AR-MGNET: ADAPTIVE MACRO–MICRO RECONCILIATION MULTI-GRID NETWORK FOR GENERAL SPATIO-TEMPORAL FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatio-temporal forecasting underpins numerous real-world applications, yet existing methods struggle to jointly achieve global-scale efficiency and local-scale fidelity, particularly when system states shift or abrupt events occur. Most approaches rely on fixed spatial partitions or uniform refinement strategies, limiting their ability to adapt to dynamically changing functional regions. To overcome these limitations, we propose AR-MGNet, an Adaptive Macro-Micro Reconciliation Multi-Grid Network that reformulates forecasting as a bidirectional macro-micro reconciliation process. AR-MGNet first aggregates fine-grained observations into functional macro-grids to efficiently solve a coherent macroscopic future, then selectively recovers unresolved micro-scale innovations in event-critical regions through a top-down adaptive refinement pathway. A dynamic multi-grid neighborhood learning module further captures scale-aware spatial dependencies by constructing sparse, relation-aware adjacency matrices that adapt to the current spatial regime. Across four LargeST traffic benchmarks and five diverse general spatio-temporal datasets (wind, air quality, EV charging, electricity, solar), AR-MGNet achieves the best performance in all traffic and most cross-domain datasets, with ablation studies confirming the contribution of each core mechanism. These results demonstrate that adaptive macro-micro reconciliation offers a unified principle for general spatio-temporal forecasting.

1 INTRODUCTION

Spatio-temporal forecasting infers the future states of multiple spatial locations from their observations over a historical window. It jointly characterizes temporal dynamics, spatial interactions, cross-location propagation, and the coupling among different variables (Yu et al., 2018; Bai et al., 2020). Spatio-temporal forecasting has therefore become a fundamental capability in intelligent transportation, urban computing, weather forecasting, energy scheduling, and environmental monitoring (Goswami et al., 2024; Liu et al., 2024c; Woo et al., 2024). Despite substantial progress, existing spatio-temporal forecasting methods still face severe challenges (Das et al., 2024; Zheng et al., 2025). Specifically, congestion, road restrictions, weather disturbances, and abrupt energy-demand changes require state-dependent functional regions beyond fixed adjacency matrices and static partitions (Guo et al., 2019; Bai et al., 2020; Lan et al., 2022; Chen et al., 2025; Xin et al., 2026). Dynamic graphs and global attention enlarge the receptive field while introducing quadratic interactions, adjacency-generation cost, and over-smoothing at large node counts (Fang et al., 2024; Ma et al., 2025a). Distribution shifts, data scarcity, missing sensors, unseen events, and heterogeneous spatial granularities further demand a transferable forecasting mechanism (Ma et al., 2025b; An et al., 2026; Zhang et al., 2026b; Chen & Liang, 2025).

Recently, existing studies have improved spatial relation learning, temporal pattern extraction, computational efficiency, and knowledge transfer to better characterize complex spatio-temporal dynamics. Learnable adjacency matrices, dynamic graphs, hypergraphs, and propagation operators informed by causal or physical priors are used to discover state-dependent spatial dependencies and reduce the constraints imposed by fixed topologies on information exchange (Chen et al., 2025; Zhang et al., 2026a; Xin et al., 2026; Liu et al., 2026). Sequence decomposition, frequency-domain analysis, spatial aggregation, and multi-resolution representations expand receptive fields, coordinate long-term trends, periodic regularities, and local variations, and reduce the modeling cost of long sequences and large-scale spatial networks (Wang et al., 2024; Lin et al., 2024; Fang et al., 2024; Ai et al., 2026; Qin

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.

et al., 2026; Shao et al., 2026). Collectively, these advances are moving spatio-temporal forecasting beyond the one-model-per-dataset paradigm toward models that jointly leverage structural priors, cross-scale information, and transferable knowledge.

However, the above advances are typically designed in isolation to address a particular modeling bottleneck and do not provide a unified adaptive mechanism that jointly adjusts spatial structure, representation scale, and computational resources in response to system states (Li et al., 2024; Chen & Liang, 2025; An et al., 2026). Dynamic relation learning can update connection strengths among nodes, but spatial nodes and grid cells generally remain fixed, limiting the ability to reorganize the forecasting domain when functional regions shift or abrupt events occur (Chen et al., 2025; Xin et al., 2026; Liu et al., 2026). Decomposition, aggregation, and multi-resolution modeling improve long-horizon forecasting efficiency, but often rely on predefined regions, fixed frequency bands, or uniform refinement strategies and may overlook rapidly growing local residuals in a small number of anomalous regions (Song et al., 2026; Zhang et al., 2026b).

To address these limitations, we propose **AR-MGNet**, an **Adaptive Macro–Micro Reconciliation Multi-Grid Network**. AR-MGNet consists of four core modules operating in concert: a bottom-up macro context aggregation module that compresses fine-grained observations into compact functional macro-grids; a top-down adaptive micro reconciliation module that propagates the macroscopic forecast downward and selectively restores fine-scale details; a dynamic multi-grid neighborhood learning module that constructs sparse, relation-aware multi-scale adjacencies; and a consistency-constrained optimization objective that jointly supervises the hierarchical forecast fields. AR-MGNet is applicable to spatio-temporal systems with a sparse relational structure, from road networks and sensor grids to weather stations and power plants, making it a general spatio-temporal forecasting solution. Our contributions are summarized as follows:

- We establish an adaptive macro-micro reconciliation mechanism, a bottom-up pathway aggregates fine-grained states into functional macro-grids for efficient coarse-scale prediction, while a top-down pathway selectively recovers unresolved micro-scale innovations only in event-critical regions via a learnable active mask, which ensures both global coherence and local precision.
- We develop adaptive multi-grid neighborhood learning that dynamically partitions functional grids for each input window and constructs sparse, relation-aware assignment and adjacency matrices for cross-scale restriction and prolongation. A scale-relation contraction then aggregates messages from visible grid levels, preserving coarse regional context and fine event structure in the neighborhood representation.
- We evaluate AR-MGNet on nine benchmarks spanning four LargeST traffic regions and five non-traffic domains: wind power, air quality, EV charging, electricity, and solar energy. AR-MGNet achieves the best performance across all four traffic datasets and the lowest MAE/RMSE on four of the five cross-domain datasets. Analysis results validate AR-MGNet as both accurate and practical for large-scale general spatio-temporal forecasting.

2 Related Work

Graph-based spatio-temporal modeling Graph-based spatio-temporal models represent roads, sensors, weather stations, or spatial grids as nodes and encode geographic proximity, road connectivity, statistical dependence, or physical propagation through edges. Early approaches modeled traffic propagation with diffusion convolution and spatio-temporal graph convolution, while adaptive graph learning uncovered latent dependencies beyond predefined topologies (Yu et al., 2018; Guo et al., 2019; Bai et al., 2020; Wu et al., 2020). Dynamic graphs, mixture-of-experts architectures, and models for missing or irregular observations further improved adaptation to time-varying and incomplete data (Lan et al., 2022; Lee & Ko, 2024; Marisca et al., 2024; Zhang et al., 2024). Linear graph networks, adaptive hypergraphs, causally inspired graphs, and physics-informed graphs have subsequently improved scalability, higher-order interaction modeling, and structural interpretability (Han et al., 2024; Ma et al., 2025a; Chen et al., 2025; Zhang et al., 2026a; Xin et al., 2026; Liu et al., 2026; Qi et al., 2026).

Hierarchical spatio-temporal modeling Hierarchical methods organize and fuse spatio-temporal representations across node–region–global, short–long-term, or low–high-frequency resolutions. Regional hierarchies, spatial patches, and implicit multi-resolution representations reduce the cost of large-scale spatial interaction and use macro-level states to guide fine-grained forecasting (Wang et al.,

2026; Fang et al., 2024; Jiang et al., 2025). Temporal decomposition, frequency-domain modeling, and adaptive scale pathways further balance long-term trends, periodic structures, and local variations (Wang et al., 2024; Nie et al., 2023; He et al., 2025; Qin et al., 2026; Shao et al., 2026).

Foundation-based spatio-temporal modeling Foundation-model-based approaches learn transferable representations from multi-city, multi-task, or multimodal data and adapt them through fine-tuning, prompting, or parameter-efficient modules. Spatio-temporal graph pretraining and predictive representation learning reduce dependence on downstream labels (Li et al., 2023b; Zheng et al., 2025), while general-purpose time-series foundation models and language-model reprogramming enable few-shot and zero-shot forecasting (Goswami et al., 2024; Woo et al., 2024; Das et al., 2024; Jin et al., 2024). Prompt tuning, continual learning, cross-region pretraining, and domain adaptation improve transfer across datasets and cities; multimodal prompting, retrieval, and dataset distillation broaden available knowledge and reduce training costs (Chen & Liang, 2025; An et al., 2026; Wen et al., 2026; Ruan et al., 2026; Kwon et al., 2026).

3 Preliminaries

3.1 Problem Formulation

We study general spatio-temporal forecasting over a sparse, potentially time-varying relational structure:

$$\begin{aligned} &\mathcal{G}_t = (\mathcal{V}, \mathcal{E}_t, \mathcal{R}), \quad \mathbf{X} \in \mathbb{R}^{B\times T\times N\times C_{\text{in}}}, \quad \mathbf{Y} \in \mathbb{R}^{B\times H\times N\times C_{\text{out}}}, \\ &\widehat{\mathbf{Y}} = F_\theta(\mathbf{X}, \mathcal{G}_{t-T+1:t}, \mathbf{Z}_{\text{time}}, \mathbf{Z}_{\text{node}}) \in \mathbb{R}^{B\times H\times N\times C_{\text{out}}}. \end{aligned} \tag{1}$$

where B is the batch size; T and H are the observation and forecast horizons; N is the number of spatial entities; and C_{in} and C_{out} are the input and output channel counts. The vertex set $\mathcal{V}$ and sparse edge set $\mathcal{E}_t$ describe the observed system, while $\mathcal{R}$ indexes relation types such as road direction, transport lag, functional similarity, or domain-specific connectivity. The descriptors $\mathbf{Z}_{\text{time}}$ and $\mathbf{Z}_{\text{node}}$ encode temporal identity and node attributes.

3.2 Multi-Grid Forecast Space

AR-MGNet constructs an L-level multi-grid forecast space $\mathcal{G}^{(0)} \to \cdots \to \mathcal{G}^{(L-1)}$, where level 0 is the observation grid, level $L-1$ is the macro grid, and $N = N_0 > N_1 > \cdots > N_{L-1}$. A window sparse assignment $\mathbf{P}_\ell$ defines the paired restriction and prolongation operators:

$$\big[\mathcal{R}_\ell(\mathbf{H})\big]_j = \frac{\sum_i P_{ij}^{(\ell)} \mathbf{H}_i}{\sum_i P_{ij}^{(\ell)} + \epsilon}, \qquad \big[\mathcal{U}_\ell(\mathbf{Z})\big]_i = \sum_{j\in\mathcal{C}_\ell(i)} P_{ij}^{(\ell)} \mathbf{Z}_j. \tag{2}$$

where $\mathcal{C}_\ell(i)$ is the constant-size candidate-parent set of node i, $P_{ij}^{(\ell)}$ is its normalized assignment to parent j, and ϵ ensures numerical stability. Each row of $\mathbf{P}_\ell \in \mathbb{R}^{B\times N_\ell \times N_{\ell+1}}$ has at most K nonzero entries, so aggregation and forecast transport share the same sparse cross-scale semantics without constructing a dense node-pair matrix. Meanwhile, in multi-grid forecast space, each level carries a supervised future field $\widehat{\mathbf{Y}}^{(\ell)} \in \mathbb{R}^{B\times H\times N_\ell\times C_{\text{out}}}$. Bottom-up restriction aggregates coherent macro context, while top-down prolongation transports the explicit macro future to finer grids for selective active recovery. Figure 1 summarizes this adaptive macro-micro bidirectional forecast space.

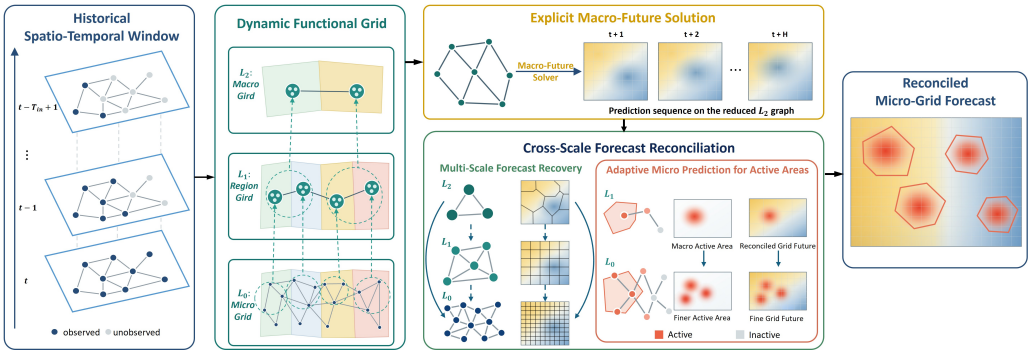


Figure 1: Adaptive macro–micro forecast reconciliation in AR-MGNet.

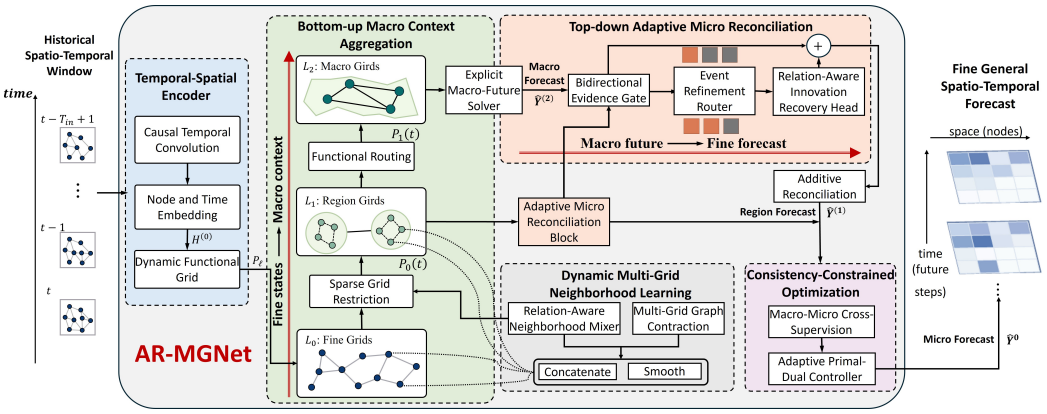


Figure 2: Overall architecture of AR-MGNet, which mainly consists of Bottom-up Macro Context Aggregation, Top-down Adaptive Micro Reconciliation, and Dynamic Multi-Grid Neighborhood Learning, respectively.

