecasting tasks (Mean ± Std over 5 seeds). **Bold** denotes the best result, and underlined denotes the second best.

Model	METR-LA → PEMS-BAY						PEMS-BAY → METR-LA					
	5 min		15 min		30 min		5 min		15 min		30 min	
	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE
HA	3.99	9.33	3.99	9.33	3.99	9.33	5.20	14.15	5.20	14.15	5.20	14.15
DCRNN	$1.64_{\pm 0.04}$	$3.66_{\pm 0.21}$	$2.03_{\pm 0.06}$	$4.45_{\pm 0.21}$	$2.65_{\pm 0.06}$	$5.79_{\pm 0.15}$	$3.12_{\pm 0.14}$	$7.87_{\pm 0.21}$	$3.55_{\pm 0.11}$	$9.42_{\pm 0.25}$	$4.30_{\pm 0.18}$	$11.90_{\pm 0.42}$
GWNet	$1.48_{\pm 0.18}$	$3.67_{\pm 0.70}$	$2.06_{\pm 0.16}$	$5.14_{\pm 0.76}$	$2.71_{\pm 0.23}$	$7.04_{\pm 1.04}$	$2.77_{\pm 0.08}$	$6.66_{\pm 0.48}$	$3.49_{\pm 0.15}$	$9.82_{\pm 0.80}$	$4.35_{\pm 0.21}$	$12.77_{\pm 1.07}$
STFGNN	$1.48_{\pm 0.11}$	$3.21_{\pm 0.50}$	$1.92_{\pm 0.06}$	$4.05_{\pm 0.17}$	$2.65_{\pm 0.13}$	$5.68_{\pm 0.18}$	$3.39_{\pm 0.32}$	$7.57_{\pm 0.46}$	$3.92_{\pm 0.20}$	$9.78_{\pm 0.42}$	$4.51_{\pm 0.25}$	$12.49_{\pm 0.65}$
DLinear	$1.44_{\pm 0.14}$	$2.64_{\pm 0.21}$	$2.08_{\pm 0.07}$	$4.04_{\pm 0.14}$	$2.80_{\pm 0.24}$	$5.63_{\pm 0.46}$	$2.84_{\pm 0.06}$	$6.62_{\pm 0.31}$	$3.66_{\pm 0.22}$	$9.33_{\pm 0.69}$	$4.46_{\pm 0.10}$	$11.86_{\pm 0.42}$
PatchTST	$\underline{1.00}_{\pm \underline{0.03}}$	$\underline{1.91}_{\pm \underline{0.07}}$	$1.61_{\pm 0.02}$	$3.26_{\pm 0.04}$	$2.23_{\pm 0.01}$	$4.73_{\pm 0.02}$	$\underline{2.34}_{\pm \underline{0.04}}$	$\underline{5.48}_{\pm \underline{0.18}}$	$2.93_{\pm 0.02}$	$7.44_{\pm 0.12}$	$3.55_{\pm 0.01}$	$9.56_{\pm 0.10}$
iTransformer	$1.05_{\pm 0.07}$	$2.03_{\pm 0.13}$	$\underline{1.59}_{\pm \underline{0.03}}$	$\underline{3.21}_{\pm \underline{0.05}}$	$2.16_{\pm 0.03}$	$\underline{4.64}_{\pm \underline{0.07}}$	$2.37_{\pm 0.03}$	$5.56_{\pm 0.27}$	$\underline{2.90}_{\pm \underline{0.03}}$	$\underline{7.21}_{\pm \underline{0.15}}$	$\underline{3.48}_{\pm \underline{0.02}}$	$\underline{9.20}_{\pm \underline{0.14}}$
ST-GFSL	$1.40_{\pm 0.04}$	$2.74_{\pm 0.14}$	$1.88_{\pm 0.05}$	$3.85_{\pm 0.08}$	$2.19_{\pm 0.09}$	$4.84_{\pm 0.15}$	$2.75_{\pm 0.03}$	$6.96_{\pm 0.05}$	$3.11_{\pm 0.08}$	$8.77_{\pm 0.11}$	$3.76_{\pm 0.16}$	$11.07_{\pm 0.68}$
STEP	$1.17_{\pm 0.01}$	$2.38_{\pm 0.18}$	$1.75_{\pm 0.02}$	$3.69_{\pm 0.19}$	$2.36_{\pm 0.04}$	$5.29_{\pm 0.16}$	$2.73_{\pm 0.13}$	$7.23_{\pm 1.18}$	$3.31_{\pm 0.01}$	$9.31_{\pm 0.34}$	$4.14_{\pm 0.26}$	$11.88_{\pm 0.54}$
TransGTR	$1.47_{\pm 0.03}$	$3.61_{\pm 0.14}$	$1.89_{\pm 0.02}$	$4.51_{\pm 0.17}$	$2.35_{\pm 0.04}$	$5.57_{\pm 0.20}$	$2.86_{\pm 0.08}$	$7.72_{\pm 0.08}$	$3.31_{\pm 0.03}$	$8.99_{\pm 0.12}$	$4.07_{\pm 0.04}$	$11.51_{\pm 0.61}$
FlashST	$1.33_{\pm 0.06}$	$3.34_{\pm 0.08}$	$1.90_{\pm 0.03}$	$4.11_{\pm 0.16}$	$2.53_{\pm 0.07}$	$5.35_{\pm 0.14}$	$2.70_{\pm 0.11}$	$6.51_{\pm 0.24}$	$3.34_{\pm 0.08}$	$9.04_{\pm 0.20}$	$4.15_{\pm 0.07}$	$11.21_{\pm 0.14}$
TPB	$1.10_{\pm 0.08}$	$2.13_{\pm 0.29}$	$1.60_{\pm 0.07}$	$3.34_{\pm 0.29}$	$\underline{2.09}_{\pm \underline{0.06}}$	$4.78_{\pm 0.31}$	$2.45_{\pm 0.16}$	$5.70_{\pm 0.11}$	$3.01_{\pm 0.10}$	$7.93_{\pm 0.64}$	$3.74_{\pm 0.09}$	$10.85_{\pm 1.11}$
FEPCross	$1.33_{\pm 0.01}$	$2.55_{\pm 0.02}$	$1.85_{\pm 0.01}$	$3.64_{\pm 0.04}$	$2.49_{\pm 0.02}$	$5.06_{\pm 0.07}$	$2.66_{\pm 0.07}$	$6.11_{\pm 0.04}$	$3.45_{\pm 0.01}$	$8.62_{\pm 0.26}$	$4.51_{\pm 0.01}$	$11.51_{\pm 0.20}$
PACT	$\mathbf{0.88}_{\pm \mathbf{0.002}}$	$\mathbf{1.67}_{\pm \mathbf{0.009}}$	$\mathbf{1.45}_{\pm \mathbf{0.01}}$	$\mathbf{2.96}_{\pm \mathbf{0.01}}$	$\mathbf{2.04}_{\pm \mathbf{0.02}}$	$\mathbf{4.39}_{\pm \mathbf{0.03}}$	$\mathbf{2.13}_{\pm \mathbf{0.004}}$	$\mathbf{4.95}_{\pm \mathbf{0.007}}$	$\mathbf{2.71}_{\pm \mathbf{0.003}}$	$\mathbf{6.84}_{\pm \mathbf{0.01}}$	$\mathbf{3.31}_{\pm \mathbf{0.01}}$	$\mathbf{8.83}_{\pm \mathbf{0.02}}$
Improv.	12.00%	12.57%	8.81%	7.79%	2.39%	5.39%	8.97%	9.67%	6.55%	5.13%	4.89%	4.02%

