CAST: STRUCTURAL INDUCTIVE BIASES FOR ZERO-INFLATED SEQUENCE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-inflated sequences pose a distinctive learning challenge: exact zeros may encode inactivity rather than ordinary low-magnitude observations, yet many sequence models process both on the same numerical scale. We introduce CAST, a structural learning framework that separates observed activity from active-state magnitude and uses the resulting temporal organization as an inductive bias for sparse sequence prediction. CAST first learns a temporally ordered graph through a hurdle-style two-part structural model; under the stated structural assumptions this graph admits a causal interpretation, while otherwise it represents directed temporal dependence. The learned graph is then embedded in hyperbolic space to obtain hierarchy-aware positional representations, which are transformed into sample-adaptive gates that modulate nonlinear basis functions in a Kolmogorov-Arnold Network. Our analysis characterizes the population target of the two-part likelihood, the score shift induced by fitting inactive zeros with a continuous objective, and complementary guarantees for hyperbolic representation, radial ordering, realized gate energy, and gate-conditioned sensitivity. Controlled synthetic studies isolate these mechanisms, while real-world benchmarks demonstrate consistent predictive gains over competitive statistical and deep sequence baselines.

1 INTRODUCTION

Sequential prediction is central to applications such as retail and e-commerce Fildes et al. (2022), where observations are often dominated by zeros and only intermittently contain informative activity Mao et al. (2024). Such sparsity may arise from cold-start behavior Schein et al. (2002), prolonged inactivity Fader et al. (2005), or other forms of intermittent engagement. The challenge is therefore not only limited information: an exact zero may represent a qualitatively different observation state from a small active value.

Existing approaches capture only part of this distinction. Statistical methods, including Croston's approach Croston (1972) and ZIP models Lambert (1992), explicitly distinguish occurrence from magnitude, while deep temporal models such as DeepAR Salinas et al. (2020), DeepSurv Katzman et al. (2018), and TimeMIL Chen et al. (2024) provide substantially greater nonlinear capacity. Classical intermittent-demand models, however, are not designed to learn task-adaptive temporal structure, whereas generic sequence architectures typically process the observed numerical sequence unless activity is modeled separately. We instead treat exact zeros as observed inactivity and nonzero values as realizations of an active-magnitude process. CAST therefore does not infer competing latent explanations for an observed zero; it asks whether structural learning can become more reliable when inactivity is modeled as an observation state rather than forced onto the continuous scale of active responses Seeger et al. (2016).

From a modeling perspective, this distinction matters because sparsity is not merely a reduction in sample size. It changes which observations should contribute to the continuous response objective, how temporal dependence is estimated, and how downstream capacity should be allocated across positions whose predictive relevance differs across samples and time.

This perspective leads to **CAST**, illustrated in Figure 1. CAST carries the activity–magnitude distinction from structural learning to representation and prediction. First, a hurdle-style ZI-SEM jointly models observed activity and active-state magnitude while estimating a temporally ordered

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.

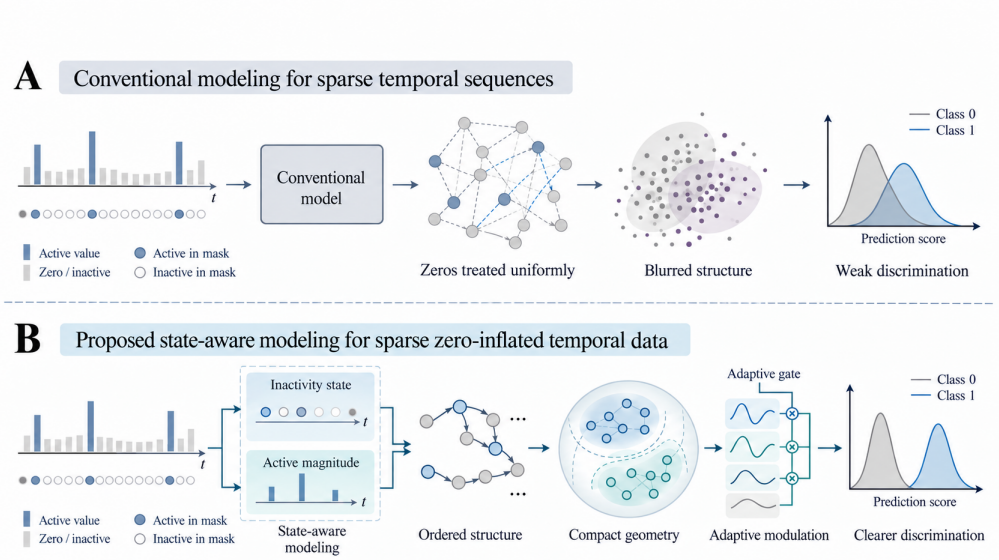


Figure 1: **Motivation of CAST. A.** Modeling inactive zeros on the active response scale can distort structural learning. **B.** CAST separates activity from magnitude and uses learned temporal structure to guide nonlinear prediction.

directed graph. Because the graph is learned from observational sequences, we interpret it by default as a directed temporal dependency structure; causal interpretation requires Assumption 3.1. Second, the learned graph is embedded in hyperbolic space to obtain structure-aware positional representations. Third, these representations generate sample-adaptive gates that modulate KAN basis activations across temporal positions. Since KAN weights remain shared, the gates regulate realized activation rather than remove stored parameters. Structure-conditioned attention and temporal max-pooling then aggregate the sequence representation for prediction.

