FORECASTING AS CONDITIONAL TRANSPORT: DOOB-REVERSIBLE GIBBS DYNAMICS ON SPATIO-TEMPORAL GRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Observed variation in a spatio-temporal graph mixes history-predictable evolution with conditional surprise. Their distinction depends on the same available history, motivating coordinates that expose both the learned predictable reference and its residual. We frame forecasting around the construction and transport of predictive coordinates: represent variation relative to available history, propagate the resulting evidence through graph-informed interactions, and retain access to it across forecast horizons. DR-GIBBSFORMER realizes this view with a causal multi-offset Student-t projector, dual-stream temporal–spatial Gibbs attention, triangular exchange, and horizon-specific full-history functionals. The analysis characterizes coordinate bias and residual-score behavior, content–graph factorization, operator sensitivity, recoverability under exchange, and linear readout expressivity. Across ten highway and urban-flow tasks, the framework improves 18 of 20 aggregate MAE/RMSE comparisons, with gains reaching 8.26% in MAE and 7.61% in RMSE. Component and projector ablations support the value of learned conditional coordinates, structural interaction, and access to the terminal history. These findings connect the choice of predictive coordinates to the way historical evidence is exchanged and used for forecasting.

1 INTRODUCTION

The predictive meaning of a change depends on its history: the same traffic increase can reflect recurrence, spatial propagation, or an unexpected deviation. These sources can require different forecasting responses, even when the observed increments have similar magnitudes. A learned reference conditioned on preceding observations makes this distinction explicit. We therefore ask which history-conditioned coordinates should be constructed before their information is exchanged across space and time.

Diffusion networks (Li et al., 2018), graph convolutions (Yu et al., 2018; Wu et al., 2019), and Transformers (Jiang et al., 2023; Liu et al., 2023) provide powerful interactions across nodes and timestamps. Their success motivates a complementary question about the evidence entering those interactions: how should observed levels and deviations from historical expectations be represented together? Our coordinate-centered view connects the evidence constructed from history to its structured interaction and horizon-specific use.

Three linked challenges follow. **C1: construct conditional coordinates** by measuring change relative to historical predictability. **C2: transport their information** through global temporal–spatial interaction informed by the directed graph. **C3: retain access through prediction** so that interacting level and innovation evidence remains available to distinct horizon-specific functionals.

DR-GIBBSFORMER implements this sequence through conditional coordinate construction, graph-informed Gibbs transport, recoverable cross-stream exchange, and horizon-specific reading of the terminal history (Figure 1). Here transport is the row-normalized aggregation of latent values; recoverability concerns the joint state during triangular exchange. This organization gives the theoretical analysis and targeted ablations a shared focus. Our contributions are:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

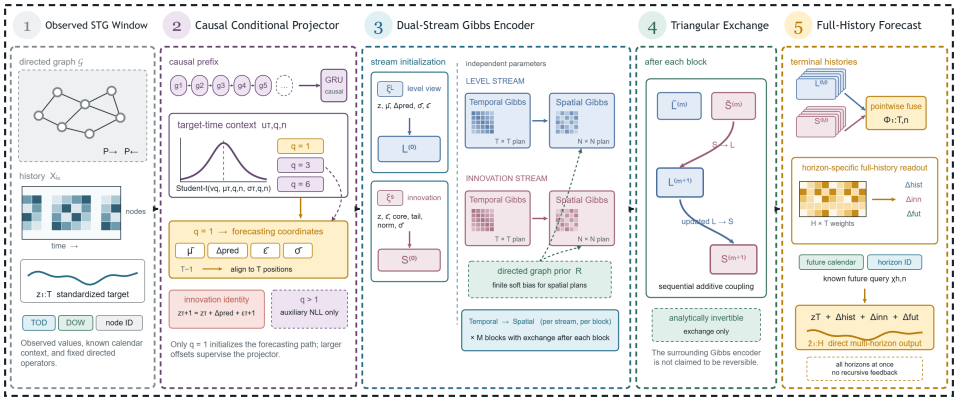


Figure 1: Architecture of DR-GIBBSFORMER.

1. We construct history-conditioned level and innovation-oriented coordinates, relating residual bias to conditional-mean error and Student-t estimation to bounded residual scores.
2. We combine content–graph Gibbs interaction with bijective cross-stream exchange, characterizing structural weighting, sensitivity, and joint-state recoverability.
3. We give each horizon access to the complete encoded history, characterize the shared linear readout bottleneck, and test the construction–interaction–readout design across ten forecasting tasks and targeted ablations.

2 RELATED WORK

Spatio-temporal graph forecasting. DCRNN (Li et al., 2018) and STGCN (Yu et al., 2018) combine temporal modeling with graph propagation. Graph WaveNet (Wu et al., 2019) and AGCRN (Bai et al., 2020) learn graph structure, while STGODE (Fang et al., 2021) and STSGCN (Song et al., 2020) use continuous-time and synchronous interactions. PDFormer (Jiang et al., 2023) and STAEformer (Liu et al., 2023) extend attention-based forecasting. Our focus is the conditional coordinate system preceding these interactions.

Innovations and conditional forecasting. Doob decomposition separates predictable and martingale components (Doob, 1953); state-space innovations similarly depend on the available history (Durbin & Koopman, 2012). We characterize learned residuals relative to conditional means while using Student-t likelihood for its robust residual-score behavior (Lange et al., 1989).

Gibbs attention and reversible coupling. Entropy-regularized transport yields Gibbs plans (Cuturi, 2013; Peyré & Cuturi, 2019), with relaxed marginals in unbalanced formulations (Chizat et al., 2018). Our one-marginal formulation is scaled dot-product attention (Vaswani et al., 2017) with an explicit structural prior. Triangular coupling follows reversible-network constructions (Dinh et al., 2017; Gomez et al., 2017), applied only to cross-stream exchange.

3 METHOD

DR-GIBBSFORMER constructs conditional coordinates, transports their latent values through Gibbs plans, exchanges the streams recoverably, and reads the terminal history through horizon-specific functionals. The propositions characterize these stages, with proofs in Appendix B. LN denotes layer normalization; Appendix A specifies implementation conventions.

3.1 PROBLEM FORMULATION

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \omega)$ be a directed graph with N nodes and nonnegative edge costs. At time s, $\mathbf{x}_s \in \mathbb{R}^N$ contains node targets and $\mathbf{c}_s \in \mathbb{R}^{d_c}$ is known context. Given T observations $\mathbf{X}_{\text{in}} = [\mathbf{x}_{t-T}^\top; \ldots; \mathbf{x}_{t-1}^\top]$

and contexts $\mathbf{C}_{\text{in}}$ over the same times, predict $\mathbf{X}_{\text{out}} = [\mathbf{x}_t^\top; \dots; \mathbf{x}_{t+H-1}^\top]$ by

$$\widehat{\mathbf{X}}_{\text{out}} = \mathcal{F}_\Theta(\mathbf{X}_{\text{in}}, \mathbf{C}_{\text{in}}; \mathcal{G}) \in \mathbb{R}^{H \times N}. \tag{1}$$

Training-set standardization maps x to internal coordinates z; its inverse restores forecast units.

3.2 Conditional Coordinate Construction

Reindex $z_{\tau,n} = \text{Scale}(x_{n,t-T+\tau-1})$ for $\tau = 1, \dots, T$. Let $\mathbf{z}_\tau$ collect all nodes and $e_\tau^{\text{cal}} = \phi_c(\mathbf{c}_{t-T+\tau-1})$; future calendar embeddings share ϕ_c. Batch indices are suppressed.

Directed graph operators. We construct fixed forward and backward row-stochastic operators $P^{\rightarrow}$ and $P^{\leftarrow}$ from directed one- and two-hop path affinities. Both the conditional projector and the spatial encoder use these operators; Appendix A.1 gives the graph construction and stabilization rules.

Graph-conditioned history input. At node n, the projector represents each available input position by its level, two directed graph aggregates, and backward increment: Define $\nabla^- z_{1,n} = 0$ and $\nabla^- z_{\tau,n} = z_{\tau,n} - z_{\tau-1,n}$ for $\tau > 1$. The projector input is

$$\mathbf{g}_{\tau,n} = [z_{\tau,n}, (P^{\rightarrow}\mathbf{z}_\tau)_n, (P^{\leftarrow}\mathbf{z}_\tau)_n, \nabla^- z_{\tau,n}]. \tag{2}$$

For each offset $q \in \mathcal{Q}$ with $1 \in \mathcal{Q}$, auxiliary supervision uses $1 \le \tau \le T - q$. A shared GRU summarizes the causal prefix $\mathbf{g}_{1:\tau,n}$. Let f_{proj} map this summary to the latent space, and let e_n and o_q be learned node and offset embeddings. The target-time causal context is

$$\mathbf{u}_{\tau,q,n} = \text{LN}_{\text{proj}}\big(f_{\text{proj}}(\text{GRU}(\mathbf{g}_{1:\tau,n})) + e_{\tau+q}^{\text{cal}} + e_n + o_q\big). \tag{3}$$

Let $\mathcal{F}_\tau$ contain observations $\mathbf{z}_{1:\tau}$, the directed graph operators, and calendar context known at position τ, including the target-time embedding $e_{\tau+q}^{\text{cal}}$. Its signal-dependent history ends at τ.