Model	Chengdu → Shenzhen						Shenzhen → Chengdu					
	10 min		30 min		60 min		10 min		30 min		60 min	
	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE
HA	4.43	17.32	4.43	17.32	4.43	17.32	3.77	16.27	3.77	16.27	3.77	16.27
DCRNN	$2.18_{\pm 0.05}$	$8.25_{\pm 0.15}$	$2.69_{\pm 0.04}$	$10.66_{\pm 0.09}$	$2.95_{\pm 0.06}$	$11.38_{\pm 0.21}$	$2.55_{\pm 0.02}$	$10.38_{\pm 0.24}$	$3.23_{\pm 0.03}$	$14.06_{\pm 0.31}$	$3.55_{\pm 0.04}$	$15.25_{\pm 0.23}$
GWNet	$2.25_{\pm 0.04}$	$8.62_{\pm 0.29}$	$2.87_{\pm 0.08}$	$11.46_{\pm 0.44}$	$3.15_{\pm 0.11}$	$12.21_{\pm 0.55}$	$2.57_{\pm 0.07}$	$10.48_{\pm 0.63}$	$3.35_{\pm 0.11}$	$14.73_{\pm 0.99}$	$3.69_{\pm 0.06}$	$15.81_{\pm 0.83}$
STFGNN	$2.55_{\pm 0.08}$	$9.96_{\pm 0.54}$	$3.09_{\pm 0.10}$	$12.60_{\pm 0.81}$	$3.35_{\pm 0.09}$	$13.20_{\pm 0.47}$	$2.90_{\pm 0.11}$	$11.79_{\pm 0.86}$	$3.59_{\pm 0.09}$	$15.77_{\pm 0.98}$	$3.95_{\pm 0.07}$	$16.95_{\pm 0.85}$
DLinear	$2.34_{\pm 0.01}$	$8.79_{\pm 0.06}$	$2.85_{\pm 0.04}$	$11.07_{\pm 0.08}$	$3.00_{\pm 0.05}$	$11.19_{\pm 0.12}$	$2.71_{\pm 0.02}$	$10.75_{\pm 0.08}$	$3.39_{\pm 0.07}$	$14.15_{\pm 0.27}$	$3.45_{\pm 0.09}$	$13.98_{\pm 0.32}$
PatchTST	$1.90_{\pm 0.01}$	$7.36_{\pm 0.02}$	$2.44_{\pm 0.00}$	$9.69_{\pm 0.01}$	$2.86_{\pm 0.00}$	$11.46_{\pm 0.01}$	$2.22_{\pm 0.00}$	$\underline{9.01}_{\pm \underline{0.02}}$	$2.88_{\pm 0.00}$	$12.05_{\pm 0.03}$	$3.41_{\pm 0.00}$	$14.45_{\pm 0.02}$
iTransformer	$\underline{1.88}_{\pm \underline{0.03}}$	$\underline{7.35}_{\pm \underline{0.09}}$	$\underline{2.36}_{\pm \underline{0.02}}$	$\underline{9.37}_{\pm \underline{0.05}}$	$2.74_{\pm 0.02}$	$11.02_{\pm 0.06}$	$2.19_{\pm 0.04}$	$\underline{9.01}_{\pm \underline{0.22}}$	$2.71_{\pm 0.02}$	$\underline{11.51}_{\pm \underline{0.14}}$	$3.14_{\pm 0.04}$	$13.49_{\pm 0.19}$
ST-GFSL	$2.03_{\pm 0.03}$	$8.12_{\pm 0.55}$	$2.46_{\pm 0.07}$	$10.11_{\pm 0.89}$	$2.84_{\pm 0.10}$	$11.82_{\pm 1.00}$	$2.21_{\pm 0.06}$	$10.11_{\pm 0.48}$	$\underline{2.61}_{\pm \underline{0.11}}$	$12.16_{\pm 0.88}$	$2.97_{\pm 0.11}$	$13.94_{\pm 1.03}$
STEP	$1.95_{\pm 0.01}$	$7.61_{\pm 0.09}$	$2.43_{\pm 0.02}$	$9.91_{\pm 0.18}$	$2.72_{\pm 0.07}$	$10.88_{\pm 0.14}$	$2.31_{\pm 0.03}$	$9.21_{\pm 0.12}$	$2.88_{\pm 0.04}$	$12.05_{\pm 0.01}$	$3.16_{\pm 0.04}$	$13.30_{\pm 0.11}$
TransGTR	$2.03_{\pm 0.04}$	$7.94_{\pm 0.09}$	$2.45_{\pm 0.04}$	$9.83_{\pm 0.08}$	$\underline{2.63}_{\pm \underline{0.03}}$	$\underline{10.60}_{\pm \underline{0.04}}$	$2.27_{\pm 0.04}$	$9.51_{\pm 0.13}$	$2.79_{\pm 0.02}$	$12.23_{\pm 0.18}$	$3.02_{\pm 0.03}$	$13.37_{\pm 0.17}$
FlashST	$2.07_{\pm 0.02}$	$8.00_{\pm 0.07}$	$2.47_{\pm 0.02}$	$9.71_{\pm 0.04}$	$2.76_{\pm 0.03}$	$10.82_{\pm 0.04}$	$2.76_{\pm 0.01}$	$9.89_{\pm 0.18}$	$2.71_{\pm 0.03}$	$12.25_{\pm 0.11}$	$3.06_{\pm 0.01}$	$13.79_{\pm 0.03}$
TPB	$2.13_{\pm 0.07}$	$9.46_{\pm 0.13}$	$2.60_{\pm 0.08}$	$11.35_{\pm 0.16}$	$3.01_{\pm 0.08}$	$13.39_{\pm 0.20}$	$\underline{2.18}_{\pm \underline{0.02}}$	$9.18_{\pm 0.13}$	$2.75_{\pm 0.04}$	$12.09_{\pm 0.19}$	$\underline{2.90}_{\pm \underline{0.04}}$	$\underline{12.82}_{\pm \underline{0.27}}$
FEPCross	$2.04_{\pm 0.03}$	$7.74_{\pm 0.05}$	$2.63_{\pm 0.04}$	$10.09_{\pm 0.08}$	$2.88_{\pm 0.03}$	$11.00_{\pm 0.07}$	$2.19_{\pm 0.01}$	$9.19_{\pm 0.09}$	$2.69_{\pm 0.02}$	$11.93_{\pm 0.12}$	$\underline{2.90}_{\pm \underline{0.01}}$	$13.25_{\pm 0.05}$
PACT	$\mathbf{1.80}_{\pm \mathbf{0.003}}$	$\mathbf{7.04}_{\pm \mathbf{0.01}}$	$\mathbf{2.24}_{\pm \mathbf{0.005}}$	$\mathbf{9.09}_{\pm \mathbf{0.02}}$	$\mathbf{2.54}_{\pm \mathbf{0.01}}$	$\mathbf{10.46}_{\pm \mathbf{0.05}}$	$\mathbf{2.08}_{\pm \mathbf{0.004}}$	$\mathbf{8.55}_{\pm \mathbf{0.02}}$	$\mathbf{2.58}_{\pm \mathbf{0.003}}$	$\mathbf{11.18}_{\pm \mathbf{0.02}}$	$\mathbf{2.88}_{\pm \mathbf{0.005}}$	$\mathbf{12.66}_{\pm \mathbf{0.03}}$
Improv.	4.26%	4.22%	5.08%	2.99%	3.42%	1.32%	4.59%	5.11%	1.15%	2.87%	0.69%	1.25%

initialization on target observations, while Local Only retains only temporal encoding and local prediction. For global aggregation, w/o Phase uses only the encoded state to construct the condition, w/o State uses only the future cycle position, and Unconditional removes both conditioning inputs. Shared Query replaces the history-dependent node queries with a learned query shared across nodes at each forecast step.

Table 2: Ablation results on SZ → CD. MAPE is in percentages.

Variant	10 min		30 min		60 min	
	MAE	MAPE	MAE	MAPE	MAE	MAPE
Target Only	2.15	8.79	2.67	11.32	3.08	13.18
Local Only	2.18	8.90	2.82	11.97	3.34	14.33
w/o Phase	2.10	8.60	2.60	11.25	2.99	13.22
w/o State	2.09	8.62	2.58	11.19	2.92	13.01
Unconditional	2.09	8.59	2.59	11.21	2.98	13.14
Shared Query	2.11	8.71	2.60	11.25	2.91	12.82
PACT	**2.08**	**8.55**	**2.58**	**11.18**	**2.88**	**12.66**

Source pretraining and global information both improve forecasting performance: at 60 minutes, Target Only and Local Only yield MAEs of 3.08 and 3.34, respectively, compared with 2.88 for PACT. Removing phase or state conditioning increases the corresponding MAE to 2.99 or 2.92, showing that future cycle position and recent observations both contribute to constructing a useful global summary. Unconditional aggregation similarly performs worse at the longest horizon, while differences at shorter horizons are smaller. Shared Query increases both MAE and MAPE across all three horizons, supporting the use of each node's history to determine how it retrieves global information.

5.4 Robustness to Source Selection

To support multi-source pretraining, we train a shared PACT by alternating batches from different source domains, with each batch containing observations from one domain. Updates are distributed approximately equally across source domains under a fixed total training budget, and the pretrained model is then adapted to the target domain. We evaluate all seven nonempty subsets of the three available source cities for each target. Figure 3 shows that forecasting performance varies only

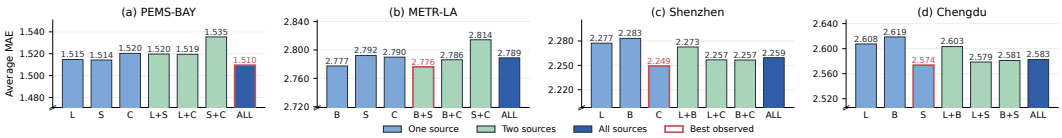

Figure 3: Average MAE over six forecast steps across source combinations. L/B/S/C denote METR-LA/PEMS-BAY/Shenzhen/Chengdu; ALL includes all non-target cities. Red outlines mark the best results. Axes are truncated.

slightly across source combinations: for every target, the MAE range is below 1.8% of the mean across combinations. This limited variation supports the robustness of PACT to source-domain selection across the evaluated settings.

5.5 SCALABILITY ANALYSIS

We evaluate scalability on four regional subsets of LargeST (Liu et al., 2023a): San Diego (SD), Orange County (OC), Los Angeles (LA), and Greater Bay Area (GBA), containing 716, 953, 1,859, and 2,352 nodes, respectively. Unlike the traffic speed prediction tasks above, this benchmark forecasts traffic flow. We consider SD ↔ GBA and LA ↔ OC, with forecasting horizons extending to 60 minutes at 5-minute intervals.

Table 3: Large-scale forecasting results on LargeST at horizons up to 60 minutes. MAPE is reported in percentages. Bold and underlined values indicate the best and second-best results, respectively. Improv. reports the relative reduction over the best baseline in each column.

Model	SD → GBA										GBA → SD									
	5 min		15 min		30 min		45 min		60 min		5 min		15 min		30 min		45 min		60 min	
	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE
PatchTST	20.10	12.20	23.88	14.68	28.84	18.01	34.18	21.83	39.80	25.99	18.20	11.22	22.42	13.40	27.48	16.12	33.17	19.35	39.35	22.87
ST-GFSL	20.78	13.43	23.34	15.09	26.57	17.72	29.45	20.01	32.74	23.63	19.31	12.57	21.82	14.07	24.62	16.13	27.45	18.89	30.55	22.33
TPB	24.34	28.05	25.99	27.67	28.66	26.73	31.40	26.77	34.22	28.26	19.39	12.63	21.77	12.98	24.93	15.01	27.94	17.93	31.93	22.33
FEPCross	20.60	14.28	24.58	17.66	29.24	20.23	33.80	23.16	37.32	27.44	21.15	22.03	26.31	30.98	29.00	28.48	31.99	21.41	37.50	28.83
PACT	**19.30**	**11.70**	**22.16**	**13.72**	**25.51**	**16.21**	**28.86**	**18.99**	**31.96**	**21.73**	**16.90**	**10.64**	**19.88**	**12.09**	**23.36**	**14.04**	**26.75**	**16.15**	**30.00**	**18.26**
Improv.	3.98%	4.10%	5.06%	6.54%	3.99%	8.52%	2.00%	5.10%	2.38%	8.04%	7.14%	5.17%	8.68%	6.86%	5.12%	6.46%	2.55%	9.93%	1.80%	18.23%

Model	LA → OC										OC → LA									
	5 min		15 min		30 min		45 min		60 min		5 min		15 min		30 min		45 min		60 min	
	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE
PatchTST	18.03	14.58	22.39	17.28	27.78	20.81	33.68	24.91	40.06	29.61	20.06	8.48	24.72	10.39	30.71	12.88	37.07	15.62	43.87	18.64
ST-GFSL	19.72	20.20	22.06	20.74	25.17	24.84	28.34	29.54	32.00	38.09	21.11	9.47	24.08	10.40	27.76	11.73	31.23	13.39	34.21	14.92
TPB	22.88	46.43	26.00	45.79	27.56	44.49	29.97	40.03	32.45	57.35	22.64	10.35	26.26	11.54	30.56	12.96	34.39	15.30	38.67	18.44
FEPCross	20.86	20.52	22.99	28.02	28.91	35.45	36.20	54.91	37.03	34.20	23.83	14.92	25.60	12.00	31.54	17.18	36.44	19.27	41.50	18.12
PACT	**17.03**	**14.44**	**20.33**	**16.42**	**23.99**	**19.30**	**27.41**	**22.24**	**30.85**	**25.91**	**19.11**	**8.07**	**22.48**	**9.45**	**26.35**	**11.05**	**30.06**	**12.66**	**33.72**	**14.47**
Improv.	5.55%	0.96%	7.84%	4.98%	4.69%	7.26%	3.28%	10.72%	3.59%	12.50%	4.74%	4.83%	6.64%	9.05%	5.08%	5.80%	3.75%	5.45%	1.43%	3.02%

Table 3 shows that PACT achieves the lowest MAE and MAPE across all four transfer directions and five reported horizons. On SD → GBA, the target contains more than three times as many nodes as the source, yet PACT achieves MAEs of 19.30 and 31.96 at 5 and 60 minutes, improving over the best baselines by 3.98% and 2.38%, respectively. It also maintains its advantage in the reverse direction and on LA ↔ OC. These results demonstrate that PACT can transfer effectively between regions with substantially different node counts, using shared parameters without requiring an adjacency matrix or cross-region node correspondence.