4 METHOD

4.1 MODEL ARCHITECTURE

AR-MGNet proposes a bottom-up then top-down bidirectional hierarchy: macro context is first aggregated for efficient coarse forecasting (Section 4.2), then adaptively reconciled to micro-scale forecasts with selective refinement in event-critical regions (Section 4.3), supported by dynamic multi-grid neighborhood modeling (Section 4.4). At level ℓ, $\mathbf{H}^{(\ell)}$ is the N_ℓ-node state, $\mathbf{P}_\ell$ is the sparse fine-to-coarse assignment, $\mathcal{G}^{(\ell)}$ is the typed graph, and $\widehat{\mathbf{Y}}^{(\ell)}$ is the forecast field; $\widehat{\mathbf{Y}}^{(0)}$ is deployed. Figure 2 presents the complete architecture of AR-MGNet.

4.2 BOTTOM-UP MACRO CONTEXT AGGREGATION

Sparse Grid Restriction The encoder produces the fine state $\mathbf{H}^{(0)}$, local change $\boldsymbol{\delta}^{(0)}$, and future anchor $\mathbf{A}^{(0)}_{\text{future}}$. At each level, the shared sparse assignment restricts states and observable anchor evidence before a relation-aware smoother updates the coarser graph:

$$\mathbf{H}^{(\ell+1)} = \mathcal{S}_{\ell+1}\Big(\mathcal{R}_\ell(\mathbf{H}^{(\ell)}), \mathcal{G}^{(\ell+1)}\Big), \qquad \ell = 0, \dots, L-2. \tag{3}$$

where $\mathcal{S}_{\ell+1}$ is the sparse direction-aware smoother. The same $\mathcal{R}_\ell$ restricts event evidence and the anchor, so the receptive field expands while predictable level-specific structure is retained.

Explicit Macro-Future Solver In macro context aggregation, the smallest graph receives direct full-horizon supervision:

$$\widehat{\mathbf{Y}}^{(L-1)} = \mathbf{A}^{(L-1)}_{\text{future}} + f_{\text{macro}}\Big(\mathbf{H}^{(L-1)}, \mathcal{G}^{(L-1)}\Big). \tag{4}$$

where f_{macro} is a Macro Graph Mixer followed by horizon projection. Global evolution is therefore solved on $N_{L-1} \ll N$ functional regions and then reused by every finer level. Appendix C details routing and contraction of explicit macro-future solver.

4.3 TOP-DOWN ADAPTIVE MICRO RECONCILIATION

Bidirectional Evidence Gate At each finer level, the prolonged macro future provides a complete forecast while the encoder retains local observations that may contradict it. Let $\mathbf{E}^{(\ell)}_{\uparrow} = \phi^{(\ell)}_{\uparrow}(\mathbf{H}^{(\ell)}, \boldsymbol{\delta}^{(\ell)})$ denote upward local evidence and $\mathbf{E}^{(\ell)}_{\downarrow} = \phi^{(\ell)}_{\downarrow}(\mathcal{U}_\ell(\mathbf{H}^{(\ell+1)}), \overline{\mathbf{Y}}^{(\ell)})$ downward forecast evidence; define their discrepancy as $\Delta\mathbf{E}^{(\ell)} = |\mathbf{E}^{(\ell)}_{\uparrow} - \mathbf{E}^{(\ell)}_{\downarrow}|$. The gate uses this cross-scale signal:

$$\begin{aligned}
&\overline{\mathbf{Y}}^{(\ell)} = \mathbf{A}^{(\ell)}_{\text{future}} + \mathcal{U}_\ell\Big(\widehat{\mathbf{Y}}^{(\ell+1)} - \mathbf{A}^{(\ell+1)}_{\text{future}}\Big), \qquad \mathbf{G}^{(\ell)} = \sigma\Big(\mathbf{W}^{(\ell)}_g[\mathbf{E}^{(\ell)}_{\uparrow} \| \mathbf{E}^{(\ell)}_{\downarrow} \| \Delta\mathbf{E}^{(\ell)}] + \mathbf{b}^{(\ell)}_g\Big), \\
&\mathbf{Z}^{(\ell)} = \mathbf{G}^{(\ell)} \odot \mathbf{E}^{(\ell)}_{\uparrow} + (\mathbf{1} - \mathbf{G}^{(\ell)}) \odot \mathbf{E}^{(\ell)}_{\downarrow}.
\end{aligned} \tag{5}$$

where $\overline{\mathbf{Y}}^{(\ell)}$ is the anchor-preserving inherited forecast, $\phi_{\uparrow}^{(\ell)}$ encodes fine-scale state and recent change, and $\phi_{\downarrow}^{(\ell)}$ encodes prolonged parent state and forecast context. Thus $\mathbf{G}^{(\ell)}$ resolves evidence direction channel by channel and supplies the fused state for event-selective innovation recovery, which concentrates fine computation on hotspots.

Event-Selective Innovation Recovery The residual disagreement $|\mathbf{E}_{\uparrow}^{(\ell)} - \mathbf{E}_{\downarrow}^{(\ell)}|$, together with the learned uncertainty score and neighborhood context, determines where fine computation is necessary. Under budget β_ℓ, routed innovations reconcile the inherited field:

$$\begin{aligned} \mathbf{Q}_\ell &= \mathrm{TopBudget}(\mathbf{s}^{(\ell)}, \beta_\ell), \\ \widehat{\mathbf{Y}}^{(\ell)} &= \overline{\mathbf{Y}}^{(\ell)} + \mathbf{Q}_\ell \odot f_{\mathrm{innovation}}^{(\ell)}(\mathbf{Z}^{(\ell)}, \overline{\mathbf{Y}}^{(\ell)}, \mathbf{N}^{(\ell)}). \end{aligned} \tag{6}$$

where $\mathbf{s}^{(\ell)}$ is the event score, $\mathbf{Q}_\ell$ is the active-region mask, $\mathbf{N}^{(\ell)}$ is neighborhood context, and $f_{\mathrm{innovation}}^{(\ell)}$ is the Innovation Recovery Head. By searching for hotspots or active spatio-temporal events, the Recovery Head enables AR-MGNet to quickly and accurately identify key spatio-temporal ranges for prediction, greatly improving the model's inference and generalization capabilities.

4.4 Dynamic Multi-Grid Neighborhood Learning

For each input window $\mathbf{X}$, the router produces a sparse assignment $\mathbf{P}_\ell^\star(\mathbf{X})$, allowing macro cells to follow the current spatial regime rather than a fixed partition. The resulting Multi-Grid Graph Contraction and its Relation-Aware Neighborhood Mixer are summarized by:

$$\begin{aligned} \mathbf{A}_r^{(\ell+1)}(\mathbf{X}) &= \mathcal{N}_{D_g}\left(\mathbf{P}_\ell^\star(\mathbf{X})^\top \mathbf{A}_r^{(\ell)} \mathbf{P}_\ell^\star(\mathbf{X})\right), \\ \mathbf{N}_i^{(\ell)} &= \sum_{q \in \mathcal{S}(\ell)} \sum_{r \in \mathcal{R}} \alpha_{i,r}^{(\ell \leftarrow q)} \left[\mathcal{T}_{\ell \leftarrow q}\left(\widetilde{\mathbf{A}}_r^{(q)} \mathbf{H}^{(q)}\right)\right]_i. \end{aligned} \tag{7}$$

where the first expression contracts every typed relation through the window-dependent grid; $\mathcal{N}_{D_g}$ retains and normalizes at most D_g coarse neighbors per node. The second expression associates relation messages from the visible scales $q \in \mathcal{S}(\ell)$: $\mathcal{T}_{\ell \leftarrow q}$ transports a message to level ℓ, while the normalized relevance $\alpha_{i,r}^{(\ell \leftarrow q)}$ selects the scale–relation pairs that agree with node i. Thus $\mathbf{N}_i^{(\ell)}$ jointly preserves coarse regional evolution and fine event structure for the innovation recovery in Equation 6. For fixed K, D_g, and $|\mathcal{R}|$ on bounded-degree graphs, the sparse routing, contraction, and relation-mixing core scales linearly with the number of nodes.

4.5 Consistency-Constrained Optimization

Let $\mathbf{Y}^{(0)} = \mathbf{Y}$ and $\mathbf{Y}^{(\ell+1)} = \mathcal{R}_\ell(\mathbf{Y}^{(\ell)})$. With the mask-normalized discrepancy $\mathcal{D}$, prediction and hierarchical risks are:

$$\begin{aligned} \mathcal{L}_{\mathrm{pred}} &= \mathcal{D}(\widehat{\mathbf{Y}}^{(0)}, \mathbf{Y}^{(0)}), \\ \mathcal{L}_{\mathrm{hier}} &= \frac{1}{2(L-1)} \sum_{\ell=0}^{L-2} \left[\mathcal{D}(\widehat{\mathbf{Y}}^{(\ell+1)}, \mathbf{Y}^{(\ell+1)}) + \mathcal{D}(\mathcal{R}_\ell \widehat{\mathbf{Y}}^{(\ell)}, \widehat{\mathbf{Y}}^{(\ell+1)})\right]. \end{aligned} \tag{8}$$

where $\mathcal{L}_{\mathrm{pred}}$ trains the deployed field and $\mathcal{L}_{\mathrm{hier}}$ supervises macro fields while aligning adjacent predictions. Macro–micro supervision separates active innovations from inherited forecasts. With $\Delta\mathbf{Y}^{(\ell),*} = \mathbf{Y}^{(\ell)} - \overline{\mathbf{Y}}^{(\ell)}$,

$$\begin{aligned} \mathcal{L}_{\mathrm{MM}} &= \mathcal{D}(\widehat{\mathbf{Y}}^{(L-1)}, \mathbf{Y}^{(L-1)}) + \frac{1}{L-1} \sum_{\ell=0}^{L-2} \mathcal{I}_\ell, \\ \mathcal{I}_\ell &= \mathcal{D}_{\mathbf{Q}_\ell}(\Delta\widehat{\mathbf{Y}}^{(\ell)}, \Delta\mathbf{Y}^{(\ell),*}) + \mathcal{D}_{\mathbf{1}-\mathbf{Q}_\ell}(\overline{\mathbf{Y}}^{(\ell)}, \mathbf{Y}^{(\ell)}). \end{aligned} \tag{9}$$

where $\Delta\widehat{\mathbf{Y}}^{(\ell)}$ is the recovered innovation and $\mathcal{I}_\ell$ supervises active recovery and inactive inheritance. We cast $\mathcal{L}_{\mathrm{hier}}$ and $\mathcal{L}_{\mathrm{MM}}$ as normalized constraints $\bar{c}_k = (\mathcal{L}_k - \varepsilon_k)/(\xi_k + \epsilon)$ and optimize an augmented primal–dual objective:

$$\begin{aligned} \mathcal{L}_{\mathrm{APD}} &= \mathcal{L}_{\mathrm{pred}} + \sum_k \lambda_k [\bar{c}_k]_+ + \frac{\mu}{2} \sum_k [\bar{c}_k]_+^2, \\ \lambda_k^{(q+1)} &= [\lambda_k^{(q)} + \eta_\lambda \bar{c}_k^{(q)}]_+, \qquad k \in \{\mathrm{hier}, \mathrm{MM}\}. \end{aligned} \tag{10}$$

where $\lambda_k \geq 0$ is an adaptive multiplier, ε_k a tolerance, ξ_k the moving violation scale, μ the penalty, and $[x]_+ = \max(x, 0)$.

5 EXPERIMENTS

Datasets We evaluate nine benchmarks spanning traffic and five additional spatio-temporal domains. LargeST (Liu et al., 2024a) provides four California road-sensor regions: SD (716 nodes), GBA (2,352), GLA (3,834), and CA (8,600). The cross-domain suite contains SDWPF wind power, KnowAir air quality, UrbanEV charging demand, Electricity, and Solar-Energy. These five datasets use their public chronological splits, and all methods within a dataset share identical windows, normalization statistics, graph inputs, and target channels. Dataset periods, intervals, and protocol summaries are reported in Appendix Table 6.

Metrics We report mean absolute error (MAE) and root mean squared error (RMSE) on all nine datasets, together with mean absolute percentage error (MAPE) on the four traffic regions. MAPE is omitted for non-traffic benchmarks because zero and near-zero wind, pollutant, charging, and energy targets make percentage errors unstable. Lower values indicate better performance, and the exact reductions are given in Appendix B.3.

Baselines We compare AR-MGNet with 16 baselines, which comprise the state-of-the-art traffic forecasting model HieraMix (Wang et al., 2026) and advanced cross-domain spatio-temporal forecasting models PatchSTG (Fang et al., 2024), as well as other competitive models including HIEST (Ma et al., 2023), PDFormer (Jiang et al., 2023), DGCRN (Li et al., 2023a), TESTAM (Lee & Ko, 2024), STID (Shao et al., 2022), GSNet (Kong et al., 2025), and LSTNN (Wang et al., 2025), FreTS (Yi et al., 2023), TSMixer (Chen et al., 2023), TimeMixer (Wang et al., 2024), RPMixer (Yeh et al., 2024), EiFormer (Sun et al., 2025), iTransformer (Liu et al., 2024b), and NEST (Ai et al., 2026) provide the strongest hierarchical comparisons.

Implementation Details AR-MGNet is implemented in PyTorch with three functional-grid levels, hidden width 96 by default, depthwise causal convolutions with dilations $1/2/4$, dropout 0.1, and Top-K routing with $K \in \{2, 3, 4\}$. Training uses sparse graph operations, automatic mixed precision, the masked fine-grid forecasting objective, and the consistency-constrained primal–dual updates in Section 4.5. Experiments run on a single NVIDIA RTX 4090 GPU, with micro-batching and gradient accumulation where required by memory.

5.1 MAIN RESULTS

Large-scale traffic forecasting As shown in Table 1, AR-MGNet achieves consistent state-of-the-art performance in all traffic benchmarks. Relative to HieraMix, the strongest baseline across the four regions, the regional averages of MAE, RMSE, and MAPE decrease by 0.9%, 1.2%, and 5.7%, respectively. The gain is especially clear on GBA, where AR-MGNet lowers RMSE by 2.6% and MAPE by 13.8%. These consistent improvements across traffic networks ranging from 716 to 8,600 nodes demonstrate that explicit macro forecasting and micro reconciliation retain the accuracy of AR-MGNet as spatial scale grows.