Student-t conditional prediction. The router $\alpha_{\tau,q,n} = \text{softmax}(f_\alpha(\mathbf{u}_{\tau,q,n}))$, with components indexed by $(0, \rightarrow, \leftarrow)$, mixes the local level and the two directed graph aggregates. Together with a scalar location-correction head f_μ, a context-dependent positive scale $\sigma_{\tau,q,n}$ produced from the same projector context $\mathbf{u}_{\tau,q,n}$, and a learned offset-specific degree of freedom $\nu_q > 2$, it defines

$$\mu_{\tau,q,n} = \alpha_{\tau,q,n}^{(0)} z_{\tau,n} + \alpha_{\tau,q,n}^{(\rightarrow)} (P^{\rightarrow}\mathbf{z}_\tau)_n + \alpha_{\tau,q,n}^{(\leftarrow)} (P^{\leftarrow}\mathbf{z}_\tau)_n + f_\mu(\mathbf{u}_{\tau,q,n}), \tag{4a}$$

$$z_{\tau+q,n} \mid \mathcal{F}_\tau \sim \text{StudentT}(\nu_q, \mu_{\tau,q,n}, \sigma_{\tau,q,n}). \tag{4b}$$

The $q = 1$ outputs initialize streams; $q > 1$ supplies auxiliary supervision. Define

$$\Delta_{\tau,n}^{\text{pred}} = \mu_{\tau,1,n} - z_{\tau,n}, \qquad \varepsilon_{\tau+1,n} = z_{\tau+1,n} - \mu_{\tau,1,n}, \qquad z_{\tau+1,n} = z_{\tau,n} + \Delta_{\tau,n}^{\text{pred}} + \varepsilon_{\tau+1,n}. \tag{5}$$

Proposition 1 (Conditional bias of learned innovations)**.** *Let $z_{\tau+1}$ and a fixed, $\mathcal{F}_\tau$-measurable projector μ_θ be integrable. Define $m_\tau = \mathbb{E}[z_{\tau+1} \mid \mathcal{F}_\tau]$ and $\varepsilon_{\tau+1}^\theta = z_{\tau+1} - \mu_\theta$. Then*

$$\mathbb{E}[\varepsilon_{\tau+1}^\theta \mid \mathcal{F}_\tau] = m_\tau - \mu_\theta. \tag{6}$$

If both variables lie in L^2, let $\varepsilon^* = z_{\tau+1} - m_\tau$. Then $\mathbb{E}[\varepsilon^* \mid \mathcal{F}_\tau] = 0$ and $\mathbb{E}[(\varepsilon^\theta)^2] = \mathbb{E}[(\varepsilon^*)^2] + \mathbb{E}[(m_\tau - \mu_\theta)^2]$.

This identity gives the learned coordinate its conditional meaning: its bias equals the projector's conditional-mean error. Appendix B.1 proves the result; Appendix B.2 analyzes its residual-score behavior.

For robust coordinate estimation, fix $\nu > 2$ and $\sigma > 0$ and define the residual loss and score $\ell(r) = \frac{\nu+1}{2}\log(1 + \frac{r^2}{\nu\sigma^2})$ and $\psi(r) = \ell'(r) = \frac{(\nu+1)r}{\nu\sigma^2 + r^2}$.

Proposition 2 (Bounded residual score)**.** *For the fixed parameters above,* $\sup_{r\in\mathbb{R}} |\psi(r)| = \frac{\nu+1}{2\sigma\sqrt{\nu}}$ *and* $\lim_{|r|\to\infty} \psi(r) = 0$.

Proof: Appendix B.2.

Alignment and stream initialization. The one-step quantities are aligned back to the T observed positions as

$$\bar{\mu}_{:,n} = [z_{1,n}; \mu_{1,1,n}; \ldots; \mu_{T-1,1,n}], \qquad \overline{\Delta}^{\mathrm{pred}}_{:,n} = [0; \Delta^{\mathrm{pred}}_{1,n}; \ldots; \Delta^{\mathrm{pred}}_{T-1,n}], \tag{7}$$

$$\bar{\varepsilon}_{:,n} = [z_{1,n}; \ldots; z_{T,n}] - \bar{\mu}_{:,n}, \qquad \bar{\sigma}_{:,n} = [\sigma_{1,1,n}; \sigma_{1,1,n}; \ldots; \sigma_{T-1,1,n}]. \tag{8}$$

At the first position, setting $\bar{\mu}_{1,n} = z_{1,n}$ gives zero drift and innovation; $\bar{\sigma}_{1,n} = \sigma_{1,1,n}$ reuses the first available scale. This padding aligns all fields to length T. For threshold $\kappa_{\mathrm{core}} > 0$, define the scale-bounded core $\varepsilon^{\mathrm{core}}_{\tau,n} = \mathrm{clip}(\bar{\varepsilon}_{\tau,n}, -\kappa_{\mathrm{core}}\bar{\sigma}_{\tau,n}, \kappa_{\mathrm{core}}\bar{\sigma}_{\tau,n})$, its remainder $\varepsilon^{\mathrm{tail}}_{\tau,n} = \bar{\varepsilon}_{\tau,n} - \varepsilon^{\mathrm{core}}_{\tau,n}$, and normalized view $\varepsilon^{\mathrm{norm}}_{\tau,n} = \tanh(\bar{\varepsilon}_{\tau,n}/\bar{\sigma}_{\tau,n})$. We then form level- and innovation-oriented coordinates

$$\boldsymbol{\xi}^{L}_{\tau,n} = [z_{\tau,n}, \bar{\mu}_{\tau,n}, \overline{\Delta}^{\mathrm{pred}}_{\tau,n}, \bar{\sigma}_{\tau,n}, \bar{\varepsilon}_{\tau,n}], \qquad \boldsymbol{\xi}^{S}_{\tau,n} = [z_{\tau,n}, \bar{\varepsilon}_{\tau,n}, \varepsilon^{\mathrm{core}}_{\tau,n}, \varepsilon^{\mathrm{tail}}_{\tau,n}, \varepsilon^{\mathrm{norm}}_{\tau,n}, \bar{\sigma}_{\tau,n}]. \tag{9}$$

The coordinates overlap: $\boldsymbol{\xi}^{L}$ retains the aligned residual, and $\boldsymbol{\xi}^{S}$ retains the observed level. Learned maps $\mathrm{Emb}_L, \mathrm{Emb}_S$ and stream-specific position–node embeddings $E^{L,\mathrm{pos}}, E^{S,\mathrm{pos}}$ initialize

$$\begin{aligned} L^{(0)}_{\tau,n} &= \mathrm{LN}_L\Big(\mathrm{Emb}_L(\boldsymbol{\xi}^{L}_{\tau,n}) + e^{\mathrm{cal}}_{\tau} + e_n + E^{L,\mathrm{pos}}_{\tau,n}\Big), \\ S^{(0)}_{\tau,n} &= \mathrm{LN}_S\Big(\mathrm{Emb}_S(\boldsymbol{\xi}^{S}_{\tau,n}) + e^{\mathrm{cal}}_{\tau} + e_n + E^{S,\mathrm{pos}}_{\tau,n}\Big). \end{aligned} \tag{10}$$

For a future target $Y \in L^2$, write $\mathsf{H} = (L, S)$ for a history-measurable stream pair and $V = f(\mathsf{H})$ for deterministic fusion. Define $\mathcal{R}(U) = \mathbb{E}[(Y - \mathbb{E}[Y \mid U])^2]$.

Proposition 3 (Prediction cost of a coarser representation)**.**

$$\mathcal{R}(V) - \mathcal{R}(\mathsf{H}) = \mathbb{E}[(\mathbb{E}[Y \mid \mathsf{H}] - \mathbb{E}[Y \mid V])^2] \geq 0. \tag{11}$$

Equality holds exactly when the conditional mean given H *is measurable with respect to* V*, up to null sets.*

Proof: Appendix B.3. The fusion identity motivates retaining joint predictive evidence during cross-stream interaction.