The advantage remains at longer forecasting horizons, although its magnitude varies across tasks and metrics. At 60 minutes, PACT reduces MAE relative to the best baseline by 2.38%, 1.80%, 3.59%, and 1.43% on SD → GBA, GBA → SD, LA → OC, and OC → LA, respectively, with corresponding MAPE reductions of 8.04%, 18.23%, 12.50%, and 3.02%. Thus, its forecasting accuracy is retained across both larger sensing scales and longer prediction horizons.

5.6 ZERO- AND FEW-SHOT GENERALIZATION

We examine how PACT performs with zero, one, or two days of target observations on BAY → LA and CD → SZ. In the zero-shot setting, the source model is applied without target adaptation or target normalization statistics, while historical observations remain available as forecasting inputs.

One-day and two-day adaptation use only the corresponding initial target observations. All three settings share the same test interval, and we average MAE over the three horizons reported in the main table.

Figure 4 shows that target adaptation improves average MAE in both directions. On BAY → LA, MAE decreases from 2.857 without adaptation to 2.736 with one day and 2.717 with two days, with most of the improvement obtained from the first day. On CD → SZ, one-day adaptation produces a smaller change, from 2.239 to 2.234, while two-day adaptation further reduces MAE to 2.194. These results indicate that limited target observations can improve the transferred model, although the benefit of an additional observation day differs between the two tasks. Results for all four transfer directions are provided in the appendix.

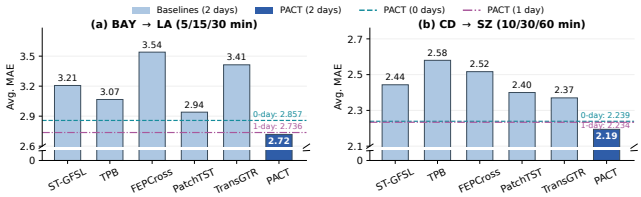

Figure 4: Average MAE over the three reported horizons. Bars show two-day adaptation; dashed lines show zero-shot and one-day results of PACT.

5.7 Efficiency Analysis

We compare computational efficiency across eight target cities with 207 to 2,352 nodes. All methods use the same GPU and batch size of 4, while retaining their native history lengths. We measure adaptation time over 500 target updates, inference latency per batch, and peak GPU memory during target adaptation, excluding source pretraining.

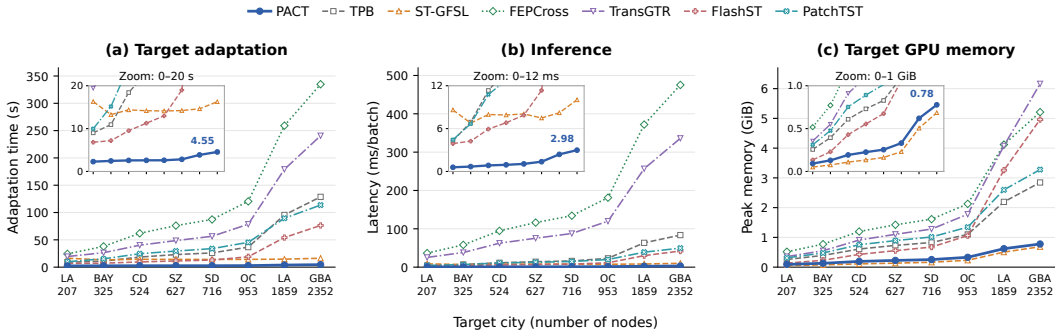

Figure 5: Target adaptation time, inference latency, and peak adaptation memory across eight target scales. Insets enlarge the lower ranges using the same node ordering. Lower values are better.

Figure 5 reports the full ranges, with insets highlighting differences among the lower-cost methods. PACT achieves the fastest adaptation and inference at every target scale. On GBA with 2,352 nodes, it completes 500 adaptation updates in 4.55 seconds and inference in 2.98 milliseconds per batch, corresponding to 3.57× and 3.36× speedups over the next-fastest method, ST-GFSL. Its peak adaptation memory is the second lowest across all eight scales, reaching 0.78 GiB on GBA compared with 0.69 GiB for ST-GFSL. As the target node count increases, PACT exhibits smaller absolute increases in runtime and memory than all baselines. These results complement the complexity analysis: compact global aggregation avoids constructing pairwise node interactions, enabling efficient adaptation and prediction at larger target scales.

6 Conclusion

We presented PACT, a graph-free framework for transferable spatio-temporal forecasting from a periodic perspective. By combining normalized relative-change encoding with global aggregation conditioned on future cycle positions and recent observations, PACT uses recurring temporal patterns to inform node-specific forecasts. Its shared parameters accommodate different node sets, while compact aggregation and retrieval yield computational complexity linear in the number of nodes at fixed model dimensions. Experiments on real-world traffic datasets demonstrate accurate cross-region forecasting, effective transfer across sensing scales, and efficient target adaptation and inference. Future work will expand our proposal to more diverse spatio-temporal domains and tasks.

AI USE STATEMENT

Generative AI assisted with language editing, LaTeX formatting, and consistency checks of author-provided materials. It did not generate experimental data or results. The authors are responsible for verifying the manuscript and its claims.

ETHICS STATEMENT

This work uses aggregate traffic sensor data and involves no human subjects or personally identifiable information. The results are intended for research and should not be interpreted as establishing reliability for safety-critical deployment.

REPRODUCIBILITY STATEMENT

The appendix provides the model architecture, training settings, data splits, and checkpoint-selection criteria. The model implementation and running instructions are available at https://anonymous.4open.science/r/PACT-F67B/.

REFERENCES

Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 922–929, 2019.

Lu Han, Xu-Yang Chen, Han-Jia Ye, and De-Chuan Zhan. Softs: Efficient multivariate time series forecasting with series-core fusion. *Advances in Neural Information Processing Systems*, 37: 64145–64175, 2024.

Yue Jiang, Yile Chen, Xiucheng Li, Qin Chao, Shuai Liu, and Gao Cong. Fstllm: Spatio-temporal llm for few shot time series forecasting. In *Forty-second International Conference on Machine Learning*, 2025.

Yilun Jin, Kai Chen, and Qiang Yang. Transferable graph structure learning for graph-based traffic forecasting across cities. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 1032–1043, 2023.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International conference on learning representations*, 2021.

Mengzhang Li and Zhanxing Zhu. Spatial-temporal fusion graph neural networks for traffic flow forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 4189–4196, 2021.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018.

Zhonghang Li, Lianghao Xia, Yong Xu, and Chao Huang. Flashst: A simple and universal prompt-tuning framework for traffic prediction. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 28978–28988, 2024.

Zhonghang Li, Long Xia, Lei Shi, Yong Xu, Dawei Yin, and Chao Huang. Open spatio-temporal foundation models for traffic prediction. *ACM Transactions on Intelligent Systems and Technology*, 2025.

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. *Advances in Neural Information Processing Systems*, 36:75354–75371, 2023a.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 11116–11140, 2024a.

Zhanyu Liu, Guanjie Zheng, and Yanwei Yu. Cross-city few-shot traffic forecasting via traffic pattern bank. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pp. 1451–1460, 2023b.

Zhanyu Liu, Jianrong Ding, and Guanjie Zheng. Frequency enhanced pre-training for cross-city few-shot traffic forecasting. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 35–52. Springer, 2024b.

Bin Lu, Xiaoying Gan, Weinan Zhang, Huaxiu Yao, Luoyi Fu, and Xinbing Wang. Spatio-temporal graph few-shot learning with cross-city knowledge transfer. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 1162–1172, 2022.

Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023.

Zezhi Shao, Zhao Zhang, Fei Wang, and Yongjun Xu. Pre-training enhanced spatial-temporal graph neural network for multivariate time series forecasting. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 1567–1577, 2022.

Leye Wang, Xu Geng, Xiaojuan Ma, Feng Liu, and Qiang Yang. Cross-city transfer learning for deep spatio-temporal prediction. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 1893–1899, 2019.

Senzhang Wang, Jiannong Cao, and S Yu Philip. Deep learning for spatio-temporal data mining: A survey. *IEEE transactions on knowledge and data engineering*, 34(8):3681–3700, 2020.

Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven Hoi. Cost: Contrastive learning of disentangled seasonal-trend representations for time series forecasting. In *International Conference on Learning Representations*, 2022.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 1907–1913, 2019.

Huaxiu Yao, Yiding Liu, Ying Wei, Xianfeng Tang, and Zhenhui Li. Learning from multiple cities: A meta-learning approach for spatial-temporal prediction. In *2019 World Wide Web Conference, WWW 2019*, pp. 2181–2191. Association for Computing Machinery, 2019.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianzhong Qi. Gman: A graph multi-attention network for traffic prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 1234–1241, 2020.

Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):1–55, 2014.

