TOPODUET: ONE TOPOLOGY, TWO RELATIONAL FIELDS FOR SPATIO-TEMPORAL FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatio-temporal graph forecasters typically encode observations and temporal variations in a single latent tensor, although locations that look alike need not change alike. We introduce TOPODUET, which separates these relational roles into a *Configuration Field* $\mathbf{L}$ induced by absolute observations and a *Transition Field* $\mathbf{I}$ induced by signed first- and second-order variations. The Dual-Field Spatio-Temporal Encoder learns independent temporal and spatial relations for the two fields. Field-Adaptive Road Attention supplies both fields with forward–backward road support while learning field- and layer-specific prior strengths, and Progressive Cross-Field Coupling exchanges their evidence after each block without collapsing their identities. A horizon-adaptive dual-path readout then balances nonlinear correction with historical persistence. Our structural analysis quantifies the distortion caused by forcing one shared relation, identifies the minimal direction-complete closure of a directed road prior, and characterizes when a common soft prior yields identical field relations. Experiments on eight traffic forecasting datasets show that TOPODUET ranks first on all 12 MAE/RMSE comparisons across six graph-based highway datasets, reducing the best competing error by up to 4.9% in MAE and 7.3% in RMSE. The same architecture also transfers effectively across the T-Drive and CHIBike grid benchmarks, extending the dual-field formulation from sensor graphs to urban-flow fields. Controlled ablations further show that removing field separation increases MAE/RMSE by $0.10/0.62$ on PeMS08 and $0.28/0.95$ on T-Drive, supporting the value of modeling configuration and transition relations separately.

INTRODUCTION

Spatio-temporal forecasting learns *where* observations interact and *when* dependencies unfold, but usually leaves implicit *which relation is being learned*. At a given time, absolute observations induce a configuration relation among locations with comparable values, whereas signed variations induce a transition relation among locations that change together. On road networks, these relations share the same physical topology but need not agree: similar measurements can move in opposite directions, while different measurements can respond synchronously. Diffusion-based models capture directed propagation Li et al. (2018). Graph-convolutional models learn local spatial dependencies Yu et al. (2018). Adaptive-graph models learn latent relations beyond the supplied topology Wu et al. (2019). Node-adaptive graph convolution further specializes these relations by location Bai et al. (2020). Attention-based models capture dynamic and long-range interactions Zheng et al. (2020). Propagation-aware attention additionally models delayed spatial dependencies Jiang et al. (2023). Adaptive spatio-temporal embeddings offer another route to dynamic dependency learning Liu et al. (2023). However, most compress values, variations, node identities, and calendar context into one hidden tensor, causing configuration similarity and transition synchrony to be inferred from the same carrier. Even when explicit differences are appended as input channels, their relational identity disappears after the first projection. This motivates modeling two interacting node-time fields rather than one shared spatio-temporal relation.

This single-carrier convention leaves three fundamental challenges. **C1: Spatio-Temporal Relational Aliasing.** One latent field must simultaneously answer which locations look alike and which locations change alike. When the two answers disagree, a shared representation conflates configuration similarity with transition synchrony, obscuring precisely the cases in which similar observations

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.

lead to different futures. **C2: The Shared-Topology Support Dilemma.** Both relations inhabit the same directed network, yet forward-only support omits reverse-aligned structural associations. Giving the fields different physical supports would confound topology with relational identity, while forcing one common bias strength would assume that both relations rely on topology equally. They should receive the same forward–backward candidate support but retain separate control over its influence. **C3: The Cross-Field Coupling Dilemma.** The relations cannot be isolated because transitions act on observed configurations and continually reshape them. Yet early fusion erases their identities before either relation is formed, whereas output-only fusion postpones interaction until after abstraction. Their exchange must occur without collapsing the distinction that made it useful.

To address these challenges, we propose TOPODUET with three corresponding designs. (1) A *Dual-Field Spatio-Temporal Encoder* (DFSTE) lifts absolute observations and signed variations into configuration and transition fields, giving each a complete temporal–spatial pathway to prevent relational aliasing (C1). (2) *Field-Adaptive Road Attention* (FARA) expands the directed direct/two-hop affinity with fixed backward support, supplies the same candidate topology to both fields, and learns a separate prior strength for every field and layer (C2). (3) *Progressive Cross-Field Coupling* (PCFC) performs gated residual exchange after every block, preserving distinct pathways while allowing their evidence to refine one another. A horizon-adaptive readout carries this coordination into prediction by balancing nonlinear correction and historical persistence (C3).

The resulting model, illustrated in Figure 1, is intentionally compact. Its contributions are:

- We introduce a *dual-field relational formulation* for spatio-temporal forecasting that distinguishes configuration relations, induced by absolute observations, from transition relations, induced by signed first- and second-order temporal variations. This formulation explicitly addresses relational aliasing caused by representing these two forms of dependency within a single latent carrier.
- We develop TOPODUET, a unified dual-field forecasting architecture consisting of the Dual-Field Spatio-Temporal Encoder (DFSTE), Field-Adaptive Road Attention (FARA), Progressive Cross-Field Coupling (PCFC), and a horizon-adaptive dual-path readout. DFSTE independently models temporal and spatial dependencies in the two fields; FARA provides a common forward–backward road support with field- and layer-specific prior strengths; and PCFC enables progressive information exchange without prematurely collapsing the two field representations.
- We provide a structural analysis of the proposed dual-field design. Specifically, we quantify the representation distortion incurred when two learned field-specific attention relations are collapsed into one shared relation, establish the minimal direction-complete closure of a directed road support, and derive the exact condition under which two fields receiving the same soft road prior produce identical spatial attention relations.
- Extensive experiments on eight spatio-temporal forecasting datasets demonstrate the effectiveness and generality of TOPODUET. The model achieves the best MAE and RMSE on all six graph-based highway datasets and consistently strong performance across the T-Drive and CHIBike grid benchmarks. Controlled ablations further verify the contributions of field separation, road-guided structural support, field adaptation, and progressive cross-field coupling.

RELATED WORK

SPATIO-TEMPORAL TRAFFIC FORECASTING

Spatio-temporal graph forecasting models node signals over irregular spatial supports. DCRNN Li et al. (2018) combines directed diffusion with recurrence, whereas STGCN Yu et al. (2018) and STSGCN Song et al. (2020) use graph-based spatio-temporal convolutions. Graph WaveNet Wu et al. (2019) and AGCRN Bai et al. (2020) learn adaptive graph relations. Attention-based models further expand the receptive field: GMAN Zheng et al. (2020) applies spatial and temporal attention, PDFormer Jiang et al. (2023) models dynamic and propagation-aware dependencies, and STAEformer Liu et al. (2023) and STID Shao et al. (2022) emphasize learned spatial, temporal, node, and

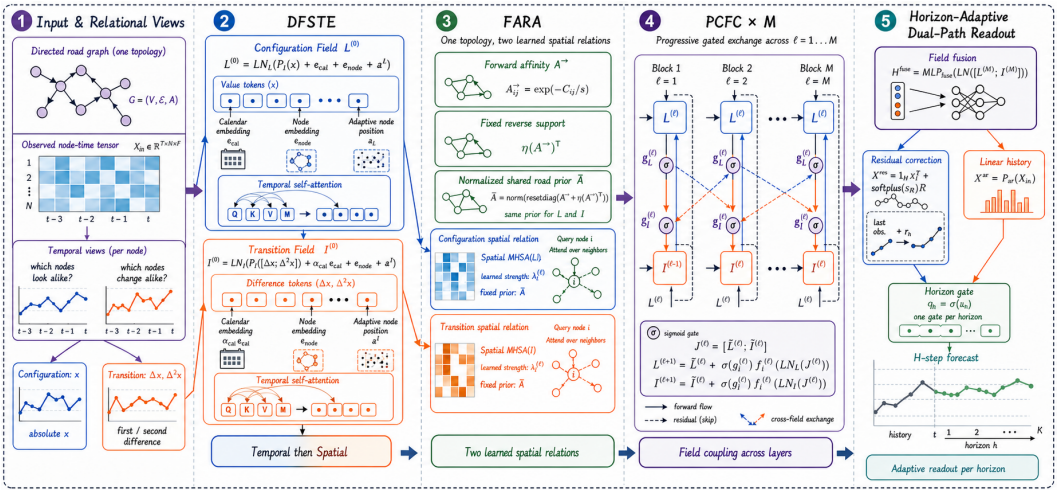

Figure 1: Overview of TOPODUET. DFSTE models configuration and transition fields separately, FARA injects field-adaptive road priors, PCFC enables progressive cross-field exchange, and the dual-path readout produces horizon-adaptive forecasts.

calendar identities. These methods improve where and when information is aggregated, but generally encode configuration similarity and transition synchrony in one hidden carrier; TOPODUET instead assigns them separate relational fields.