Table 1: Prediction performance on large-scale traffic forecasting benchmarks. Lower is better.

Model	SD			GBA			GLA			CA		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
HIEST (Ma et al., 2023)	17.94	29.03	10.33	22.50	35.75	14.02	22.37	35.26	11.84	22.06	35.03	14.18
PDFormer (Jiang et al., 2023)	15.84	25.91	9.70	–	–	–	–	–	–	–	–	–
DGCRN (Li et al., 2023a)	15.60	25.90	9.93	19.11	31.21	14.17	–	–	–	–	–	–
TESTAM (Lee & Ko, 2024)	18.51	30.42	10.42	19.57	32.32	13.94	–	–	–	–	–	–
STID (Shao et al., 2022)	18.21	29.97	10.18	20.82	34.43	13.19	19.83	32.33	10.38	19.11	31.45	11.68
FreTS (Yi et al., 2023)	23.95	37.12	14.89	26.87	40.55	18.39	27.76	42.00	15.85	27.58	42.50	18.47
TSMixer (Chen et al., 2023)	16.81	29.79	10.52	19.93	34.83	16.56	17.86	30.89	10.96	16.78	29.42	12.43
TimeMixer (Wang et al., 2024)	17.15	28.30	9.74	21.74	35.15	13.85	19.76	31.96	10.81	18.16	30.11	11.61
RPMixer (Yeh et al., 2024)	16.72	29.64	10.63	19.05	32.75	15.26	17.57	30.42	10.56	16.75	29.28	12.36
PatchSTG (Fang et al., 2024)	15.93	28.19	9.69	18.94	33.02	13.17	17.53	29.72	10.02	16.69	28.58	11.50
EiFormer (Sun et al., 2025)	17.90	32.06	10.79	20.39	35.35	15.58	18.91	33.05	11.00	18.22	32.07	12.89
GSNet (Kong et al., 2025)	16.65	27.33	9.62	19.37	31.98	13.15	18.81	30.65	10.09	17.61	29.15	11.15
LSTNN (Wang et al., 2025)	15.83	27.08	9.59	18.48	31.59	13.11	17.48	29.71	9.95	16.58	28.24	11.05
HieraMix (Wang et al., 2026)	14.80	25.06	9.22	17.73	30.67	12.71	16.45	28.03	9.53	15.55	27.05	10.55
NEST (Ai et al., 2026)	24.41	39.09	17.11	23.26	37.04	21.49	28.38	45.38	18.11	27.16	43.87	21.19
AR-MGNet	**14.64**	**24.94**	**8.94**	**17.62**	**29.87**	**10.95**	**16.24**	**27.95**	**9.32**	**15.44**	**26.76**	**10.40**

Cross-domain spatio-temporal forecasting AR-MGNet also shows great performance in cross-domain spatio-temporal forecasting. As shown in Table 2, AR-MGNet achieves the best result in 8 of 10 metric cells across the five non-traffic datasets. On UrbanEV and Solar-Energy, AR-MGNet reduces the strongest MAE/RMSE values by 10.7%/4.0%, and 19.0%/16.4%, respectively. Thus, the results of various universal spatio-temporal prediction tasks across irregular transport fields, periodic energy series, sparse charging demand, and wind generation have demonstrated the generalization and robustness of AR-MGNet.

Table 2: Prediction performance on non-traffic spatio-temporal benchmarks. Lower is better.

Model	SDWPF		KnowAir		UrbanEV		Electricity		Solar-Energy	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
STID (Shao et al., 2022)	0.2348	0.4092	16.66	25.91	4.89	11.28	0.2355	0.3819	0.2182	0.4429
PatchSTG (Fang et al., 2024)	0.2295	**0.3926**	16.08	24.70	5.16	11.53	0.2295	0.3699	0.2245	0.4465
iTransformer (Liu et al., 2024b)	0.2250	0.4000	17.03	26.63	5.44	12.20	0.2400	0.3847	0.2370	0.4506
TimeMixer (Wang et al., 2024)	**0.2226**	0.4036	16.76	25.65	6.86	15.96	0.2470	0.3912	0.2590	0.4347
HieraMix (Wang et al., 2026)	0.2335	0.3995	15.90	24.38	5.19	11.24	0.2437	0.3909	0.2322	0.4554
NEST (Ai et al., 2026)	0.2342	0.3992	16.02	24.43	5.35	11.25	0.2350	0.3755	0.2110	0.4483
AR-MGNet	0.2243	0.3929	**14.84**	**22.89**	**4.36**	**10.80**	**0.1990**	**0.3286**	**0.1709**	**0.3634**

5.2 Model Analysis

Ablations Figure 3 evaluates four controlled variants of AR-MGNet on CA, GLA, SDWPF, and Electricity. Removing each of the four submodules of AR-MGNet increases forecasting errors. Averaged over applicable metrics, removing Macro Aggregation, Dynamic Multi-Grid Learning, Adaptive Micro Reconciliation, and Consistency Constraints increases error by 14.9%, 15.4%, 14.9%, and 14.5%. Specifically, in traffic datasets, the weakening of Macro Aggregation and Dynamic Multi-Grid Learning is more pronounced; cross-domain spatio-temporal prediction benefits from Adaptive Micro Reconciliation in key hotspot regions.

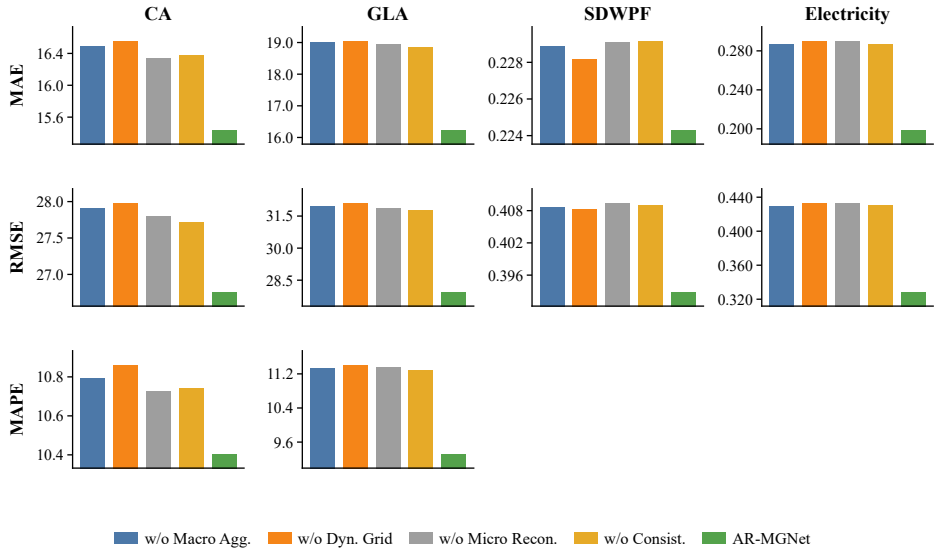


Figure 3: Ablation of the four central AR-MGNet components on CA, GLA, SDWPF, and Electricity.

Spatio-temporal prediction visualization Figure 4 compares system trajectories, high-change nodes, prediction fields, and errors of traffic forecasts on LargeST-SD. It can be seen that the trajectory of AR-MGNet is closer to the ground truth. AR-MGNet also yields lower errors than NEST and HieraMix. The visualization results of non-traffic dataset predictions are also shown in Appendix Figure 9, which all demonstrate the accuracy and stability of AR-MGNet in various general spatio-temporal forecasting tasks.

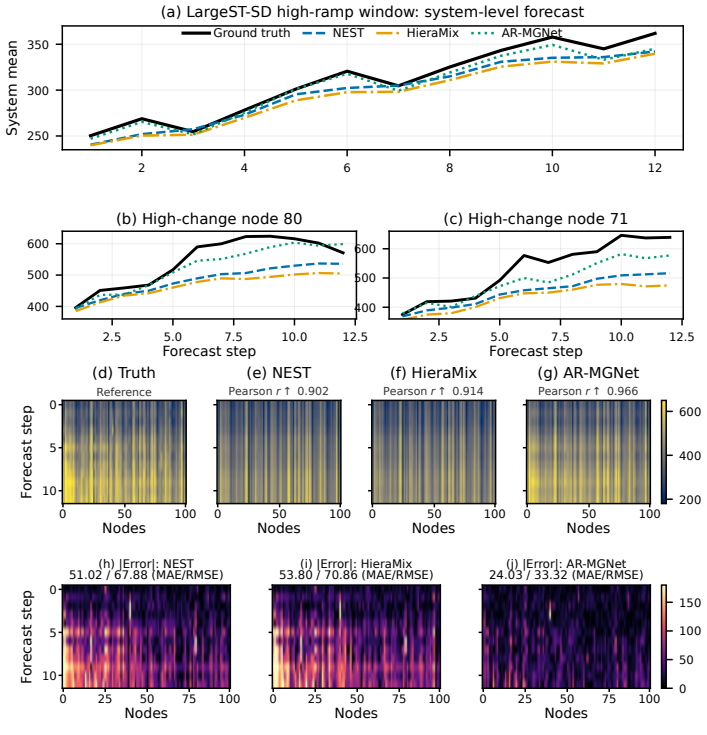


Figure 4: Traffic forecasts on a high-change LargeST-SD window.

Efficiency analysis Table 3 shows the training and inference time of three strong methods. The results show AR-MGNet has the lowest inference time on GBA (14.2 s) and GLA (17.0 s) and the lowest GLA training time (105.8 s/epoch). NEST trains in 75.0 s/epoch on GBA, while AR-MGNet trains in 92.2 s/epoch and achieves lower forecasting errors. Therefore, Table 3 demonstrates that AR-MGNet combines high forecasting accuracy with competitive efficiency.

Table 3: Training and inference time on GBA and GLA (Batch Size = 64, single RTX 4090 GPU).

Model	GBA		GLA	
	Training Time (s/epoch)	Inference Time (s)	Training Time (s/epoch)	Inference Time (s)
HieraMix	94.6	26.4	155.3	38.7
NEST	**75.0**	20.0	137.0	37.0
AR-MGNet	92.2	**14.2**	**105.8**	**17.0**

Analysis on the number of grid scales Table 4 shows the metrics of AR-MGNet for traffic and cross-domain prediction tasks using different grid scales. $L = 3$ gives the best CA MAE/RMSE and the best or tied-best KnowAir RMSE. Increasing to $L = 4$ lowers KnowAir MAE only from 14.84 to 14.71 but raises RMSE from 22.89 to 23.00. Thus, $L = 3$ is the balanced default across three of the four dataset–metric cells, rather than the optimum of every individual cell.

Table 4: Effect of the number of grid levels of AR-MGNet on CA and KnowAir.

	CA		KnowAir	
Levels	MAE	RMSE	MAE	RMSE
1	35.25	55.28	15.31	23.43
2	<u>22.98</u>	<u>35.86</u>	14.95	23.21
3	**15.44**	**26.76**	<u>14.84</u>	**22.89**
4	23.44	37.30	**14.71**	<u>23.00</u>

Noise robustness We evaluate spatio-temporal forecasting under different noise conditions on Electricity. Figure 5 and Appendix Table 7 show that AR-MGNet reduces NEST's pinball loss/median MAE by $11.5\%/9.2\%$ at $\sigma = 0.05$ and $12.8\%/10.7\%$ at $\sigma = 0.20$. Its high-noise coverage error is

also lower (0.043 versus 0.078), demonstrating that AR-MGNet can enhance the expression of spatial intrinsic relationships through neighborhood learning, thereby maintaining strong robustness even when signals are disturbed.

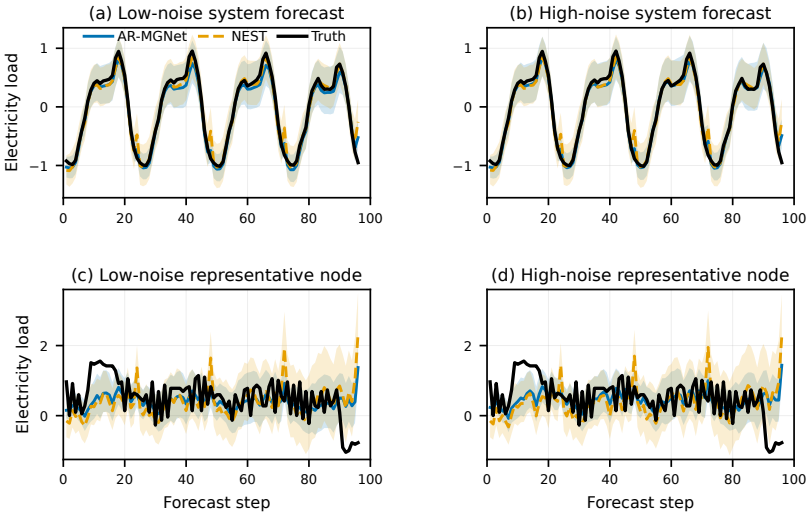


Figure 5: Electricity quantile forecasts under low and high input noise.

Forecast trajectory embeddings Figure 6 compares all three forecast-trajectory embeddings in one standardized space. AR-MGNet ranks first in silhouette, Davies–Bouldin, and 7-NN diagnostics. These scores demonstrate that AR-MGNet has strong spatiotemporal prediction ability for both macro trends and micro active events.

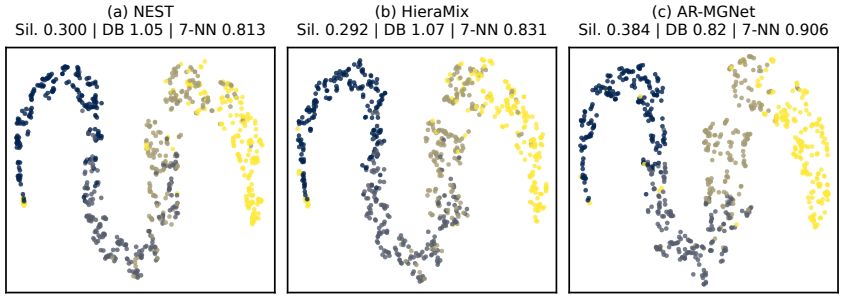


Figure 6: Forecast-trajectory embeddings on the SD case from (a) NEST, (b) HieraMix, and (c) AR-MGNet. Colors denote ground-truth traffic-flow quartiles. Silhouette (higher), Davies–Bouldin (lower), and 7-NN accuracy (higher) are computed in the common standardized high-dimensional trajectory space.