CONTENTS

1 Introduction **1**

2 Related Work **2**

3 Problem Definition **3**

4 Methodology **3**

4.1 Overview 3

4.2 Relative-Change Temporal Encoding 4

4.3 Conditional Global Aggregation 4

4.4 Node-Specific Retrieval 5

4.5 Training and Complexity 6

5 Experiments **6**

5.1 Experimental setup 6

5.2 Transferable Forecasting Performance 6

5.3 Ablation Studies 6

5.4 Robustness to Source Selection 7

5.5 Scalability Analysis 8

5.6 Zero- and Few-Shot Generalization 8

5.7 Efficiency Analysis 9

6 Conclusion **9**

A Additional Related Work **13**

B Data and Evaluation Protocols **13**

C Implementation and Training Details **14**

D Baseline Methods **14**

E Zero-Shot and Few-Shot Transfer **16**

F Additional Motivation Diagnostics **17**

F.1 Observed levels and daily variation 17

F.2 Matched histories at different periodic positions 17

G Interpretability and Mechanism Analysis **17**

G.1 Global corrections and local prediction errors 18

G.2 Corrections across periodic positions and recent trends 18

G.3 Structure of the global summaries 18

G.4 Fixed-checkpoint interventions 19

H Efficiency Measurement Protocol **19**

I Analysis of Relative Changes and Conditional Global Aggregation **21**

I.1 Notation and scope 21

I.2 Signal level, scale, and relative-change encoding 21

I.3 Concentration and sensitivity of the conditional summary 22

I.4 The numerical floor and stable perturbations 23

I.5 Interaction between phase and observed state 23

I.6 Node composition and summary stability 24

I.7 Symmetry and factored cross-node communication 25

I.8 A limited argument for conditioning before compression 25

A Additional Related Work

Normalization and relative prediction. RevIN (Kim et al., 2021) normalizes each input instance and restores its statistics at the output to address distribution shifts in time-series forecasting. PACT uses fixed domain-level statistics and represents each node's history through deviations from its latest observation and first-order differences. Predictions combine this observed level with learned future changes, connecting the transferred temporal model to the target region's current measurements.

Compact cross-node interaction. SOFTS (Han et al., 2024) introduces a STar Aggregate-Redistribute module that summarizes multiple series into a shared core and combines it with individual series representations, enabling efficient cross-series interaction. PACT constructs a separate global summary for each forecast step, with aggregation weights conditioned jointly on the future daily-cycle position and the current state summarized from encoded node histories. Each node retrieves a correction from the corresponding summary to refine its local forecast. The temporal encoder and aggregation and retrieval projections are shared across nodes and cities, supporting transfer between regions with different node sets.

Pretraining and transferable forecasting. OpenCity (Li et al., 2025) combines Transformer and graph-based modeling with pretraining on heterogeneous traffic datasets for zero-shot prediction across urban environments. FSTLLM (Jiang et al., 2025) uses pretrained language models for few-shot forecasting and supports integration with existing forecasting models. PACT learns shared temporal and cross-node computations through source pretraining and adapts them to target observations. Its aggregation is conditioned on recent encoded histories and future daily-cycle positions, allowing the transferred model to construct global summaries for individual forecast steps.

B Data and Evaluation Protocols

Datasets. The small-scale benchmark uses traffic-speed observations from METR-LA, PEMS-BAY, Chengdu, and Shenzhen. Table 4 summarizes their dimensions and the four LargeST subsets used to evaluate larger target scales. METR-LA and the Los Angeles subset of LargeST are different datasets; we distinguish them by their node counts. The small-scale transfer tasks are METR-LA $\leftrightarrow$ PEMS-BAY and Chengdu $\leftrightarrow$ Shenzhen, while the LargeST tasks are SD $\leftrightarrow$ GBA and LA $\leftrightarrow$ OC.

Small-scale transfer. For PACT, source observations are split chronologically into 70% for training and the next 10% for validation. Target adaptation uses the first two days of observations, with no additional target validation set. A target checkpoint is selected by its mean MAE across the six forecast steps on this adaptation set. Training labels remain within the adaptation interval; test labels

Table 4: Dataset configurations. LargeST observations measure traffic flow; the four small-scale datasets measure traffic speed.

Collection	Dataset	Nodes	Interval (min)
Small-scale	METR-LA	207	5
	PEMS-BAY	325	5
	Chengdu	524	10
	Shenzhen	627	10
LargeST	SD	716	5
	OC	953	5
	LA	1,859	5
	GBA	2,352	5

begin on the third day, and their preceding input history may cross this boundary. The input length is 12 steps and the output length is six steps. We report steps 1, 3, and 6, corresponding to 5, 15, and 30 minutes for METR-LA and PEMS-BAY and 10, 30, and 60 minutes for Chengdu and Shenzhen.

Large-scale transfer. The LargeST experiments use 12 historical steps to predict 12 future steps. For PACT, the first two target days provide training observations, with no additional target validation set. Input and output windows lie wholly within each partition. Full-test evaluation uses stride one over the remaining observations, yielding 103,945 windows per direction. We report forecast steps 1, 3, 6, 9, and 12.

Metrics and normalization. All errors are computed after returning predictions to the original measurement scale. For a fixed forecast step t, let $\mathcal{I}_t$ contain all evaluated window–node pairs and let $\mathcal{I}_t^+ = \{a \in \mathcal{I}_t : y_{a,t} \neq 0\}$. We compute

$$\mathrm{MAE}_t = \frac{1}{|\mathcal{I}_t|} \sum_{a \in \mathcal{I}_t} |\widehat{y}_{a,t} - y_{a,t}|, \qquad \mathrm{MAPE}_t = \frac{100}{|\mathcal{I}_t^+|} \sum_{a \in \mathcal{I}_t^+} \frac{|\widehat{y}_{a,t} - y_{a,t}|}{|y_{a,t}|}. \tag{12}$$

Thus, MAE retains zero observations and MAPE excludes exactly zero ground-truth values, without a further low-value threshold. For PACT, each domain's scalar normalization statistics are estimated from its training observations and fixed during evaluation. Zero-shot evaluation instead retains the source statistics, as specified in the corresponding experiment. When reporting an average over the three displayed horizons, we average their MAEs; this differs from averaging all six forecast steps.

C Implementation and Training Details

Model configuration. The small-scale implementation uses a shared GRU with representation width $d_z = 64$ and global-summary width $d_m = 16$. All nodes share encoder and projection parameters, and the target model is initialized from the source checkpoint. The source and target optimizers are AdamW with weight decay 10^{-5}. Table 5 summarizes the training configurations. Source checkpoints in the small-scale experiments are selected by validation MAE averaged over all six forecast steps; all trainable parameters are updated during target adaptation.

Training objective. The implementation applies a fixed forecast-step weighting to the normalized absolute error: the final step receives weight two and the other steps receive weight one. Evaluation reports each forecast step separately.

D Baseline Methods

The comparison includes a statistical predictor, supervised spatio-temporal and time-series models, and methods designed for pretraining or cross-city transfer. The following descriptions summarize their forecasting architectures and transfer mechanisms.

Table 5: Training configurations of PACT

Setting	Small-scale	LargeST
History / forecast length	12 / 6	12 / 12
Batch size	64	16
Source learning rate	0.002	0.002
Target learning rate	0.001	0.001
Source epochs	20	20
Target epochs	50	50
Target training	2	2

HA. Historical Average predicts from averages of past observations without fitting a neural forecasting model. It provides a statistical reference for assessing the value of learned temporal and spatial representations.

DCRNN. DCRNN (Li et al., 2018) incorporates diffusion convolution into recurrent units to capture directed spatial dependencies and temporal dynamics. Its diffusion operator uses bidirectional random walks on the graph, while an encoder–decoder architecture with scheduled sampling generates multi-step predictions. It represents a recurrent graph-based forecasting baseline.

GWNet. Graph WaveNet (Wu et al., 2019) combines graph convolution with stacked dilated temporal convolutions. A learned dependency matrix, parameterized by node embeddings, supplements the spatial relationships represented by a predefined graph. This baseline captures both adaptive spatial dependencies and temporal patterns with an expanding receptive field.

STFGNN. STFGNN (Li & Zhu, 2021) constructs a data-driven temporal graph to capture relationships that may be absent from spatial adjacency. It fuses spatial and temporal graphs across time steps and combines the resulting graph module with gated temporal convolutions. It provides a comparison based on joint spatial-temporal graph construction.

DLinear. DLinear (Zeng et al., 2023) decomposes each input sequence into a moving-average trend and a residual component. Separate linear mappings project the two components into the forecasting horizon, and their outputs are added. It tests how far a simple temporal decomposition and linear predictor can account for the forecasting task.

PatchTST. PatchTST (Nie et al., 2023) divides each univariate sequence into patches and processes them using a Transformer whose parameters are shared across channels. Its channel-independent architecture models temporal patterns without explicitly mixing node representations.

iTransformer. iTransformer (Liu et al., 2024a) embeds the historical observations of each variable into a token and applies self-attention across variable tokens. Its feedforward layers transform each token's temporal representation, exchanging the roles of the time and variable axes relative to temporal-token Transformers. With nodes treated as variables, it provides an attention-based comparison for learning cross-node relationships.

ST-GFSL. ST-GFSL (Lu et al., 2022) learns node-level meta-knowledge and uses it to generate node-specific forecasting parameters, allowing nodes with similar spatio-temporal characteristics to share knowledge across cities. An auxiliary graph-reconstruction objective makes this meta-knowledge sensitive to spatial structure.

STEP. STEP (Shao et al., 2022) pretrains TSFormer on masked segments of long historical sequences to obtain segment-level temporal representations. These representations enrich the short-history inputs of a downstream spatio-temporal graph network and support the learning of dependencies among time series. It is a pretraining-enhanced forecasting baseline, without a dedicated cross-city adaptation mechanism in its original formulation.

Table 6: Zero-shot and few-shot results of PACT. $H_1/H_3/H_6$ correspond to 5/15/30 minutes for LA and BAY and 10/30/60 minutes for CD and SZ. Avg. averages these three steps.