We analyze the mechanisms connecting these stages and evaluate CAST through controlled synthetic studies and real-world sparse-sequence benchmarks. Our main contributions are:

- We formulate sparse sequence learning with a hurdle-style observation model that separates observed inactivity from active-state magnitude, avoiding the treatment of inactive zeros as ordinary continuous responses.
- We develop CAST, a three-stage framework combining zero-inflation-aware structural learning, hyperbolic representation of the learned temporal graph, and sample-adaptive modulation of KAN basis activations.
- We characterize the principal interfaces of CAST theoretically, covering population targeting under zero inflation, geometric and radial organization of the learned graph, realized gate energy, and gate-conditioned sensitivity.
- Controlled synthetic studies examine these mechanisms, while real-world benchmarks show consistent predictive improvements over the compared statistical and deep temporal baselines.

2 Related Work

We review three lines of work that are highly relevant to CAST: sparse sequence modeling, causal structure learning, and structural deep architectures. We postpone this section to **Appendix A** due to space limitation.

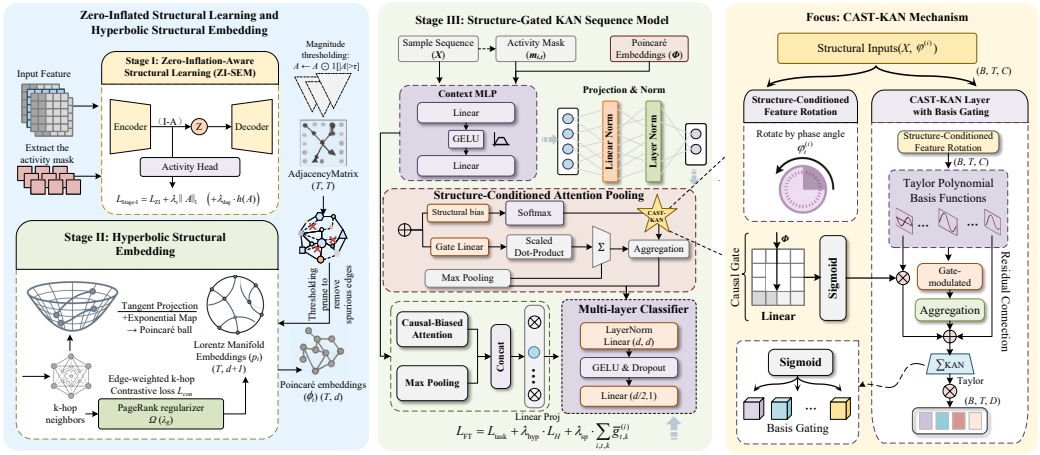


Figure 2: **The framework of CAST.** (A) Stage I separates activity from active magnitude and learns a temporally ordered structural graph. (B) Stage II embeds the graph in hyperbolic space with topology-aware contrastive learning and PageRank regularization. (C) Stage III uses the learned geometry to modulate CAST-KAN basis activations adaptively. Weak edges are removed by magnitude thresholding.

3 Methodology

3.1 Preliminary and Framework of CAST

Let $\mathcal{V} = \{v_t\}_{t=1}^{T}$ denote T temporally ordered features observed across N samples, collected in $\boldsymbol{X} \in \mathbb{R}^{N\times T}$. Exact zeros define the observed activity mask

$$m_{i,t} = \mathbb{1}[X_{i,t} \neq 0],$$

which reduces to $\mathbb{1}[X_{i,t} > 0]$ for the nonnegative real-world sequences considered here. CAST separates observed activity from active-state magnitude without introducing a latent label for zero.

Assumption 3.1 (Structural interpretation)**.** *A causal interpretation of the learned temporal graph requires a structural equation model consistent with the prescribed order, independent exogenous noises, no unobserved confounding among modeled variables, and identifiable parent structure within the stated functional family. Otherwise, the learned graph represents directed temporal dependence only.*

We associate this two-part observation model with a temporally ordered directed graph $\mathcal{G}^* = ([T], \boldsymbol{A}^*)$ Didelez & Pigeot (2001), where $A^*_{s,t} \neq 0$ only if $s < t$. For each position t,

$$\begin{aligned} m_{i,t} &\sim \mathrm{Ber}\big(1 - \pi_t(\boldsymbol{x}_{i,\mathrm{pa}(t)})\big)\,, \\ x_{i,t} \mid m_{i,t} &\sim \begin{cases} \mathcal{N}\big(f_t(\boldsymbol{x}_{i,\mathrm{pa}(t)}), \sigma_t^2\big)\,, & m_{i,t} = 1, \\ 0, & m_{i,t} = 0, \end{cases} \end{aligned} \tag{1}$$

where $\mathrm{pa}(t) = \{s : A^*_{s,t} \neq 0\}$, π_t is the inactivity probability, f_t is the active-state structural response, and $\sigma_t^2 > 0$ is the conditional variance. Here, "parent" denotes a directed structural predecessor and is causal only under Assumption 3.1. Equation (1) has hurdle semantics: zero mass is discrete and the continuous density applies only to active observations. For nonnegative magnitudes, the Gaussian is a working likelihood; other positive-support densities preserve the factorization. Section 4 gives the corresponding population target.

Figure 2 summarizes CAST: Stage I learns $\mathcal{G} = ([T], \boldsymbol{A})$, Stage II maps it to $\{\boldsymbol{\phi}_t\}_{t=1}^{T}$, and Stage III uses these representations to modulate CAST-KAN basis activations. Here $\mathcal{G}^*$ is the data-generating graph and $\mathcal{G}$ its learned counterpart.