6 Conclusion and Future Work

In this work, we proposed AR-MGNet, which introduces a bidirectional macro-micro reconciliation paradigm that explicitly separates efficient coarse-scale evolution from selective fine-scale innovation recovery. Meanwhile, AR-MGNet constructs a dynamic multi-grid neighborhood learning mechanism that adaptively partitions functional grids and constructs sparse cross-scale adjacencies, and a consistency-constrained optimization that jointly supervises the hierarchical forecast fields. Across nine datasets spanning traffic, wind, air quality, EV charging, electricity, and solar energy, AR-MGNet achieves the lowest errors on all four traffic datasets and the best result in most non-traffic metric cells, with competitive efficiency. Ablations and analysis experiments validate the necessity of each designed component. These results demonstrate that AR-MGNet offers a strong cross-domain framework for general spatio-temporal forecasting. Future work will extend it to multimodal settings and explore its potential as a foundation model for rapid domain adaptation.

7 AI USE STATEMENT

In this study, we employed a Large Language Model (LLM) solely as a writing-assistance tool to improve the readability and overall presentation of the manuscript. All substantive research, data analysis, and intellectual reasoning were conducted exclusively by the authors. Every suggestion generated by the LLM was carefully examined, adapted where necessary, and validated by us to guarantee its consistency with our research objectives. The authors take full and exclusive responsibility for all content presented herein. The LLM contributed to the following specific tasks:

• **Text Refinement:** LLM helped correct grammatical issues, refine sentence structures, and maintain uniform terminology and stylistic tone throughout the paper.

• **Coherence and Structure:** LLM assisted in restructuring the flow of arguments, improving the coherence between sections and paragraphs, and ensuring smoother transitions that enhance the overall narrative clarity.

8 ETHICS STATEMENT

Our work only focuses on the scientific problem, so there is no potential ethical risk.

9 REPRODUCIBILITY STATEMENT

We involve the implementation details in Appendix B, including dataset descriptions, metric calculation and experiment configuration. The source code is provided in supplementary materials for reproducibility.

REFERENCES

Yinghao Ai, Yukai Zhou, Ruoxi Jiang, Junyi An, Chao Qu, Zhijian Zhou, Shiyu Wang, Fenglei Cao, Zenglin Xu, Furao Shen, and Yuan Qi. Nested spatio-temporal time series forecasting, 2026.

Rui An, Yifeng Zhang, Ziran Liang, Wenqi Fan, Yuxuan Liang, Xuequn Shang, and Qing Li. Dambast: Domain-adaptive mamba for efficient urban spatio-temporal prediction. In *2026 IEEE 42nd International Conference on Data Engineering (ICDE)*, 2026. URL https://arxiv.org/abs/2506.18939. Accepted paper; arXiv:2506.18939.

Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. Adaptive graph convolutional recurrent network for traffic forecasting. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17804–17815, 2020.

Jiawen Chen, Qi Shao, Duxin Chen, and Wenwu Yu. Decoupling spatio-temporal prediction: When lightweight large models meet adaptive hypergraphs. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pp. 167–178. Association for Computing Machinery, 2025. doi: 10.1145/3711896.3736904. URL https://doi.org/10.1145/3711896.3736904.

Si-An Chen, Chun-Liang Li, Nate Yoder, Sercan O. Arik, and Tomas Pfister. TSMixer: An all-MLP architecture for time series forecasting. *Transactions on Machine Learning Research*, 2023.

Wei Chen and Yuxuan Liang. Expand and compress: Exploring tuning principles for continual spatio-temporal graph forecasting. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/cb2266111eadcfa2c02187ace64e2183-Paper-Conference.pdf.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 10148–10167. PMLR, 2024.

Yuchen Fang, Yuxuan Liang, Bo Hui, Zezhi Shao, Liwei Deng, Xu Liu, Xinke Jiang, and Kai Zheng. Efficient large-scale traffic forecasting with transformers: A spatial data management perspective, 2024.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML '24. JMLR.org, 2024.

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.

Jindong Han, Weijia Zhang, Hao Liu, Tao Tao, Naiqiang Tan, and Hui Xiong. Bigst: Linear complexity spatio-temporal graph neural network for traffic forecasting on large-scale road networks. *Proceedings of the VLDB Endowment*, 17(5):1081–1090, 2024. doi: 10.14778/3641204.3641217. URL https://doi.org/10.14778/3641204.3641217.

Sicheng He, Junzhong Ji, and Minglong Lei. Decomposed spatio-temporal mamba for long-term traffic prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39:11772–11780, 2025. doi: 10.1609/aaai.v39i11.33281.

Jiawei Jiang, Chengkai Han, Wayne Xin Zhao, and Jingyuan Wang. PDFormer: Propagation delay-aware dynamic long-range transformer for traffic flow prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 4365–4373, 2023. doi: 10.1609/aaai.v37i4.25556.

Ruoxi Jiang, Xiao Zhang, Karan Jakhar, Peter Y. Lu, Pedram Hassanzadeh, Michael Maire, and Rebecca Willett. Hierarchical implicit neural emulators. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=90qt3kVYaE.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-llm: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/680b2a8135b9c71278a09cafb605869e-Paper-Conference.pdf.

Weiyang Kong, Kaiqi Wu, Sen Zhang, and Yubao Liu. GraphSparseNet: A novel method for large scale traffic flow prediction, 2025.

Taehyung Kwon, Yeonje Choi, Yeongho Kim, and Kijung Shin. Effective dataset distillation for spatio-temporal forecasting with bi-dimensional compression, 2026.

Shiyong Lan, Yitong Ma, Weikang Huang, Wenwu Wang, Hongyu Yang, and Pyang Li. Dstagnn: Dynamic spatial-temporal aware graph neural network for traffic flow forecasting. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 11906–11917. PMLR, 2022. URL https://proceedings.mlr.press/v162/lan22a.html.

Hyunwook Lee and Sungahn Ko. Testam: A time-enhanced spatio-temporal attention model with mixture of experts. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/0cb310ed8121549488fea8e8c2056096-Paper-Conference.pdf.

Fuxian Li, Jie Feng, Huan Yan, Guangyin Jin, Fan Yang, Funing Sun, Depeng Jin, and Yong Li. Dynamic graph convolutional recurrent network for traffic prediction: Benchmark and solution. *ACM Transactions on Knowledge Discovery from Data*, 17(1):9:1–9:21, 2023a.

Zhonghang Li, Lianghao Xia, Yong Xu, and Chao Huang. Gpt-st: Generative pre-training of spatio-temporal graph neural networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 70229–70246, 2023b. doi: 10.52202/075280-3077.

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*, ICML '24. JMLR.org, 2024.

Shengsheng Lin, Weiwei Lin, Wentai Wu, Haojun Chen, and Junjie Yang. Sparsetsf: Modeling long-term time series forecasting with 1k parameters. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 30211–30226. PMLR, 2024. URL https://proceedings.mlr.press/v235/lin24n.html.

Aoyu Liu, Liming Wei, and Yaying Zhang. Decoupling universal laws and environmental heterogeneity: A physics-inspired framework for robust spatio-temporal forecasting. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=qiy68EEwhh.

Xu Liu, Yutong Xia, Yuxuan Liang, Junfeng Hu, Yiwei Wang, Lei Bai, Chao Huang, Zhenguang Liu, Bryan Hooi, and Roger Zimmermann. Largest: A benchmark dataset for large-scale traffic forecasting. In *Advances in Neural Information Processing Systems*, volume 36, 2024a.

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*, 2024b.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML '24. JMLR.org, 2024c.

Jiaming Ma, Binwu Wang, Guanjun Wang, Kuo Yang, Zhengyang Zhou, Pengkun Wang, Xu Wang, and Yang Wang. Less but more: Linear adaptive graph learning empowering spatiotemporal forecasting. In *Advances in Neural Information Processing Systems*, volume 38, pp. 58382–58418, 2025a.

Jiaming Ma, Binwu Wang, Pengkun Wang, Zhengyang Zhou, Xu Wang, and Yang Wang. Robust spatio-temporal centralized interaction for ood learning. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=q6ba0bZfpl.

Qian Ma, Zijian Zhang, Xiangyu Zhao, Haoliang Li, Hongwei Zhao, Yiqi Wang, Zitao Liu, and Wanyu Wang. Rethinking sensors modeling: Hierarchical information enhanced traffic forecasting. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 1756–1765, 2023.

Ivan Marisca, Cesare Alippi, and Filippo Maria Bianchi. Graph-based forecasting with missing data through spatiotemporal downsampling. In *Proceedings of the 41st International Conference on Machine Learning*, ICML '24. JMLR.org, 2024.

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

Ji Qi, Mingxiao Liu, VIET HO TAM THUC DO, Yuzhe Li, Zhuoshi Pan, Gene Cheung, and H. Vicky Zhao. Lightweight and interpretable transformer via mixed graph algorithm unrolling for traffic forecast. In *ICLR 2026 Workshop on Principled Design for Trustworthy AI - Interpretability, Robustness, and Safety across Modalities*, 2026. URL https://openreview.net/forum?id=W2mvNLVYkQ.

Yanjun Qin, Yuchen Fang, Xinke Jiang, Hao Miao, and Xiaoming Tao. Adaptive frequency pathways for spatiotemporal forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40 (18):15653–15661, 2026. doi: 10.1609/aaai.v40i18.38595.

Weilin Ruan, Xilin Dang, Ziyu Zhou, Sisuo Lyu, and Yuxuan Liang. A retrieval augmented spatio-temporal framework for traffic prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(46):39163–39172, 2026. doi: 10.1609/aaai.v40i46.41264.

Minlan Shao, Zijian Zhang, Yili Wang, Yiwei Dai, Xu Shen, and Xin Wang. Hyperd: Hybrid periodicity decoupling framework for traffic forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(18):15689–15697, 2026. doi: 10.1609/aaai.v40i18.38599.

Zezhi Shao, Zhao Zhang, Fei Wang, Wei Wei, and Yongjun Xu. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, pp. 4454–4458, 2022.

Dongjian Song, Yunhao Meng, Songjun Huang, and Jiayi Han. Tf-face: Time-frequency fusion learning via frequency-domain adaptive and controllable enhancement for trajectory prediction. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=XCqnmajjW5.

Jiarui Sun, Chin-Chia Michael Yeh, Yujie Fan, Xin Dai, Xiran Fan, Zhimeng Jiang, Uday Singh Saini, Vivian Lai, Junpeng Wang, Huiyuan Chen, Zhongfang Zhuang, Yan Zheng, and Girish Chowdhary. Towards efficient large scale spatial-temporal time series forecasting via improved inverted transformers, 2025.

Dongjing Wang, Gangming Guo, Tianpei Ouyang, Dongjin Yu, Haiping Zhang, Bao Li, Rong Jiang, Guandong Xu, and Shuiguang Deng. A lightweight spatio-temporal neural network with sampling-based time series decomposition for traffic forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 26(6):8682–8693, 2025. doi: 10.1109/TITS.2025.3552010.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations*, 2024.

Yongyao Wang, Xie Yu, Jingyuan Wang, Jiahao Ji, and Chao Li. Hieramix: A hierarchical mlp-mixer for large-scale traffic forecasting, 2026.

Ya Wen, Jixuan Cai, Qiyao Ma, Linyan Li, Xinhuan Chen, Chris Webster, and Yulun Zhou. Mora: Mobility as the backbone for geospatial representation learning at scale. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/3f3f2d552b22bf985f8d3f91caa0b610-Paper-Conference.pdf.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, ICML '24. JMLR.org, 2024.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 753–763, 2020. doi: 10.1145/3394486.3403118.

Weilin Xin, Chenyu Huang, Peilin Li, Jing Zhong, and Jiawei Yao. Urbangraph: Physics-informed spatio-temporal dynamic heterogeneous graphs for urban microclimate prediction. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ckjNF94cIi.

Chin-Chia Michael Yeh, Yujie Fan, Xin Dai, Uday Singh Saini, Vivian Lai, Prince Osei Aboagye, Junpeng Wang, Huiyuan Chen, Yan Zheng, Zhongfang Zhuang, Liang Wang, and Wei Zhang. RPMixer: Shaking up time series forecasting with random projections for large spatial-temporal data. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3919–3930, 2024.

Kun Yi, Qi Zhang, Wei Fan, Shoujin Wang, Pengyang Wang, Hui He, Ning An, Defu Lian, Longbing Cao, and Zhendong Niu. Frequency-domain MLPs are more effective learners in time series forecasting. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 3634–3640. AAAI Press, 2018.

Shiqi Zhang, Pan Mu, Hanting Yan, Yuchao Zhu, Jinglin Zhang, and Cong Bai. Causalx: A unified and causally-interpretable plug-and-play model for multi-modal spatio-temporal forecasting. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=qu17SFBMAO.

Weijia Zhang, Chenlong Yin, Hao Liu, Xiaofang Zhou, and Hui Xiong. Irregular multivariate time series forecasting: A transformable patching graph neural networks approach. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=UZlMXUGI6e.

Zhiwei Zhang, Xinyi Du, Weihao Wang, Xuanchi Guo, and Wenjuan Han. Visifold: Long-term traffic forecasting via temporal folding graph and node visibility. In *2026 IEEE 42nd International Conference on Data Engineering (ICDE)*, 2026b. URL https://arxiv.org/abs/2603.11816. Accepted paper; arXiv:2603.11816.

Qi Zheng, Zihao Yao, and Yaying Zhang. ST-ReP: Learning predictive representations efficiently for spatial-temporal forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

APPENDIX

A NOTATIONS

Table 5 summarizes the symbols used by the forecasting and reconciliation modules in the main paper.

Table 5: Summary of key notations.

Symbol	Description
$\mathbf{H}^{(\ell)}$	Latent node representation at grid level ℓ.
$\boldsymbol{\delta}^{(\ell)}$	Window-conditioned temporal-change evidence at level ℓ.
$\mathbf{A}^{(\ell)}_{\text{future}}$	Level-ℓ future anchor obtained from the observed input window.
$\mathbf{C}_\ell$	Candidate-parent list constructed from training data for the restriction from level ℓ to $\ell+1$.
$\mathbf{W}_\ell$	Sparse routing weights over the candidate parents in $\mathbf{C}_\ell$.
$\mathbf{P}_\ell$	Logical sparse assignment induced by $(\mathbf{C}_\ell, \mathbf{W}_\ell)$.
$\mathcal{R}_\ell, \mathcal{U}_\ell$	Sparse restriction and anchor-preserving prolongation operators.
$\mathbf{Y}^{(\ell)}, \widehat{\mathbf{Y}}^{(\ell)}$	Restricted target and prediction at grid level ℓ.
$\overline{\mathbf{Y}}^{(\ell)}$	Coarse forecast transported to level ℓ before innovation recovery.
$\mathbf{Q}_\ell, \mathbf{Q}_\ell^+$	Event-selection mask and its context-expanded counterpart.
$\Delta\widehat{\mathbf{Y}}^{(\ell)}$	Fine-scale innovation predicted beyond the transported macro forecast.
$\rho^{(\ell)}, u^{(\ell)}, \nu^{(\ell)}$	Routing discrepancy, learned uncertainty score, and novelty used to select event nodes.
$\mathbf{A}_r^{(\ell)}$	Sparse level-ℓ adjacency for relation type r.