Direction	Target days	H_1 MAE	H_3 MAE	H_6 MAE	Avg. MAE
LA → BAY	0	0.959 ± 0.005	1.499 ± 0.004	2.052 ± 0.009	1.503 ± 0.004
	1	0.924 ± 0.015	1.565 ± 0.048	2.236 ± 0.097	1.575 ± 0.051
	2	0.887 ± 0.002	1.454 ± 0.012	2.043 ± 0.021	1.493 ± 0.010
BAY → LA	0	2.227 ± 0.012	2.830 ± 0.026	3.514 ± 0.061	2.857 ± 0.029
	1	2.138 ± 0.003	2.728 ± 0.002	3.343 ± 0.007	2.736 ± 0.003
	2	2.129 ± 0.001	2.712 ± 0.003	3.312 ± 0.004	2.717 ± 0.002
CD → SZ	0	1.810 ± 0.001	2.279 ± 0.006	2.628 ± 0.012	2.239 ± 0.006
	1	1.811 ± 0.002	2.279 ± 0.004	2.611 ± 0.015	2.234 ± 0.007
	2	1.800 ± 0.001	2.245 ± 0.003	2.538 ± 0.006	2.194 ± 0.003
SZ → CD	0	2.100 ± 0.006	2.659 ± 0.010	3.036 ± 0.019	2.598 ± 0.010
	1	2.095 ± 0.002	2.617 ± 0.006	2.948 ± 0.008	2.553 ± 0.003
	2	2.085 ± 0.001	2.579 ± 0.004	2.878 ± 0.006	2.514 ± 0.003

TransGTR. TransGTR (Jin et al., 2023) transfers both graph structures and forecasting models across cities. A node-feature network learns from knowledge distillation, and a shared structure generator uses the resulting representations to construct city-specific graphs; temporal-decoupled regularization encourages the spatial features extracted from these graphs to have similar distributions. This baseline evaluates transfer through learned spatial relationships.

FlashST. FlashST (Li et al., 2024) adapts pretrained forecasting models with a lightweight spatio-temporal prompt network. Its distribution-mapping mechanism supports transfer between pretraining data and downstream observations, making it a prompt-based adaptation baseline.

TPB. TPB (Liu et al., 2023b) encodes source-city traffic patches and clusters their representations into a traffic pattern bank. Target observations query this bank to aggregate reusable pattern information and construct an adjacency matrix for the downstream spatio-temporal predictor. Its meta-training framework learns an initialization for the trainable modules. It provides a comparison based on explicit retrieval of shared traffic patterns.

FEPCross. FEPCross (Liu et al., 2024b) combines temporal and frequency-domain information in a cross-domain spatio-temporal encoder. Reconstruction and contrastive pretraining objectives learn representations that can be reused in a data-scarce target city. Its fine-tuning stage enriches training samples and maintains a momentum-updated graph to support forecasting with limited target observations. It provides a frequency-based transfer comparison for PACT.

E ZERO-SHOT AND FEW-SHOT TRANSFER

We evaluate PACT with zero, one, or two days of target adaptation, using a common test interval beginning on target day 3. Zero-shot inference retains the source checkpoint and source normalization statistics; one-day and two-day adaptation estimate statistics from their respective target training intervals. No additional target validation observations are used. Zero-shot inference still receives the historical input window required for forecasting. Table 6 reports the mean and sample standard deviation across five seeds. For Avg. MAE, we first average the three reported forecast steps within each seed and then calculate the mean and standard deviation across seeds.

Adaptation reduces the final-step MAE in BAY → LA, CD → SZ, and SZ → CD, with further gains from a second target day. LA → BAY behaves differently: the source checkpoint achieves a final-step MAE of 2.052, compared with 2.236 and 2.043 after one-day and two-day adaptation. The benefit of adaptation thus depends on both the transfer direction and the forecast horizon.

Table 7: Descriptive statistics after normalization using the first two days. Daily range denotes mean node-day P90–P10.

Dataset	Mean	Standard deviation	Daily range
PEMS-BAY	−0.448	1.459	2.451
METR-LA	0.088	0.973	1.493
Shenzhen	−0.070	0.992	1.038
Chengdu	−0.148	1.054	1.671

Table 8: Future-change differences for matched histories, in support-period standard-deviation units. Coverage is the fraction of eligible queries retained.

Dataset	Queries	Coverage	Same phase	Different phase
PEMS-BAY	1,939	97.0%	0.09608	0.15241
METR-LA	1,855	99.1%	0.07267	0.10276
Shenzhen	769	81.5%	0.03260	0.07895
Chengdu	498	52.8%	0.04661	0.10165

F Additional Motivation Diagnostics

F.1 Observed levels and daily variation

We estimate each dataset's scalar mean and standard deviation using its first two days, and describe subsequent complete days. To measure within-day variation, we compute the difference between the 90th and 10th percentiles for each node and day, then average these differences. Table 7 reports values after normalization. Variation remains different across datasets after removing the support-period location and scale, motivating the use of recent observations alongside periodic information. This diagnostic does not establish that normalization removes all amplitude differences or that a particular model handles them better.

F.2 Matched histories at different periodic positions

We examine whether similar recent histories are followed by more similar future changes when they occur at nearby positions in the daily cycle. We match windows using a 14-dimensional descriptor containing the 12 normalized city-mean historical observations, the latest normalized city mean, and the mean within-node historical standard deviation normalized by the support-period standard deviation. Candidates are selected using these historical descriptors and relative daily phase, without future values. Same-phase matches have circular phase distance at most $1/16$ of a day; different-phase matches are separated by at least $1/4$ of a day. Within-city matches are drawn from different days. We retain queries whose root-mean-square descriptor distance is at most 0.25 in both comparison groups and whose two match distances differ by at most 0.1.

Table 8 compares the absolute difference in city-mean future change at the sixth forecast step. For each dataset, different-phase matches have larger mean future differences than same-phase matches under this rule. Similar recent aggregate histories can thus precede different future changes, with relative periodic position providing additional descriptive context. These matches do not hold complete node histories fixed, and candidates may be reused. The results establish an association rather than an independent causal effect of phase.

G Interpretability and Mechanism Analysis

We examine two-day adaptation checkpoints through visualizations and inference-time interventions. The visualizations use seed 0 and 2,048 uniformly spaced test windows per direction, retaining all nodes. A separate intervention study uses five seeds and 4,096 uniformly spaced windows per checkpoint. Both are fixed-checkpoint, exploratory analyses of the existing test data; no parameters are updated.

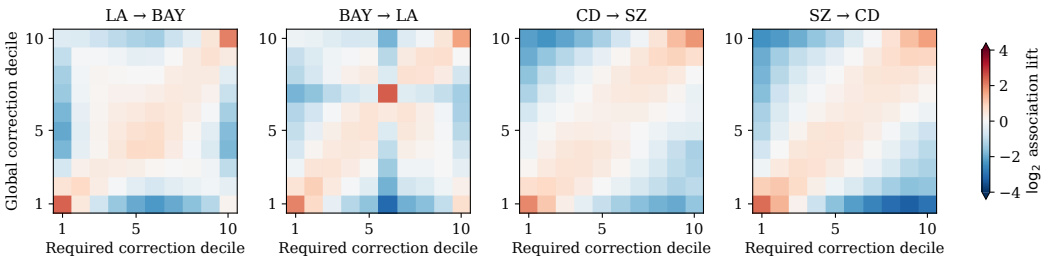

Figure 6: Required and predicted global corrections at the final forecast step, grouped into deciles. Color indicates log-scaled association lift.

Table 9: Fixed-checkpoint correction analysis at step six, using seed 0 and 2,048 windows per direction. Sign agreement compares required and predicted corrections.

Direction	Local MAE	Full MAE	Sign agreement	Spearman ρ
LA → BAY	2.2341	2.0715	58.3%	0.314
BAY → LA	3.5429	3.3155	61.2%	0.261
CD → SZ	2.9505	2.5311	68.1%	0.483
SZ → CD	3.5882	2.8853	69.4%	0.541

G.1 Global corrections and local prediction errors

Let $\ell_{i,t}$ denote the prediction obtained from the latest observation and the jointly trained local branch, expressed in the original data scale. Define the required correction $e_{i,t} = y_{i,t} - \ell_{i,t}$ and the predicted correction $c_{i,t} = \widehat{y}_{i,t} - \ell_{i,t} = \sigma_x \gamma r_{i,t}$, where σ_x restores the normalized correction to the original scale. For each direction and forecast step, we independently divide both quantities into deciles and measure their joint frequency relative to the product of their marginal frequencies. Figure 6 displays the base-two logarithm of this lift at step six. A half-count is added to each cell for visualization, and the color range is fixed to $[-4, 4]$ across directions.

The enrichment toward the diagonal indicates that larger required corrections tend to receive larger predicted corrections, particularly on the two Chinese transfer directions. Table 9 also shows lower MAE after adding the correction in every direction. The correction does not improve every prediction: on LA → BAY, 48.6% of node-window predictions improve, while the aggregate MAE still decreases. The local output here is part of the jointly trained model, not the independently trained Local Only baseline.

G.2 Corrections across periodic positions and recent trends

We group the final prediction's daily phase into 24 equal-width bins and the least-squares slope over the 12 historical observations into ten quantile groups. Figure 7 compares the mean required and predicted corrections within these groups. Each pair uses identical samples and a shared color scale; cells with fewer than 50 node-window observations are left blank. The CD → SZ panels show alternating upward and downward corrections across phase groups, with the model smoothing some larger required corrections. Other directions exhibit different patterns, so these maps should not be interpreted as a universal daily schedule. Grouping by slope does not equate complete histories or node identities.