3.2 Zero-Inflation-Aware Structural Learning (Stage I)

ZI-SEM learns temporally ordered dependence under Eq. (1); its coefficients are causal only under Assumption 3.1. Following Yu et al. (2019), we define a graph-gated context:

$$\boldsymbol{x}_i^{(<t)} := \boldsymbol{x}_i \odot \boldsymbol{u}_t, \qquad (\boldsymbol{u}_t)_s = \mathbb{1}[s < t], \qquad \boldsymbol{c}_{i,t}(\boldsymbol{A}) := \boldsymbol{A}_{:,t} \odot \boldsymbol{x}_i^{(<t)}, \qquad \hat{x}_{i,t} = f_{\theta,t}(\boldsymbol{c}_{i,t}(\boldsymbol{A}))\,. \tag{2}$$

Here $f_{\theta,t}$ predicts active magnitude from the learned predecessor context, and $\boldsymbol{A} \in \mathbb{R}^{T\times T}$ is temporally masked so admissible edges point forward:

$$\boldsymbol{A} \leftarrow \boldsymbol{A} \odot \boldsymbol{U}, \qquad U_{s,t} = \mathbb{1}[s < t].$$

Two-part likelihood. Observed activity and active magnitude enter distinct likelihood components:

$$p(x_{i,t}, m_{i,t} \mid \hat{x}_{i,t}, \hat{\pi}_{i,t}, \sigma_t) = \Big[(1-\hat{\pi}_{i,t})\mathcal{N}(x_{i,t} \mid \hat{x}_{i,t}, \sigma_t^2)\Big]^{m_{i,t}} \hat{\pi}_{i,t}^{\,1-m_{i,t}}, \tag{3}$$

$$\ell_{i,t}^{\mathrm{ZI}} = m_{i,t}\big[\log(1-\hat{\pi}_{i,t}) + \log\mathcal{N}(x_{i,t} \mid \hat{x}_{i,t}, \sigma_t^2)\big] + (1-m_{i,t})\log\hat{\pi}_{i,t}. \tag{4}$$

Here $\hat{x}_{i,t}$ follows Eq. (2). The activity branch shares this graph-gated predecessor context:

$$\hat{\pi}_{i,t} = \sigma(g_{\psi,t}(\boldsymbol{c}_{i,t}(\boldsymbol{A})))\,, \tag{5}$$

where $g_{\psi,t}$ is the inactivity head. Thus $\boldsymbol{A}$ enters both likelihood components through $\boldsymbol{c}_{i,t}(\boldsymbol{A})$, giving its coefficients gradients directly from magnitude and activity. The temporal mask excludes the target and future coordinates, preserving the ordered dependence assumed in Eq. (1). The continuous density is evaluated only when $m_{i,t} = 1$.

Structural pretraining objective. Let Θ_{ZI} collect the magnitude heads, activity heads, and likelihood parameters. Stage I minimizes

$$\mathcal{L}_{\mathrm{ZI}} := -\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\ell_{i,t}^{\mathrm{ZI}}, \qquad h(\boldsymbol{A}) := \mathrm{tr}\big(e^{\boldsymbol{A}\odot\boldsymbol{A}}\big) - T, \tag{6}$$

$$(\widehat{\Theta}_{\mathrm{ZI}}, \widehat{\boldsymbol{A}}) \in \arg\min_{\Theta_{\mathrm{ZI}},\,\boldsymbol{A}} \underbrace{\Big\{\mathcal{L}_{\mathrm{ZI}} + \lambda_s\,\|\boldsymbol{A}\|_1 + \lambda_{\mathrm{dag}}\Big[\frac{\rho}{2}h(\boldsymbol{A})^2 + \alpha_{\mathrm{dag}}h(\boldsymbol{A})\Big]\Big\}}_{\mathcal{L}_{\mathrm{DAG}}}. \tag{7}$$

For unrestricted $\boldsymbol{A}$, $h(\boldsymbol{A}) = 0$ is the differentiable acyclicity constraint of Zheng et al. (2018). Under the temporal mask, $\boldsymbol{A}$ is upper triangular and $h(\boldsymbol{A}) = 0$ identically; the augmented-Lagrangian block applies only to the optional unmasked implementation (Appendix D).

After convergence, λ_s-regularized weak edges are removed by

$$\boldsymbol{A} \leftarrow \boldsymbol{A} \odot \mathbb{1}[|\boldsymbol{A}| > \tau], \qquad \tau > 0. \tag{8}$$

The graph and Stage I likelihood parameters are then frozen.

3.3 Hyperbolic Structural Embedding (Stage II)

Stage II embeds the frozen graph $\mathcal{G} = ([T], \boldsymbol{A})$ in the d-dimensional hyperboloid model $\mathcal{L}^d = (\mathcal{H}^d, \boldsymbol{g}_l)$ Nickel & Kiela (2018), with

$$\boldsymbol{g}_l = \mathrm{diag}(-1, 1, \ldots, 1), \qquad \mathcal{H}^d = \Big\{\boldsymbol{p} \in \mathbb{R}^{d+1} : \langle \boldsymbol{p}, \boldsymbol{p}\rangle_l = -1,\ p^{(0)} > 0\Big\}\,,$$

and origin $\boldsymbol{p}_o = (1, \boldsymbol{0}_d)^\top$. For $\boldsymbol{p} = (p^{(0)}, \widetilde{\boldsymbol{p}})$, the Lorentzian inner product and induced distance are

$$\langle \boldsymbol{p}_s, \boldsymbol{p}_t\rangle_l = -p_s^{(0)}p_t^{(0)} + \widetilde{\boldsymbol{p}}_s^\top\widetilde{\boldsymbol{p}}_t, \qquad d_l(\boldsymbol{p}_s, \boldsymbol{p}_t) = \mathrm{arcosh}(-\langle \boldsymbol{p}_s, \boldsymbol{p}_t\rangle_l)\,.$$

The node embeddings $\{\boldsymbol{p}_t\}_{t=1}^T$ are learned with

$$\min_{\boldsymbol{p}_1,\ldots,\boldsymbol{p}_T \in \mathcal{H}^d} \mathcal{L}_{\mathcal{H}} = \frac{1}{T}\sum_{t=1}^{T}\left[\mathcal{L}_{\mathrm{con}}(\boldsymbol{p}_t) + \lambda_g\Omega(\boldsymbol{p}_t)\right], \tag{9}$$

where $\mathcal{L}_{\mathrm{con}}$ preserves graph proximity and Ω organizes structural centrality radially.