B EXPERIMENTAL DETAILS

B.1 DATASETS DETAILS

Table 6 summarizes the spatial scale, sampling interval, and temporal coverage of every benchmark. Each experiment uses a fixed sparse input graph: $\mathcal{G}_\tau = \mathcal{G}^{(0)}$ for $\tau = t-T+1, \ldots, t$ in Equation 1. Window-dependent assignments generate the dynamic coarser graphs inside AR-MGNet.

Table 6: Summary of the spatio-temporal forecasting datasets.

Datasets	Time Steps	Time Interval	Time Range
SD (716 sensors)	35,040	15 min	2019-01-01–2019-12-31
GBA (2,352 sensors)	35,040	15 min	2019-01-01–2019-12-31
GLA (3,834 sensors)	35,040	15 min	2019-01-01–2019-12-31
CA (8,600 sensors)	35,040	15 min	2019-01-01–2019-12-31
SDWPF	35,280	10 min	Day 1–245
KnowAir	11,688	3 h	2015-01-01–2018-12-31
UrbanEV	8,760	1 h	2022-09-01–2023-08-31
Electricity	26,304	1 h	2016-07-01–2019-07-02
Solar-Energy	52,560	10 min	2006

LargeST traffic SD, GBA, GLA, and CA use a 12-step input and 12-step forecast protocol with chronological 6:2:2 splits. The target is traffic flow, normalization statistics are computed from the training split, and sparse road-graph candidates are constructed from topology and training data only.

SDWPF SDWPF uses a 12-step input and 12-step forecast protocol with a chronological 6:2:2 split and normalized turbine power as the target. Normalization and the sparse proximity/functional graph are constructed from the training split only.

KnowAir KnowAir uses a $24 \rightarrow 24$ protocol with a chronological 6:2:2 split and $PM_{2.5}$ concentration as the target. Training-set statistics normalize the observations, and monitoring locations form a sparse geographic graph.

UrbanEV UrbanEV uses a $24 \rightarrow 24$ protocol with a chronological 6:2:2 split and charging demand as the target. Training-set statistics normalize the series, and charging stations form the sparse input graph.

Electricity and Solar-Energy Electricity and Solar-Energy use $96 \rightarrow 96$ forecasting with chronological 7:1:2 splits and load or solar-power targets, respectively. Both use training-set standardization and sparse functional graphs constructed only from training-series similarities.

B.2 Baselines details

In the realm of time series and spatio-temporal forecasting, numerous models have surfaced in recent years. For every main-table dataset or LargeST region, all models use fresh initialization and independent training on that dataset or region's training split. Initialization and training data remain local to each region. Within each benchmark, all models share the same input and forecast lengths, split boundaries, normalization statistics, and target channels. Dashes in Table 1 denote unavailable results. We compared SOTA models from different years. The explanation of important models is as follows:

STID A compact identity-based predictor that represents spatial and temporal identities without explicit message passing (Shao et al., 2022).

PatchSTG A scalable traffic transformer that groups spatial observations into patches before global interaction (Fang et al., 2024).

HIEST A hierarchical traffic forecaster that learns sensor- and region-level representations at multiple spatial scales (Ma et al., 2023).

PDFormer A propagation-delay-aware transformer that models dynamic long-range spatial dependencies in traffic networks (Jiang et al., 2023).

iTransformer An inverted transformer that treats variables as tokens and learns cross-variable relations through attention (Liu et al., 2024b).

TimeMixer A decomposable multiscale temporal mixer evaluated with the same input and prediction lengths (Wang et al., 2024).

HieraMix A hierarchical all-MLP traffic model that mixes features across progressively coarser resolutions (Wang et al., 2026).

NEST A nested macro-to-micro forecaster whose regional future representation guides fine prediction (Ai et al., 2026).

B.3 Metric details

Let $\mathcal{I}$ be the valid target entries, $m = |\mathcal{I}|$, and $\epsilon > 0$ prevent division by zero. The three forecasting errors reported in the main tables are

$$\begin{aligned} \mathrm{MAE} &= \frac{1}{m}\sum_{i\in\mathcal{I}} |\widehat{y}_i - y_i|, \qquad \mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i\in\mathcal{I}} (\widehat{y}_i - y_i)^2}, \\ \mathrm{MAPE} &= \frac{100}{m}\sum_{i\in\mathcal{I}} \frac{|\widehat{y}_i - y_i|}{\max(|y_i|, \epsilon)}. \end{aligned} \tag{11}$$

Lower values indicate better forecasts for all three metrics.

For the noise experiment, let $\mathcal{Q} = \{0.1, 0.5, 0.9\}$, $u_{i,q} = y_i - \widehat{y}_i^{(q)}$, and $\rho_q(u) = u\big(q - \mathbb{I}[u < 0]\big)$. The averaged pinball loss is

$$\mathcal{L}_{\mathrm{pin}} = \frac{1}{m|\mathcal{Q}|}\sum_{i\in\mathcal{I}}\sum_{q\in\mathcal{Q}} \rho_q(u_{i,q}). \tag{12}$$

The nominal 80% interval is $[\widehat{y}_i^{(0.1)}, \widehat{y}_i^{(0.9)}]$. Its empirical coverage, mean width, coverage error, and median MAE are

$$\begin{aligned} \mathrm{Cov}_{80} &= \frac{1}{m}\sum_i \mathbb{I}[\widehat{y}_i^{(0.1)} \le y_i \le \widehat{y}_i^{(0.9)}], \qquad \mathrm{Width}_{80} = \frac{1}{m}\sum_i (\widehat{y}_i^{(0.9)} - \widehat{y}_i^{(0.1)}), \\ \mathrm{CE}_{80} &= |\,\mathrm{Cov}_{80} - 0.80|, \qquad\qquad \mathrm{MedMAE} = \frac{1}{m}\sum_i |\widehat{y}_i^{(0.5)} - y_i|. \end{aligned} \tag{13}$$

Lower pinball loss, interval width, coverage error, and median MAE are better; coverage is interpreted by its distance from 0.80. The noise study activates the quantile head, while the main tables use the point-forecasting objective.

For the forecast-field panels, let $\mathcal{J}$ index the raw entries of the displayed high-change nodes over the displayed prediction window, let $n = |\mathcal{J}|$, and let $\bar{y}$ and $\bar{\widehat{y}}$ denote the corresponding means. Their spatio-temporal shape agreement is measured by Pearson correlation:

$$r = \frac{\sum_{i\in\mathcal{J}}(y_i - \bar{y})(\widehat{y}_i - \bar{\widehat{y}})}{\sqrt{\sum_{i\in\mathcal{J}}(y_i - \bar{y})^2}\sqrt{\sum_{i\in\mathcal{J}}(\widehat{y}_i - \bar{\widehat{y}})^2}}. \tag{14}$$

Pearson $r \in [-1, 1]$, and a higher value indicates that the predicted field better preserves the temporal and cross-node variation of the reference field. The values shown in the forecast-field panels are computed on exactly the displayed nodes and window and rounded to three decimals; the truth panel is marked *Reference* because its self-correlation is trivially one.

For the representation analysis, each complete predicted trajectory is flattened and standardized with the same feature-wise statistics before any diagnostic is computed. With a_i denoting the mean within-cluster distance and b_i the smallest mean distance from sample i to another cluster, the Silhouette score is

$$\mathrm{Sil} = \frac{1}{m}\sum_{i=1}^{m} \frac{b_i - a_i}{\max(a_i, b_i)}. \tag{15}$$

For K_{cl} clusters with centroid $\boldsymbol{\mu}_k$, within-cluster scatter S_k, and centroid distance $M_{kj} = \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_j\|_2$, the Davies–Bouldin index is

$$\mathrm{DB} = \frac{1}{K_{\mathrm{cl}}}\sum_{k=1}^{K_{\mathrm{cl}}} \max_{j\neq k} \frac{S_k + S_j}{M_{kj}}. \tag{16}$$

Finally, if $\mathcal{N}_7(i)$ contains the seven nearest training trajectories and mode returns their majority class, the 7-NN accuracy is

$$\mathrm{Acc}_{7\mathrm{NN}} = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}[y_i = \mathrm{mode}\{y_j : j \in \mathcal{N}_7(i)\}]\,. \tag{17}$$

Higher Silhouette and 7-NN accuracy are better, whereas lower Davies–Bouldin is better. These three diagnostics are evaluated in the common standardized high-dimensional trajectory space; the two-dimensional visualization is used only for display and does not determine the reported scores.

B.4 Implementation details

AR-MGNet uses three grid levels, one Adaptive Macro–Micro Reconciliation Cycle, two typed graph blocks per graph smoother, dropout 0.1, and at most eight outgoing coarse edges. The temporal encoder applies depthwise causal convolutions with kernel size three and dilations 1, 2, and 4. Training uses AdamW and automatic mixed precision. Validation data select one deployable checkpoint, which is then evaluated unchanged once on the complete test split.

The efficiency comparison among AR-MGNet, HieraMix, and NEST uses identical GBA/GLA $12 \to 12$ tasks, chronological splits, normalization, batch size 64, one RTX 4090, mixed precision, ten warm-up batches, CUDA synchronization, and 30 timed batches, with the same full-epoch and full-test timing scopes.

C Model Details

C.1 Temporal–Spatial Encoding and Future Anchor

This subsection expands the encoder in Section 4.1 of the main paper; it produces $\mathbf{H}^{(0)}$, $\boldsymbol{\delta}^{(0)}$, and $\mathbf{A}^{(0)}_{\text{future}}$ for sparse restriction, evidence gating, and anchor-preserving prolongation. For observed history $\mathbf{X} \in \mathbb{R}^{B\times T\times N_0\times C_{\text{in}}}$,

$$\begin{aligned}\mathbf{E}_{\text{in}} &= \phi_x(\mathbf{X}) + \mathbf{E}_{\text{node}} + \mathbf{E}_{\text{time}} + \mathbf{E}_{\text{cal}},\\ \mathbf{H}^{(0)} &= \text{GraphTemporalEncode}(\mathbf{E}_{\text{in}}, \mathcal{G}^{(0)}), \qquad \boldsymbol{\delta}^{(0)} = \psi_\delta(\mathbf{X}),\\ \mathbf{A}^{(0)}_{\text{future}} &= \text{Repeat}_H(\mathbf{P}_o\mathbf{X}_{:,T,:,:}) + f_{\text{anc}}(\mathbf{H}^{(0)}, \mathbf{E}_{\text{cal}}).\end{aligned} \tag{18}$$

Here $\mathbf{P}_o$ extracts the forecasting target from the last observed step, Repeat_H broadcasts it across the horizon, and the shared f_{anc} adds a learned temporal/calendar residual used for every dataset. The representation supplies content evidence, $\boldsymbol{\delta}^{(0)}$ summarizes recent change, and coarser states are formed by the same restriction operator used for the anchor.

C.2 Sparse Grid Restriction and Explicit Macro-Future Solver

This subsection details Section 4.2 of the main paper; it converts the fine representation into sparse coarse grids and supplies the explicit macro forecast consumed by top-down reconciliation. The following reference construction specifies the candidate support and static prior.

For a coarsening ratio $r_\ell > 1$, set $N_{\ell+1} = \max(1, \lceil N_\ell/r_\ell \rceil)$. A topology traversal groups the nodes into balanced parent cells with sizes $\lfloor N_\ell/N_{\ell+1} \rfloor$ or $\lceil N_\ell/N_{\ell+1} \rceil$. With coordinates, the seed order follows a Morton ordering of normalized, quantized locations; otherwise it follows a breadth-first traversal of the input graph. Each group grows by breadth-first expansion until its capacity is reached, filling any remaining slots from unassigned nodes in seed order. This defines the primary parent $p_\ell(i)$ of each node. The same graph traversal applies to functional graphs built from training-series similarities.

Each $\mathcal{C}_\ell(i)$ first contains $p_\ell(i)$, then distinct parent groups of outgoing neighbors ranked by their total edge weight. Cyclic parent-index order fills remaining slots up to $K_\ell = \min(K, N_{\ell+1})$ candidates. A symbolic primary mass $p_{\text{pri}} \in (0,1)$ defines $\pi^{(\ell)}_{i,p_\ell(i)} = p_{\text{pri}}$ and $\pi^{(\ell)}_{ic} = (1-p_{\text{pri}})/(K_\ell - 1)$ for the other candidates; a single candidate receives unit mass. Candidate lists and priors are fixed after construction and reused for every input window. Parent prototypes and window-dependent scores are

$$\begin{aligned}\boldsymbol{\mu}^{(\ell+1)}_c &= \frac{\sum_i \pi^{(\ell)}_{ic}\mathbf{h}^{(\ell)}_i}{\epsilon + \sum_i \pi^{(\ell)}_{ic}}, \quad \mathbf{q}^{(\ell)}_i = \mathbf{W}_q[\mathbf{h}^{(\ell)}_i \| \boldsymbol{\delta}^{(\ell)}_i], \quad \mathbf{k}^{(\ell)}_c = \mathbf{W}_k\boldsymbol{\mu}^{(\ell+1)}_c,\\ a^{(\ell)}_{ic} &= \frac{\langle \mathbf{q}^{(\ell)}_i, \mathbf{k}^{(\ell)}_c\rangle}{\sqrt{D}} + b^{(\ell)}_{ic} + \log(\pi^{(\ell)}_{ic} + \epsilon), \qquad c \in \mathcal{C}_\ell(i),\\ w^{(\ell)}_{ic} &= \begin{cases}\text{GumbelSoftmax}_c(\mathbf{a}^{(\ell)}_i/\tau), & \text{training},\\ \text{softmax}_c(\mathbf{a}^{(\ell)}_i/\tau), & \text{inference}.\end{cases}\end{aligned} \tag{19}$$

The candidate mask is applied before either normalization. We mix in the static prior, $\widetilde{w}_{ic} \leftarrow 0.95w_{ic} + 0.05\pi_{ic}$, and perform three capacity-balancing iterations:

$$\begin{aligned}v^{(a)}_{ic} &= \frac{(N_\ell/N_{\ell+1})\widetilde{w}^{(a)}_{ic}}{\max(\epsilon, \sum_j \widetilde{w}^{(a)}_{jc})},\\ \widetilde{w}^{(a+1)}_{ic} &= \frac{v^{(a)}_{ic}}{\sum_{c'\in\mathcal{C}_\ell(i)} v^{(a)}_{ic'}}, \qquad a = 0,1,2.\end{aligned} \tag{20}$$

Column scaling adjusts the parent masses toward $N_\ell/N_{\ell+1}$, and row normalization restores unit mass on each node's candidate support. The three iterations produce an approximately balanced assignment. Scattering these candidate weights gives $P_{ic} = \widetilde{w}_{ic}$ on $c \in \mathcal{C}_\ell(i)$ and zero otherwise.