TEMPORAL REPRESENTATIONS AND SPATIAL PRIORS

Classical forecasting uses differencing to expose temporal change Box et al. (2015), while DLinear Zeng et al. (2023), PatchTST Nie et al. (2023), and RevIN Kim et al. (2022) demonstrate the importance of temporal representation and scale. Our distinction is to lift observations and innovations into separate node-time fields so that each can induce its own spatial relation. Spatially, fixed graphs preserve physical reachability, whereas adaptive graphs capture latent dependencies Wu et al. (2019). FARA combines these roles through a shared soft road prior with field- and layer-specific strengths, and PCFC exchanges information only after each field completes independent spatio-temporal processing. The dual-path readout follows the mixture-of-experts principle Jacobs et al. (1991) by balancing nonlinear correction and linear persistence at each horizon.

METHOD

TOPODUET treats absolute observations and signed variations as two sources of relational evidence. The configuration field asks which nodes exhibit similar observed levels, whereas the transition field asks which nodes exhibit similar directions and changes of motion. Both fields retain complete temporal and spatial processing; they differ in their input coordinates and parameters, not in whether they model time or space. A forward pass has four stages. DFSTE initializes and separately encodes the two fields, FARA injects the same road-derived support into their spatial attention with field-specific strength, PCFC exchanges the resulting evidence after every block, and the readout maps the terminal fields to all forecast horizons. Figure 1 follows these stages from left to right.

PROBLEM FORMULATION

We represent the road network as a directed graph $G = (V, E)$, where V is a set of $N = |V|$ sensor nodes and E encodes the directed road connections. At each discrete time step t, the graph state is represented by a node feature matrix $\mathbf{X}_t = [\mathbf{x}_{1,t}, \mathbf{x}_{2,t}, \ldots, \mathbf{x}_{N,t}]^\top \in \mathbb{R}^{N \times F}$, where $\mathbf{x}_{n,t} \in \mathbb{R}^F$ denotes the feature vector observed at node n and time step t, and F is the number of node features. We further define $\mathbf{c}_t \in \mathbb{R}^{d_c}$ as the calendar vector at time step t, where d_c denotes the number of calendar features, such as time-of-day and day-of-week information.

STG Forecasting Problem. Given the historical graph states $\mathcal{X}_{\text{in}} = [\mathbf{X}_{t-T}; \ldots; \mathbf{X}_{t-1}] \in \mathbb{R}^{T\times N\times F}$ and their corresponding calendar vectors $\mathcal{C}_{\text{in}} = [\mathbf{c}_{t-T}^{\top}; \ldots; \mathbf{c}_{t-1}^{\top}] \in \mathbb{R}^{T\times d_c}$ over the preceding T time steps, together with the graph G, we learn the mapping

$$\mathcal{X}_{\text{out}} = \mathcal{F}\left(\mathcal{X}_{\text{in}}, \mathcal{C}_{\text{in}}; G, \Theta\right), \tag{1}$$

where $\mathcal{X}_{\text{out}} = [\mathbf{X}_t; \ldots; \mathbf{X}_{t+H-1}] \in \mathbb{R}^{H\times N\times F}$ represents the future graph states over the subsequent H time steps, and Θ denotes all learnable parameters.

Dual-Field Spatio-Temporal Encoder

The Dual-Field Spatio-Temporal Encoder (DFSTE) constructs two complementary node–time fields from the historical observations. The configuration, or level, field, denoted by $\mathbf{L}$, represents the observed system states, whereas the transition, or innovation, field, denoted by $\mathbf{I}$, represents their first- and second-order temporal variations. Each field is subsequently processed through an independent temporal–spatial pathway. The following paragraphs define the observed variation coordinates, field initialization, and factorized axis attention. The road-guided spatial interaction and progressive cross-field exchange are detailed in the subsequent FARA and PCFC subsections, respectively.

Observed variation coordinates. For each position $\tau \in \{1, \ldots, T\}$ in the historical input window and each node $n \in \{1, \ldots, N\}$, let $\mathbf{x}_{\tau,n} := (\mathcal{X}_{\text{in}})_{\tau,n,:} \in \mathbb{R}^F$ denote the corresponding node feature vector. The transition field uses first- and second-order temporal differences computed entirely within the observed window:

$$\Delta\mathbf{x}_{\tau,n} = \begin{cases} \mathbf{0}_F, & \tau = 1, \\ \mathbf{x}_{\tau,n} - \mathbf{x}_{\tau-1,n}, & \tau > 1, \end{cases} \tag{2}$$

$$\Delta^2\mathbf{x}_{\tau,n} = \begin{cases} \mathbf{0}_F, & \tau \le 2, \\ \Delta\mathbf{x}_{\tau,n} - \Delta\mathbf{x}_{\tau-1,n}, & \tau > 2. \end{cases} \tag{3}$$

The leading zero vectors reproduce the boundary convention used by the implementation and do not involve target-time padding. The first-order variation describes the direction and magnitude of change, whereas the second-order variation characterizes how that change is strengthening, weakening, or reversing.

Field initialization. Let d denote the dimensionality of the learned hidden representation associated with each node–time pair. For each input-window position τ, let $\mathbf{c}_\tau := \left((\mathcal{C}_{\text{in}})_{\tau,:}\right)^{\top} \in \mathbb{R}^{d_c}$ denote the corresponding calendar vector. Among its components, c_τ^{tod} and c_τ^{dow} denote the time-of-day and day-of-week indices used by the model. Their calendar representation is

$$\mathbf{e}_\tau^{\text{cal}} = E_{\text{tod}}\left(c_\tau^{\text{tod}}\right) + E_{\text{dow}}\left(c_\tau^{\text{dow}}\right) \in \mathbb{R}^d. \tag{4}$$

Let $\mathbf{e}_n^{\text{node}} \in \mathbb{R}^d$ denote the learned embedding of node n, and let $\mathbf{a}_{\tau,n}^L, \mathbf{a}_{\tau,n}^I \in \mathbb{R}^d$ denote independent field-specific node–position embeddings. We use two learned affine projections $P_L : \mathbb{R}^F \to \mathbb{R}^d$ and $P_I : \mathbb{R}^{2F} \to \mathbb{R}^d$ for the configuration and transition inputs, respectively. The initial fields are constructed as

$$\mathbf{L}_{\tau,n}^{(0)} = \text{LN}_L^{\text{in}}\left(P_L(\mathbf{x}_{\tau,n}) + \mathbf{e}_\tau^{\text{cal}} + \mathbf{e}_n^{\text{node}} + \mathbf{a}_{\tau,n}^L\right), \tag{5}$$

$$\mathbf{I}_{\tau,n}^{(0)} = \text{LN}_I^{\text{in}}\left(P_I\left([\Delta\mathbf{x}_{\tau,n}; \Delta^2\mathbf{x}_{\tau,n}]\right) + \alpha_{\text{cal}}\mathbf{e}_\tau^{\text{cal}} + \mathbf{e}_n^{\text{node}} + \mathbf{a}_{\tau,n}^I\right). \tag{6}$$

Consequently, $\mathbf{L}^{(0)}, \mathbf{I}^{(0)} \in \mathbb{R}^{T\times N\times d}$. The two fields share calendar and node identities but use distinct input projections, field-specific node–position embeddings, and input normalization layers. Here, α_{cal} is a fixed coefficient controlling the contribution of calendar information to the transition field.

Factorized axis attention. Let M denote the number of stacked encoder blocks. For each block $\ell \in \{0, \ldots, M-1\}$, $\mathbf{L}^{(\ell)}, \mathbf{I}^{(\ell)} \in \mathbb{R}^{T\times N\times d}$ denote the level-field and innovation-field representations entering that block, respectively. For $\ell = 0$, they are the initial fields defined above; for $\ell > 0$, they are produced by the preceding PCFC operation.

For field $f \in \{L, I\}$ and axis $a \in \{\mathrm{temp}, \mathrm{spat}\}$, the corresponding axis operator maps an input $\mathbf{H} \in \mathbb{R}^{T\times N\times d}$ as

$$\begin{aligned}\mathbf{U}^{(\ell)}_{a,f}(\mathbf{H}) &= \mathbf{H} + \mathrm{MHSA}^{(\ell)}_{a,f}\left(\mathrm{LN}^{(\ell)}_{a,f}(\mathbf{H})\right),\\ \mathrm{Axis}^{(\ell)}_{a,f}(\mathbf{H}) &= \mathbf{U}^{(\ell)}_{a,f}(\mathbf{H}) + \mathrm{FFN}^{(\ell)}_{a,f}\left(\mathrm{LN}'^{(\ell)}_{a,f}\left(\mathbf{U}^{(\ell)}_{a,f}(\mathbf{H})\right)\right).\end{aligned} \tag{7}$$

Each feed-forward network independently transforms every node–time representation through a $d \to 3d \to d$ mapping with GELU activation and dropout. The temporal operator compares the T historical representations belonging to the same node, whereas the spatial operator compares the N node representations observed at the same input-window position. These comparisons are implemented through multi-head self-attention. The spatial operator additionally incorporates the road-derived guidance defined in the FARA subsection.