Weighted contrastive loss. For node v_t, let $\mathcal{P}_{t\to s}^{(\le k)}$ be the set of directed paths from t to s containing at most k edges, and define

$$\omega_{t,s}^{(k)} := \max_{p\in\mathcal{P}_{t\to s}^{(\le k)}} \prod_{(u,v)\in p} |A_{u,v}|\,, \qquad \max\varnothing := 0. \tag{10}$$

We set

$$\mathcal{N}_k^+(t) = \{s \neq t : \omega_{t,s}^{(k)} > 0\}, \qquad \mathcal{N}_k^-(t) = [T] \setminus \mathcal{N}_k^+(t) \setminus \{t\}.$$

The topology-aware contrastive term is

$$\mathcal{L}_{\text{con}}(\boldsymbol{p}_t) = -\sum_{s\in\mathcal{N}_k^+(t)} \omega_{t,s}^{(k)} \log \frac{e^{-d_l(\boldsymbol{p}_t,\boldsymbol{p}_s)}}{e^{-d_l(\boldsymbol{p}_t,\boldsymbol{p}_s)} + \sum_{s'\in\mathcal{N}_k^-(t)} e^{-d_l(\boldsymbol{p}_t,\boldsymbol{p}_{s'})}}. \tag{11}$$

For $k = 1$, $\omega_{t,s}^{(1)} = |A_{t,s}|$; for $k > 1$, it gives nonzero path-based weight to reachable indirect neighbors.

PageRank hierarchy regularizer. We compute PageRank on reverse edges, so walks follow learned predecessors. Define

$$d_t^{\text{in}} = \sum_{s=1}^{T} |A_{s,t}|\,,$$

and

$$P_{t,s} = \begin{cases} \dfrac{|A_{s,t}|}{d_t^{\text{in}}}, & d_t^{\text{in}} > 0, \\ \dfrac{1}{T}, & d_t^{\text{in}} = 0. \end{cases} \tag{12}$$

With teleportation parameter $w \in (0,1)$,

$$\widehat{\boldsymbol{P}} = (1-w)\boldsymbol{P} + \frac{w}{T}\mathbf{1}_{T\times T}, \qquad \widehat{\boldsymbol{P}}^\top\boldsymbol{\pi} = \boldsymbol{\pi}, \qquad \boldsymbol{\pi}^\top\mathbf{1}_T = 1. \tag{13}$$

We regularize predecessor centrality by

$$\Omega(\boldsymbol{p}_t) = \pi_t^{\text{PR}} d_l(\boldsymbol{p}_t, \boldsymbol{p}_o), \tag{14}$$

where π_t^{PR} is distinct from inactivity probability π_t. Proposition 4.3 gives the conditional radial ordering.

Optimization and projection. Coordinates are optimized by RSGD on $\mathcal{H}^d$ Bonnabel (2013) (Appendix E) and mapped to the Poincaré ball by

$$\boldsymbol{e}_t = f_d(\boldsymbol{p}_t) = \frac{(p_t^{(1)}, \ldots, p_t^{(d)})^\top}{p_t^{(0)} + 1}, \qquad \boldsymbol{\phi}_t := \boldsymbol{e}_t \in \mathcal{B}^d. \tag{15}$$

The graph $\boldsymbol{A}$ remains fixed, whereas $\{\boldsymbol{p}_t\}$ remain trainable during Stage III fine-tuning.

3.4 Structure-Gated KAN Sequence Model (Stage III)

Stage III uses the learned geometry for sample-dependent nonlinear modulation.

Sample-adaptive structural representation. For sample i, the positional representation is adapted to its activity pattern:

$$\boldsymbol{\phi}_t^{(i)} = \boldsymbol{\phi}_t \odot \sigma(\boldsymbol{W}_s\,\text{MLP}(\boldsymbol{m}_i) + \boldsymbol{b}_s)\,, \tag{16}$$

where $\boldsymbol{m}_i = (m_{i,1}, \ldots, m_{i,T})^\top$. This adapts graph-level position to each sample's sparsity pattern.

Structure-conditioned feature rotation. We map $\phi_t^{(i)}$ to phase angles

$$\theta_t^{(i)} = c\,\phi_t^{(i)}, \qquad c = \pi/4,$$

and apply blockwise $\mathrm{SO}(2)$ rotations to consecutive hidden channels:

$$\begin{pmatrix} [\tilde{h}_t^{(\ell)}]_{2m} \\ [\tilde{h}_t^{(\ell)}]_{2m+1} \end{pmatrix} = \begin{pmatrix} \cos\theta_{t,m}^{(i)} & -\sin\theta_{t,m}^{(i)} \\ \sin\theta_{t,m}^{(i)} & \cos\theta_{t,m}^{(i)} \end{pmatrix} \begin{pmatrix} [h_t^{(\ell)}]_{2m} \\ [h_t^{(\ell)}]_{2m+1} \end{pmatrix}, \quad m = 0, \ldots, M-1, \tag{17}$$

where $M = \min\{d, \lfloor C/2 \rfloor\}$; unmatched channels are left unchanged. Each rotation is orthogonal, so the complete operation is norm-preserving.

Adaptive structure-conditioned gate. Each CAST-KAN layer produces per-basis activation weights:

$$g_t^{(i,\ell)} = \sigma\left(W_g^{(\ell)} \phi_t^{(i)} + b_g^{(\ell)}\right) \in (0,1)^K, \tag{18}$$

where $W_g^{(\ell)} \in \mathbb{R}^{K \times d}$ and $b_g^{(\ell)} \in \mathbb{R}^K$. The gates modulate realized activity without deleting shared KAN parameters.

