Sparse-Aware Spatio-Temporal Demand Forecasting with Snapshot-Based Global Attention

1 Input Representation

1.1 Static Region Features

Each region corresponds to an irregular spatial area. Static features describe functional and geographic characteristics and remain constant over time.

Land-use composition. Let C_{land} denote the number of land-use categories.

$$\boldsymbol{x}_r^{\text{land}} \in \mathbb{R}^{C_{\text{land}}}, \qquad \sum_{k=1}^{C_{\text{land}}} x_{r,k}^{\text{land}} = 1.$$

Example:

$$\boldsymbol{x}_r^{\text{land}} = [0.6\ (\text{res}),\ 0.3\ (\text{comm}),\ 0.1\ (\text{office}),\ 0,\ 0].$$

POI statistics with learnable importance. Let C_{POI} denote the number of POI categories. We use raw POI counts without explicit area normalization, since region area is separately encoded as a geographic attribute.

$$\boldsymbol{x}_r^{\text{POI}} \in \mathbb{R}_{\geq 0}^{C_{\text{POI}}}.$$

We decompose POI information into presence and magnitude components:

$$z_{r,c} \triangleq \mathbb{I}(x_{r,c}^{\text{POI}} > 0), \qquad s_{r,c} \triangleq \log(1 + x_{r,c}^{\text{POI}}), \quad c = 1, \ldots, C_{\text{POI}}.$$

To account for heterogeneous impact across POI categories, we introduce learnable non-negative importance weights:

$$w_c^{(z)} = \text{softplus}(\theta_c^{(z)}), \qquad w_c^{(s)} = \text{softplus}(\theta_c^{(s)}).$$

The final POI feature is defined as:

$$\tilde{x}_{r,c}^{\text{POI}} = w_c^{(z)} z_{r,c} + w_c^{(s)} s_{r,c}.$$

$$\tilde{\boldsymbol{x}}_r^{\text{POI}} = [\tilde{x}_{r,1}^{\text{POI}}, \ldots, \tilde{x}_{r,C_{\text{POI}}}^{\text{POI}}].$$

Geographic attributes. Geographic features encode region location and size:

$$\boldsymbol{x}_r^{\text{geo}} = [\text{lat}_r,\ \text{lon}_r,\ \log(1 + \text{area}_r)].$$

Static encoder.

$$\boldsymbol{u}_r^{\text{stat}} = f_{\text{stat}}(\boldsymbol{x}_r^{\text{land}} \| \tilde{\boldsymbol{x}}_r^{\text{POI}} \| \boldsymbol{x}_r^{\text{geo}}), \qquad \boldsymbol{u}_r^{\text{stat}} \in \mathbb{R}^{d_s},\ d_s = 32.$$

1.2 Temporal Context Embedding

We model calendar periodicity using learnable embedding tables.

Day-of-week embedding. Let $\text{dow}(t) \in \{1, \ldots, 7\}$ denote the day of week at time t. We define a learnable embedding matrix:

$$\boldsymbol{E}_{\text{dow}} \in \mathbb{R}^{7 \times d_t}, \qquad d_t = 16.$$

The day-of-week embedding is obtained by lookup:

$$\boldsymbol{u}_t^{\text{dow}} = \boldsymbol{E}_{\text{dow}}[\text{dow}(t)].$$

Hour-of-day embedding. Similarly, let $\text{hod}(t) \in \{0, \ldots, 23\}$. We define:

$$\boldsymbol{E}_{\text{hod}} \in \mathbb{R}^{24 \times d_t},$$

$$\boldsymbol{u}_t^{\text{hod}} = \boldsymbol{E}_{\text{hod}}[\text{hod}(t)].$$

Holiday embedding. We introduce a learnable holiday embedding vector:

$$\boldsymbol{e}^{\text{hol}} \in \mathbb{R}^{d_t}.$$

Temporal context representation. The final temporal context embedding is defined as:

$$\boldsymbol{u}_t^{\text{time}} = \boldsymbol{u}_t^{\text{dow}} + \boldsymbol{u}_t^{\text{hod}} + \mathbb{I}(\text{holiday}(t)) \cdot \boldsymbol{e}^{\text{hol}}.$$

1.3 Dynamic Demand Features (Local Lag Encoding)

We distinguish local lag encoding from the recurrent temporal horizon.