The two fields follow separate temporal–spatial pathways:

$$\begin{aligned}\mathbf{L}^{(\ell)}_{\mathrm{temp}} &= \mathrm{Axis}^{(\ell)}_{\mathrm{temp},L}\left(\mathbf{L}^{(\ell)}\right), \quad \widetilde{\mathbf{L}}^{(\ell)} = \mathrm{Axis}^{(\ell)}_{\mathrm{spat},L}\left(\mathbf{L}^{(\ell)}_{\mathrm{temp}}\right),\\ \mathbf{I}^{(\ell)}_{\mathrm{temp}} &= \mathrm{Axis}^{(\ell)}_{\mathrm{temp},I}\left(\mathbf{I}^{(\ell)}\right), \quad \widetilde{\mathbf{I}}^{(\ell)} = \mathrm{Axis}^{(\ell)}_{\mathrm{spat},I}\left(\mathbf{I}^{(\ell)}_{\mathrm{temp}}\right).\end{aligned} \tag{8}$$

The resulting representations $\widetilde{\mathbf{L}}^{(\ell)}$ and $\widetilde{\mathbf{I}}^{(\ell)}$ are subsequently passed to PCFC for gated cross-field information exchange.

Field-Adaptive Road Attention

FARA supplies structural guidance to the spatial axis blocks without replacing data-dependent attention. Its construction follows three decisions. First, road distance produces a forward affinity. Second, a fixed transpose term gives reverse-aligned pairs candidate support that a forward-only prior cannot provide. Third, the resulting support is supplied to both fields because they occupy the same physical network; each field still learns its own content relation and its own reliance on the prior.

Intuition. Consider two sensors u and v for which the road file contains only the directed connection $u \to v$. The forward affinity $A^{\rightarrow}$ gives the attention from query node v to source node u a road-based preference, because u can reach v in the recorded direction. It gives the reverse ordered pair no such preference. Reverse-aligned support is useful when measurements upstream and downstream co-vary during queue formation, where it represents a predictive structural association. The augmented affinity $A^{\leftrightarrow}$ retains the forward preference and adds an attenuated transpose, so both ordered pairs are available as candidate structural support. In short, $A^{\rightarrow}$ describes affinity following the recorded direct/two-hop direction, whereas $A^{\leftrightarrow}$ supplements it with fixed reverse-aligned coverage.

Road-derived candidate support. The implementation stores costs using the convention $C_{\mathrm{target,source}}$. After duplicate ordered edges are removed, a directed edge $j \to i$ of recorded length $\ell_{j\to i}$ defines

$$C^{(1)}_{ij} = \begin{cases} 0, & i = j,\\ \log(1 + \max(\ell_{j\to i}, 0)), & (j \to i) \in E,\\ +\infty, & \text{otherwise.}\end{cases} \tag{9}$$

One min-plus composition adds direct and two-hop connectivity:

$$C_{ij} = \min_{k\in V}\left\{C^{(1)}_{ik} + C^{(1)}_{kj}\right\}. \tag{10}$$

Thus C_{ij} contains either a direct route or the best route through exactly one intermediate node. Applying the logarithmic transform before composition yields the compressed connectivity cost C used by the road-prior construction.

Let s be the median positive finite entry of C. With a small constant $\varepsilon > 0$, the forward affinity is

$$A^{\rightarrow}_{ij} = \begin{cases} \exp(-C_{ij}/\max(s, \varepsilon)), & C_{ij} < +\infty,\\ 0, & C_{ij} = +\infty.\end{cases} \tag{11}$$

The diagonal is therefore one. With a fixed reverse-support hyperparameter $\eta > 0$, the code augments only off-diagonal entries and then max-normalizes:

$$A^{\leftrightarrow}_{ij} = \begin{cases} 1, & i = j, \\ A^{\rightarrow}_{ij} + \eta A^{\rightarrow}_{ji}, & i \neq j, \end{cases} \qquad \overline{A} = \frac{A^{\leftrightarrow}}{\max_{u,v} A^{\leftrightarrow}_{uv}}. \tag{12}$$

The transpose term provides a fixed structural bonus to pairs supported in the reverse recorded direction, and resetting the diagonal before normalization completes the implemented road-prior construction.

Additive spatial attention bias. For field $f \in \{L, I\}$ and encoder block $\ell \in \{0, \ldots, M-1\}$, the learnable prior strength is

$$\lambda_f^{(\ell)} = \zeta\left(\rho_f^{(\ell)}\right), \qquad f \in \{L, I\}, \tag{13}$$

where $\rho_f^{(\ell)}$ is one scalar parameter owned by that field's spatial attention module in encoder block ℓ. When $\rho_f^{(\ell)}$ becomes strongly negative, $\lambda_f^{(\ell)}$ approaches zero and the field approaches content-only spatial attention. Increasing $\lambda_f^{(\ell)}$ gives differences in road affinity more influence, without changing the prior itself.

Let R denote the number of attention heads and let $d_h = d/R$ denote the hidden dimension of each head. For observed step τ and attention head $r \in \{1, \ldots, R\}$, let $Q^{(\ell)}_{f,\tau,r}, K^{(\ell)}_{f,\tau,r}, V^{(\ell)}_{f,\tau,r} \in \mathbb{R}^{N \times d_h}$. The pre-dropout attention distribution and the corresponding head output are

$$\mathbf{S}^{(\ell)}_{f,\tau,r} = \frac{Q^{(\ell)}_{f,\tau,r}(K^{(\ell)}_{f,\tau,r})^\top}{\sqrt{d_h}} + \lambda_f^{(\ell)}\overline{A}, \tag{14}$$

$$\mathbf{P}^{(\ell)}_{f,\tau,r} = \mathrm{softmax}_{\mathrm{row}}(\mathbf{S}^{(\ell)}_{f,\tau,r}), \tag{15}$$

$$\mathbf{O}^{(\ell)}_{f,\tau,r} = \mathrm{Dropout}(\mathbf{P}^{(\ell)}_{f,\tau,r})V^{(\ell)}_{f,\tau,r}. \tag{16}$$

The standard multi-head concatenation and output projection then complete the spatial MHSA module. The analysis below concerns the deterministic pre-dropout distribution $\mathbf{P}^{(\ell)}_{f,\tau,r}$, because the road prior enters before attention dropout. The representational consequences of field separation and the shared road prior, together with the corresponding propositions and proofs, are analyzed in Appendix C.

Progressive Cross-Field Coupling

After each field independently completes its temporal and spatial attention, Progressive Cross-Field Coupling (PCFC) allows the two resulting representations to exchange information. For encoder block $\ell \in \{0, \ldots, M-1\}$, let

$$\mathbf{J}^{(\ell)} = [\widetilde{\mathbf{L}}^{(\ell)}; \widetilde{\mathbf{I}}^{(\ell)}] \in \mathbb{R}^{T \times N \times 2d} \tag{17}$$

denote their channel-wise concatenation. We use two independent LayerNorm modules and two independent update networks $f_L^{(\ell)}, f_I^{(\ell)} : \mathbb{R}^{2d} \to \mathbb{R}^d$. Each update network applies a $2d \to d$ affine projection, GELU activation, and a $d \to d$ affine projection. The coupled fields are updated as

$$\mathbf{L}^{(\ell+1)} = \widetilde{\mathbf{L}}^{(\ell)} + \sigma(g_L^{(\ell)}) f_L^{(\ell)}\left(\mathrm{LN}_L^{(\ell)}(\mathbf{J}^{(\ell)})\right), \tag{18}$$

$$\mathbf{I}^{(\ell+1)} = \widetilde{\mathbf{I}}^{(\ell)} + \sigma(g_I^{(\ell)}) f_I^{(\ell)}\left(\mathrm{LN}_I^{(\ell)}(\mathbf{J}^{(\ell)})\right). \tag{19}$$

Consequently, $\mathbf{L}^{(\ell+1)}, \mathbf{I}^{(\ell+1)} \in \mathbb{R}^{T \times N \times d}$. Because each update network receives the concatenated representation, it can combine information from both the level and innovation fields, while the residual connection preserves the field representation obtained before coupling. The two gate parameters $g_L^{(\ell)}$ and $g_I^{(\ell)}$ are learned independently, allowing different information-exchange strengths in the two directions. Each gate is a block-wise scalar shared across observed steps and nodes. After the final encoder block, PCFC produces the terminal fields $\mathbf{L}^{(M)}$ and $\mathbf{I}^{(M)}$, which are passed to the readout module.