G.3 Structure of the global summaries

We visualize the final-step summary $\mathbf{m}_T$ for each of the 2,048 windows using t-SNE. Before embedding, PCA reduces the summary to at most 30 dimensions without coordinate-wise standardization. The embedding uses random seed 0 and perplexity 30; perplexities 15 and 50 were also examined as sensitivity checks. The same coordinates in Figure 8 are colored by relative daily phase and by the node-averaged historical slope. CD → SZ shows an ordered phase pattern and regional differences

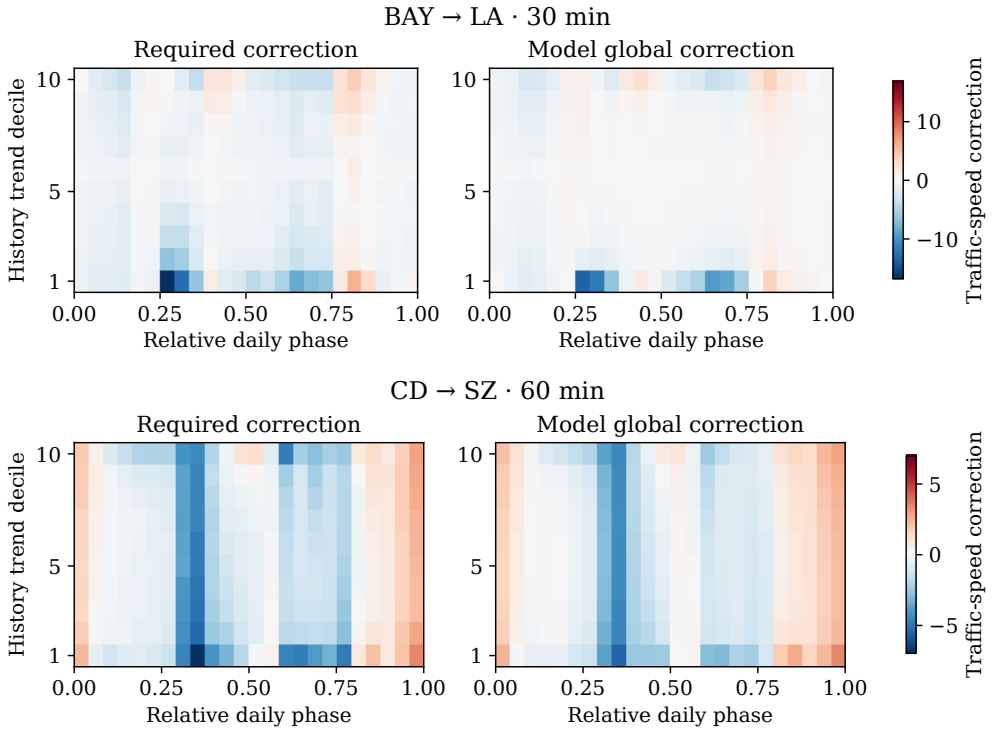

Figure 7: Required and predicted corrections by relative daily phase and recent trend on BAY → LA and CD → SZ. Color scales are shared within each direction.

in recent trends. These plots describe the learned representation. Since phase is an input to aggregation and t-SNE distorts distances, visual separation alone does not establish predictive utility or discrete traffic categories.

G.4 Fixed-checkpoint interventions

To examine whether these internal quantities affect predictions, we modify individual forward-pass components while keeping parameters and input histories fixed. Phase offsets shift the periodic inputs by the stated hours. Previous-day state substitutes the state from the preceding day at the same relative phase. Source state and source condition use averages over 256 source-training windows, computed with the frozen target checkpoint. Removing the global correction retains only the jointly trained local prediction. These are inference-time interventions rather than retrained ablations.

Phase offsets increase final-step MAE for every direction and seed, supporting the predictive role of correctly specified periodic inputs. A six-hour forward shift increases MAE by 0.227 ± 0.044 on BAY → LA and 0.566 ± 0.065 on CD → SZ, where uncertainty is the sample standard deviation across seeds. Removing the global correction increases MAE by 0.199 ± 0.018 and 0.527 ± 0.265, respectively. The contribution of the current domain state is smaller and depends on the reference: replacing it with the preceding day's state has much less effect than replacing it with a source average. LA → BAY improves under several constant-condition interventions, so the results do not imply that every learned condition is optimal. Removing the numerical floor measurably affects CD → SZ, consistent with the attenuation in equation 23. The observed mechanism therefore includes both relative weighting and, when active, floor-induced scaling.

H Efficiency Measurement Protocol

Hardware and workload. We measure PACT, ST-GFSL, TPB, FEPCross, TransGTR, FlashST, and PatchTST on an NVIDIA A100 GPU. All runs use batch size four, FP32 arithmetic, disabled

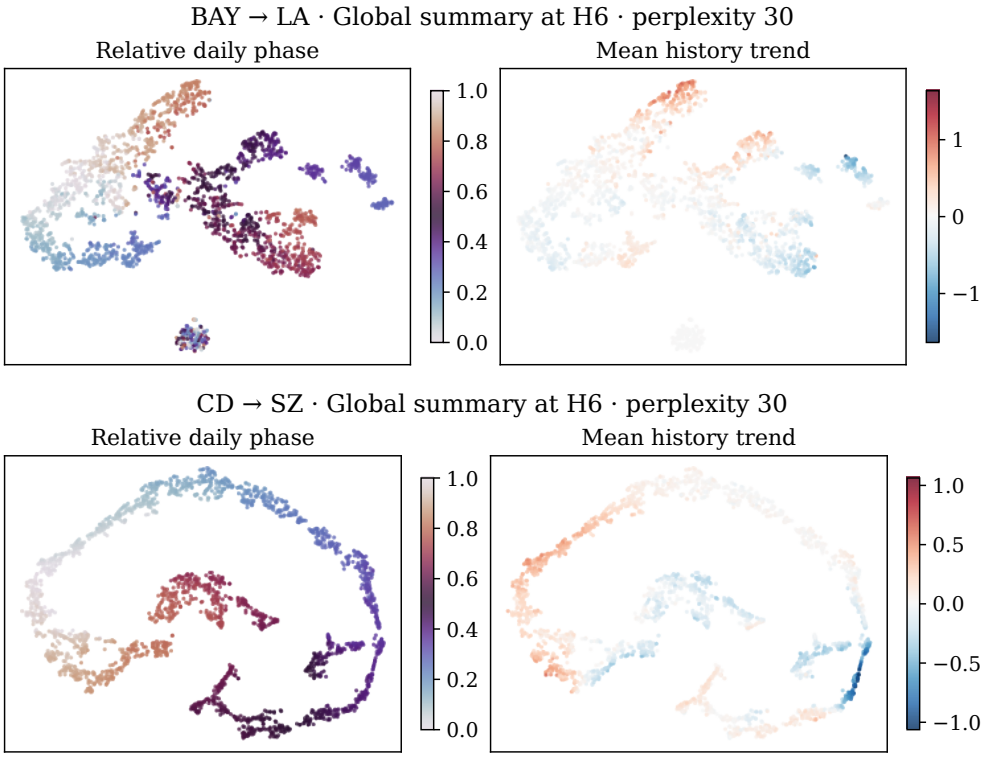

Figure 8: Global-summary t-SNE on BAY → LA and CD → SZ, colored by relative daily phase and mean recent trend. Each point represents one window at the final forecast step.

TF32, and two CPU threads. Target adaptation samples windows from the first two days and performs 500 optimizer updates. This measures a fixed adaptation workload, not training time to convergence. Adaptation timing includes CPU batch construction, host-to-device transfer, forward computation, loss evaluation, backpropagation, and the optimizer step; it excludes initial data preparation, source pretraining, validation, and checkpoint I/O. Runs are executed sequentially without concurrent training workloads.

Context and initialization. All efficiency models output six steps. PACT, ST-GFSL, FlashST, and PatchTST use 12 historical steps; TPB and FEPCross use 288, and TransGTR uses 864. Native long-context encoders are executed during inference without a hidden-state cache. Unavailable long context follows the corresponding adapter's zero-padding convention. Existing weights are reused where available, and resource-only random initialization is used when the required weights are unavailable. These measurements characterize implementation cost and do not establish the accuracy of a randomly initialized model.

Timing and memory. Adaptation uses ten warmup updates, after which model and optimizer states are reset before measurement. GPU synchronization brackets the timed workload. Peak allocated GPU memory is reset after warmup and measured during adaptation, including gradients and optimizer states. Inference uses GPU-resident batches, 20 warmup iterations, and 100 CUDA-event measurements; its latency excludes data transfer. We report medians over three fresh-process repetitions.

I Analysis of Relative Changes and Conditional Global Aggregation

I.1 Notation and scope

For clarity, we analyze one scalar observed variable. The parameters and normalization statistics are fixed during a prediction. The raw prediction is $\widehat{y}_{i,t} = \mu + \sigma_x(\widetilde{x}_{i,L} + u_{i,t} + \gamma r_{i,t})$, with $\widetilde{x} = (x - \mu)/\sigma_x$. The GRU and local branch receive only centered histories and their consecutive differences. The condition uses the mean encoded state $\mathbf{s} = N^{-1}\sum_i \mathbf{z}_i$ and the Fourier phase vector,

$$\boldsymbol{\beta}(\mathbf{s}, \boldsymbol{\phi}) = \mathbf{W}_\phi \boldsymbol{\phi} + (\mathbf{W}_s \mathbf{s}) \odot (\mathbf{W}_{s\phi}\boldsymbol{\phi}). \tag{13}$$

For a fixed step t and coordinate k, define $\xi_i = (\mathbf{W}_K \mathbf{z}_i + \mathbf{b}_K)_k$, $v_i = (\mathbf{W}_t^V \mathbf{z}_i)_k$, and $p_i(\beta) = \sigma(\xi_i + \beta)$. The ideal and implemented summaries are

$$\begin{aligned} w_i(\beta) &= \frac{p_i(\beta)}{\sum_j p_j(\beta)}, \qquad m(\beta) = \sum_i w_i(\beta) v_i, \\ d(\beta) &= \frac{\lambda}{N}\sum_i p_i(\beta), \quad m_\epsilon(\beta) = \frac{\lambda N^{-1}\sum_i p_i(\beta) v_i}{\max\{d(\beta), \epsilon\}}. \end{aligned} \tag{14}$$

The manuscript's ideal operator is independent of the common multiplier $\lambda > 0$. The implementation uses $\lambda = 2$ and $\epsilon = 10^{-6}$. Results for m concern exact normalization; results for m_ϵ retain the floor. Throughout the analysis, all logits and learned parameters are finite. These are architectural properties and conditional stability statements, not guarantees of target forecasting accuracy or of successful optimization.

I.2 Signal level, scale, and relative-change encoding

Define the linear feature map $\mathbf{D}$ by $D^{(1)}_{i,\ell} = x_{i,\ell} - x_{i,L}$, $D^{(2)}_{i,1} = 0$, and $D^{(2)}_{i,\ell} = x_{i,\ell} - x_{i,\ell-1}$ for $\ell > 1$. Repeated entries used by the direct branch can be included without changing the following arguments. Let $R_{i,t}$ be the normalized learned correction $u_{i,t} + \gamma r_{i,t}$, including the state-dependent aggregation and its numerical floor.