3.3 Temporal–Spatial Gibbs Transport

Directed graph prior. Let $K_{\mathcal{G}}$ be the diffusion order, ϖ the hop index, and $\lambda_u \in (0, 1)$ a uniform-mixing weight. The spatial prior averages forward and backward diffusion powers and adds finite global support:

$$R_0 = \frac{1}{1 + 2K_{\mathcal{G}}}\left[I_N + \sum_{\varpi=1}^{K_{\mathcal{G}}}\{(P^{\rightarrow})^{\varpi} + (P^{\leftarrow})^{\varpi}\}\right], \tag{12}$$

$$R = \mathrm{RowNorm}\left((1 - \lambda_u)R_0 + \frac{\lambda_u}{N}\mathbf{1}_N\mathbf{1}_N^{\top}\right). \tag{13}$$

Here I_N is the identity, $\mathbf{1}_N$ is the all-ones vector, and RowNorm normalizes each row. The result R is a positive row-stochastic graph prior; $\bar{R}$ denotes its row-standardized log form used below.

One-axis Gibbs operator. For either stream $Z \in \mathbb{R}^{T\times N\times d}$, temporal attention uses $Z_{:,n,:} \in \mathbb{R}^{T\times d}$ at fixed node n and forms a $T \times T$ plan; spatial attention uses $Z_{\tau,:,:} \in \mathbb{R}^{N\times d}$ at fixed position τ and forms an $N \times N$ plan. On axis a, each head has projections $Q, K, V \in \mathbb{R}^{n_a\times d_{\mathrm{head}}}$, with $n_a = T$ or N, respectively. Suppressing head indices, the learned head-specific temperature $\vartheta > 0$ and bias $\mathcal{B}^a$ define

$$G^a = \frac{QK^{\top}}{\sqrt{d_{\mathrm{head}}}\,\vartheta} + \mathcal{B}^a, \qquad \Pi^a = \mathrm{softmax}_{\mathrm{row}}(G^a). \tag{14}$$

For each head, $\mathcal{B}^{\mathrm{time}} = \eta^{\mathrm{time}} \in \mathbb{R}^{T\times T}$ is learned and initialized from relative temporal distance; $\mathcal{B}^{\mathrm{space}} = \gamma\bar{R} \in \mathbb{R}^{N\times N}$ uses a learned strength $\gamma > 0$. Equivalently, with $\mathcal{R}_a = \{Y \geq 0 : Y\mathbf{1} = \mathbf{1}\}$, the plan solves

$$\Pi^a = \arg\min_{Y\in\mathcal{R}_a}\{-\langle Y, G^a\rangle_F + \langle Y, \log Y\rangle_F\}. \tag{15}$$

Here $s_{ij} = \langle Q_i, K_j\rangle/(\sqrt{d_{\mathrm{head}}}\vartheta)$, $a_i = \mathrm{mean}_j \log R_{ij}$, and $b_i = \max(\mathrm{std}_j \log R_{ij}, \epsilon_R)$, so $\bar{R}_{ij} = (\log R_{ij} - a_i)/b_i$.

Proposition 4 (Row-tempered structural prior). *For fixed $R_{ij} > 0$, the spatial plan is*

$$\Pi_{ij} = \frac{\exp(s_{ij}) R_{ij}^{\gamma/b_i}}{\sum_k \exp(s_{ik}) R_{ik}^{\gamma/b_i}}. \tag{16}$$

Equivalently, for $\rho_{ij} = R_{ij}^{\gamma/b_i} / \sum_k R_{ik}^{\gamma/b_i}$, each row minimizes $-\sum_j p_j s_{ij} + \mathrm{KL}(p_i \| \rho_i)$.

Proof: Appendix B.4. At fixed content, $\gamma \to 0$ recovers global attention; $\gamma \to \infty$ concentrates on row-wise prior maxima. Finite γ retains global support.

The pre-dropout plan is the Gibbs solution above; dropout is applied only during value aggregation. Restoring the head index r for this step, define the full-width aggregation

$$\mathcal{A}_a(Z) = \mathrm{Concat}_{r=1}^{n_{\mathrm{head}}} \left[\mathrm{Dropout}(\Pi^{a,r}) V^r \right],$$

where n_{head} is the number of heads. Each block, stream, and axis owns its projection maps, biases, temperatures, normalization, feed-forward network, and residual gains. Suppressing these module indices, attention gain λ_{tr} and feed-forward gain λ_{ff} give

$$U = Z + \lambda_{\mathrm{tr}} W_O(\mathcal{A}_a(Z)), \qquad \mathrm{Axis}_a(Z) = U + \lambda_{\mathrm{ff}} \, \mathrm{FFN}_a(\mathrm{LN}(U)). \tag{17}$$

Proposition 5 (Sensitivity of a sequential axis operator). *In deterministic evaluation with fixed bias, row-softmax satisfies $\|\Pi(G') - \Pi(G)\|_F \le \frac{1}{2}\|G' - G\|_F$. On a domain where $T_a = W_O \mathcal{A}_a$ and $F_a = \mathrm{FFN}_a \circ \mathrm{LN}$ have Lipschitz constants L_T, L_F, respectively,*

$$\mathrm{Lip}(\mathrm{Axis}_a) \le (1 + |\lambda_{\mathrm{ff}}| L_F)(1 + |\lambda_{\mathrm{tr}}| L_T). \tag{18}$$

Appendix B.5 proves the bound and derives L_T from bounded projected features, projection Lipschitz constants, and $\vartheta \ge \vartheta_{\min} > 0$.

Temporal–spatial composition. For each block $m = 0, \ldots, M-1$ and stream $X \in \{L, S\}$, apply time then space attention:

$$\widetilde{X}^{(m)} = \mathrm{Axis}_{\mathrm{space}}^{m,X}\big(\mathrm{Axis}_{\mathrm{time}}^{m,X}(X^{(m)})\big). \tag{19}$$

The pair $(\widetilde{L}^{(m)}, \widetilde{S}^{(m)})$ enters the exchange.

3.4 Recoverable Cross-Stream Exchange

Let $\mathcal{C}_m^{S \to L}$ and $\mathcal{C}_m^{L \to S}$ be learned directional update networks and $\lambda_m^L, \lambda_m^S > 0$ their learned gates. The exchange updates the two streams sequentially:

$$L^{(m+1)} = \widetilde{L}^{(m)} + \lambda_m^L \mathcal{C}_m^{S \to L} \left(\mathrm{LN}_m^S(\widetilde{S}^{(m)}) \right), \tag{20}$$

$$S^{(m+1)} = \widetilde{S}^{(m)} + \lambda_m^S \mathcal{C}_m^{L \to S} \left(\mathrm{LN}_m^L(L^{(m+1)}) \right). \tag{21}$$

For fixed parameters, absorb gates and normalization into f and g, writing the exchange as $L' = L + f(S)$, $S' = S + g(L')$.

Proposition 6 (Bijective exchange without deterministic collisions). *The inverse is $S = S' - g(L')$, $L = L' - f(S)$. Thus distinct input stream pairs have distinct output pairs. Wherever differentiable, the exchange Jacobian determinant equals one.*

Proof: Appendix B.6. Each stream thus conditions the other while the exchange preserves joint-state recoverability and local volume.