CAST-KAN layer. For output channel j at temporal position t,

$$h_{j,t}^{(\ell+1)} = \sum_{c=1}^{C} \sum_{k=1}^{K} w_{j,c,k}^{(\ell)} g_{t,k}^{(i,\ell)} \varphi_k\left([\tilde{h}_t^{(\ell)}]_c\right) + b_j^{(\ell)}, \tag{19}$$

where $w_{j,c,k}^{(\ell)}$ and $b_j^{(\ell)}$ are shared KAN parameters and φ_k is the k-th basis function; structure therefore changes realized basis contributions, not stored parameters.

Dual-path structure-conditioned pooling. After L CAST-KAN layers, attention and temporal max-pooling aggregate the sequence:

$$h_{\mathrm{attn}}^{(i)} = \sum_{t=1}^{T} \alpha_t^{(i)} h_t^{(L)}, \qquad \alpha_t^{(i)} = \frac{\exp\left((q_{\mathrm{cls}})^\top k_t / \sqrt{C} + w_b^\top \phi_t^{(i)}\right)}{\sum_{t'} \exp\left((q_{\mathrm{cls}})^\top k_{t'} / \sqrt{C} + w_b^\top \phi_{t'}^{(i)}\right)}, \tag{20}$$

where $k_t = W_k h_t^{(L)}$ and $w_b^\top \phi_t^{(i)}$ is an unconstrained structural bias.

The final representation is

$$h_{\mathrm{final}}^{(i)} = W_p \left[h_{\mathrm{attn}}^{(i)} \parallel \max_t h_t^{(L)}\right], \tag{21}$$

followed by an MLP classifier producing $\hat{y}_i$.

Training objective. To address class imbalance, we combine focal reweighting with label smoothing:

$$\begin{aligned} y_i^{\mathrm{ls}} &= 0.97 y_i + 0.015, & p_i &= \sigma(\hat{y}_i), \\ p_{i,\mathrm{true}} &= y_i p_i + (1-y_i)(1-p_i), & \alpha_{i,\mathrm{true}} &= \alpha y_i + (1-\alpha)(1-y_i), \end{aligned} \tag{22}$$

with focal weight

$$\omega_i = \alpha_{i,\mathrm{true}} \left(1 - p_{i,\mathrm{true}}\right)^\gamma. \tag{23}$$

The task loss is

$$\mathcal{L}_{\mathrm{task}} = -\frac{1}{N} \sum_{i=1}^{N} \omega_i \left[y_i^{\mathrm{ls}} \log p_i + (1 - y_i^{\mathrm{ls}}) \log(1 - p_i)\right]. \tag{24}$$

CAST is trained in two phases. Stage I learns $(\Theta_{\mathrm{ZI}}, A)$ with Eq. (7) and freezes the thresholded graph and likelihood parameters. Stages II–III then optimize

$$\mathcal{L}_{\mathrm{FT}} = \mathcal{L}_{\mathrm{task}} + \lambda_{\mathrm{hyp}} \mathcal{L}_{\mathcal{H}} + \frac{\lambda_{\mathrm{sp}}}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{k=1}^{K} \bar{g}_{t,k}^{(i)}, \qquad \bar{g}_{t,k}^{(i)} := \frac{1}{L} \sum_{\ell=1}^{L} g_{t,k}^{(i,\ell)}. \tag{25}$$

During fine-tuning, hyperbolic coordinates and downstream modules remain trainable while the graph is fixed; thus neither $\mathcal{L}_{\mathrm{ZI}}$ nor $\lambda_s \|A\|_1$ enters $\mathcal{L}_{\mathrm{FT}}$. Appendix S gives the full optimization schedule.

4 THEORETICAL CHARACTERIZATION OF CAST

Scope. We characterize five interfaces of CAST: hurdle population targeting, hyperbolic representation, PageRank-induced radial ordering, realized gate energy, and gate-conditioned sensitivity. The results are modular rather than an end-to-end recovery guarantee; causal interpretation still requires the structural conditions stated earlier. See Appendix C for the proof roadmap.

Assumption 4.1 (Standing conditions)**.** *The hurdle law satisfies $\mathbb{P}(m_t = 0 \mid \boldsymbol{z}) = \pi^*(\boldsymbol{z})$ and $x_t \mid (m_t = 1, \boldsymbol{z}) \sim f^*(\cdot \mid \boldsymbol{z})$, with non-degenerate activity probabilities. The candidate models obey standard differentiability and identification conditions. For the downstream analysis, the basis functions are L_φ-Lipschitz, $\|\boldsymbol{W}^{(\ell)}\|_2 \le B_w$, and preceding transformations remain non-expansive.*

Population targeting.

Proposition 4.1 (Hurdle target)**.** *Under the standing conditions,*

$$\Delta\mathcal{R}_{\mathrm{ZI}} = D_{\mathrm{KL}}\big(\mathrm{Ber}(1-\pi^*)\|\mathrm{Ber}(1-\pi_\psi)\big) + (1-\pi^*)D_{\mathrm{KL}}\big(f^*\|f_\theta\big) \ge 0, \tag{26}$$

$$\mathbb{E}[S_{\mathrm{G}}(\theta^*) \mid \boldsymbol{z}] = -\frac{\pi^*\mu_{\theta^*}\partial_\theta\mu_{\theta^*}}{\sigma_t^2}, \qquad \mathbb{E}[S_{\mathrm{ZI}}(\psi^*, \theta^*) \mid \boldsymbol{z}] = 0. \tag{27}$$

Accordingly, the hurdle objective targets the mixed observation law, whereas direct Gaussian fitting can shift the population score. See Appendix D.

Hyperbolic representation. Let $D_{\mathcal{M}}^{\max}(\mathcal{G})$ denote worst-case multiplicative metric distortion as defined in Appendix G.