Restriction and prolongation are the normalized sparse scatter/gather operations

$$[\mathcal{R}_\ell(\mathbf{Z})]_c = \frac{\sum_i P_{ic}^{(\ell)} \mathbf{Z}_i}{\epsilon + \sum_i P_{ic}^{(\ell)}}, \qquad [\mathcal{U}_\ell(\mathbf{Z})]_i = \sum_{c \in \mathcal{C}_\ell(i)} P_{ic}^{(\ell)} \mathbf{Z}_c. \tag{21}$$

Candidate sparsity avoids materializing $N_\ell \times N_{\ell+1}$ dense assignments. Each restricted state is updated by the graph smoother $\mathcal{S}_{\ell+1}$ in Equation 3; the same restriction transports the anchor and temporal-change evidence. This bottom-up sequence yields $\mathbf{H}^{(L-1)}$ and $\mathbf{A}_{\text{future}}^{(L-1)}$. The coarsest solver then predicts the macro future directly:

$$\widehat{\mathbf{Y}}^{(L-1)} = \mathbf{A}_{\text{future}}^{(L-1)} + f_{\text{macro}}\Big(\text{GraphMix}(\mathbf{H}^{(L-1)}, \mathcal{G}^{(L-1)})\Big). \tag{22}$$

The coarse output is therefore a target-space forecast consumed directly by finer reconciliation levels.

C.3 Bidirectional Evidence Gate and Event-Selective Innovation Recovery

This subsection expands Section 4.3 of the main paper; it combines bottom-up node evidence with the transported macro future, then produces the routed innovation added at the current level. The anchor-preserving coarse forecast and bidirectional gate are

$$\begin{aligned} \overline{\mathbf{Y}}^{(\ell)} &= \mathcal{U}_\ell\Big(\widehat{\mathbf{Y}}^{(\ell+1)} - \mathbf{A}_{\text{future}}^{(\ell+1)}\Big) + \mathbf{A}_{\text{future}}^{(\ell)}, \\ \mathbf{z}_i^{(\ell)} &= \mathbf{G}_i^{(\ell)} \odot \mathbf{E}_{\uparrow,i}^{(\ell)} + (\mathbf{1} - \mathbf{G}_i^{(\ell)}) \odot \mathbf{E}_{\downarrow,i}^{(\ell)}, \\ \mathbf{G}_i^{(\ell)} &= \sigma\Big(\mathbf{W}_g^{(\ell)}[\mathbf{E}_{\uparrow,i}^{(\ell)} \| \mathbf{E}_{\downarrow,i}^{(\ell)} \| \Delta\mathbf{E}_i^{(\ell)}] + \mathbf{b}_g^{(\ell)}\Big). \end{aligned} \tag{23}$$

Here $\mathbf{z}_i^{(\ell)} = \mathbf{Z}_i^{(\ell)}$ is the node-wise fused state in Equation 5, using the same upward and downward evidence maps $\phi_\uparrow^{(\ell)}$ and $\phi_\downarrow^{(\ell)}$. Fine evidence preserves local changes, and coarse evidence supplies their macro context. Routing discrepancy, learned uncertainty score, temporal novelty, and the resulting score are

$$\begin{aligned} \rho_i^{(\ell)} &= \text{stopgrad}\left(\frac{\|\mathbf{h}_i^{(\ell)} - \mathcal{U}_\ell(\mathbf{H}^{(\ell+1)})_i\|_1}{D}\right), \quad & u_i^{(\ell)} &= \text{softplus}(\mathbf{w}_u^\top \mathbf{h}_i^{(\ell)} + b_u), \\ \nu_i^{(\ell)} &= \|\boldsymbol{\delta}_i^{(\ell)}\|_1 / D, & s_i^{(\ell)} &= f_s([\mathbf{z}_i^{(\ell)} \| \rho_i^{(\ell)} \| u_i^{(\ell)} \| \nu_i^{(\ell)}]). \end{aligned} \tag{24}$$

For $m_\ell = \max(1, \text{round}(\beta_\ell N_\ell))$, inference uses the hard mask $Q_i = \mathbb{I}[i \in \text{Top}_{m_\ell}(\mathbf{s})]$. During training, Gumbel noise gives $\widetilde{s}_i = s_i + g_i$ and a Top-m_ℓ threshold ϑ; the straight-through mask is

$$q_i^{\text{soft}} = \sigma((\widetilde{s}_i - \vartheta)/\tau_q), \qquad Q_i = q_i^{\text{hard}} + q_i^{\text{soft}} - \text{stopgrad}(q_i^{\text{soft}}), \qquad q_i^{\text{hard}} = \mathbb{I}[i \in \text{Top}_{m_\ell}(\widetilde{\mathbf{s}})]. \tag{25}$$

$\mathbf{Q}_\ell^+$ adds one-hop neighbors of active nodes as context; if the budget covers the level, both masks are all ones. The innovation head then applies

$$\Delta\widehat{\mathbf{Y}}^{(\ell)} = f_{\text{innovation}}^{(\ell)}(\mathbf{z}^{(\ell)}, \overline{\mathbf{Y}}^{(\ell)}, \mathbf{n}^{(\ell)}), \qquad \widehat{\mathbf{Y}}^{(\ell)} = \overline{\mathbf{Y}}^{(\ell)} + \mathbf{Q}_\ell \odot \Delta\widehat{\mathbf{Y}}^{(\ell)}. \tag{26}$$

Inactive nodes therefore inherit the explicit macro future, whereas active nodes recover only the unresolved local innovation.

C.4 Dynamic Multi-Grid Neighborhood Learning

This subsection supports Section 4.4 of the main paper; it updates the window-conditioned hierarchy and supplies the multiscale relational context $\mathbf{n}^{(\ell)} = \mathbf{N}^{(\ell)}$ used by the innovation head. Equation 19 parameterizes the routing weights, and Equation 20 normalizes them on the candidate support:

$$\begin{aligned} \mathbf{P}_\ell^\star(\mathbf{X}) &= \mathbf{P}_{\ell,\theta}(\mathbf{X}) \in \Pi_\ell, \\ \mathbf{A}_r^{(\ell+1)} &= \mathcal{N}_{D_g}\Big(\mathbf{P}_{\ell,\theta}(\mathbf{X})^\top \mathbf{A}_r^{(\ell)} \mathbf{P}_{\ell,\theta}(\mathbf{X})\Big). \end{aligned} \tag{27}$$

Here $\Pi_\ell = \{\mathbf{P} \geq 0 : \sum_c P_{ic} = 1,\ P_{ic} = 0 \text{ for } c \notin \mathcal{C}_\ell(i)\}$ describes the row-normalized candidate support. The superscript $\star$ denotes the learned window-conditioned routing output. Gradients of

$\mathcal{L}_{\mathrm{APD}}$ update the router parameters together with the forecasting modules. The three balancing passes regulate parent mass in the forward computation, and $\mathcal{N}_{D_g}$ retains and normalizes at most D_g outgoing contracted neighbors.

For level ℓ, the visible scales are $\mathcal{S}(\ell) = \{\ell, \dots, L-1\}$. Their states $\{\mathbf{H}^{(q)}\}_{q\in\mathcal{S}(\ell)}$ are available after the bottom-up pass. Transport is the identity for $q = \ell$ and $\mathcal{T}_{\ell\leftarrow q} = \mathcal{U}_\ell \circ \cdots \circ \mathcal{U}_{q-1}$ for $q > \ell$. With adjacency rows indexing receiving nodes, $\widetilde{A}^{(q)}_{r,ij} = A^{(q)}_{r,ij} / \max(\epsilon, \sum_j A^{(q)}_{r,ij})$ normalizes each relation message. Define $\mathbf{m}^{(\ell\leftarrow q)}_{i,r} = [\mathcal{T}_{\ell\leftarrow q}(\widetilde{\mathbf{A}}^{(q)}_r \mathbf{H}^{(q)})]_i$. The mixer in Equation 7 then becomes

$$\begin{aligned} \mathbf{n}^{(\ell)}_i &= \sum_{q\in\mathcal{S}(\ell)} \sum_{r\in\mathcal{R}} \alpha^{(\ell\leftarrow q)}_{i,r} \mathbf{m}^{(\ell\leftarrow q)}_{i,r}, \\ \alpha^{(\ell\leftarrow q)}_{i,r} &= \mathrm{softmax}_{q,r}\Big(\kappa(\mathbf{z}^{(\ell)}_i, \mathbf{m}^{(\ell\leftarrow q)}_{i,r}) + \eta_{q-\ell} + b_r\Big). \end{aligned} \tag{28}$$

The current gated state supplies the query, and the transported bottom-up messages supply the scale–relation context. The affinity is $\kappa(\mathbf{z}, \mathbf{m}) = \langle \mathbf{W}_Q \mathbf{z}, \mathbf{W}_K \mathbf{m}\rangle / \sqrt{D}$; $\eta_{q-\ell}$ and b_r are learned scale and relation biases. Algorithm 1 passes these multiscale tensors to the mixer during coarse-to-fine recovery.

Let $|E_\ell| = \sum_{r\in\mathcal{R}} |E_{\ell,r}|$. Each fine edge contributes at most K^2 candidate parent pairs during graph contraction. For a fixed number of levels, the scalar contraction cost and the feature-routing/message cost are

$$\mathcal{O}\left(BK^2 \sum_{\ell=0}^{L-2} |E_\ell|\right) + \mathcal{O}\left(BD|\mathcal{R}| \sum_{\ell=0}^{L-1} N_\ell (K + D_g)\right). \tag{29}$$

The input graph has bounded degree d_0, and each intermediate relation graph retains at most D_g outgoing edges per node, giving $|E_\ell| = \mathcal{O}(|\mathcal{R}| N_\ell \max(d_0, D_g))$. Thus this core is linear in node count for fixed L, B, D, K, d_0, D_g, and $|\mathcal{R}|$. The temporal encoder, candidate construction, TopBudget selection, horizon projection, and learned feature projections contribute their respective costs to the complete model.

C.5 Consistency-Constrained Optimization

This subsection expands Section 4.5 of the main paper; it trains the macro solver, the intermediate forecasts, and event-selective recovery as one coherent hierarchy. Let $\mathbf{V}$ mark finite targets and let a node mask $\mathbf{M}$ broadcast over horizon and channel axes. The common masked discrepancy is

$$\mathcal{D}_{\mathbf{M}}(\widehat{\mathbf{Z}}, \mathbf{Z}) = \frac{\sum \mathbf{V} \odot \mathbf{M} \odot |\widehat{\mathbf{Z}} - \mathbf{Z}|}{\epsilon + \sum \mathbf{V} \odot \mathbf{M}}, \qquad \mathcal{D} = \mathcal{D}_{\mathbf{1}}. \tag{30}$$

Targets are restricted recursively by $\mathbf{Y}^{(\ell+1)} = \mathcal{R}_\ell(\mathbf{Y}^{(\ell)})$, and the innovation target is $\Delta\mathbf{Y}^{(\ell),*} = \mathbf{Y}^{(\ell)} - \overline{\mathbf{Y}}^{(\ell)}$. These quantities instantiate the supervised and Macro–Micro risks in Equations 8 and 9 of the main paper. For each auxiliary risk $k \in \{\mathrm{hier}, \mathrm{MM}\}$, let b index consecutive training mini-batches and $d_\xi \in (0, 1)$ be the EMA decay. A positive initial scale $\xi^{(0)}_k$ is updated from the absolute risk violation, and the current risk is normalized as

$$\begin{aligned} \xi^{(b)}_k &= d_\xi \xi^{(b-1)}_k + (1 - d_\xi) \left|\mathrm{stopgrad}(\mathcal{L}^{(b)}_k - \varepsilon_k)\right|, \\ \bar{c}^{(b)}_k &= \frac{\mathcal{L}^{(b)}_k - \varepsilon_k}{\mathrm{stopgrad}(\xi^{(b)}_k) + \epsilon}. \end{aligned} \tag{31}$$

The updated EMA acts as a fixed scale during each gradient step. Let $\mathcal{B}_q$ contain the training batches of epoch q and define $\bar{c}^{(q)}_k = |\mathcal{B}_q|^{-1} \sum_{b\in\mathcal{B}_q} \mathrm{stopgrad}(\bar{c}^{(b)}_k)$. The batch objective and epoch-level multiplier update are

$$\mathcal{L}_{\mathrm{APD}} = \mathcal{L}_{\mathrm{pred}} + \sum_k \lambda_k [\bar{c}_k]_+ + \frac{\mu}{2} \sum_k [\bar{c}_k]^2_+, \qquad \lambda^{(q+1)}_k = [\lambda^{(q)}_k + \eta_\lambda \bar{c}^{(q)}_k]_+. \tag{32}$$

Here ε_k is the risk tolerance, μ is the fixed quadratic penalty, and λ_k adapts to the mean normalized training violation. In Equation 10, q indexes this epoch-level dual update. The scales persist across epochs, and the batch accumulators are reset at each epoch boundary.

C.6 Cross-Scale Forecast Reconciliation Algorithms

Algorithm 1 connects the detailed operators above in the same order as the main architecture. The candidate-list representation realizes $\mathbf{P}_\ell$ without constructing a dense assignment matrix.