3.5 Horizon-Specific Full-History Readout

After M blocks, a learned map f_{fuse} fuses the terminal streams at every observed position. For horizon h, define the known future calendar embedding $e_h^{\mathrm{fut}} = e_{T+h}^{\mathrm{cal}} = \phi_c(\mathbf{c}_{t+h-1})$ using the same calendar map as the projector, and let e_h^{hor} be a learned horizon embedding. The fused history and future query are

$$\Phi_{\tau,n} = f_{\mathrm{fuse}} \left(\mathrm{LN}_{\mathrm{fuse}}([L_{\tau,n}^{(M)}; S_{\tau,n}^{(M)}]) \right), \qquad \chi_{h,n} = e_h^{\mathrm{fut}} + e_n + e_h^{\mathrm{hor}}. \tag{22}$$

The vectors $\Phi_{\tau,n}, S^{(M)}_{\tau,n}, W^{\text{hist}}_{h,\tau}, W^{\text{inn}}_{h,\tau}$ lie in $\mathbb{R}^d$. Positive scalar gains $\lambda_{\text{hist}}, \lambda_{\text{inn}}, \lambda_{\text{fut}}$ weight the fused-history, direct-innovation, and future-context terms. With horizon bias β^{hist} and future head f_{fut}, the forecast is

$$\Delta^{\text{hist}}_{h,n} = \lambda_{\text{hist}} \left(\sum_{\tau=1}^{T} \langle \Phi_{\tau,n}, W^{\text{hist}}_{h,\tau} \rangle + \beta^{\text{hist}}_h \right), \qquad \Delta^{\text{inn}}_{h,n} = \lambda_{\text{inn}} \sum_{\tau=1}^{T} \langle S^{(M)}_{\tau,n}, W^{\text{inn}}_{h,\tau} \rangle, \tag{23a}$$

$$\Delta^{\text{fut}}_{h,n} = \lambda_{\text{fut}} f_{\text{fut}}(\text{LN}_{\text{fut}}(\chi_{h,n})), \qquad \widehat{z}_{h,n} = z_{T,n} + \Delta^{\text{hist}}_{h,n} + \Delta^{\text{inn}}_{h,n} + \Delta^{\text{fut}}_{h,n}. \tag{23b}$$

The direct innovation term provides a horizon-specific path from $S^{(M)}$ to the forecast alongside the nonlinear fused-history path, preserving separate access to residual features as motivated by Proposition 3. Inverse standardization yields $\widehat{\mathbf{X}}_{\text{out}}$. With future context fixed, flatten the history and innovation features into $u \in \mathbb{R}^D$. The readout is $f_h(u) = w_h^\top u + b_h$, with scales absorbed into w_h.

Proposition 7 (Representability through a shared linear contraction). *Let W have rows $w_h^\top$. A shared contraction $v = Cu \in \mathbb{R}^k$ followed by horizon-specific linear heads Av represents Wu for every $u \in \mathbb{R}^D$ if and only if $W = AC$. For fixed C, this is equivalent to* $\text{row}(W) \subseteq \text{row}(C)$. *Some width-$k$ contraction can represent W if and only if* $\text{rank}(W) \leq k$.

Proof: Appendix B.7. Each horizon can therefore express its own dependence on the complete terminal history.

3.6 Training Objective and Complexity

Let $\mathcal{L}_{\text{fore}}$ denote the masked MAE/Smooth-L1 forecast loss evaluated in the original target units. The causal projector is jointly regularized by a multi-offset Student-t likelihood and a lag-one innovation-product penalty:

$$\mathcal{L}_{\text{NLL}} = -\sum_{q \in \mathcal{Q}} w_q \, \text{mean}_{b,n,\, \tau+q \leq T} \log \text{StudentT}(z_{b,\tau+q,n}; \nu_q, \mu_{b,\tau,q,n}, \sigma_{b,\tau,q,n}), \tag{24}$$

$$\mathcal{L}_{\text{lag}} = \left(\text{mean}_{b,n,\, 2 \leq \tau < T} \, \widetilde{\varepsilon}_{b,\tau+1,n} \widetilde{\varepsilon}_{b,\tau,n} \right)^2, \tag{25}$$

where $w_q \propto q^{-1/2}$ is normalized over $\mathcal{Q}$, and $\widetilde{\varepsilon}$ denotes the clipped scale-normalized innovation $\bar{\varepsilon}/\bar{\sigma}$; the exact clipping convention is given in Appendix A.1. The complete objective is

$$\mathcal{L} = \mathcal{L}_{\text{fore}} + \lambda_{\text{NLL}} \mathcal{L}_{\text{NLL}} + \lambda_{\text{lag}} \mathcal{L}_{\text{lag}}. \tag{26}$$

The standardized-coordinate losses provide auxiliary conditional and lag-product supervision; λ_{NLL} and λ_{lag} set their weights relative to the original-unit forecast loss (Appendix A.1).

Multi-offset population interpretation. For shared history features $U_\eta(\mathcal{F}_\tau)$, let $p_{\eta,q}$ be the offset density and p^*_q its true conditional counterpart. When the expectations are finite, the population NLL satisfies

$$\mathcal{L}_{\text{proj}} = C + \sum_{q \in \mathcal{Q}} w_q \mathbb{E} \, \text{KL}(p^*_q \| p_{\eta,q}), \qquad w_q > 0, \tag{27}$$

with C independent of η. Joint realizability and zero divergence imply matching every offset; otherwise supervision balances the offset-specific conditional divergences. Appendix B.2 derives the identity.

For latent width d, M encoder blocks cost $O(Md(NT^2 + TN^2))$ per sample for attention, and the full-history contractions cost $O(HTNd)$; projection and feed-forward costs are additional.

4 Experiments

4.1 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. Appendix A.3 provides the detailed optimization configuration.

Datasets. We evaluate six highway benchmarks: PeMS03 (Song et al., 2020), PeMS04 (Song et al., 2020), PeMS07 (Song et al., 2020), PeMS08 (Song et al., 2020), PeMSD7(M) (Yu et al., 2018), and PeMSD7(L) (Yu et al., 2018). These sensor-network benchmarks assess forecasting on road graphs. We additionally evaluate T-Drive (Pan et al., 2019) and CHIBike (Wang et al., 2021), with separate inflow and outflow targets for each urban grid, yielding ten forecasting tasks in total. All datasets use chronological splits and normalization fitted only on the training split. We report MAE and RMSE in the original data units; lower values indicate better forecasts.

Baselines. The comparisons cover statistical and recurrent models, ARIMA (Box et al., 2015), FC-LSTM (Hochreiter & Schmidhuber, 1997), and GRU-ED (Cho et al., 2014); graph-based models, 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), and STGNCDE (Choi et al., 2022); and attention-based or enhanced forecasting methods, TFormer (Yan et al., 2022), PDFormer (Jiang et al., 2023), STAEformer (Liu et al., 2023), CaST (Xia et al., 2023), RAST (Ruan et al., 2026), and MFormer (Hu et al., 2026). The tables retain the available benchmark comparisons for each task. *Improve* denotes $(e_{\text{best}} - e_{\text{ours}})/e_{\text{best}} \times 100\%$, measured against the strongest listed baseline for each metric.

4.2 Forecasting Accuracy

Tables 1 and 2 assess whether the conditional-coordinate design transfers across road graphs and urban grids. DR-GibbsFormer improves 18 of 20 aggregate MAE/RMSE comparisons, covering 11 of 12 highway and seven of eight urban-flow metrics. Highway gains include 6.44% in PeMS03 RMSE, 5.75% in PeMSD7(L) MAE, and 4.79% in PeMS08 MAE, suggesting that retaining history-conditioned variation alongside traffic levels is useful across different network scales. T-Drive Inflow and Outflow improve by $7.13\%/7.61\%$ and $6.84\%/6.87\%$ in MAE/RMSE, respectively: lower errors under both criteria support the combined representation and interaction design across both error criteria. PeMS04 RMSE is slightly worse than the strongest baseline (30.562 versus 30.55), while CHIBike Inflow improves MAE by 8.26% and has RMSE 5.333 versus the best baseline's 5.09. This MAE–RMSE difference indicates a stronger improvement in average absolute error than in the squared-error criterion. These aggregate results motivate the following ablations, which examine coordinate construction, graph-informed interaction, and preservation of historical access (C1–C3).

Table 1: Highway forecasting (MAE/RMSE). Red: best; blue: second best.

	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
DR-GibbsFormer (ours)	**14.539 ± 0.058**	**24.149 ± 0.112**	**18.053 ± 0.082**	**30.562 ± 0.313**	**18.831 ± 0.021**	**32.798 ± 0.038**	**13.206 ± 0.010**	**22.875 ± 0.032**	**2.564 ± 0.007**	**5.283 ± 0.024**	**2.705 ± 0.013**	**5.667 ± 0.040**
Improve	**+1.89%**	**+6.44%**	**+0.04%**	–	**+3.53%**	**+0.52%**	**+4.79%**	**+1.95%**	**+4.32%**	**+1.99%**	**+5.75%**	**+1.61%**

4.3 Main Component Ablation

Table 3 connects the architectural choices for C1–C3 to controlled removals on PeMS08 and T-Drive Inflow. Every removal raises both MAE and RMSE. For C1, removing the innovation stream causes

Table 2: Urban-flow forecasting (MAE/RMSE). Red: best; blue: second best.