Proposition I.1 (Level and scale equivariance)**.** *The predictor has the form*

$$F_{\mu,\sigma_x}(\mathbf{X}, \boldsymbol{\phi})_{i,t} = x_{i,L} + \sigma_x R_{i,t}\left(\frac{\mathbf{D}(\mathbf{X})}{\sigma_x}, \boldsymbol{\phi}\right). \tag{15}$$

Consequently, it is independent of μ at fixed σ_x. For constants b_i that do not vary across the historical steps, fixed σ_x gives $F(\mathbf{X} + \mathbf{b}, \boldsymbol{\phi})_{i,t} = F(\mathbf{X}, \boldsymbol{\phi})_{i,t} + b_i$. For a common $a > 0$, common b, and $\sigma'_x = a\sigma_x$,

$$F_{\mu',\sigma'_x}(a\mathbf{X} + b, \boldsymbol{\phi}) = aF_{\mu,\sigma_x}(\mathbf{X}, \boldsymbol{\phi}) + b \tag{16}$$

for any μ'.

Proof. Centering the normalized history gives $(x_{i,\ell} - x_{i,L})/\sigma_x$, and differencing gives $(x_{i,\ell} - x_{i,\ell-1})/\sigma_x$. Neither contains μ. All hidden states, local changes, weights, values, queries, and global corrections are functions of these features and the phase. In the final prediction, $\mu + \sigma_x\widetilde{x}_{i,L} = x_{i,L}$, proving equation 15. Adding b_i leaves $\mathbf{D}$ unchanged. A common positive rescaling satisfies $\mathbf{D}(a\mathbf{X} + b) = a\mathbf{D}(\mathbf{X})$, so division by $a\sigma_x$ preserves every input to R. Only the final anchor and output scale change, giving equation 16. □

If a common affine transformation is also applied to the observations used to estimate normalization statistics, their empirical standard deviation scales by a, provided no variance floor or clipping changes this relation. By contrast, recomputing a pooled standard deviation after different node-specific offsets need not preserve that standard deviation. The fixed-scale node-offset statement does not cover this operation. Likewise, source statistics retained in zero-shot evaluation need not match an unknown target rescaling. The proposition describes a known affine change with corresponding statistics, not removal of arbitrary domain shifts, time-varying amplitudes, or different node-specific multiplicative factors.

Proposition I.2 (Effect of scale mismatch). *Fix $\mathbf{X}$, ϕ and one output coordinate. Suppose R is K_R-Lipschitz between $\mathbf{D}(\mathbf{X})/\sigma_x$ and $a\mathbf{D}(\mathbf{X})/\sigma'_x$ in a specified norm, and $|R(\mathbf{D}(\mathbf{X})/\sigma_x, \phi)| \le M_R$. Then*

$$\begin{aligned} &\left|F_{\mu',\sigma'_x}(a\mathbf{X}+b) - aF_{\mu,\sigma_x}(\mathbf{X}) - b\right| \\ &\qquad \le |\sigma'_x - a\sigma_x| \left(M_R + K_R \frac{\|\mathbf{D}(\mathbf{X})\|}{\sigma_x}\right). \end{aligned} \tag{17}$$

The phase is held fixed and omitted on the left.

Proof. The anchors cancel. Add and subtract $\sigma'_x R(\mathbf{D}(\mathbf{X})/\sigma_x, \phi)$ and use the Lipschitz bound. The feature difference is $|a/\sigma'_x - 1/\sigma_x|\|\mathbf{D}(\mathbf{X})\|$, and multiplication by σ'_x yields $|\sigma'_x - a\sigma_x|\|\mathbf{D}(\mathbf{X})\|/\sigma_x$. □

This bound identifies the role of scale estimation without asserting that K_R is small or independent of the full predictor's dimensions.

I.3 Concentration and sensitivity of the conditional summary

Proposition I.3 (Ordered flattening of node weights). *Order the logits as $\xi_1 \ge \cdots \ge \xi_N$. If $\beta_2 \ge \beta_1$, then*

$$\sum_{i=1}^{j} w_i(\beta_2) \le \sum_{i=1}^{j} w_i(\beta_1), \qquad 1 \le j < N. \tag{18}$$

Thus the weights at β_2 are majorized by those at β_1. Their entropy is nondecreasing, and their effective number of contributing nodes, $N_{\text{eff}} = (\sum_i w_i^2)^{-1}$, is nondecreasing. Their limits are softmax weights at negative infinity and uniform weights at positive infinity.

Proof. Sigmoid preserves the logit ordering. Since $p'_i = p_i(1-p_i)$,

$$w'_i = w_i(\overline{p} - p_i), \qquad \overline{p} = \sum_i w_i p_i. \tag{19}$$

For the highest j nodes let $S_j = \sum_{i \le j} w_i$. The weighted means of p inside and outside this set satisfy $\overline{p}_{\text{in}} \ge \overline{p}_{\text{out}}$, and

$$S'_j = S_j(1-S_j)(\overline{p}_{\text{out}} - \overline{p}_{\text{in}}) \le 0. \tag{20}$$

Integration proves the cumulative inequalities. For any two sequences f_i, g_i,

$$\operatorname{Cov}_w(f,g) = \frac{1}{2}\sum_{i,j} w_i w_j (f_i - f_j)(g_i - g_j). \tag{21}$$

Hence covariances of equally ordered sequences are nonnegative. Direct differentiation gives $\frac{\mathrm{d}}{\mathrm{d}\beta}[-\sum_i w_i \log w_i] = \operatorname{Cov}_w(\log w, p) \ge 0$ and $\frac{\mathrm{d}}{\mathrm{d}\beta}\sum_i w_i^2 = -2\operatorname{Cov}_w(w,p) \le 0$. Finally, $e^{-\beta}p_i \to e^{\xi_i}$ as $\beta \to -\infty$, while $p_i \to 1$ as $\beta \to +\infty$, giving the limits. □

Proposition I.4 (Response to a shared condition). *For fixed logits and values, the derivative of the ideal summary is*

$$\frac{\partial m}{\partial \beta} = -\operatorname{Cov}_w(v,p), \qquad \left|\frac{\partial m}{\partial \beta}\right| \le \frac{\max_i v_i - \min_i v_i}{4}, \tag{22}$$

where Cov_w is covariance over the nodes with probabilities w_i.

Proof. The ordering, cumulative inequalities, and limits follow from Proposition I.3. With values fixed, equation 19 gives $m' = -\operatorname{Cov}_w(v,p)$. A variable with range length A has variance at most $A^2/4$: subtract its range midpoint, use the bound $A/2$ on its absolute deviation, and note that the mean minimizes squared deviations. Cauchy–Schwarz consequently bounds a covariance by one quarter of the product of the two range lengths. Since $p_i \in (0,1)$, equation 22 follows. □

The condition varies concentration within a fixed ranking; it does not arbitrarily exchange the highest- and lowest-weight nodes in that coordinate. Different coordinates can have different rankings. If all v_i are equal, the ideal summary cannot respond to β, even when its weights change. If values increase with p_i, increasing β decreases the summary; the reverse association reverses this response. A shared additive offset inside node-wise softmax cancels exactly, whereas its placement before sigmoid permits this response. This comparison concerns that specific additive softmax construction, not every form of conditional attention.

I.4 The numerical floor and stable perturbations

Proposition I.5 (Attenuation and perturbation bound)**.** *For m_ϵ in equation 14,*

$$m_\epsilon = c_\epsilon m, \qquad c_\epsilon = \frac{d}{\max\{d, \epsilon\}} \in (0, 1]. \tag{23}$$

Its effective weights are nonnegative and sum to c_ϵ. Suppose $|v_i|, |v_i'| \le V$ and let $\ell_i = \xi_i + \beta$. For two logit and value arrays of the same size,

$$|m_\epsilon(\boldsymbol{\ell}', \mathbf{v}') - m_\epsilon(\boldsymbol{\ell}, \mathbf{v})| \le \|\mathbf{v}' - \mathbf{v}\|_\infty + V\|\boldsymbol{\ell}' - \boldsymbol{\ell}\|_\infty. \tag{24}$$

This bound holds for all $N \ge 1$, all $\lambda > 0$, and $\epsilon \ge 0$; when $\epsilon = 0$, finite logits keep the denominator positive.

Proof. The attenuation identity follows by factoring the denominator. Effective weights with total mass at most one give the value perturbation bound. To change logits at fixed values, interpolate $\ell_i(s) = \ell_i + sh_i$, where $\|h\|_\infty = H$. On an interval with inactive floor, differentiation gives $\frac{\mathrm{d}m}{\mathrm{d}s} = \mathrm{Cov}_w(v, (1-p)h)$. The ranges of these variables are at most $2V$ and $2H$, respectively, so the covariance bound is VH. With active floor, the derivative equals $\lambda(N\epsilon)^{-1}\sum_i p_i(1-p_i)h_i v_i$, whose magnitude is at most $VHd/\epsilon \le VH$. The summary is continuous and piecewise differentiable, with an absolutely continuous restriction to this finite line segment. Integrating yields the logit perturbation bound. Combining the two perturbations proves equation 24. □

The stronger ideal common-offset bound is $(\max_i v_i - \min_i v_i)/4$, whereas the implemented operator admits the common-offset bound V even when its floor is active. Both are independent of N. In particular, active-floor sensitivity does not require dividing this logit perturbation bound by ϵ. However, the effective weights no longer form a probability distribution: their entropy and effective sample size should be computed after renormalizing their relative weights, and reported separately from c_ϵ. As $\beta \to -\infty$, the implemented summary tends to zero for fixed $\epsilon > 0$, not to the softmax-weighted value. The factor λ affects floor activation and must be retained when matching the implementation.

I.5 Interaction between phase and observed state

Proposition I.6 (State-dependent phase contrast)**.** *For two states and two phase encodings, the mixed contrast of equation 13 is*

$$\begin{aligned}&\boldsymbol{\beta}(\mathbf{s}, \boldsymbol{\phi}) - \boldsymbol{\beta}(\mathbf{s}, \boldsymbol{\phi}') - \boldsymbol{\beta}(\mathbf{s}', \boldsymbol{\phi}) + \boldsymbol{\beta}(\mathbf{s}', \boldsymbol{\phi}') \\ &\quad = (\mathbf{W}_s(\mathbf{s} - \mathbf{s}')) \odot (\mathbf{W}_{s\phi}(\boldsymbol{\phi} - \boldsymbol{\phi}')).\end{aligned} \tag{25}$$

If this vector is nonzero, no additively separable conditioner $g(\mathbf{s}) + h(\boldsymbol{\phi})$ matches these four condition vectors exactly.