HORIZON-ADAPTIVE DUAL-PATH READOUT

After M encoder blocks, the terminal level and innovation fields $\mathbf{L}^{(M)}, \mathbf{I}^{(M)} \in \mathbb{R}^{T\times N\times d}$ are concatenated and mapped into a fused representation:

$$\mathbf{H}^{\text{fuse}} = \text{MLP}_F\left(\text{LN}[\mathbf{L}^{(M)}; \mathbf{I}^{(M)}]\right) \in \mathbb{R}^{T\times N\times d}. \tag{20}$$

The readout contains two complementary prediction paths. For each node n, the residual path uses the fused historical representation to predict horizon-specific corrections around the last observed state:

$$\mathbf{R}_n = P_R\left(\text{vec}(\mathbf{H}^{\text{fuse}}_{:,n,:})\right), \tag{21}$$

$$\mathbf{X}^{\text{res}}_n = \mathbf{1}_H \mathbf{x}^{\top}_{T,n} + \zeta(s_R)\mathbf{R}_n, \tag{22}$$

where $\mathbf{x}_{T,n} = (\mathcal{X}_{\text{in}})_{T,n,:} \in \mathbb{R}^F$ is the last observed state of node n, and $\zeta(s_R)$ is a learnable positive scale controlling the magnitude of the residual correction.

In parallel, the linear path directly maps the observed history of each node to the future horizons:

$$\mathbf{X}^{\text{ar}}_n = P_{\text{ar}}\left((\mathcal{X}_{\text{in}})_{:,n,:}\right), \tag{23}$$

where P_{ar} is shared across nodes.

Finally, a horizon-specific gate $q_h = \sigma(u_h)$ adaptively combines the two prediction paths:

$$(\mathcal{X}_{\text{out}})_{h,n,:} = (1-q_h)(\mathbf{X}^{\text{res}}_n)_{h,:} + q_h(\mathbf{X}^{\text{ar}}_n)_{h,:}, \qquad h \in \{1,\dots,H\}. \tag{24}$$

The gate is learned separately for each forecast horizon and shared across nodes, allowing the model to adapt the balance between nonlinear correction and historical persistence over the prediction range.

EXPERIMENTS

DATASETS

We use PeMS03, PeMS04, PeMS07, and PeMS08 Song et al. (2020), PeMSD7(M) and PeMSD7(L) Yu et al. (2018), and T-Drive and CHIBike Hu et al. (2026). Table 1 summarizes their domains, graph sizes, sampling intervals, time ranges, and chronological splits.

BASELINE

We compare with ARIMA Box et al. (2015), FC-LSTM Hochreiter & Schmidhuber (1997), GRU-ED Cho et al. (2014), DCRNN Li et al. (2018), ASTGCN Guo et al. (2019), STGCN Yu et al. (2018), STGODE Fang et al. (2021), STSGCN Song et al. (2020), MSTGCN Guo et al. (2019), TFormer Cai et al. (2020), PDFormer Jiang et al. (2023), STAEformer Liu et al. (2023), STGNCDE Choi et al. (2022), CaST Xia et al. (2023), RAST Ruan et al. (2026), and MFormer Hu et al. (2026).

EXPERIMENTAL SETTINGS

All experiments were conducted on a Linux server with a single NVIDIA GeForce RTX 4090 GPU (24 GB), using Python 3.10.8, PyTorch 2.1.2, CUDA 11.8, and cuDNN 8.7. Models were optimized with AdamW using linear warm-up, cosine learning-rate decay, gradient clipping, exponential moving-average weights, and validation-based early stopping. Checkpoint selection follows validation MAE, and the main results report the mean and standard deviation over three independent seeds.

EXPERIMENTAL RESULTS

Tables 2 and 3 demonstrate strong and consistent forecasting performance across graph-based highway networks and grid-based urban-flow systems. On all six highway datasets, TOPODUET achieves

Table 1: Dataset statistics.

Dataset	Domain	#Nodes	#Edges	Interval	Time Range	Split
PeMS03	Traffic Flow	358	547	5 min	09/2018–11/2018	6:2:2
PeMS04	Traffic Flow	307	340	5 min	01/2018–02/2018	6:2:2
PeMS07	Traffic Flow	883	866	5 min	05/2017–08/2017	6:2:2
PeMS08	Traffic Flow	170	303	5 min	07/2016–08/2016	6:2:2
PeMSD7(M)	Traffic Speed	228	1,132	5 min	05/2012–06/2012	6:2:2
PeMSD7(L)	Traffic Speed	1,026	10,150	5 min	05/2012–06/2012	6:2:2
CHIBike	Urban Flow	270	1,966	30 min	07/2020–09/2020	7:1:2
T-Drive	Urban Flow	1,024	7,812	60 min	02/2015–06/2015	7:1:2

Table 2: Performance comparison on graph-based highway datasets (MAE / RMSE; lower is better). Red and blue boldface indicate the best and second-best mean results.

	PEMS03		PeMS04		PeMS07		PeMS08		PeMSD7(M)		PeMSD7(L)	
Model	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
ARIMA	34.92	46.73	32.91	47.42	37.56	58.17	30.67	43.51	7.27	13.20	7.51	12.39
FC-LSTM	20.85	34.18	25.93	39.27	29.12	44.82	22.31	34.16	4.16	7.51	4.66	8.20
GRU-ED	18.67	31.84	22.89	38.15	26.94	42.53	21.38	35.27	4.78	9.05	3.98	7.71
DCRNN	17.73	29.97	20.94	32.91	24.89	38.02	16.58	25.93	3.83	7.18	4.33	8.33
ASTGCN	16.98	29.01	22.41	34.52	23.51	37.15	17.92	27.78	3.61	6.87	4.09	7.64
STGCN	17.16	29.78	20.67	34.11	24.78	38.52	17.16	26.58	4.01	7.55	4.84	8.28
STGODE	16.18	27.16	20.36	32.08	22.04	36.71	16.35	25.37	2.97	5.66	3.22	5.98
STSGCN	17.09	28.56	20.69	32.91	23.72	38.14	16.76	26.22	3.01	5.93	3.61	6.88
MSTGCN	19.12	31.24	23.39	36.41	28.31	42.78	18.67	28.64	3.54	6.14	3.58	6.43
TFormer	**14.82**	26.18	19.47	32.35	21.39	35.07	15.73	25.69	2.88	5.84	3.02	5.97
PDFormer	15.13	26.78	18.73	31.24	20.47	33.94	14.03	24.29	2.81	5.60	2.92	5.90
STAEformer	14.83	25.97	**18.06**	30.97	19.87	33.18	**13.87**	24.01	2.72	5.56	2.93	5.81
STGNCDE	16.47	27.83	19.21	31.74	22.17	36.08	15.94	25.48	**2.68**	**5.39**	**2.87**	**5.76**
CaST	15.62	26.94	18.84	31.38	21.03	34.57	14.76	24.83	3.21	6.03	3.37	6.30
RAST	15.20	**25.81**	18.39	**30.55**	**19.52**	**32.97**	14.16	**23.33**	2.96	5.99	3.21	6.33
MFormer	15.88	26.61	19.12	30.98	20.45	33.41	14.70	24.49	3.22	6.13	3.44	6.51
TopoDuet	**14.49** ±0.04	**23.93** ±0.15	**18.04** ±0.03	**30.41** ±0.08	**18.97** ±0.08	**32.96** ±0.15	**13.27** ±0.06	**22.92** ±0.21	**2.55** ±0.00	**5.30** ±0.02	**2.73** ±0.00	**5.72** ±0.03

the best MAE and RMSE, outperforming graph-based and Transformer-based competitors including PDFormer, STAEformer, and RAST. The gains hold across both traffic-flow and traffic-speed benchmarks and across networks of substantially different sizes, showing that the dual-field relational formulation remains effective across signal types and graph scales. The same architecture retains strong accuracy on the urban-grid benchmarks, attaining leading results on T-Drive and stable performance on CHIBike. Together, these results support the central premise that distinguishing how locations *look* from how they *change* provides a more informative representation when spatial configurations and temporal transitions jointly determine future observations.

Ablation Study

Table 4 evaluates the contribution of each component on PeMS08 and T-Drive Inflow. For C1, removing the transition field or collapsing field separation increases prediction error on both datasets, with field collapse producing the largest degradation and directly validating the independent temporal–spatial pathways of DFSTE. For C2, the road-prior, backward-support, and field-adaptation variants consistently demonstrate the value of FARA: a shared direction-complete candidate topology is most effective when each field learns its own reliance on that support. For C3, removing PCFC raises MAE on both datasets and RMSE on T-Drive, confirming that progressive cross-field interaction contributes predictive information beyond isolated field processing. Taken together, the results validate the separation–guidance–coupling design of TopoDuet and motivate the relational analysis that follows.