	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
DR-GIBBSFORMER (ours)	**15.936 ± 0.032**	**28.493 ± 0.093**	**15.977 ± 0.028**	**28.506 ± 0.126**	**3.495 ± 0.009**	5.333 ± 0.010	**3.468 ± 0.008**	**5.261 ± 0.020**
Improve	**+7.13%**	**+7.61%**	**+6.84%**	**+6.87%**	**+8.26%**	–	**+7.02%**	**+0.17%**

the largest increases, $0.424/0.675$ and $0.674/1.037$, while removing conditional innovation or the level stream also degrades accuracy. The pattern supports retaining the overlapping level and residual views as complementary forecasting inputs. For C2, removing the Gibbs graph prior raises MAE by $0.204/0.284$, consistent with a useful structural preference within global attention. For C3, removing triangular exchange raises MAE by $0.174/0.244$, whereas removing full-history readout raises it by $0.314/0.484$. Access to the complete terminal history is therefore particularly useful among these two tested mechanisms, complementing the exchange's information-recoverability property. Removing Student-t NLL has a larger effect than removing the lag penalty, supporting conditional supervision as the stronger auxiliary signal in these comparisons. Across both tasks, the ablations support the sequence from conditional coordinate construction through structured interaction to horizon-specific access to the complete history.

Table 3: Component ablation. Δ: increase over the full model.

	PeMS08				T-Drive Inflow			
Variant	MAE	ΔMAE	RMSE	ΔRMSE	MAE	ΔMAE	RMSE	ΔRMSE
Full DR-GIBBSFORMER	**13.206**	–	**22.875**	–	**15.936**	–	**28.493**	–
w/o Conditional Innovation	13.57	+0.364	23.43	+0.555	16.49	+0.554	29.34	+0.847
w/o Level Stream	13.47	+0.264	23.26	+0.385	16.34	+0.404	29.11	+0.617
w/o Innovation Stream	13.63	+0.424	23.55	+0.675	16.61	+0.674	29.53	+1.037
w/o Gibbs Graph Prior	13.41	+0.204	23.18	+0.305	16.22	+0.284	28.91	+0.417
w/o Triangular Exchange	13.38	+0.174	23.12	+0.245	16.18	+0.244	28.84	+0.347
w/o Full-History Readout	13.52	+0.314	23.36	+0.485	16.42	+0.484	29.23	+0.737
w/o Student-t NLL	13.31	+0.104	23.02	+0.145	16.10	+0.164	28.72	+0.227
w/o Lag Loss	13.25	+0.044	22.94	+0.065	16.00	+0.064	28.59	+0.097

4.4 CONDITIONAL COORDINATE STUDY

Table 4 isolates the distributional, temporal, and coordinate choices underlying C1. The Student-t projector with $\mathcal{Q} = \{1, 3, 6\}$ performs best on both datasets. Replacing it with a Gaussian raises MAE/RMSE by $0.104/0.165$ on PeMS08 and $0.164/0.257$ on T-Drive Inflow. This pattern is consistent with the bounded residual-score mechanism in Proposition 2, which reduces the location gradient assigned to large residuals at fixed scale. Restricting supervision to $q = 1$ raises the same metrics by $0.184/0.285$ and $0.284/0.447$. Because only the one-step output initializes the streams, this result supports the benefit of supervising the shared history representation at additional

offsets while keeping the inference inputs unchanged. First-difference coordinates produce the largest increases, $0.384/0.585$ and $0.614/0.957$: a raw change cannot distinguish history-predictable movement from conditional surprise as encoded by the learned history-conditioned location. The larger loss from first differences is particularly informative: in these comparisons, change measured relative to learned historical predictability provides more useful forecasting coordinates than raw change.

Table 4: Projector ablation. Δ: increase over Student-t with $\{1, 3, 6\}$.

	PeMS08				T-Drive Inflow			
Projector variant	MAE	ΔMAE	RMSE	ΔRMSE	MAE	ΔMAE	RMSE	ΔRMSE
Student-t, offsets $\{1, 3, 6\}$	**13.206**	–	**22.875**	–	**15.936**	–	**28.493**	–
Gaussian, offsets $\{1, 3, 6\}$	13.31	+0.104	23.04	+0.165	16.10	+0.164	28.75	+0.257
Student-t, offset $\{1\}$	13.39	+0.184	23.16	+0.285	16.22	+0.284	28.94	+0.447
First-difference coordinates	13.59	+0.384	23.46	+0.585	16.55	+0.614	29.45	+0.957

4.5 Horizon-wise Behavior

Figure 2 examines how the direct full-history readout behaves across 12 forecast steps. PeMS04 MAE increases from 16.36 at 5 minutes to 19.18 at 60 minutes, while PeMSD7(M) increases from 1.272 to 3.170; the three-run standard-deviation bands remain narrow. The different error-growth profiles characterize the task-dependent difficulty faced by the horizon-specific functionals in Section 3.5. Each horizon uses the complete encoded history and a direct innovation-stream contribution, so later predictions retain access to observed evidence instead of consuming earlier predicted values. The curves complement the full-history ablation: the ablation tests the value of that access, while the profiles characterize the remaining difficulty as forecast distance grows. Complete per-horizon MAE and RMSE values appear in Appendix C.

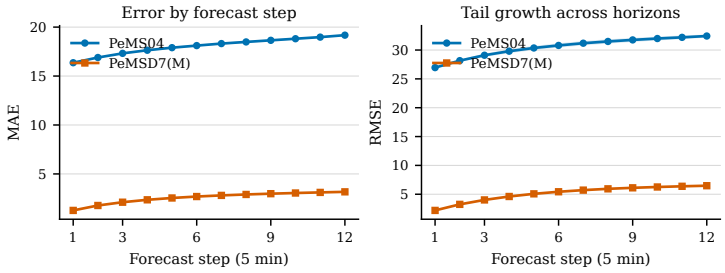


Figure 2: Horizon-wise error (mean and standard deviation over three seeds).

5 Conclusion

This work connects history-conditioned predictive coordinates with the exchange and horizon-specific use of their information. DR-GibbsFormer builds overlapping level and innovation-oriented coordinates, propagates their latent values through content- and graph-dependent Gibbs interaction, coordinates the streams through recoverable triangular exchange, and lets each horizon read the complete terminal history. The theoretical results establish the conditional meaning of the residual, the structural weighting and sensitivity of interaction, joint-state recoverability during exchange, and the expressivity of horizon-specific linear functionals. Across ten forecasting tasks, the framework improves 18 of 20 aggregate MAE/RMSE comparisons, with gains reaching 8.26% and 7.61%, respectively. Ablations connect these improvements to conditional rather than raw-difference coordinates, multi-offset supervision, the graph prior, exchange, and full-history access. Together, the results support conditional coordinate transport: construct history-conditioned evidence, exchange it through graph-informed and recoverable interactions, and read the resulting history separately for each horizon.

Reproducibility statement

Appendix A lists the end-to-end tensor trace, model dimensions, optimization settings, data splits, and seeds. Every reported run uses random initialization, validation-only checkpoint selection, a sealed test split during training, and a single frozen test pass.

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.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018. doi: 10.1090/mcom/3303.

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.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HkpbnH9lx.

Joseph L. Doob. *Stochastic Processes*. Wiley, 1953.

James Durbin and Siem Jan Koopman. *Time Series Analysis by State Space Methods*. Oxford University Press, 2 edition, 2012. doi: 10.1093/acprof:oso/9780199641178.001.0001.

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.

Aidan N. Gomez, Mengye Ren, Raquel Urtasun, and Roger B. Grosse. The reversible residual network: Backpropagation without storing activations. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/f9be311e65d81a9ad8150a60844bb94c-Paper.pdf.

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.

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.

Kenneth L. Lange, Roderick J. A. Little, and Jeremy M. G. Taylor. Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84(408):881–896, 1989. doi: 10.1080/01621459.1989.10478852.

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.

Zheyi Pan, Yuxuan Liang, Weifeng Wang, Yong Yu, Yu Zheng, and Junbo Zhang. Urban traffic prediction from spatio-temporal data using deep meta learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1720–1730, 2019. doi: 10.1145/3292500.3330884.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019. doi: 10.1561/2200000073.

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.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Jingyuan Wang, Jiawei Jiang, Wenjun Jiang, Chao Li, and Wayne Xin Zhao. LibCity: An open library for traffic prediction. In *Proceedings of the 29th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '21, pp. 145–148. Association for Computing Machinery, 2021. doi: 10.1145/3474717.3483923. URL https://doi.org/10.1145/3474717.3483923.

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.

Haoyang Yan, Xiaolei Ma, and Ziyuan Pu. Learning dynamic and hierarchical traffic spatiotemporal features with transformer. *IEEE Transactions on Intelligent Transportation Systems*, 23(11): 22386–22399, 2022. doi: 10.1109/TITS.2021.3102983.

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.