Proposition 4.2 (Tree-approximation transfer)**.** *If a tree metric satisfies*

$$d_{\mathcal{G}}(u,v) \le d_{\mathcal{T}}(u,v) \le d_{\mathcal{G}}(u,v) + \Delta_T, \tag{28}$$

then, under the stated graph-growth condition,

$$D_{\mathbb{H}}^{\max} \le (1+\varepsilon)(1+\Delta_T), \qquad D_{\mathbb{E}}^{\max} \ge \left[\frac{(c_b T)^{1/d} - 1}{2r_T}\right]^{1/2}. \tag{29}$$

This is a representation, not optimization, guarantee. See Appendix G.

PageRank–radius ordering.

Proposition 4.3 (Conditional PageRank ordering)**.** *Under matched strictly decreasing outward radial forces,*

$$\pi_s^{\mathrm{PR}} > \pi_t^{\mathrm{PR}} \Longrightarrow r_s^* < r_t^* \Longrightarrow \|\boldsymbol{e}_s\|_2 < \|\boldsymbol{e}_t\|_2. \tag{30}$$

The claim is pairwise and conditional, rather than a global hierarchy. See Appendix H.

Gate energy. Let $q = r/T$, and let E, E^{in}, and η^2 denote total, on-support, and off-support gate energies, respectively.

Proposition 4.4 (Support–leakage)**.**

$$E = qE^{\mathrm{in}} + (1-q)\eta^2 \le q + (1-q)\eta^2, \qquad 1 - E \ge (1-q)(1-\eta^2). \tag{31}$$

Moreover, at most $K(T-r)\eta^2/\tau^2$ off-support gates exceed τ.

The result quantifies realized gate energy, not parameter removal. See Appendix I.

Gate-conditioned sensitivity. Let $\gamma_{i,t,\ell} = \|\boldsymbol{g}_t^{(i,\ell)}\|_2/\sqrt{K}$.

Proposition 4.5 (Gate-conditioned sensitivity)**.** *Under the standing conditions,*

$$\mathrm{Lip}(G_{i,t}^{(\ell)} \mid g) \le B_w L_\varphi \sqrt{K}\,\gamma_{i,t,\ell}, \qquad \mathrm{Lip}(H_{i,t}^{(1:L)} \mid g) \le \prod_{\ell=1}^{L}\Big(1 + B_w L_\varphi \sqrt{K}\,\gamma_{i,t,\ell}\Big). \tag{32}$$

Further composition and local-stability results appear in Appendix J.

Summary. Supporting derivations appear in Appendix D and Appendices G–J, with a consolidated summary in Appendix K and matched synthetic evaluation in Appendix L.

5 Numerical Experiments

5.1 Validation with Synthetic Datasets

We use controlled synthetic experiments to examine the mechanisms isolated in Section 4. Each construction is matched to the assumptions of its corresponding proposition. When a simulator introduces an additional effect, notably predictor masking in the hurdle stress test, we separate that effect from the theorem-level mechanism. Complete generators, protocols, and numerical results are reported in Appendix L.

5.2 Experimental Results with Real-World Datasets

Datasets. We evaluate CAST on five retail and e-commerce benchmarks. CDNOW and Retail include future-value and three-month churn prediction Chen et al. (2012); Fader et al. (2005); Instacart Instacart (2017) includes future activity and four-week churn prediction; Sales-Weekly Makridakis et al. (2022) covers future sales-decline risk and seasonality; and Ta-Feng Rendle et al. (2010) covers future risk and repurchase. After temporal aggregation, input sparsity ranges from 28.22% to 84.53% (Table 5). Every label is computed from a non-overlapping future target window. The 70:15:15 partitions are entity-disjoint and ordered by prediction cutoff whenever possible. Label thresholds and preprocessing statistics are estimated using training entities only (Appendix M). The resulting binary tasks have an average positive ratio of approximately 35.0%; infrastructure details are reported in Appendix S.

Baselines. We compare CAST with 12 baselines spanning statistical, neural, and modern sequence modeling strategies. The statistical group contains Croston Croston (1972), SBA Syntetos & Boylan (2005), ZIP Lambert (1992), and Hurdle Mullahy (1986). Neural baselines include DSN Xiao et al. (2022), SoftShape Liu et al. (2025a), TimeMIL Chen et al. (2024), and InterpGN Wen et al. (2025). We further compare with UniTS Gao et al. (2024), Mamba Gu & Dao (2024), MambaSL Jung & Kim (2026), and GRU-D Che et al. (2018), covering sequence, state-space, and missingness-aware architectures. Baseline descriptions and complexity analyses are provided in Appendices N and P.

Main Results. Table 1 reports Test F1 across ten task configurations. CAST achieves the highest overall average at 88.25%, exceeding the strongest competing average, MambaSL at 85.94%, by 2.31 points. The statistical baselines remain near the low-82% range, while stronger deep sequence models narrow the gap without matching the complete model. UniTS obtains an average F1 of 77.88%, suggesting that broad time-series representation learning alone is insufficient for these sparse observation regimes. Together with the controlled ablations, the results support the design sequence underlying CAST: separate inactivity from active magnitude, learn temporally ordered structure under that observation model, and use the learned structure to regulate the downstream nonlinear predictor. Complementary Test Accuracy results are reported in Appendix O.

Ablation Study. To isolate the contribution of the major architectural components, we restrict the ablation study to the five primary risk-classification tasks. Table 2 compares full CAST with four variants obtained by removing or replacing one component at a time.

Table 2: **Ablation results on primary tasks (F1 %).** Best results in **bold**; second-best underlined.