Table 3: Performance comparison on grid-based urban-flow datasets (MAE / RMSE; lower is better). Red and blue boldface indicate the best and second-best mean results.

	T-Drive Inflow		T-Drive Outflow		CHIBike Inflow		CHIBike Outflow	
Model	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
DCRNN	21.67	37.84	21.29	37.42	4.15	5.86	4.12	5.71
ASTGCN	18.34	33.21	18.36	33.31	3.98	**5.09**	3.89	5.49
STGCN	20.93	37.24	20.48	36.85	4.12	5.83	4.07	5.66
STGODE	21.07	37.45	22.23	39.47	4.07	5.80	4.04	5.59
STSGCN	23.34	40.37	23.78	41.41	4.18	5.82	4.18	5.77
MSTGCN	21.70	37.95	21.63	37.58	4.12	5.83	4.13	5.74
TFormer	18.43	33.78	18.48	34.51	3.98	5.76	3.95	5.52
PDFormer	17.46	30.97	**17.15**	**30.61**	**3.81**	**5.31**	**3.76**	**5.29**
STAEformer	17.28	**30.84**	17.36	30.99	**3.84**	5.38	**3.73**	**5.27**
RAST	17.67	31.22	17.82	31.33	3.90	5.65	3.84	5.61
MFormer	**17.16**	30.92	17.20	31.09	3.87	5.47	**3.76**	5.35
TopoDuet	**16.89** ±0.20	**30.75** ±0.21	**17.09** ±0.18	**30.97** ±0.05	3.92 ±0.02	5.53 ±0.03	3.89 ±0.02	5.51 ±0.04

	PeMS08				T-Drive Inflow			
	MAE	ΔMAE	RMSE	ΔRMSE	MAE	ΔMAE	RMSE	ΔRMSE
Full	**13.23**	—	22.73	—	**16.78**	—	**30.66**	—
w/o Transition Field	13.28	+0.05	23.06	+0.33	16.97	+0.19	31.18	+0.52
w/o Field Separation	13.33	+0.10	23.36	+0.62	17.06	+0.28	31.61	+0.95
w/o Road Prior	13.29	+0.06	23.02	+0.28	16.99	+0.21	31.14	+0.48
w/o Backward Support	13.30	+0.07	23.00	+0.26	17.01	+0.23	31.11	+0.45
w/o Field Adaptation	13.23	—	23.02	+0.28	16.95	+0.17	31.07	+0.41
w/o PCFC	13.33	+0.10	**22.61**	−0.12	17.04	+0.26	31.02	+0.36

Table 4: Ablation results on PeMS08 and T-Drive Inflow.

Relational and Mechanism Analysis

We examine whether the two fields learn distinct spatial relations and whether the adaptive components behave as intended. Comparing their attention matrices over the complete test set gives non-zero disagreement in every encoder block (0.0080 overall on PeMS08 and 0.0342 on T-Drive Inflow), showing that the fields attend to different node pairs; the larger T-Drive value also agrees with its larger field-separation ablation gap. Disagreement decreases but remains non-zero with depth, consistent with progressive coordination rather than relational collapse. In addition, the fields learn different road-prior strengths, both PCFC gates remain active, and the readout weights vary across horizons, respectively supporting field-adaptive topology use, bidirectional cross-field exchange, and horizon-adaptive prediction. Definitions, layer-wise results, and all visualizations are provided in Appendices A and B.

Conclusion

We introduced TopoDuet, a multi-relational framework for spatio-temporal forecasting that replaces a single latent relation with two complementary fields defined over a common physical topology. By assigning absolute observations to a Configuration Field and signed first- and second-order variations to a Transition Field, the model preserves distinct relational roles while allowing them to exchange evidence. Independent temporal–spatial encoding, field-adaptive direction-complete road guidance, progressive cross-field coupling, and horizon-adaptive dual-path readout together turn this principle into an end-to-end forecasting architecture. Across eight datasets spanning highway sensor graphs and urban-flow grids, TopoDuet achieves the best MAE and RMSE on all six graph-based benchmarks, improving over the strongest competing results by up to 4.9% and 7.3%, respectively, while transferring competitively to T-Drive and CHIBike without changing the core design. Controlled ablations and test-set relational analyses provide complementary evidence for the mechanism: collapsing the two fields causes the largest degradation, removing topology guidance or progressive coupling weakens accuracy, and the learned attention relations remain distinct across encoder depth even as their disagreement decreases. The structural analysis further explains these observations by quantifying the cost of relation collapse, establishing the minimal direction-complete support, and characterizing when a shared soft prior can or cannot erase field-specific relations, thereby motivating multi-relational forecasting.

Reproducibility statement

Appendices A and B provide the relational and learned-mechanism analyses. Appendix C gives the structural analysis of the dual-field design, and Appendix D specifies the implementation, optimization, checkpoint selection, and evaluation procedures.

AI use statement

Generative AI tools were used to assist with reasoning, language editing, and manuscript organization. The authors reviewed and verified all AI-assisted text, mathematical statements, citations, implementation details, and experimental claims, and remain responsible for the complete content of the submission.

References

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.

George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time Series Analysis: Forecasting and Control*. Wiley, 5 edition, 2015.

Ling Cai, Krzysztof Janowicz, Gengchen Mai, Bo Yan, and Rui Zhu. Traffic transformer: Capturing the continuity and periodicity of time series for traffic forecasting. *Transactions in GIS*, 24(3): 736–755, 2020. doi: 10.1111/tgis.12644.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014. doi: 10.3115/v1/D14-1179.

Jeongwhan Choi, Hwangyong Choi, Jeehyun Hwang, and Noseong Park. Graph neural controlled differential equations for traffic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6367–6374, 2022. doi: 10.1609/aaai.v36i6.20587.

Zheng Fang, Qingqing Long, Guojie Song, and Kunqing Xie. Spatial-temporal graph ODE networks for traffic flow forecasting. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 364–373, 2021. doi: 10.1145/3447548.3467430.

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. doi: 10.1609/aaai.v33i01.3301922.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Puqing Hu, Chunjiang Wu, Chen Wang, Xin Yang, Zhibin Li, Tinghui Chen, and Shijie Zhou. Memory-augmented spatio-temporal transformer for robust traffic flow forecasting. *Biomimetics*, 11(3):170, 2026. doi: 10.3390/biomimetics11030170.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

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.

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*, 2022.

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.

Hangchen Liu, Zheng Dong, Renhe Jiang, Jiewen Deng, Jinliang Deng, Quanjun Chen, and Xuan Song. Spatio-temporal adaptive embedding makes vanilla transformer SOTA for traffic forecasting. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 4125–4129, 2023. doi: 10.1145/3583780.3615160.

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.

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

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

Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 914–921, 2020. doi: 10.1609/aaai.v34i01.5438.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph WaveNet for deep spatial-temporal graph modeling. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 1907–1913, 2019. doi: 10.24963/ijcai.2019/264.

Yutong Xia, Yuxuan Liang, Haomin Wen, Xu Liu, Kun Wang, Zhengyang Zhou, and Roger Zimmermann. Deciphering spatio-temporal graph forecasting: A causal lens and treatment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 37068–37088, 2023. doi: 10.52202/075280-1611.

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

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. doi: 10.1609/aaai.v37i9.26317.

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. doi: 10.1609/aaai.v34i01.5477.

A Relational Field Analysis

To examine whether the two fields learn genuinely different spatial relations, we collect the pre-dropout attention distributions $\mathbf{P}_L^{(\ell)}$ and $\mathbf{P}_I^{(\ell)}$ over the complete test set and define

$$D_{\text{rel}}^{(\ell)} = \mathbb{E}\left[\frac{1}{2N}\left\|\mathbf{P}_L^{(\ell)} - \mathbf{P}_I^{(\ell)}\right\|_F^2\right]. \tag{25}$$

A non-zero value indicates that the configuration and transition fields retain distinct learned spatial relations.

Table 5: Normalized relational disagreement across encoder blocks.

Dataset	Block 1	Block 2	Block 3	Overall
PeMS08	0.0098	0.0088	0.0055	0.0080
T-Drive Inflow	0.0535	0.0398	0.0093	0.0342

Table 5 shows that the two fields remain relationally distinct throughout the encoder, directly supporting the dual-field separation introduced by DFSTE. T-Drive Inflow exhibits substantially stronger disagreement than PeMS08, which is consistent with its larger degradation when field separation is removed in Table 4. Meanwhile, the decreasing disagreement across depth suggests increasing coordination between the two fields while preserving non-zero relational differences, which is consistent with the progressive coupling behavior of PCFC.