A Reproducibility Details

A.1 Exact Operator and Loss Definitions

Graph construction and stabilization. Let $i, j, k \in \mathcal{V}$ index target, source, and intermediate nodes. Writing $\mathcal{D}_{ij}^{(1)} = \log(1 + \omega_{j\to i})$ on an edge, zero on the diagonal, and $+\infty$ otherwise, the direct/two-hop graph closure and its affinity are

$$\mathcal{D}_{ij} = \min\left(\mathcal{D}_{ij}^{(1)}, \min_k \mathcal{D}_{ik}^{(1)} + \mathcal{D}_{kj}^{(1)}\right), \qquad A_{ij}^{\rightarrow} = \exp(-\mathcal{D}_{ij}/\bar{\mathcal{D}})\,\mathbb{I}[\mathcal{D}_{ij} < +\infty], \tag{28}$$

where $\bar{\mathcal{D}}$ is the median positive finite path cost and $\mathbb{I}$ is the indicator. Row-normalizing the off-diagonal entries of $A^{\rightarrow}$ and $(A^{\rightarrow})^{\top}$ yields the fixed forward and backward operators $P^{\rightarrow}$ and $P^{\leftarrow}$, respectively. They condition both the projector and the spatial encoder.

Let $\epsilon_{\mathcal{G}}, \epsilon_R > 0$ be numerical stabilizers. In Equation (28), the implementation replaces $\bar{\mathcal{D}}$ by $\max(\bar{\mathcal{D}}, \epsilon_{\mathcal{G}})$ and sets unreachable affinities to zero. After Equation (13), it standardizes the finite log-prior row by row:

$$\bar{R}_{ij} = \frac{\log R_{ij} - \mathrm{mean}_{j'} \log R_{ij'}}{\max(\mathrm{std}_{j'} \log R_{ij'}, \epsilon_R)}. \tag{29}$$

Here j' is a dummy source-node index. The backward increment is zero at the first input position. Isolated rows of both directed graph operators receive self-loops.

Innovation views and lag-one product. Equation (8) uses $z_{1,n}$ as the first aligned predicted level, sets the first aligned drift and innovation to zero, and reuses $\sigma_{1,1,n}$ at the first two positions because no preceding one-step prediction exists at $\tau = 1$. For a scalar residual v and transition $\delta > 0$, the forecast loss uses $\mathrm{SL1}_\delta(v) = v^2/(2\delta)$ when $|v| < \delta$ and $|v| - \delta/2$ otherwise.

For core threshold $\kappa_{\mathrm{core}} > 0$, the coordinate vectors in Equation (9) use

$$\varepsilon_{\tau,n}^{\mathrm{core}} = \mathrm{clip}(\bar{\varepsilon}_{\tau,n}, -\kappa_{\mathrm{core}}\bar{\sigma}_{\tau,n}, \kappa_{\mathrm{core}}\bar{\sigma}_{\tau,n}), \qquad \varepsilon_{\tau,n}^{\mathrm{tail}} = \bar{\varepsilon}_{\tau,n} - \varepsilon_{\tau,n}^{\mathrm{core}}, \tag{30}$$

$$\varepsilon_{\tau,n}^{\mathrm{norm}} = \tanh(\bar{\varepsilon}_{\tau,n}/\bar{\sigma}_{\tau,n}). \tag{31}$$

The historical implementation key `shock_autocorrelation` refers exactly to the lag-one innovation-product penalty $\mathcal{L}_{\mathrm{lag}}$ in Equation (25).

A.2 End-to-end tensor trace

Let n_{head} denote the number of attention heads and $d_{\mathrm{head}} = d/n_{\mathrm{head}}$ their width.

Inside either axis call, the code reshapes each of Q, K, V to $B \times T \times N \times n_{\mathrm{head}} \times d_{\mathrm{head}}$. For head r, query ι, and source $\jmath$, Equation (14) is evaluated as $G_{\iota\jmath}^{a,r} = \langle Q_\iota^r, K_\jmath^r\rangle/(\sqrt{d_{\mathrm{head}}}\,\vartheta_r) + \eta_{\iota\jmath}^{r,\mathrm{time}}$ on the temporal axis and as $G_{\iota\jmath}^{a,r} = \langle Q_\iota^r, K_\jmath^r\rangle/(\sqrt{d_{\mathrm{head}}}\,\vartheta_r) + \gamma_r \bar{R}_{\iota\jmath}$ on the spatial axis. Here $\vartheta_r = 0.35 + \mathrm{softplus}(\theta_r^{\mathrm{temp}})$ and $\gamma_r = \mathrm{softplus}(\theta_r^{\mathrm{road}})$, where θ_r^{temp} and θ_r^{road} are learned raw parameters. Every block owns four separate axis modules—level temporal, level spatial, innovation temporal, and innovation spatial—and each module has its own projections, biases, temperatures, residual gains, normalization layers, and feed-forward network. Attention dropout is applied after row-softmax and before multiplication by V.

A.3 Implementation Details

The implementation specializes the general STG task to one traffic target and two calendar features. Each prepared mini-batch has shape $B \times 12 \times N \times 3$, with channels for traffic, normalized time of day, and day of week. The traffic channel is standardized using a mean and standard deviation fitted only

Table 5: Forward-pass tensor shapes for batch size B.

Stage	Input	Output
Model input	$\mathbf{X}_{\text{in}}, \mathbf{C}_{\text{in}}$, and $\mathcal{G}$	$z \in \mathbb{R}^{B\times T\times N}$, calendar embeddings, $P^{\rightarrow}, P^{\leftarrow}$
Projector	z, graph operators, calendar, node identities	$q = 1$ fields $\mu, \Delta^{\text{pred}}, \varepsilon, \sigma \in \mathbb{R}^{B\times(T-1)\times N}$; scalar $\mathcal{L}_{\text{NLL}}$ from all $q \in \mathcal{Q}$
Alignment	raw $q = 1$ fields and $z_{:,1,:}$	$\bar{\mu}, \overline{\Delta}^{\text{pred}}, \bar{\varepsilon}, \bar{\sigma} \in \mathbb{R}^{B\times T\times N}$
Initialization	$\boldsymbol{\xi}^L, \boldsymbol{\xi}^S$ and embeddings	$L^{(0)}, S^{(0)} \in \mathbb{R}^{B\times T\times N\times d}$
Temporal Gibbs	one stream in $\mathbb{R}^{B\times T\times N\times d}$	same-shape stream; plan in $\mathbb{R}^{B\times N\times n_{\text{head}}\times T\times T}$
Spatial Gibbs	temporal-axis output	same-shape stream; plan in $\mathbb{R}^{B\times T\times n_{\text{head}}\times N\times N}$
Triangular exchange	two spatial-axis outputs	$L^{(m+1)}, S^{(m+1)} \in \mathbb{R}^{B\times T\times N\times d}$
Readout	$L^{(M)}, S^{(M)}$, future calendar queries	$\Delta^{\text{hist}}, \Delta^{\text{inn}}, \Delta^{\text{fut}} \in \mathbb{R}^{B\times H\times N}$
Model output	readout terms and the last standardized target value	$\widehat{\mathbf{Z}}_{\text{out}} \in \mathbb{R}^{B\times H\times N}$; inverse-transformed $\widehat{\mathbf{X}}_{\text{out}}$

on the training split; the two calendar channels are not standardized. The network returns $\widehat{\mathbf{Z}}_{\text{out}}$ at 12 horizons with an implementation-only singleton target axis, giving API shape $B\times 12\times N\times 1$. The data scaler inverse-transforms these values; omitting the singleton target axis gives $\widehat{\mathbf{X}}_{\text{out}}$ in Equation (1). Both the forecast loss and reported metrics use the valid set $\Omega = \{(b,h,n) : x^{+}_{b,h,n} > 0\}$. Future time-of-day and day-of-week indices are advanced deterministically from the final input timestamp.

Table 6: Multi-seed experiment configuration.