Algorithm 1 One AR-MGNet Adaptive Macro–Micro Reconciliation Cycle

Require: $\mathcal{I} = \{\mathbf{X}, \mathcal{G}^{(0)}, \{\mathcal{C}_\ell\}, \mathbf{E}_{\text{node}}, \mathbf{E}_{\text{time}}, \mathbf{E}_{\text{cal}}\}$; budgets $\{\beta_\ell\}$

$(\mathbf{H}^{(0)}, \boldsymbol{\delta}^{(0)}, \mathbf{A}^{(0)}_{\text{future}}) \leftarrow \text{TemporalSpatialEncoder}(\mathcal{I})$
for $\ell = 0$ to $L-2$ **do**
  $(\mathbf{W}_\ell, \mathbf{P}_\ell) \leftarrow \text{SparseGridRestriction}(\mathbf{H}^{(\ell)}, \boldsymbol{\delta}^{(\ell)}, \mathcal{C}_\ell)$
  $\mathcal{G}^{(\ell+1)} \leftarrow \text{MultiGridGraphContraction}(\mathcal{G}^{(\ell)}, \mathbf{P}_\ell)$
  $\mathbf{H}^{(\ell+1)} \leftarrow \mathcal{S}_{\ell+1}(\mathcal{R}_\ell \mathbf{H}^{(\ell)}, \mathcal{G}^{(\ell+1)})$
  $(\boldsymbol{\delta}^{(\ell+1)}, \mathbf{A}^{(\ell+1)}_{\text{future}}) \leftarrow (\mathcal{R}_\ell \boldsymbol{\delta}^{(\ell)}, \mathcal{R}_\ell \mathbf{A}^{(\ell)}_{\text{future}})$
end for
$\widehat{\mathbf{Y}}^{(L-1)} \leftarrow \text{ExplicitMacroFutureSolver}(\mathbf{H}^{(L-1)}, \mathbf{A}^{(L-1)}_{\text{future}}, \mathcal{G}^{(L-1)})$
for $\ell = L-2$ down to 0 **do**
  $\overline{\mathbf{Y}}^{(\ell)} \leftarrow \text{ForecastProlongation}(\widehat{\mathbf{Y}}^{(\ell+1)}, \mathbf{A}^{(\ell+1)}_{\text{future}}, \mathbf{A}^{(\ell)}_{\text{future}}, \mathbf{P}_\ell)$
  $\mathbf{z}^{(\ell)} \leftarrow \text{BidirectionalEvidenceGate}(\mathbf{H}^{(\ell)}, \mathbf{H}^{(\ell+1)}, \overline{\mathbf{Y}}^{(\ell)}, \boldsymbol{\delta}^{(\ell)}, \mathbf{P}_\ell)$
  $(\rho^{(\ell)}, u^{(\ell)}, \nu^{(\ell)}, \mathbf{Q}_\ell, \mathbf{Q}^+_\ell) \leftarrow \text{EventSelectiveRouter}(\mathbf{z}^{(\ell)}, \boldsymbol{\delta}^{(\ell)}, \mathcal{G}^{(\ell)}, \beta_\ell)$
  $\mathcal{M}_\ell \leftarrow \{(\mathbf{H}^{(q)}, \mathcal{G}^{(q)}, \mathcal{T}_{\ell \leftarrow q}) : q \in \mathcal{S}(\ell)\}$
  $\mathbf{n}^{(\ell)} \leftarrow \text{RelationAwareNeighborhoodMixer}(\mathbf{z}^{(\ell)}, \mathcal{M}_\ell, \mathbf{Q}^+_\ell)$
  $\Delta\widehat{\mathbf{Y}}^{(\ell)} \leftarrow \text{InnovationRecoveryHead}(\mathbf{z}^{(\ell)}, \overline{\mathbf{Y}}^{(\ell)}, \mathbf{n}^{(\ell)})$
  $\widehat{\mathbf{Y}}^{(\ell)} \leftarrow \overline{\mathbf{Y}}^{(\ell)} + \mathbf{Q}_\ell \odot \Delta\widehat{\mathbf{Y}}^{(\ell)}$
end for
return $\widehat{\mathbf{Y}}^{(0)}$, level forecasts, assignments, masks, and learned uncertainty scores

Algorithm 2 governs the main systems and controlled ablations. Each benchmark starts from a fresh initialization, validation data select the checkpoint, and test predictions use only the historical window and declared graph inputs.

Algorithm 2 Registered training, selection, and reporting protocol

Require: training/validation/test splits; configuration c

Construct candidate parents and typed sparse edges from training data only
Initialize the model and optimizer from scratch for this benchmark
Initialize positive scales ξ_k and nonnegative multipliers λ_k
for epoch $e = 1$ to E_c **do**
    Reset normalized-violation sums $v_k \leftarrow 0$ and batch count $n_e \leftarrow 0$
    Enable dynamic assignment after the configured grid warm-up
    Use straight-through differentiable masks in training and hard masks in evaluation
    for mini-batch b from the training-window set **do**
        $\mathcal{O} \leftarrow \mathrm{ReconcileCycle}(b.\mathbf{X}, \mathcal{G}; c)$
        Compute the risks and update $\xi_k, \bar{c}_k^{(b)}$ with Equation 31
        Update θ with $\mathcal{L}_{\mathrm{APD}}(\mathcal{O}, b.\mathbf{Y}; c)$
        $v_k \leftarrow v_k + \mathrm{stopgrad}(\bar{c}_k^{(b)})$; $n_e \leftarrow n_e + 1$
    end for
    Update λ_k with Equation 32 using $\bar{c}_k^{(e)} = v_k / n_e$ and fixed μ
    Evaluate the declared validation objective and update candidate checkpoints
    if minimum epochs are met and every tracked validation metric is stale for the patience window **then**
        break
    end if
end for
Select one deployable checkpoint using validation data only
Evaluate it unchanged once on the complete test split
Save predictions, metrics, and configuration; retain all declared test outputs

D Discussions

D.1 Compare With HieraMix

HieraMix is the strongest traffic baseline because its hierarchical MLP mixing captures broad spatial structure efficiently. Nevertheless, Table 1 shows that AR-MGNet lowers the four-region average MAE, RMSE, and MAPE by 0.9%, 1.2%, and 5.7%, respectively; on GBA alone, RMSE and MAPE fall by 2.6% and 13.8%. The long-horizon gap is larger: Table 8 reports GLA/CA MAE of 20.00/19.03 at 24 steps for AR-MGNet versus 38.78/36.55 for HieraMix. HieraMix coarsens and mixes representations, but it does not explicitly solve the future at the macro level or selectively restore unresolved event innovations. AR-MGNet adds both operations, retaining global trend accuracy while focusing fine computation where the hierarchy remains inconsistent.

D.2 Compare With NEST

NEST provides a strong conceptual baseline for nested macro-to-micro forecasting and has the fastest GBA training time in Table 3 (75.0 versus 92.2 s/epoch). Its inference and accuracy trade-off is weaker: on GBA and GLA, NEST requires 20.0 and 37.0 s and obtains MAE 23.26 and 28.38, whereas AR-MGNet requires 14.2 and 17.0 s and obtains MAE 17.62 and 16.24. The CA event window further yields AR-MGNet MAE/RMSE reductions of 57.9%/55.1% relative to NEST. For Table 8, all models use the same split, normalization, and autoregressive rollout; a single trained model generates every reported horizon by appending each predicted block to the history. NEST passes regional future information downward through a fixed nested structure, while AR-MGNet's dynamic grid and event-selective innovation path allocate neighborhoods and fine recovery according to the current event.

D.3 Compare With PatchSTG

PatchSTG is a scalable patch-based competitor and remains competitive on SDWPF, where its RMSE is 0.3926 versus 0.3929 for AR-MGNet in Table 2. Across the remaining cross-domain

tasks, however, AR-MGNet is stronger: it lowers the best MAE/RMSE on UrbanEV by 10.7%/4.0% and on Solar-Energy by 19.0%/16.4%. Table 8 also shows AR-MGNet below PatchSTG at every reported 16–48-step GLA/CA horizon; for example, GLA@48 falls from 32.43 to 23.90. Static patches reduce global interaction cost but can bind together nodes whose relevance changes with the forecast window. AR-MGNet instead contracts a sparse dynamic graph and restores innovations through typed multiscale neighborhoods, which better preserves local events without abandoning the macro trajectory.

E Additional Results

Across the four LargeST regions, AR-MGNet averages 15.99 MAE, 27.38 RMSE, and 9.90 MAPE, improving on HieraMix at 16.13, 27.70, and 10.50, respectively. In the non-traffic main table, AR-MGNet obtains the best result in 8 of 10 metric cells; the two exceptions are SDWPF MAE and RMSE.

E.1 Sensitivity and Robustness

Figure 7 jointly reports MAE, RMSE, and latency for two accuracy–cost controls on CA. The routing-degree panel pairs forecast errors with routing-only latency, and the refinement-budget panel pairs them with end-to-end refinement latency. Together, they quantify the accuracy–cost trade-offs of routing degree and refinement coverage.

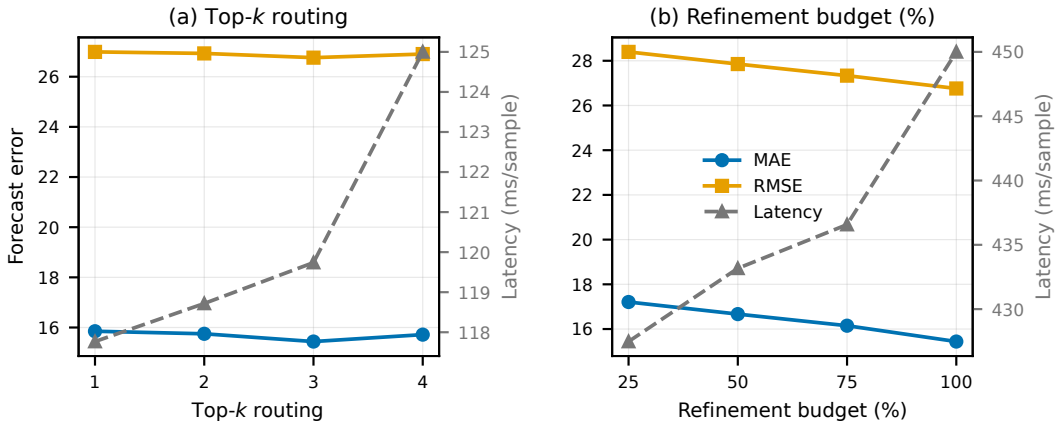


Figure 7: Sensitivity of MAE and RMSE to routing degree and refinement budget on CA. (a) pairs the errors with routing-only latency; (b) pairs them with end-to-end refinement latency.

Table 7 measures robustness on Electricity under standardized input noise. At $\sigma = 0.05$, AR-MGNet reduces NEST pinball loss from 0.087 to 0.077 and median MAE from 0.251 to 0.228; at $\sigma = 0.20$, the corresponding values fall from 0.092 to 0.080 and from 0.267 to 0.238. Its coverage error is also smaller at both noise levels, supporting the claim that relational context and selective innovation recovery of AR-MGNet remain stable when observations are perturbed.

Table 7: Electricity quantile forecasting under standardized input noise; coverage targets 0.80.

σ	Model	Pinball ↓	Coverage	Width ↓	Median MAE ↓
0.05	AR-MGNet	0.077	0.777	0.660	0.228
	NEST	0.087	0.756	0.710	0.251
0.20	AR-MGNet	0.080	0.757	0.667	0.238
	NEST	0.092	0.722	0.710	0.267

E.2 Long-Horizon Stability

Table 8 reports autoregressive forecasting from 16 to 48 steps under the shared protocol described in Section D.2. AR-MGNet is best at every reported GLA and CA horizon; at 24/48 steps it obtains 20.00/23.90 on GLA and 19.03/22.78 on CA. The advantage persists as rollout length grows because the explicit macro future anchors the trajectory while event recovery limits fine-scale corrections to unresolved deviations.

Table 8: Autoregressive 16–48-step MAE on GLA and CA under a shared rollout protocol.

Dataset	Model	MAE@16	MAE@20	MAE@24	MAE@36	MAE@48
GLA	PatchSTG	25.63	27.08	27.92	30.42	32.43
	HieraMix	33.14	36.22	38.78	44.77	49.33
	NEST	23.63	25.11	26.28	28.51	30.03
	AR-MGNet	**18.00**	**19.08**	**20.00**	**22.30**	**23.90**
CA	PatchSTG	24.17	25.52	26.54	28.10	29.15
	HieraMix	30.93	33.97	36.55	42.65	47.10
	NEST	22.05	23.46	24.56	26.60	27.92
	AR-MGNet	**17.17**	**18.16**	**19.03**	**21.22**	**22.78**

E.3 Accuracy–Efficiency Trade-off

Figure 8 combines MAE with inference time for AR-MGNet, HieraMix, and NEST under the common timing protocol in Section B.4. On GBA, AR-MGNet reaches 17.62 MAE in 14.2 s, compared with 17.73 in 26.4 s for HieraMix and 23.26 in 20.0 s for NEST. On GLA, it reaches 16.24 in 17.0 s, compared with 16.45 in 38.7 s and 28.38 in 37.0 s. Its lower-left position in both panels shows that sparse dynamic contraction improves accuracy without paying the dense multiscale inference cost.

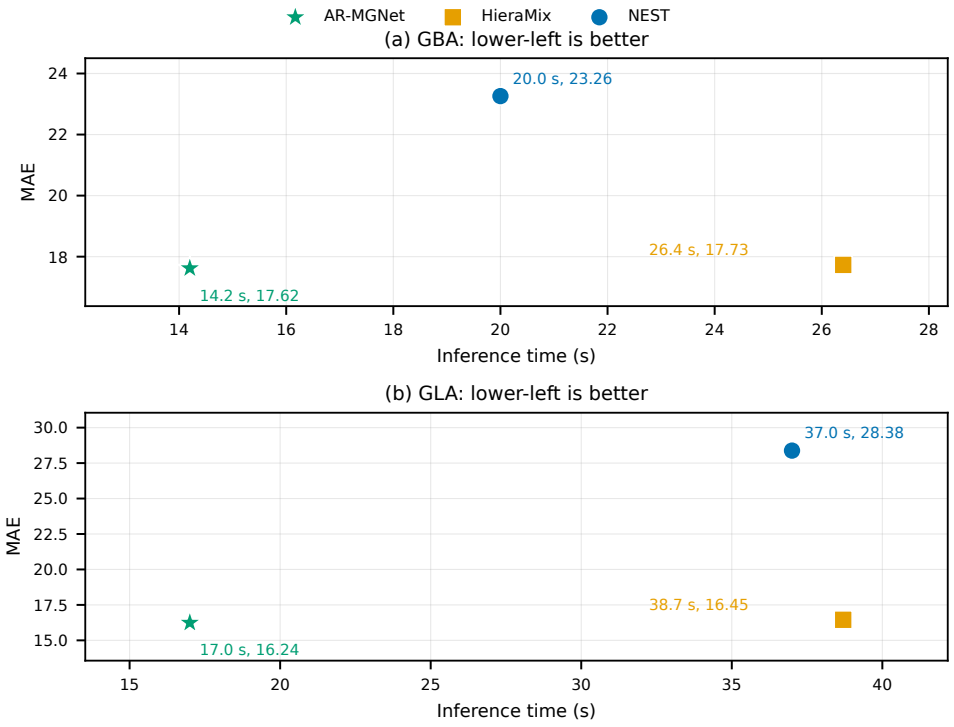


Figure 8: Accuracy–efficiency comparison on GBA and GLA. Lower MAE and shorter inference time are better.