Figure 2 further shows that the two fields share broad topology-related structure but differ in many node-pair dependencies. To make this difference spatially explicit, Figure 3 projects the final-layer PeMS08 relations onto a common layout derived from the road network. Although both fields receive the same physical support, their strongest learned relations are organized over different node pairs and network branches. The Configuration Field forms a more concentrated set of strong dependencies around several densely connected regions, whereas the Transition Field emphasizes a different combination of branches and cross-node interactions. The contrast view in Figure 3(c) further reveals spatially coherent configuration-dominant and transition-dominant regions, rather than uniformly distributed numerical differences. Together with the non-zero relational disagreement in Table 5, these results show that the common road prior preserves shared physical structure without forcing the two fields into one identical spatial relation. This directly supports FARA's design of sharing one road-derived candidate support while retaining field-specific content relations and prior

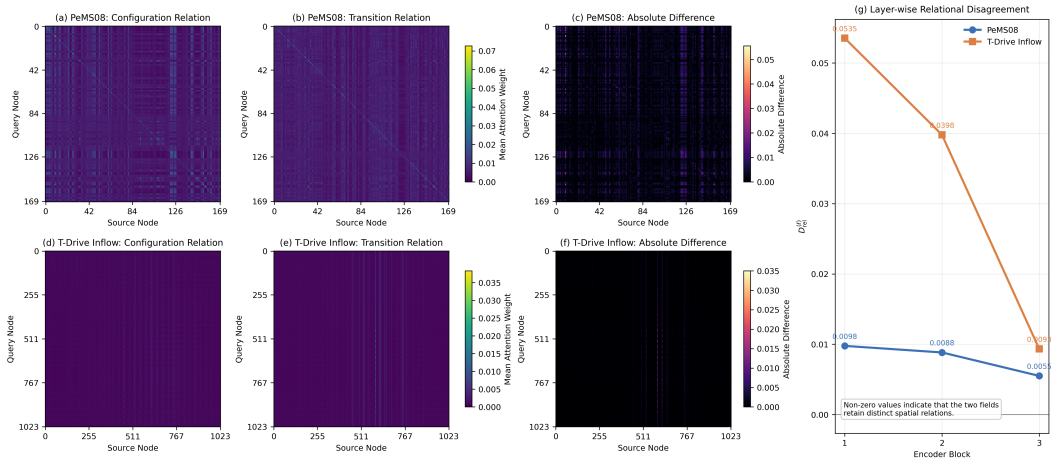

Figure 2: Learned configuration and transition relations on PeMS08 and T-Drive Inflow. (a)–(c) Configuration relation, transition relation, and their absolute difference on PeMS08. (d)–(f) Corresponding results on T-Drive Inflow. (g) Layer-wise normalized relational disagreement.

strengths. Because the two pathways originate from absolute observations and signed temporal variations, the separated spatial relations instantiate the intended configuration and transition semantics.

B Learned Mechanism Analysis

Figure 4 provides mechanism-level evidence for the three central design hypotheses of TopoDuet. First, the clearly different road-prior strengths learned by the Configuration and Transition Fields in Fig. 4(a) support the motivation of FARA and the shared-topology argument in our structural analysis: although both fields are defined on the same physical road support, they need not rely on that support to the same extent. This empirically supports the use of a common topology with field-specific prior strengths rather than either enforcing identical spatial relations or assigning different physical graphs to the two fields. Second, Fig. 4(b) shows that both PCFC gates remain active and increase above their initialization, supporting our solution to the cross-field coupling dilemma: the two relational fields benefit from progressive information exchange while retaining separate processing pathways instead of being either fully isolated or prematurely fused. Finally, Fig. 4(c) shows that the learned readout weights vary across forecast horizons, supporting the horizon-adaptive readout hypothesis that the relative importance of nonlinear correction and historical persistence should not be fixed for all prediction steps. Together with the relational disagreement analysis, these results support the central premise of TopoDuet that configuration and transition relations should share one

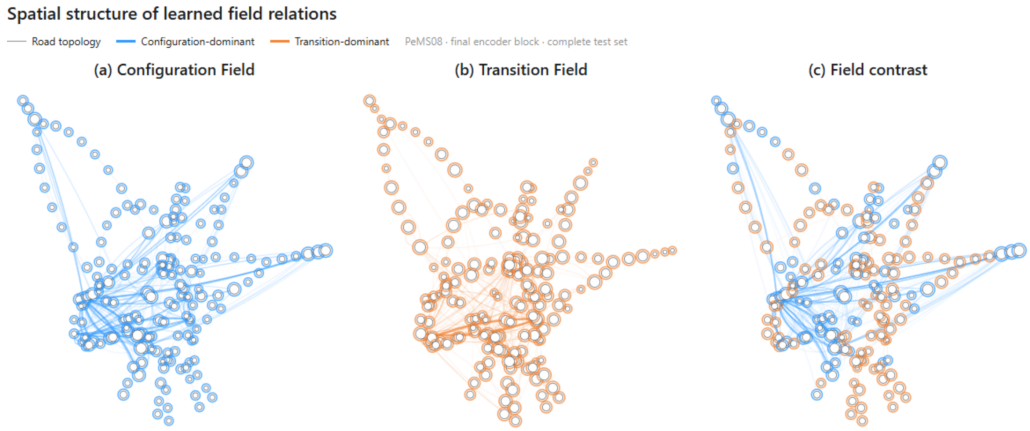

Figure 3: Topology-aware visualization of the final-layer relations on PeMS08. (a) Configuration Field, (b) Transition Field, and (c) field contrast. Gray edges denote road topology and colored curves show the 180 strongest off-diagonal relations. Node positions are non-geographic.

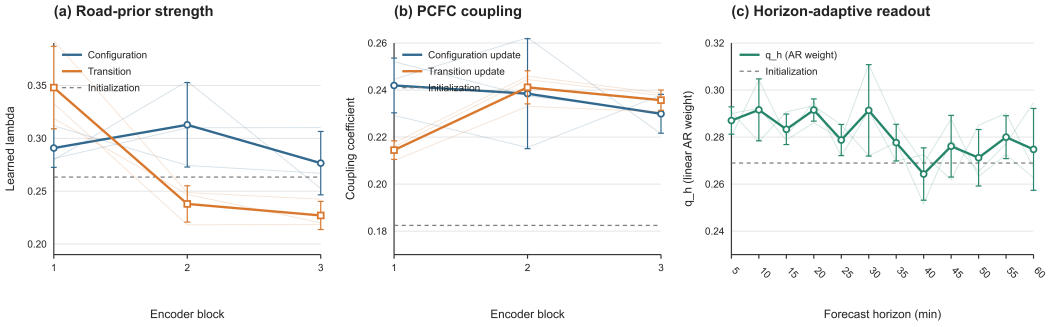

Figure 4: Learned adaptive mechanisms of TopoDuet on PeMS08. (a) Field- and layer-specific road-prior strengths. (b) Learned PCFC coupling coefficients. (c) Horizon-specific AR-path weights. Dashed lines indicate the corresponding initialization values, and error bars show variation across three independent runs.

physical topology while maintaining distinct relational behavior and interacting adaptively across encoder depth and forecast horizon.

C Structural Analysis of the Dual-Field Design

The following analysis concerns the spatial relations computed by DFSTE and FARA. It characterizes three properties of the preceding design: the representation distortion caused by collapsing two learned field relations into one, the minimal direction-complete extension of the directed road support, and the condition under which two fields receiving the same road prior produce identical attention relations. Together with the forecasting and ablation results, these properties explain the predictive role of preserving field-specific relations.

Condition 1: distinct learned field relations. Fix one input $(\mathcal{X}_{\mathrm{in}}, \mathcal{C}_{\mathrm{in}}; G)$, one encoder block $\ell \in \{0, \dots, M-1\}$, one observed step $\tau \in \{1, \dots, T\}$, and one attention head $r \in \{1, \dots, R\}$. For brevity, write

$$\mathbf{P}_L := \mathbf{P}^{(\ell)}_{L,\tau,r}, \qquad \mathbf{P}_I := \mathbf{P}^{(\ell)}_{I,\tau,r}, \tag{26}$$

where the two matrices are the pre-dropout spatial attention distributions defined in the preceding subsection. Both belong to

$$\mathcal{P}_N = \left\{ \mathbf{P} \in \mathbb{R}_+^{N \times N} : \mathbf{P}\mathbf{1}_N = \mathbf{1}_N \right\}, \tag{27}$$

the set of row-stochastic matrices. TopoDuet computes $\mathbf{P}_L$ and $\mathbf{P}_I$ through independently parameterized spatial-attention modules, while a shared-relation constraint would replace them by one common matrix $\mathbf{P} \in \mathcal{P}_N$.