Model setting	Value	Optimization setting	Value
Input/output steps	12/12	Optimizer	AdamW
Model width	128	Peak learning rate	5×10^{-4}
Causal GRU width	48	Weight decay	10^{-3}
Heads / blocks	4 / 3	Warm-up	5 epochs
Graph hops	2	EMA decay	0.999
Dropout	0.12	Gradient clip	3.0
Projector offsets	$\{1, 3, 6\}$	Max epochs / patience	120 / 20
Scale floor / DoF floor	0.03 / 2.1	Time slots/day	288
Core / lag-product clips	1.2 / 8	Forecast weights	0.9 / 0.1
Temperature floor / graph mix	0.35 / 0.9:0.1	Smooth-L1 transition	8
FFN widths	$d \to 3d \to d$	Student-t NLL weight	0.005
Exchange widths	$d \to 2d \to d$	Lag-product weight	0.001

The graph stabilizers are $\epsilon_{\mathcal{G}} = 10^{-6}$ and $\epsilon_R = 10^{-5}$. The spatial-prior mixture is $0.9R_0 + 0.1\mathbf{1}_N\mathbf{1}_N^{\top}/N$, and temporal bias is initialized as $\eta^{r,\text{time}}_{\iota\jmath} = -|\iota - \jmath|/4$. The projector GRU produces a width-48 causal trace that is projected to d before adding calendar, node, and offset embeddings. Its router has width $d \to 3$, and its location and scale heads use $d \to d/2 \to 1$; writing $s_{\tau,q,n}$ for the raw scale-head output and ρ_q for the learned offset logit, the code uses $\sigma_{\tau,q,n} = 0.03 + \text{softplus}(s_{\tau,q,n})$ and $\nu_q = 2.1 + \text{softplus}(\rho_q)$. The level/innovation input maps use $5 \to d$ and $6 \to d$. Decoder fusion uses $2d \to d \to d$, while the future head uses $d \to d \to 1$. The projector core threshold is $\kappa_{\text{core}} = 1.2$, and the loss weights are $(\lambda_{\text{mae}}, \lambda_{\text{sl1}}, \lambda_{\text{NLL}}, \lambda_{\text{lag}}) = (0.9, 0.1, 0.005, 0.001)$. PeMS04 uses batch size 8 and seeds 2035, 2036, and 2037. PeMSD7(M) uses batch size 16 and seeds 2040, 2041, and 2042. The parameter counts are 3,544,414 and 3,291,614, respectively; the difference is due to node-dependent embeddings.

B Conditional Forecasting Analysis and Proofs

This appendix connects the conditional coordinates, graph-biased interaction, triangular exchange, and horizon-specific readout. The results characterize the conditional bias, structural regularization,

and information retained by the individual operators. Parameters are fixed during evaluation unless otherwise stated.

B.1 Population and learned conditional innovations

Let $\mathcal{F}_\tau$ contain the graph, observations through input position τ, and the calendar information known at that position. For a fixed node, suppress its index and assume $z_{\tau+1} \in L^2$. Write $m_\tau = \mathbb{E}[z_{\tau+1} \mid \mathcal{F}_\tau]$. A fixed trained projector $\mu_\theta(\mathcal{F}_\tau) \in L^2$ is measurable with respect to this information. If training is treated as random, the following statements instead condition additionally on the training information, with m_τ defined under that same conditioning.

Proof of Proposition 1. Conditional expectation preserves the integrable, $\mathcal{F}_\tau$-measurable projector, so

$$\mathbb{E}[z_{\tau+1} - \mu_\theta \mid \mathcal{F}_\tau] = \mathbb{E}[z_{\tau+1} \mid \mathcal{F}_\tau] - \mu_\theta = m_\tau - \mu_\theta.$$

This proves the main-text statement under integrability alone. □

Squared-error extension. Under the stronger L^2 assumptions above, let $\varepsilon^*_{\tau+1} = z_{\tau+1} - m_\tau$. Then $\mathbb{E}[\varepsilon^*_{\tau+1} \mid \mathcal{F}_\tau] = 0$ and

$$\mathbb{E}[(\varepsilon^\theta_{\tau+1})^2] = \mathbb{E}[(\varepsilon^*_{\tau+1})^2] + \mathbb{E}[(m_\tau - \mu_\theta)^2]. \tag{32}$$

To prove this identity, expand $\varepsilon^\theta = \varepsilon^* + (m_\tau - \mu_\theta)$; the cross term has expectation zero after conditioning on $\mathcal{F}_\tau$. The vector-valued version uses squared Euclidean norms. Proposition 1 also gives

$$\|\mathbb{E}[\varepsilon^\theta_{\tau+1} \mid \mathcal{F}_\tau]\|_{L^2} = \|m_\tau - \mu_\theta\|_{L^2}.$$

The predictable change $m_\tau - z_\tau$ and innovation $\varepsilon^*_{\tau+1}$ give the discrete predictable–martingale decomposition (Doob, 1953). Replacing m_τ by a learned location preserves the algebraic identity in Equation (5), while its conditional bias is exactly the estimation error above. If a sequence of projectors converges to m_τ in L^2, that bias converges to zero in L^2. The conditioning information here is the observed input window, graph, and known context.

Location versus conditional mean. Under a correctly specified conditional Student-t law with $\nu_q > 2$, the location is the conditional mean and the second moment exists. Under misspecification, the fitted location is a Student-t location parameter, and its difference from m_τ is exactly the conditional residual bias in Proposition 1. The projector is jointly optimized with forecasting, likelihood, and lag-product losses; the proposition applies to the resulting fitted location.

B.2 Robust residual gradients and multi-offset supervision

At fixed $\nu > 2$ and $\sigma > 0$, the residual-dependent Student-t negative log likelihood is

$$\ell(r) = \frac{\nu+1}{2} \log\left(1 + \frac{r^2}{\nu\sigma^2}\right), \qquad \psi(r) = \frac{d\ell}{dr} = \frac{(\nu+1)r}{\nu\sigma^2 + r^2}. \tag{33}$$

The full likelihood additionally includes the scale and degree-of-freedom terms used in training. Here these parameters are held fixed to isolate the location residual response, following the robust Student-t perspective of Lange et al. (1989).

Proof of Proposition 2.

Proof. For $r \geq 0$, the derivative of $r/(\nu\sigma^2 + r^2)$ has the sign of $\nu\sigma^2 - r^2$. Its maximum occurs at $r = \sigma\sqrt{\nu}$. Odd symmetry and division by r^2 establish the bound and limit. □

Thus the loss grows logarithmically in the magnitude of large residuals, and its location gradient decreases in the tails. Gaussian likelihood instead has residual score r/σ^2. The residual score is bounded and redescending at fixed scale. Parameter gradients additionally involve network Jacobians, and a uniform bound over learned scales requires a positive scale floor. This gradient behavior provides the mechanism examined by the Gaussian comparison in Table 4.

What multiple offsets constrain. For a shared history representation $U_\eta(\mathcal{F}_\tau)$ and conditional density heads $p_{\eta,q}$, consider the population objective with $w_q > 0$:

$$\mathcal{L}_{\mathrm{proj}}(\eta) = \sum_{q \in \mathcal{Q}} w_q \mathbb{E}[-\log p_{\eta,q}(z_{\tau+q} \mid U_\eta(\mathcal{F}_\tau))]. \tag{34}$$

Where the expectations exist, writing the true conditional density as $p_q^*(\cdot \mid \mathcal{F}_\tau)$ gives

$$\mathcal{L}_{\mathrm{proj}}(\eta) = C + \sum_{q \in \mathcal{Q}} w_q \mathbb{E}\big[\mathrm{KL}\big(p_q^*(\cdot \mid \mathcal{F}_\tau) \,\|\, p_{\eta,q}(\cdot \mid U_\eta(\mathcal{F}_\tau))\big)\big], \tag{35}$$

where C is independent of η. This follows by adding and subtracting $\log p_q^*$ inside each expectation. If all true conditional densities are jointly representable and a global minimum with zero divergence is attained, every positive-weight offset is matched. Otherwise the objective is a weighted compromise. Additional offsets penalize representations that fit the one-step distribution while losing information useful for the other conditional prediction tasks. In our model only $q = 1$ initializes the forecasting streams, so the offset comparison tests auxiliary supervision with the same inference information.

B.3 Joint stream information and lossy fusion

Let $Y = z_{T+h,n} \in L^2$, let $H = (L, S)$ denote any encoded stream pair measurable from the observed history, and let $V = f(H)$ be a deterministic fusion. Define the optimal squared-error risk from a representation U as $\mathcal{R}(U) = \mathbb{E}[(Y - \mathbb{E}[Y \mid U])^2]$.

Proof of Proposition 3.

Proof. Because V is a function of H, the tower property gives $\mathbb{E}[\mathbb{E}[Y \mid H] \mid V] = \mathbb{E}[Y \mid V]$. Expand $Y - \mathbb{E}[Y \mid V]$ as the sum of $Y - \mathbb{E}[Y \mid H]$ and $\mathbb{E}[Y \mid H] - \mathbb{E}[Y \mid V]$. The cross term vanishes after conditioning on H, yielding Equation (11). □

The identity quantifies fusion loss through the target conditional mean available in the encoded pair. Equality holds when fusion preserves that conditional mean. Our overlapping initial coordinates, separate histories, and innovation readout give the decoder access to complementary learned views. The identity concerns optimal squared-error prediction; Tables 3–4 assess the trained model with the reported forecasting metrics.