Lag window. Let ℓ denote the local lag window length (e.g., $\ell = 4$):

$$\boldsymbol{y}_{r,t}^{(\ell)} = [y_{r,t-\ell+1}, \ldots, y_{r,t}].$$

Boundary handling. For $t - k < 0$, we apply zero-padding and introduce missing indicators:

$$y_{r,t-k} = 0, \qquad m_{r,t,k} = \mathbb{I}(t - k \geq 0).$$

Sparsity descriptors.

$$c_{r,t} = \sum_{i=0}^{\ell-1} \mathbb{I}(y_{r,t-i} > 0),$$

$$\Delta t_{r,t}^{\text{last}} = \min\left(t - \max\{\tau \le t : y_{r,\tau} > 0\},\ \Delta_{\max}\right).$$

The descriptor $\Delta t_{r,t}^{\text{last}}$ captures the global recency of demand events and is conceptually independent of the local lag window length ℓ. While local lag features encode short-term demand dynamics, $\Delta t_{r,t}^{\text{last}}$ summarizes how long the region has remained inactive since its most recent non-zero demand.

Importantly, $\Delta t_{r,t}^{\text{last}}$ is computed with respect to the *global timeline* of the dataset rather than the recurrent sequence boundary. Therefore, the last demand event may lie outside the current RNN input window. To ensure numerical stability and to reflect diminishing temporal relevance, we cap the value using a fixed threshold $\Delta_{\max}$. In practice, $\Delta_{\max}$ is chosen according to dominant temporal cycles (e.g., 24 or 48 hours for daily patterns).

Dynamic encoder.

$$\boldsymbol{u}_{r,t}^{\text{dyn}} = f_{\text{dyn}}([\boldsymbol{y}_{r,t}^{(\ell)} \| \boldsymbol{m}_{r,t}^{(\ell)} \| c_{r,t} \| \Delta t_{r,t}^{\text{last}}]), \qquad \boldsymbol{u}_{r,t}^{\text{dyn}} \in \mathbb{R}^{d_d},\ d_d = 32.$$

2 Initial Region-Time Embedding

$$\boldsymbol{e}_{r,t} = \phi\left(\boldsymbol{W}_e[\boldsymbol{u}_r^{\text{stat}} \| \boldsymbol{u}_{r,t}^{\text{dyn}} \| \boldsymbol{u}_t^{\text{time}}] + \boldsymbol{b}_e\right), \qquad \boldsymbol{e}_{r,t} \in \mathbb{R}^{64}.$$

3 Temporal State Update

We maintain a region-specific recurrent state:

$$\boldsymbol{h}_{r,t} = \text{GRU}(\boldsymbol{h}_{r,t-1}, \boldsymbol{e}_{r,t}), \qquad \boldsymbol{h}_{r,0} = \boldsymbol{0}.$$

To combine instantaneous and history-aware representations, we apply gated fusion:

$$\boldsymbol{g}_{r,t} = \sigma\left(\boldsymbol{W}_g[\boldsymbol{e}_{r,t} \| \boldsymbol{h}_{r,t}] + \boldsymbol{b}_g\right),$$

$$\boldsymbol{s}_{r,t} = \boldsymbol{g}_{r,t} \odot \boldsymbol{h}_{r,t} + (1 - \boldsymbol{g}_{r,t}) \odot \boldsymbol{e}_{r,t}.$$

4 Spatial Modeling via Snapshot-Based Global Attention

At each time step, we conceptually induce a time-varying region interaction graph. Instead of explicitly constructing such graphs, we employ global self-attention as a differentiable mechanism to capture dynamic spatial dependencies.

Attention projections. With $H = 4$ heads and $d_h = 16$:

$$\boldsymbol{q}_{r,t}^{(h)} = \boldsymbol{W}_q^{(h)} \boldsymbol{s}_{r,t}, \quad \boldsymbol{k}_{j,t}^{(h)} = \boldsymbol{W}_k^{(h)} \boldsymbol{s}_{j,t}, \quad \boldsymbol{v}_{j,t}^{(h)} = \boldsymbol{W}_v^{(h)} \boldsymbol{s}_{j,t}.$$