Proposition 1 (Cost of collapsing the learned field relations). Suppose that the two field-specific attention distributions $\mathbf{P}_L$ and $\mathbf{P}_I$ are replaced by one shared relation $\mathbf{P} \in \mathcal{P}_N$. Define the resulting relation-sharing distortion as

$$\mathcal{D}_{\mathrm{share}}(\mathbf{P}) = \|\mathbf{P} - \mathbf{P}_L\|_F^2 + \|\mathbf{P} - \mathbf{P}_I\|_F^2. \tag{28}$$

Its unique minimizer and minimum value are

$$\mathbf{P}^\star = \frac{\mathbf{P}_L + \mathbf{P}_I}{2}, \tag{29}$$

$$\min_{\mathbf{P} \in \mathcal{P}_N} \mathcal{D}_{\mathrm{share}}(\mathbf{P}) = \frac{1}{2}\|\mathbf{P}_L - \mathbf{P}_I\|_F^2. \tag{30}$$

Consequently, collapsing the two learned relations causes zero distortion if and only if $\mathbf{P}_L = \mathbf{P}_I$.

Proof. Because $\mathcal{P}_N$ is convex, $(\mathbf{P}_L + \mathbf{P}_I)/2$ is nonnegative and row-stochastic, and therefore belongs to $\mathcal{P}_N$. For every $\mathbf{P} \in \mathcal{P}_N$,

$$\mathcal{D}_{\mathrm{share}}(\mathbf{P}) = 2\left\|\mathbf{P} - \frac{\mathbf{P}_L + \mathbf{P}_I}{2}\right\|_F^2 + \frac{1}{2}\|\mathbf{P}_L - \mathbf{P}_I\|_F^2. \tag{31}$$

The first term is nonnegative and vanishes only at $\mathbf{P}^\star = (\mathbf{P}_L + \mathbf{P}_I)/2$, proving the stated minimum and uniqueness. The minimum is zero exactly when $\mathbf{P}_L = \mathbf{P}_I$. □

Proposition 1 operates directly on the attention distributions produced by TopoDuet. The configuration and transition fields use separate content projections and separate prior strengths $\lambda_L^{(\ell)}$ and $\lambda_I^{(\ell)}$, enabling field-specific spatial relations. The diagnostic $\mathcal{D}_{\mathrm{share}}$ measures the relational structure removed when the trained pair is replaced by a single shared relation. Combined with the shared-relation ablation under the same forecasting objective, this quantity connects measured relational divergence to predictive advantage.

Condition 2: the physical prior requires direction-complete coverage. Define the off-diagonal forward support as

$$\mathcal{S}^{\rightarrow} = \left\{ (i,j) : i \neq j,\ A^{\rightarrow}_{ij} > 0 \right\}. \tag{32}$$

For any set of ordered node pairs $\mathcal{T} \subseteq V \times V$, define its reversal by

$$\mathrm{rev}(\mathcal{T}) = \{(j,i) : (i,j) \in \mathcal{T}\}. \tag{33}$$

We call $\mathcal{T}$ *direction-complete* when

$$\mathcal{T} = \mathrm{rev}(\mathcal{T}). \tag{34}$$

Thus, whenever an ordered pair is supported by the road prior, its reversed ordered pair is also available to that prior. This definition formalizes direction-complete candidate support for predictive associations in the directed network.

Proposition 2 (Minimal direction-complete closure). Among all direction-complete supports containing $\mathcal{S}^{\rightarrow}$, the unique smallest one under set inclusion is

$$\mathcal{S}^{\mathrm{sym}} = \mathcal{S}^{\rightarrow} \cup \mathrm{rev}(\mathcal{S}^{\rightarrow}). \tag{35}$$

Moreover, for any $\eta > 0$,

$$\left\{(i,j) : i \neq j,\ \overline{A}_{ij} > 0\right\} = \mathcal{S}^{\mathrm{sym}}. \tag{36}$$

Hence the reverse augmentation adds at least one new ordered pair if and only if $\mathcal{S}^{\rightarrow}$ is not direction-complete.

Proof. The union $\mathcal{S}^{\mathrm{sym}}$ contains $\mathcal{S}^{\rightarrow}$ and equals its own reversal. Now let $\mathcal{T}$ be any direction-complete support satisfying $\mathcal{S}^{\rightarrow} \subseteq \mathcal{T}$. Reversing both sets gives

$$\mathrm{rev}(\mathcal{S}^{\rightarrow}) \subseteq \mathrm{rev}(\mathcal{T}) = \mathcal{T}. \tag{37}$$

Therefore $\mathcal{S}^{\mathrm{sym}} \subseteq \mathcal{T}$, establishing minimality and uniqueness.

For the reverse-augmented affinity, let

$$m_A = \max_{u,v} A^{\leftrightarrow}_{uv}. \tag{38}$$

The diagonal is reset to one before normalization, so $m_A \geq 1$ and the normalization denominator is strictly positive. For every $i \neq j$,

$$\overline{A}_{ij} = \frac{A^{\rightarrow}_{ij} + \eta A^{\rightarrow}_{ji}}{m_A}. \tag{39}$$

Because $m_A > 0$, $\eta > 0$, and both affinity terms are nonnegative,

$$\overline{A}_{ij} > 0 \iff A^{\rightarrow}_{ij} + \eta A^{\rightarrow}_{ji} > 0 \tag{40}$$

$$\iff A^{\rightarrow}_{ij} > 0 \text{ or } A^{\rightarrow}_{ji} > 0. \tag{41}$$

The off-diagonal support of $\overline{A}$ is therefore exactly $\mathcal{S}^{\mathrm{sym}}$. This union strictly enlarges $\mathcal{S}^{\rightarrow}$ precisely when $\mathcal{S}^{\rightarrow}$ is not direction-complete. □

The proposition identifies the exact support contributed by reverse augmentation: it supplies the unique minimal direction-complete closure while allowing forward and reverse affinities to retain distinct magnitudes.

Corollary 1 (Why the closure is supplied to both fields). Suppose that the configuration and transition fields are required to (i) retain $\mathcal{S}^{\rightarrow}$, (ii) use direction-complete road-prior support, and (iii) receive the same physical candidate support because both fields are defined on the same road network. Then their unique jointly minimal support assignment is

$$\mathcal{S}_L = \mathcal{S}_I = \mathcal{S}^{\mathrm{sym}}. \tag{42}$$

Proof. By Proposition 2, the support assigned to each field under conditions (i)–(ii) must contain $\mathcal{S}^{\mathrm{sym}}$. Condition (iii) requires the two assignments to be equal. Their common smallest possible assignment is therefore $\mathcal{S}^{\mathrm{sym}}$. Conversely, assigning $\mathcal{S}^{\mathrm{sym}}$ to both fields retains the forward support, is direction-complete, and supplies the same physical candidate support to both fields. □

This corollary explains the shared-support construction in TopoDuet: physical candidate pairs remain fixed across the two fields, while their content relations and learned prior strengths remain field-specific.

Condition 3: shared support preserves field-specific relations. Fix the same encoder block, observed step, attention head, and query node i. To simplify notation, suppress the fixed indices (ℓ, τ, r) and let

$$\psi_{f,ij} = \frac{q_{f,i}^\top k_{f,j}}{\sqrt{d_h}}, \qquad f \in \{L, I\}, \tag{43}$$

denote the field-specific content logit. Also write $\lambda_f := \lambda_f^{(\ell)}$. The corresponding attention probability is

$$p_{f,ij} = \frac{\exp\big(\psi_{f,ij} + \lambda_f \overline{A}_{ij}\big)}{\sum_{m=1}^{N} \exp\big(\psi_{f,im} + \lambda_f \overline{A}_{im}\big)}. \tag{44}$$

For any two source nodes j and k,

$$\log \frac{p_{f,ij}}{p_{f,ik}} = \psi_{f,ij} - \psi_{f,ik} + \lambda_f \left(\overline{A}_{ij} - \overline{A}_{ik}\right). \tag{45}$$

The first difference represents field-specific content preference, whereas the second represents the additional preference supplied by the road prior. For example, when $\overline{A}_{ij} > \overline{A}_{ik}$, the content logits override the road preference and rank k above j exactly when

$$\psi_{f,ik} - \psi_{f,ij} > \lambda_f \left(\overline{A}_{ij} - \overline{A}_{ik}\right). \tag{46}$$

Proposition 3 (Exact equality condition under a shared soft prior). Suppose that both fields receive the same finite road affinity $\overline{A}$. Their attention rows for query node i satisfy

$$\mathbf{p}_{L,i} = \mathbf{p}_{I,i} \quad \Longleftrightarrow \quad \exists\, \kappa_i \in \mathbb{R} \text{ such that } \psi_{L,ij} - \psi_{I,ij} + (\lambda_L - \lambda_I)\overline{A}_{ij} = \kappa_i \quad \text{for every } j. \tag{47}$$

Therefore, shared road-prior support preserves the capacity of the two fields to learn distinct spatial relations.