Proof. Expand the four terms. The phase-only terms cancel, leaving the displayed product. An additively separable function has zero mixed contrast for every such quadruple. □

This statement identifies the interaction introduced by the product. It does not claim that an arbitrary nonlinear network receiving concatenated state and phase cannot represent that interaction. Nonzero condition contrast also need not imply a nonzero output contrast, since aggregation or retrieval can suppress it.

Proposition I.7 (Circular continuity at a fixed forecast step)**.** *Fix all node histories and a relative forecast step t, so its value and query projections stay fixed. Write $B(\mathbf{s}) = \mathbf{W}_\phi + \mathrm{diag}(\mathbf{W}_s\mathbf{s})\mathbf{W}_{s\phi}$ and let B_k be its kth row. For the two-harmonic phase encoding and circular distance $d_{\mathrm{circ}}(\rho, \rho') = \min_{j\in\mathbb{Z}} |\rho - \rho' + j|$,*

$$|m_{\epsilon,t,k}(\rho) - m_{\epsilon,t,k}(\rho')| \le 2\pi\sqrt{5}\, V_k \|B_k(\mathbf{s})\|_2\, d_{\mathrm{circ}}(\rho, \rho'), \tag{26}$$

where $V_k = \max_i |v_{i,t,k}|$. For the ideal summary, V_k can be replaced by $(\max_i v_{i,t,k} - \min_i v_{i,t,k})/4$.

Proof. The derivative of $[\sin(2\pi\rho), \cos(2\pi\rho), \sin(4\pi\rho), \cos(4\pi\rho)]$ has constant Euclidean norm $2\pi\sqrt{5}$. Integrate it along the shorter arc and multiply by $\|B_k\|_2$ to bound the condition change. Apply the common-offset sensitivity bound from Proposition I.5, or equation 22 for the ideal operator. □

This is continuity in periodic position, including the day boundary, for a fixed relative forecast step. It does not imply smoothness between independently parameterized forecast heads, nor does it force a predicted sequence to be periodic when the observed histories change.

I.6 Node composition and summary stability

To compare different node counts without node correspondence, regard a node set as its empirical distribution $P = N^{-1}\sum_i \delta_{\mathbf{z}_i}$ over encoded histories. Here $\delta_{\mathbf{z}}$ is a point mass, unrelated to temporal differences. For fixed $t, k, \boldsymbol{\phi}$, set $\mathbf{s}_P = \mathbb{E}_P\mathbf{z}$, $v(\mathbf{z}) = (\mathbf{W}_t^V\mathbf{z})_k$, and $\kappa_P(\mathbf{z}) = \lambda\sigma((\mathbf{W}_K\mathbf{z} + \mathbf{b}_K)_k + \beta_k(\mathbf{s}_P, \boldsymbol{\phi}))$. Then

$$M_\epsilon(P) = \frac{\mathbb{E}_P[\kappa_P(\mathbf{z})v(\mathbf{z})]}{\max\{\mathbb{E}_P\kappa_P(\mathbf{z}), \epsilon\}}. \tag{27}$$

This definition retains the dependence of the condition on the node set; freezing that condition would analyze a different operator.

Proposition I.8 (Stability across node distributions)**.** *Let P, Q be supported on a common compact set $\mathcal{Z}$. Assume $|v(\mathbf{z})| \le V$ and that v is L_v-Lipschitz on $\mathcal{Z}$. Let $\mathbf{w}_{K,k}, \mathbf{w}_{s,k}, \mathbf{w}_{s\phi,k}$ denote the corresponding projection rows, and define*

$$A_k = \frac{\lambda}{4}\left(\|\mathbf{w}_{K,k}\|_2 + |\mathbf{w}_{s\phi,k}\boldsymbol{\phi}|\,\|\mathbf{w}_{s,k}\|_2\right). \tag{28}$$

Choose a uniform $\kappa_* > 0$ below both weight functions on $\mathcal{Z}$, and put $d_* = \max\{\kappa_*, \epsilon\}$. Then

$$|M_\epsilon(P) - M_\epsilon(Q)| \le \frac{\lambda L_v + 2VA_k}{d_*}\, W_1(P, Q), \tag{29}$$

where W_1 is the minimum expected Euclidean distance over couplings of P and Q.

Proof. For any coupling (Z, Z'), $\|\mathbf{s}_P - \mathbf{s}_Q\|_2 \le \mathbb{E}\|Z - Z'\|_2$. The sigmoid is $1/4$-Lipschitz, so the expected difference between $\kappa_P(Z)$ and $\kappa_Q(Z')$ is at most $A_k\mathbb{E}\|Z - Z'\|_2$. Therefore the denominator means differ by at most this amount, while the numerator means differ by at most $(\lambda L_v + VA_k)\mathbb{E}\|Z - Z'\|_2$. Clamping by ϵ is 1-Lipschitz. For numerator n_P and clamped denominator D_P, $|n_P| \le VD_P$ and $D_P \ge d_*$. Subtracting the two ratios gives

$$\left|\frac{n_P}{D_P} - \frac{n_Q}{D_Q}\right| \le \frac{|n_P - n_Q| + V|D_P - D_Q|}{d_*}. \tag{30}$$

Take the infimum over couplings. Compactness and finite parameters ensure a positive lower bound for sigmoid weights, though it may be very small. □

The constant has no explicit node-count factor. It can nevertheless be large for saturated weights or large projection norms. The bound describes stability of the aggregation map in representation space, not equality of source and target representations or forecasting errors. The empirical robustness to source selection involves training different models and is not established by a fixed-parameter result.

Adding or replacing nodes. If $Q = (1-\eta)P + \eta R$ and the representation support has diameter $D_{\mathcal{Z}}$, a coupling retaining the common mass gives $W_1(P, Q) \le \eta D_{\mathcal{Z}}$. Adding r nodes to N existing nodes corresponds to $\eta = r/(N+r)$; replacing r of N nodes admits $\eta = r/N$. Equation 29 thus yields a bound proportional to the changed fraction of nodes. This comparison holds in a shared normalized representation space with fixed parameters and statistics. Arbitrary changes in the normalization statistics are outside its assumptions.

Effect on a receiving node. If the receiver's encoded history and local branch stay fixed, then

$$|\widehat{y}_{i,t}(P) - \widehat{y}_{i,t}(Q)| \le \frac{\sigma_x|\gamma|}{\sqrt{d_m}} \sum_k |q_{i,t,k}|\,|M_{\epsilon,k}(P) - M_{\epsilon,k}(Q)|. \tag{31}$$

This follows directly from the retrieval inner product. Changes to the receiver's own history introduce additional local and query changes.

I.7 Symmetry and factored cross-node communication

Proposition I.9 (Equivariance and uniform replication)**.** *At fixed normalization statistics and timestamps, permuting input nodes permutes the output nodes in the same way. Duplicating every node the same positive integer number of times leaves each copy's prediction unchanged. Both statements hold with the implemented denominator floor.*

Proof. Shared temporal encoding, direct prediction, and query/value projections commute with permutations. The mean state is invariant, as are the numerator and denominator means of every summary coordinate. Retrieval therefore permutes with the receivers. Uniform replication preserves the empirical distribution, mean state, and both aggregation means, including whether the floor is active. Each duplicate has the same local features and query as its original node. □

Proposition I.10 (Factored contribution matrix)**.** *Fix a window and forecast step, and define $Q_{ik} = q_{i,t,k}$ and $B_{jk} = \widetilde{w}_{j,t,k} v_{j,t,k}$, where $\widetilde{w}_{j,t,k} = \kappa_{j,t,k}/(N \max\{\bar{\kappa}_{t,k}, \epsilon\})$. The global corrections admit*

$$\mathbf{r}_t = \frac{QB^\top \mathbf{1}_N}{\sqrt{d_m}}, \qquad \operatorname{rank}(QB^\top) \le d_m. \tag{32}$$

They can be evaluated as $Q(B^\top \mathbf{1}_N)/\sqrt{d_m}$ with $O(Nd_m)$ aggregation and retrieval operations after projections, without materializing $QB^\top$.

Proof. The kth entry of $B^\top \mathbf{1}_N$ is precisely the implemented global summary coordinate. Left multiplication by Q is retrieval. The rank inequality follows from the inner dimension of the two factors. Computing the two matrix-vector products requires work linear in N at fixed d_m. □

The matrix above represents signed node-to-node contributions at a fixed input. It is not a nonnegative attention matrix or a fixed learned adjacency, and its factors depend on the input. The statement does not bound the rank of the complete nonlinear predictor's Jacobian. Including the GRU, local branch, and step-specific projections gives $O(NLd_z^2 + NLT + NTd_z d_m + NTd_m)$ computation, plus condition computation independent of N. Stored node activations still grow with N; only the summary size is independent of it. The factorization explains the linear node-count dependence of the aggregation and retrieval computations.

I.8 A limited argument for conditioning before compression

Proposition I.11 (A collision in fixed sigmoid pooling)**.** *There exist two scalar node sets Z, Z' whose zero-condition normalized sigmoid summaries coincide, but whose summaries differ under a positive shared condition. Any deterministic decoder receiving only that fixed summary and the same condition incurs maximum absolute error at least half the difference between the two conditional summaries when asked to reproduce them.*

Proof. Use each scalar as both logit and value. Let $c = 2\sigma(1) - 1$, $Z = (-1, 1)$, and $Z' = (c, c)$. Then $S_0(Z) = c = S_0(Z')$. For $\beta > 0$, the ratio of the positive node's weight to the negative node's weight strictly decreases, since

$$\frac{\mathrm{d}}{\mathrm{d}\beta} \log \frac{w_+(\beta)}{w_-(\beta)} = \sigma(-1 + \beta) - \sigma(1 + \beta) < 0. \tag{33}$$

Thus $S_\beta(Z) < c = S_\beta(Z')$. The decoder receives identical inputs on both cases and returns one number g. The triangle inequality gives $\max\{|g - S_\beta(Z)|, |g - S_\beta(Z')|\} \geq |S_\beta(Z) - S_\beta(Z')|/2$. □

This example identifies information that one specified fixed scalar summary discards. It is not a separation from arbitrary higher-dimensional summaries or full forecasting architectures with local inputs. It also does not establish that the particular collision occurs in the evaluated traffic data.