Soft OD bias (robust normalization). Let

$$x_{rj,t} \triangleq \log(1 + OD_{r\to j,t}).$$

We define:

$$\mu_{\tilde{OD}}(t) = \text{mean}\{x_{rj,t}\}_{r,j}, \qquad \sigma_{\tilde{OD}}(t) = \text{std}\{x_{rj,t}\}_{r,j}.$$

$$\tilde{OD}_{rj,t} = \frac{x_{rj,t} - \mu_{\tilde{OD}}(t)}{\sigma_{\tilde{OD}}(t) + \epsilon}.$$

$$\lambda_{OD} = \text{softplus}(\theta_{OD}), \qquad g_{OD}(r,j,t) = \lambda_{OD}\tilde{OD}_{rj,t}.$$

Attention score.

$$s^{(h)}_{rj,t} = \frac{(\boldsymbol{q}^{(h)}_{r,t})^\top \boldsymbol{k}^{(h)}_{j,t}}{\sqrt{d_h}} + g_{OD}(r,j,t).$$

Aggregation.

$$\alpha^{(h)}_{rj,t} = \frac{\exp(s^{(h)}_{rj,t})}{\sum_m \exp(s^{(h)}_{rm,t})},$$

$$\tilde{\boldsymbol{s}}^{(h)}_{r,t} = \sum_j \alpha^{(h)}_{rj,t} \boldsymbol{v}^{(h)}_{j,t}.$$

$$\tilde{\boldsymbol{s}}_{r,t} = \boldsymbol{W}_o[\tilde{\boldsymbol{s}}^{(1)}_{r,t} \| \cdots \| \tilde{\boldsymbol{s}}^{(H)}_{r,t}].$$

Residual update.

$$\bar{\boldsymbol{s}}_{r,t} = \text{LayerNorm}(\boldsymbol{s}_{r,t} + \tilde{\boldsymbol{s}}_{r,t}), \qquad \epsilon = 10^{-5}.$$

5 Temporal Aggregation and Output

We construct training samples using a sliding temporal window of fixed length T (e.g., $T = 24$ hours). Specifically, for each time index t, the model consumes observations from $\{t-T+1, \ldots, t\}$ and predicts demand at the next step $t+1$. Consecutive samples are generated with a stride (up to 6 hours), resulting in overlapping sequences.

Given a temporal window $\mathcal{T}_t = \{t - T + 1, \ldots, t\}$:

$$\boldsymbol{h}_{r,\tau} = \text{GRU}(\boldsymbol{h}_{r,\tau-1}, \bar{\boldsymbol{s}}_{r,\tau}).$$

Temporal attention.

$$\beta_{r,\tau} = \frac{\exp(\boldsymbol{v}_a^\top \tanh(\boldsymbol{W}_a \boldsymbol{h}_{r,\tau}))}{\sum_{k\in\mathcal{T}_t} \exp(\boldsymbol{v}_a^\top \tanh(\boldsymbol{W}_a \boldsymbol{h}_{r,k}))},$$

$$\bar{\boldsymbol{h}}_r = \sum_{\tau\in\mathcal{T}_t} \beta_{r,\tau} \boldsymbol{h}_{r,\tau}.$$

6 Sparse-Aware Output Heads

Event probability.

$$p_{r,t+1} = \sigma(\boldsymbol{w}_p^\top \bar{\boldsymbol{h}}_r + b_p).$$

Conditional magnitude.

$$\hat{y}^+_{r,t+1} = \mathrm{softplus}(\boldsymbol{w}_+^\top \bar{\boldsymbol{h}}_r + b_+).$$

Final prediction.

$$\hat{y}_{r,t+1} = p_{r,t+1} \cdot \hat{y}^+_{r,t+1}.$$

7 Loss Function

Ground-truth event indicator.

$$\delta_{r,t+1} \triangleq \mathbb{I}(y_{r,t+1} > 0).$$

Event loss.

$$\mathcal{L}_{\mathrm{evt}} = -\sum_{r,t} \left[\delta_{r,t+1} \log p_{r,t+1} + (1 - \delta_{r,t+1}) \log(1 - p_{r,t+1})\right].$$

Magnitude loss.

$$\mathcal{L}_{\mathrm{mag}} = \sum_{r,t} \delta_{r,t+1} \left(\frac{|\hat{y}^+_{r,t+1} - y_{r,t+1}|}{1 + y_{r,t+1}} + \lambda |\hat{y}^+_{r,t+1} - y_{r,t+1}| \right).$$

OD scale regularization.

$$\mathcal{L} \leftarrow \mathcal{L} + \eta \lambda_{OD}^2, \qquad \eta = 10^{-3}.$$