Variant	Retail	CDNOW	Instacart	Sales-Weekly	Ta-Feng
w/o ZI-SEM Structure	91.33±1.12	58.86±0.46	95.35±0.31	<u>96.65</u>±2.42	81.40±0.23
w/o Hyperbolic	<u>93.02</u>±0.08	94.79±0.45	<u>97.30</u>±0.18	<u>96.65</u>±2.42	94.51±0.34
w/o Sparsity Gate	92.03±1.29	92.26±0.36	95.68±0.07	96.26±1.89	<u>95.09</u>±0.37
w/o Structural Attn	92.21±0.80	<u>95.17</u>±0.20	97.18±0.11	<u>96.65</u>±2.42	94.68±0.50
CAST (Full)	**94.14**±0.13	**95.65**±0.26	**97.52**±0.18	**97.00**±2.12	**95.20**±0.44

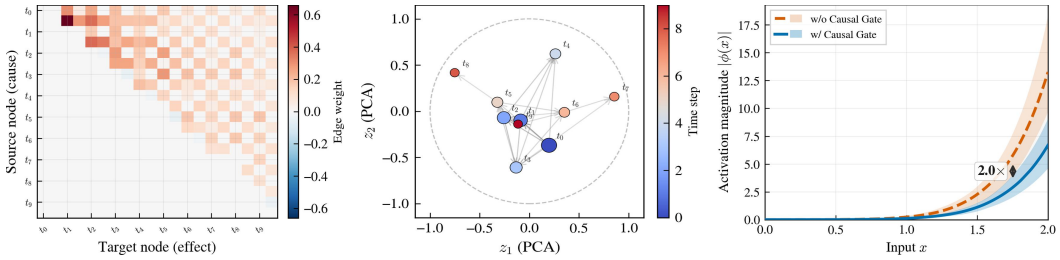
The full CAST model achieves the strongest result in every column of Table 2. Removing structural learning causes the largest degradation, most visibly on CDNOW, where F1 falls from 95.65% to 58.86%. Removing the sparsity gate also weakens performance, supporting the predictive value of sample-dependent basis modulation. Replacing hyperbolic geometry with its Euclidean counterpart or disabling structure-conditioned attention produces smaller but systematic regressions. These comparisons isolate the

Table 1: **Main results: Test F1-Score (↑) on different tasks across datasets.** All results are reported as percentages (%). We benchmark against statistical and neural models. The best results are in **bold**, and the second-best results are underlined.

Model	*Retail*		*CDNOW*		*Instacart*		*Sales-Weekly*		*Ta-Feng*		*Overall*	
	Risk	Churn	Risk	Churn	Risk	Churn	Risk	Seasonality	Risk	Repurchase	Avg (%)	Δ
Croston	88.72±0.48	72.28±0.92	77.61±1.39	91.54±0.21	81.66±0.31	64.48±0.27	96.59±1.60	96.57±0.97	72.90±0.19	72.73±0.23	81.51	-6.74
SBA	88.72±0.48	72.28±0.92	77.60±1.40	91.54±0.21	81.66±0.32	64.48±0.28	96.59±1.60	96.57±0.97	72.90±0.19	72.72±0.24	81.51	-6.74
ZIP	89.90±0.62	72.34±1.29	74.37±0.69	91.54±0.16	86.32±0.40	64.40±0.27	96.59±1.60	98.43±0.63	72.15±0.15	71.59±0.09	81.76	-6.49
Hurdle	90.82±0.94	72.53±0.92	79.16±0.92	91.58±0.18	87.88±0.11	64.41±0.27	96.26±1.89	98.70±0.97	71.19±0.12	71.67±0.04	82.42	-5.83
DSN	89.39±0.52	71.48±0.79	87.78±0.86	91.40±0.16	96.16±0.14	63.81±0.35	91.34±0.87	97.66±1.08	87.44±0.61	74.04±0.34	85.05	-3.20
SoftShape	86.19±1.94	69.71±1.07	85.23±0.58	91.17±0.17	88.88±6.03	63.24±0.46	89.13±0.85	95.28±2.42	86.73±0.84	73.91±0.32	82.95	-5.30
TimeMIL	91.54±1.08	71.89±1.14	85.71±1.09	91.49±0.13	95.33±0.40	64.29±0.22	96.57±2.48	99.22±0.64	83.25±0.45	72.92±1.55	85.22	-3.03
InterpGN	89.55±0.22	72.40±0.72	88.76±0.40	91.40±0.23	96.85±0.16	64.22±0.41	93.52±1.47	99.47±0.37	88.41±1.00	74.00±0.34	85.86	-2.39
UniTS	91.07±0.22	71.65±1.26	78.47±1.43	91.51±0.12	51.85±0.00	56.52±4.08	91.31±0.81	96.35±0.78	78.42±1.23	71.60±2.12	77.88	-10.37
Mamba	89.21±0.35	71.37±0.93	85.37±0.18	91.40±0.04	95.33±0.38	63.80±0.15	89.63±0.53	97.63±0.63	86.69±0.45	73.85±0.27	84.43	-3.82
MambaSL	91.90±0.53	72.31±1.35	88.48±0.13	91.43±0.14	97.11±0.08	64.44±0.32	93.90±2.32	98.69±0.37	87.18±0.31	73.93±0.37	85.94	-2.31
GRU-D	92.13±0.36	72.68±1.30	94.13±0.46	91.68±0.15	97.23±0.11	64.53±0.28	96.61±2.44	97.67±1.05	75.45±1.16	74.05±0.36	85.62	-2.63
CAST (Ours)	**94.14**±0.13	**72.84**±1.17	**95.65**±0.26	**91.74**±0.18	**97.52**±0.18	**64.55**±0.34	**97.00**±2.12	**99.74**±0.37	**95.20**±0.44	**74.17**±0.45	**88.25**	–

predictive contribution of the corresponding modules; they do not establish that every recovered edge is a data-generating causal effect. Parameter sensitivity is reported in Appendix Q.

Illustration of the Learned Structural Pipeline. Using Retail as a representative example, we visualize how learned temporal structure enters the predictor, from the estimated adjacency to its hyperbolic representation and the resulting modulation of KAN basis activity.