F VISUALIZATION

F.1 CROSS-DOMAIN FORECAST FIELDS

Figure 9 evaluates the unchanged Electricity validation window and the displayed high-change nodes. AR-MGNet obtains the lowest displayed system-level MAE/RMSE (0.22/0.43, versus 0.28/0.51 for NEST and 0.28/0.53 for iTransformer), while its node traces remain closer to the target around sharp changes. The smaller high-intensity regions in the absolute-error field show that the macro solver captures the shared demand trajectory and the innovation path corrects localized departures.

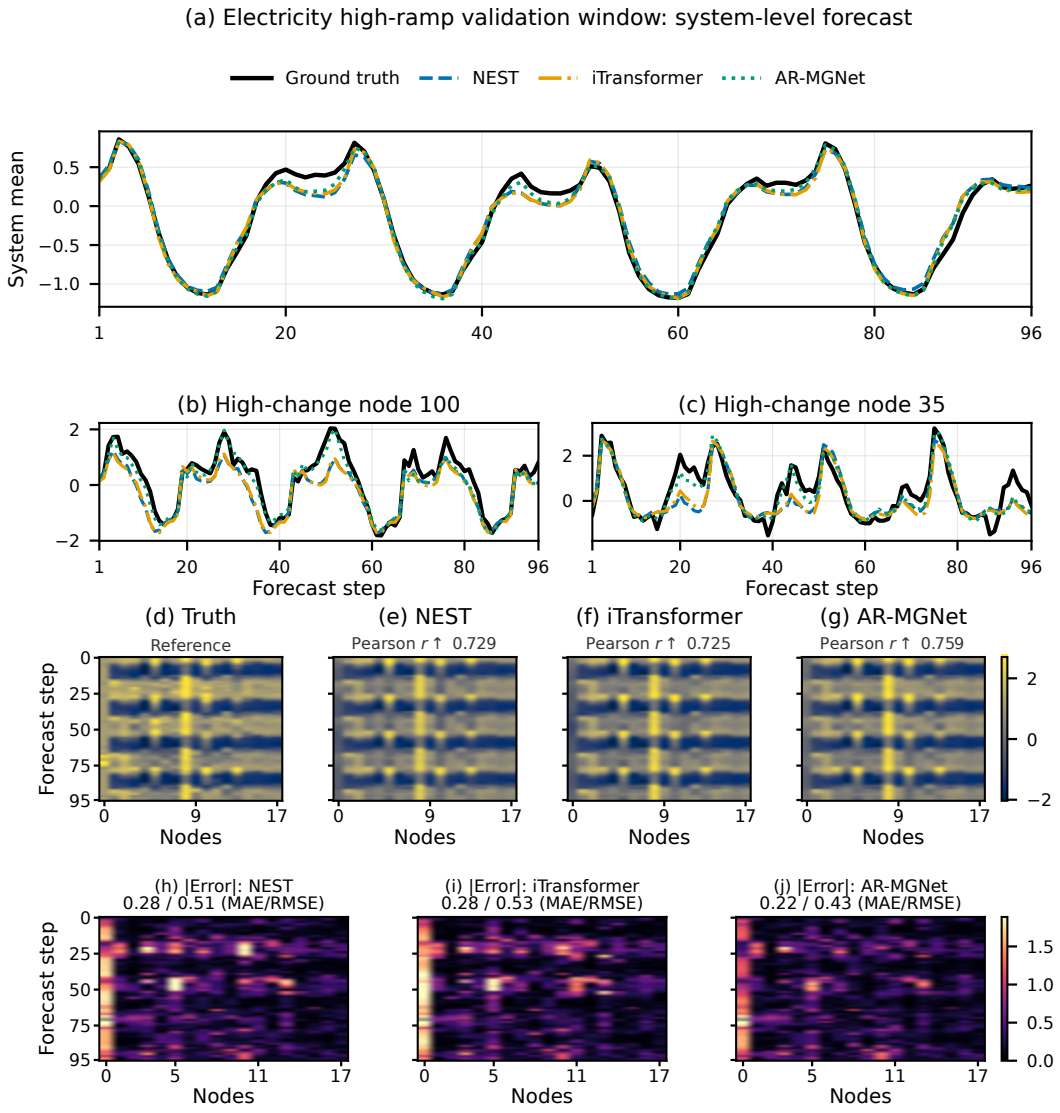


Figure 9: Electricity forecasting on a smooth high-change validation window. (a)–(j) compare NEST, iTransformer, and AR-MGNet through the system trace, two high-change nodes, prediction fields, and absolute-error fields.

F.2 Dynamic Grid Under Traffic Congestion

Figure 10 compares AR-MGNet with NEST and HieraMix on the CA congestion. AR-MGNet reaches MAE/RMSE 29.11/45.03, improving on NEST by 57.9%/55.1% and on HieraMix by 55.0%/51.9%. The window changes 9.3% of node-to-parent assignments and activates 50.0% of nodes. Along the top-down path, Level 1 represents regional corrections with mean absolute innovation 0.244, and Level 0 resolves localized departures with mean absolute innovation 0.428. The Level 1 magnitude is 43.0% lower than the Level 0 magnitude, quantifying the spatial-scale contrast between the displayed innovation fields. Together with the event-window errors, the reorganized regions and active-node maps show how the hierarchy combines regional context with localized congestion recovery.

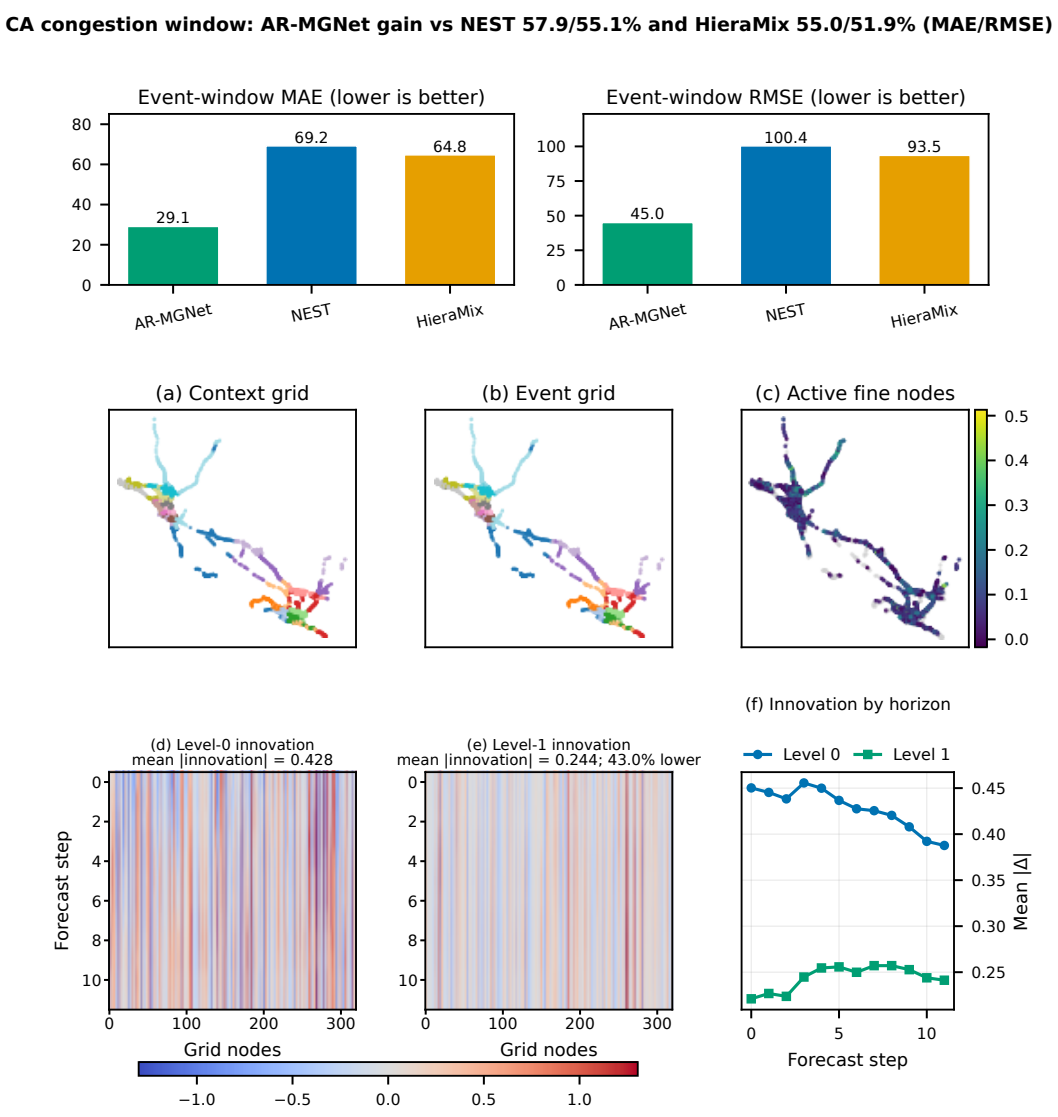


Figure 10: Dynamic functional-grid response around traffic congestion event. (a)–(b) show how the window-conditioned hierarchy reorganizes between context and event states; (c) locates the budgeted regions receiving fine computation; (d)–(e) show local Level-0 and regional Level-1 innovation fields; and (f) tracks the level-wise innovation magnitude across the forecast horizon.

F.3 Dynamic Grid Under Pollution Transport

Figure 11 reports the corresponding pollution event. AR-MGNet reaches MAE/RMSE 29.72/46.80, improving on NEST by 20.2%/16.6% and on HieraMix by 18.4%/14.9%. The event changes 23.9% of assignments and activates 50.0% of nodes. Top-down reconciliation combines Level 1 regional corrections, whose mean absolute innovation is 0.251, with Level 0 localized corrections at 0.294. The Level 1 magnitude is 14.8% lower than the Level 0 magnitude, describing the different spatial scales of the two innovation fields. The assignment changes and lower event-window forecast errors together illustrate adaptive neighborhoods for regional pollution patterns and local departures.

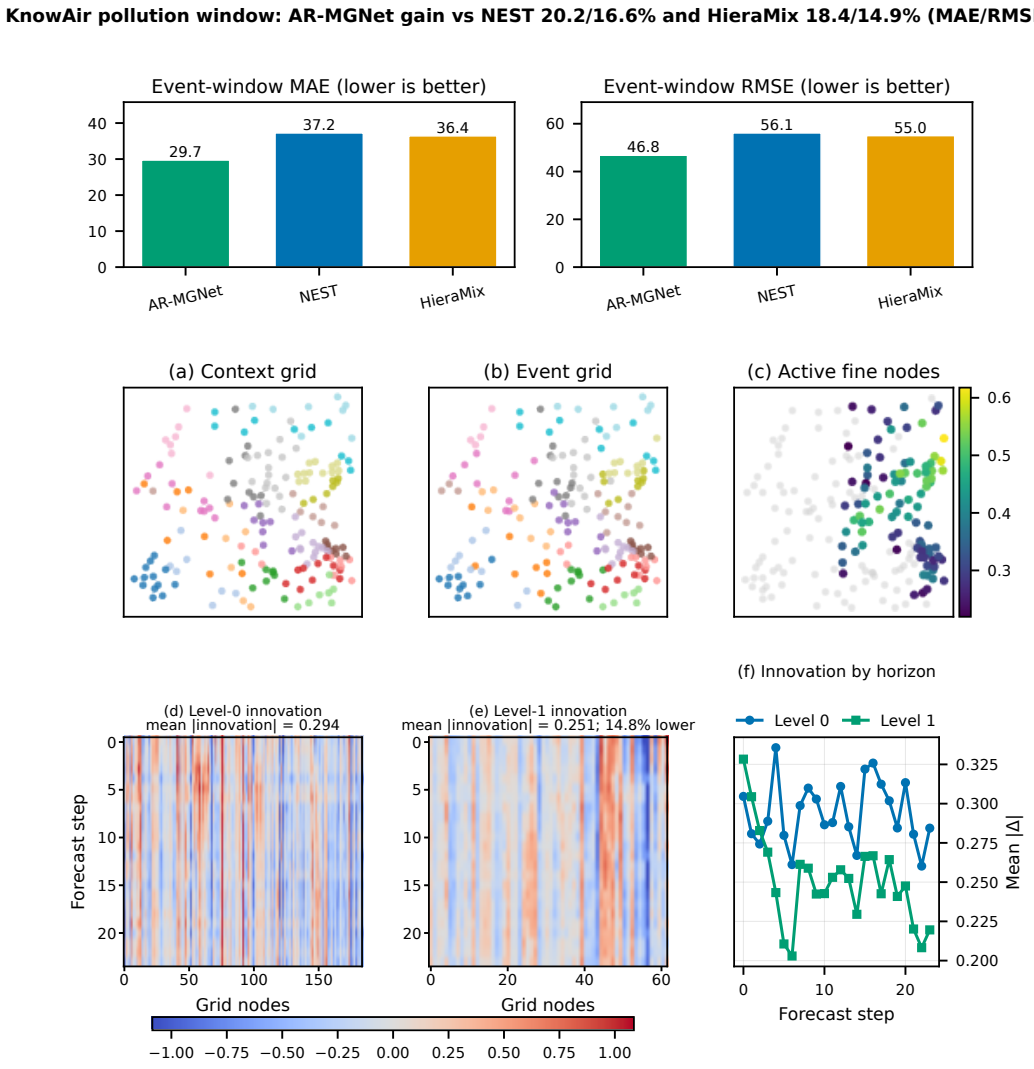


Figure 11: Dynamic functional-grid response around pollution event. (a)–(b) show how the window-conditioned hierarchy reorganizes between context and event states; (c) locates the budgeted regions receiving fine computation; (d)–(e) show local Level-0 and regional Level-1 innovation fields; and (f) tracks the level-wise innovation magnitude across the forecast horizon.