B.4 Gibbs solution and the directed graph factorization

For a fixed row, minimizing $-\sum_j p_j G_{ij} + \sum_j p_j \log p_j$ subject to $p_j \geq 0$ and $\sum_j p_j = 1$ gives the stationarity condition $-G_{ij} + \log p_j + 1 + \zeta_i = 0$. Normalizing yields $p_j = \exp(G_{ij}) / \sum_k \exp(G_{ik})$. Strict convexity establishes uniqueness. This is the standard entropy-regularized row problem; no column marginal is imposed.

Let $s_{ij} = \langle Q_i, K_j \rangle / (\sqrt{d_{\mathrm{head}}}\vartheta)$ and use the exact standardization from Appendix A.1:

$$a_i = \mathrm{mean}_j \log R_{ij}, \qquad b_i = \max(\mathrm{std}_j \log R_{ij}, \epsilon_R), \qquad \bar{R}_{ij} = (\log R_{ij} - a_i)/b_i. \tag{36}$$

Proof of Proposition 4.

Proof. Substitute $G_{ij} = s_{ij} + \gamma(\log R_{ij} - a_i)/b_i$ into softmax. The row-constant factor $\exp(-\gamma a_i / b_i)$ cancels, whereas b_i remains in the exponent. Expanding the KL term gives the same row objective up to a constant independent of p_i. □

At fixed content scores and R, $\gamma \to 0$ recovers global content attention. As $\gamma \to \infty$, mass concentrates on $J_i = \arg\max_j R_{ij}$; within this set the limiting weights are proportional to $\exp(s_{ij})$. To see this, divide the numerator and denominator by $\max_j R_{ij}^{\gamma/b_i}$: every strictly smaller prior entry vanishes in the limit. A uniform prior leaves content attention unchanged. Finite γ retains positive global support because of the uniform mixture in Equation (13). Thus the graph supplies a soft structural preference with positive global support. The KL formulation expresses the content–structure tradeoff as Gibbs regularization.

B.5 Proof of Proposition 5: axis sensitivity

All statements below concern deterministic evaluation with fixed graph bias and fixed trained parameters. Matrix norms are Frobenius unless a spectral norm $\|\cdot\|_2$ is indicated. The row-softmax Jacobian is $J(p) = \mathrm{diag}(p) - pp^\top$. Its absolute row sum is $2p_j(1-p_j) \le 1/2$; symmetry implies $\|J(p)\|_2 \le 1/2$. Integrating the Jacobian along a line segment therefore gives

$$\|\Pi(G') - \Pi(G)\|_F \le \tfrac{1}{2}\|G' - G\|_F. \tag{37}$$

For one head on a domain where $\|Q(Z)\|_F \le B_Q$, $\|K(Z)\|_F \le B_K$, $\|V(Z)\|_F \le B_V$, suppose the projected maps (including their normalization) have Lipschitz constants L_Q, L_K, L_V. If $\vartheta \ge \vartheta_{\min} > 0$, then

$$L_G \le \frac{L_Q B_K + B_Q L_K}{\sqrt{d_{\mathrm{head}}}\,\vartheta_{\min}}. \tag{38}$$

Indeed, $Q'K'^\top - QK^\top = (Q'-Q)K'^\top + Q(K'-K)^\top$, and the product norm inequality yields the bound. For an axis of length n_a, row stochasticity implies $\|\Pi\|_2 \le \sqrt{n_a}$. Expanding $\Pi'V' - \Pi V = (\Pi' - \Pi)V' + \Pi(V' - V)$ then gives

$$\mathrm{Lip}(\Pi V) \le \tfrac{1}{2}B_V L_G + \sqrt{n_a}L_V =: L_A. \tag{39}$$

After concatenating heads and applying W_O, a sufficient bound is $L_T \le \|W_O\|_2(\sum_r L_{A,r}^2)^{1/2}$. Let L_F bound the composite $\mathrm{FFN}_a \circ \mathrm{LN}$. The actual sequential residual structure in Equation (17) satisfies

$$\mathrm{Lip}(\mathrm{Axis}_a) \le (1 + |\lambda_{\mathrm{ff}}|L_F)(1 + |\lambda_{\mathrm{tr}}|L_T). \tag{40}$$

This follows by first bounding $Z \mapsto U$ and then $U \mapsto U + \lambda_{\mathrm{ff}}\,\mathrm{FFN}_a(\mathrm{LN}(U))$. The product includes the interaction between the two sequential residual updates. Positive normalization stabilizers, finite parameters, and bounded projected features suffice for these constants on the chosen domain. The implemented temperature floor 0.35 bounds the content amplification factor by $1/0.35$. Residual scales modulate the two amplification factors; the full bound also depends on projection norms, feature bounds, and the fixed graph bias.

B.6 Triangular exchange: inverse, volume, and recoverability

Absorb the fixed gates and normalization into maps $f(S)$ and $g(L')$. Equations (20)–(21) become $L' = L + f(S)$ and $S' = S + g(L')$.

Proof of Proposition 6.

Proof. Substitution recovers both original streams, proving a two-sided inverse without requiring f or g to be invertible. The Jacobians of the successive shears are

$$J_1 = \begin{bmatrix} I & Df(S) \\ 0 & I \end{bmatrix}, \qquad J_2 = \begin{bmatrix} I & 0 \\ Dg(L') & I \end{bmatrix}. \tag{41}$$

Their determinants are one, so $\det(J_2 J_1) = 1$. If two outputs coincide, applying the inverse makes their inputs coincide, proving the collision statement. □

For measurable maps with measurable inverse, the sigma-fields generated by the pre- and post-exchange pairs coincide, so their optimal conditional predictions of any integrable target coincide. If the joint pair has a density, the maps are sufficiently regular, and differential entropies are finite, the change-of-variables formula further gives $h(L', S') = h(L, S) + \mathbb{E}\log|\det J| = h(L, S)$. The deterministic exchange therefore preserves the recoverability and local volume of the joint stream pair. Its directional sensitivity is governed by the derivatives of f and g. The triangular construction follows the coupling principle of Dinh et al. (2017); Gomez et al. (2017).

B.7 Horizon-specific functionals and linear bottlenecks

Flatten the encoded history available to a linear readout into $u \in \mathbb{R}^D$. With future context fixed, the history-dependent part of Equation (23) is a family $f_h(u) = w_h^\top u + b_h$; the separate history and innovation terms can be combined by concatenating their features and absorbing their scales into w_h.

Proof of Proposition 7.

Proof. Equality of the two linear maps on all inputs is equivalent to equality of their matrices. Each row of AC lies in the row space of C, and the reverse inclusion condition supplies the coefficients forming A. Matrix factorization through a k-dimensional space is possible exactly when the rank is at most k. □

For example, two horizon targets u_1 and u_2 require rank two on $\mathbb{R}^2$, so a single shared scalar linear contraction cannot represent both. Horizon-specific full-history weights permit such distinct dependence patterns without that fixed bottleneck. For a shared linear contraction, the sufficient width is determined by the readout rank. The horizon-specific weights represent these distinct linear functionals directly, while the preceding encoder learns the features to which they are applied.

C Complete Horizon Results

Table 7: PeMS04 horizon errors (mean ± standard deviation, three seeds).

Step	MAE	RMSE	Step	MAE	RMSE
1	16.362±0.046	26.952±0.109	7	18.316±0.096	31.184±0.383
2	16.894±0.069	28.161±0.215	8	18.486±0.096	31.488±0.377
3	17.318±0.080	29.097±0.289	9	18.650±0.089	31.755±0.354
4	17.637±0.090	29.799±0.347	10	18.813±0.087	31.994±0.339
5	17.899±0.092	30.357±0.357	11	18.976±0.080	32.199±0.323
6	18.111±0.094	30.798±0.375	12	19.177±0.071	32.429±0.270

Table 8: PeMSD7(M) horizon errors (mean ± standard deviation, three seeds).

Step	MAE	RMSE	Step	MAE	RMSE
1	1.272±0.000	2.195±0.002	7	2.807±0.010	5.713±0.030
2	1.776±0.002	3.244±0.006	8	2.902±0.011	5.938±0.032
3	2.111±0.003	4.014±0.010	9	2.982±0.011	6.113±0.035
4	2.356±0.005	4.609±0.016	10	3.051±0.010	6.257±0.032
5	2.543±0.007	5.069±0.021	11	3.112±0.007	6.376±0.029
6	2.690±0.009	5.429±0.029	12	3.170±0.005	6.478±0.028