Proof. Define the total logits and their row-dependent normalization factors as

$$s_{f,ij} = \psi_{f,ij} + \lambda_f \overline{A}_{ij}, \qquad Z_{f,i} = \sum_{m=1}^{N} \exp(s_{f,im}). \tag{48}$$

If $\mathbf{p}_{L,i} = \mathbf{p}_{I,i}$, then for every j,

$$s_{L,ij} - \log Z_{L,i} = s_{I,ij} - \log Z_{I,i}. \tag{49}$$

Hence

$$s_{L,ij} - s_{I,ij} = \log Z_{L,i} - \log Z_{I,i} =: \kappa_i, \tag{50}$$

which is constant across all source nodes in row i. Substituting the definition of $s_{f,ij}$ gives the stated condition.

Conversely, if the stated condition holds, then $s_{L,ij} = s_{I,ij} + \kappa_i$ for every j. Row-wise softmax is invariant to adding the same constant to every logit in one row, so $\mathbf{p}_{L,i} = \mathbf{p}_{I,i}$. □

The same direction-complete prior can therefore be supplied to both fields while preserving their capacity to learn distinct spatial relations. Field-specific content projections and prior strengths produce different attention whenever the row-constant equality condition is not satisfied, and the equality characterization identifies precisely the exceptional case in which the two relations coincide. Combined with the shared-relation ablation and forecasting results, this establishes the predictive value of maintaining field-specific relational structure.

D Implementation Details

Data preparation. Each dataset is stored as $\mathcal{D} \in \mathbb{R}^{S \times N \times 3}$, whose channels are the scalar traffic observation, normalized time of day, and day-of-week index. For every valid start index a, we construct

$$X = \mathcal{D}_{a:a+T,:,:}, \qquad Y = \mathcal{D}_{a+T:a+T+H,:,0:1}. \tag{51}$$

All valid windows are generated first and then divided chronologically into training, validation, and test samples. Highway datasets use $T = H = 12$ and a $60/20/20$ split. T-Drive and CHIBike use

$T = 6$, $H = 1$, and a $70/10/20$ split. Training samples are shuffled after splitting; validation and test samples are not shuffled. Inflow and outflow are trained as separate scalar tasks.

Let x^{raw} denote the traffic observation stored in the first channel of $\mathcal{D}$. A single mean μ and standard deviation σ are computed over the traffic channel of all training input windows, including all samples, times, and nodes. For consistency with the notation used in the main text, x denotes the standardized traffic input supplied to the model:

$$x = \frac{x^{\mathrm{raw}} - \mu}{\sigma}. \tag{52}$$

Calendar channels and targets are left unchanged. The same training statistics are applied to validation and test inputs. Model outputs are inverse-transformed to the original traffic scale before computing the training loss or any evaluation metric.

Road prior. The road CSV contains directed `source`, `target`, and `length` entries. Duplicate ordered pairs retain their first entry. An edge $j \to i$ is stored at (i, j) with cost $\log(1+\max(\ell_{j\to i}, 0))$. One min-plus composition adds the best two-hop path. If s is the median positive finite cost, the forward affinity is $A^{\rightarrow}_{ij} = \exp(-C_{ij}/s)$ for reachable pairs and zero otherwise. We then use

$$A' = A^{\rightarrow} + 0.35(A^{\rightarrow})^{\top}, \tag{53}$$

reset its diagonal to one, and divide it by its global maximum. The resulting matrix A is fixed. Each field and encoder layer learns an independent prior strength $\lambda_f^{(m)} = \mathrm{softplus}(\rho_f^{(m)})$, where $\rho_f^{(m)}$ is initialized to -1.2, and spatial attention uses

$$\mathrm{softmax}\left(\frac{QK^{\top}}{\sqrt{d_h}} + \lambda_f^{(m)} A\right). \tag{54}$$

Model. The Configuration Field projects the standardized absolute observation x with $\mathrm{Linear}(1, 128)$. The Transition Field projects $[\Delta x; \Delta^2 x]$ with $\mathrm{Linear}(2, 128)$, where

$$\Delta x_1 = 0, \qquad \Delta x_t = x_t - x_{t-1} \quad (t \geq 2), \tag{55}$$

$$\Delta^2 x_1 = \Delta^2 x_2 = 0, \qquad \Delta^2 x_t = \Delta x_t - \Delta x_{t-1} \quad (t \geq 3). \tag{56}$$

Both fields receive time-of-day, day-of-week, shared node, and field-specific adaptive position embeddings. The transition calendar term has a fixed coefficient of 0.5. Shared node embeddings and adaptive embeddings are initialized with standard deviations 0.035 and 0.025, respectively. No precomputed periodic statistics are used.

The encoder contains three dual-field blocks with hidden size 128, four attention heads, and head size 32. Each field applies temporal attention, spatial attention, and a $128 \to 384 \to 128$ GELU feed-forward network using pre-normalization, residual connections, and dropout 0.12. After each block, PCFC concatenates the two fields and uses two independent $256 \to 128 \to 128$ GELU MLPs to exchange information. Its two scalar sigmoid gates are initialized with logits -1.5.

The readout concatenates the final fields and applies a $256 \to 128 \to 128$ fusion MLP. A nonlinear residual path projects the flattened history to all horizons and adds it to the last standardized input using a positive scale initialized as $\mathrm{softplus}(-0.7)$. A parallel linear layer maps the standardized scalar history from T to H. Horizon-specific sigmoid gates, initialized with logits -1.0 and shared across nodes, combine the two paths.

Training and checkpoint selection. Predictions are first returned to the original traffic scale. The implemented loss is

$$\mathcal{L} = 0.9\ \mathrm{MAE} + 0.1\ \mathrm{SmoothL1}_{\beta}. \tag{57}$$

For highway tasks, both terms use positions with $Y > 0$ and $\beta = 8$. For grid tasks, all targets are used during training and $\beta = 2$.

We use AdamW with learning rate 5×10^{-4}, weight decay 10^{-3}, and $\epsilon = 10^{-8}$. The first five epochs linearly warm the learning rate from 0.15 of its base value to the base value; cosine annealing then reduces it to 2×10^{-6}. The global gradient norm is clipped at 3.0. An exponential moving average with decay 0.999 is updated after every optimizer step. Validation uses EMA weights after every

Table 6: Dataset-dependent runtime settings.

Dataset	03	04	07	08	D7(M)	D7(L)	T-Drive	CHIBike
Batch size	8	8	4	16	16	2	2	16
Slots/day	288	288	288	288	288	288	24	48

Table 7: Common TopoDuet hyperparameters used in the main experiments.

Category	Hyperparameter	Value
Input	Highway input/output length	$12/12$
Input	Grid input/output length	$6/1$
Model	Hidden dim. / blocks	$128/3$
Model	Heads / head dim.	$4/32$
Model	FFN dim. / activation	384/GELU
Model	Dropout	0.12
Model	Transition calendar coeff.	0.5
Road	Reverse-support coeff.	0.35
Init.	Shared/adaptive emb. std.	$0.035/0.025$
Init.	PCFC / road raw values	$-1.5/-1.2$
Init.	Residual / horizon-gate raw	$-0.7/-1.0$
Optim.	Optimizer	AdamW
Optim.	Base/min. learning rate	$5\times10^{-4}/2\times10^{-6}$
Optim.	Weight decay / ϵ	$10^{-3}/10^{-8}$
Optim.	Warm-up / schedule	5 epochs / cosine
Optim.	Gradient-norm limit	3.0
Optim.	EMA decay	0.999
Select.	Patience / max epochs	$20/120$; P08: 150
Loss	Highway	.9 MAE $+.1$ SmoothL1
Loss	Grid	.9 MAE $+.1$ SmoothL1
Runtime	Workers / runs	$0/3$ seeds

epoch, and the EMA state with the lowest validation MAE is saved. Training stops after 20 epochs without improvement or at 120 epochs; the PeMS08 runs allowed at most 150 epochs.

Python, NumPy, and PyTorch CPU/CUDA generators are seeded. We enable deterministic cuDNN, disable cuDNN benchmarking, use zero DataLoader workers, and seed the training-shuffle generator with `seed+100000`. Each task is trained from three independent initializations. Experiments use Python 3.10.8, PyTorch 2.1.2, CUDA 11.8, cuDNN 8.7, and one 24-GB RTX 4090.

The common model and optimization settings are summarized in Table 7.

Testing. The selected EMA checkpoint is evaluated on the complete test set. Highway metrics use $Y > 0$; T-Drive and CHIBike metrics use $Y \geq 10$ and $Y \geq 5$, respectively. MAE and RMSE are accumulated over all valid test elements after inverse standardization. Results are reported as mean and standard deviation across three seeds.