(a) Learned Temporal DAG (b) Poincaré Disk Embedding (c) Gated KAN Activations

Figure 3: **Illustration of the CAST structural pipeline.** From left to right: the temporally masked learned adjacency, its Poincaré representation, and the resulting modulation of KAN basis activity. These panels visualize internal representations of the trained model; causal interpretation of the learned adjacency requires Assumption 3.1.

A panel-by-panel interpretation, together with latent-space and feature-influence analyses, is provided in Appendix R.

6 Conclusion

CAST treats zero inflation as a change in observation state rather than merely a numerical sparsity pattern. This premise is carried through the architecture: ZI-SEM separates inactivity from active magnitude while learning temporally ordered structure; hyperbolic embedding organizes that structure geometrically; and sample-adaptive gates translate the resulting representation into controlled KAN basis activity. Under Assumption 3.1, the Stage I graph admits the corresponding causal interpretation; otherwise it remains a directed temporal dependency structure. The theory follows the same progression, characterizing population targeting under the hurdle observation law, representation of tree-approximable graph metrics, PageRank-induced radial organization under matched forces, realized gate energy, and conditional sensitivity of the nonlinear branch and residual wrapper. Controlled synthetic studies probe these mechanisms under matched conditions, while real-world benchmarks evaluate the predictive utility of the complete system. Together, the results show how observation-state modeling, structural geometry, and adaptive nonlinear modulation can be combined without conflating their distinct roles. Extending these results toward finite-sample structural recovery and whole-network stability remains an important direction.

AI USE STATEMENT

Generative AI tools were used to assist with LaTeX migration, language editing, and checks of mathematical exposition and internal consistency. The authors independently verify and take responsibility for all definitions, theoretical claims, derivations, implementation choices, data processing, experiments, and reported numerical results.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024, 2024.

Nishit M Bohra, Avantika R Patil, Priyanka V Deshmukh, and Aniket K Shahade. Heterogeneous graph kolmogorov–arnold networks for spatio-temporal wind power forecasting. *Smart Grids and Sustainable Energy*, 11(1):18, 2026.

Silvere Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

Zavareh Bozorgasl and Hao Chen. Wav-kan: Wavelet kolmogorov-arnold networks, 2024. *arXiv preprint arXiv:2405.12832*, 2024.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018.

Daqing Chen, Sai Laing Sain, and Kun Guo. Data mining for the online retail industry: A case study of rfm model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, 19(3):197–208, 2012.

Wei Chen, Qingfeng Xia, and JiaHui Sun. Legendre-kan: High accuracy ka network based on legendre polynomials. *OpenReview*, 2025.

Xiwen Chen, Peijie Qiu, Wenhui Zhu, Huayu Li, Hao Wang, Aristeidis Sotiras, Yalin Wang, and Abolfazl Razi. TimeMIL: Advancing multivariate time series classification via a time-aware multiple instance learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7190–7206. PMLR, 2024.

Shao-Ting Chiu, Siu Wun Cheung, Ulisses Braga-Neto, Chak Shing Lee, and Rui Peng Li. Free-rbf-kan: Kolmogorov-arnold networks with adaptive radial basis functions for efficient function learning. *arXiv preprint arXiv:2601.07760*, 2026.

John D Croston. Forecasting and stock control for intermittent demands. *Journal of the Operational Research Society*, 23(3):289–303, 1972.

Vanessa Didelez and Iris Pigeot. Causality: models, reasoning, and inference, 2001.

Hanaa El Afia, Said Ohamouddou, Raddouane Chiheb, and Abdellatif El Afia. Dynamic graph cnn with jacobi kolmogorov-arnold networks for 3d classification of point sets. *arXiv preprint arXiv:2506.06296*, 2025.

Peter S Fader, Bruce G S Hardie, and Ka Lok Lee. "counting your customers" the easy way: An alternative to the Pareto/NBD model. *Marketing science*, 24(2):275–284, 2005.

Robert Fildes, Shaohui Ma, and Stephan Kolassa. Retail forecasting: Research and practice. *International Journal of Forecasting*, 38(4):1283–1318, 2022.

Shanghua Gao, Teddy Koker, Owen Queen, Thomas Hartvigsen, Theodoros Tsiligkaridis, and Marinka Zitnik. UniTS: A unified multi-task time series model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024.

Abdullah Al Imran and Md Farhan Ishmam. FourierKAN outperforms MLP on text classification head fine-tuning. In *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*, 2024.

S Instacart. The instacart online grocery shopping dataset, 2017, 2017.

Yoo-Min Jung and Leekyung Kim. MambaSL: Exploring single-layer mamba for time series classification. In *The Fourteenth International Conference on Learning Representations*, 2026.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):1–12, 2018.

Diane Lambert. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.

Zhen Liu, Yicheng Luo, Boyuan Li, Emadeldeen Eldele, Min Wu, and Qianli Ma. Learning soft sparse shapes for efficient time-series classification. In *Forty-second International Conference on Machine Learning*, 2025a.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y. Hou, and Max Tegmark. KAN: Kolmogorov–Arnold networks. In *International Conference on Learning Representations*, 2025b.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m5 accuracy competition: Results, findings and conclusions. *International Journal of Forecasting*, 38(4):1346–1364, 2022.

Zhibin Mao, Baoying Yang, Ke Han, and Yuchen Pan. A zero-inflated poisson integer-valued autoregressive model with time-varying coefficients covariates. *Journal of Statistical Computation and Simulation*, 94(16):3700–3722, 2024.

John Mullahy. Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):341–365, 1986.

Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. In *International Conference on Machine Learning*, pp. 3779–3788. PMLR, 2018.

Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 811–820, 2010.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International journal of forecasting*, 36(3): 1181–1191, 2020.

Rik Sarkar. Low distortion delaunay embedding of trees in hyperbolic plane. In *International Symposium on Graph Drawing*, pp. 355–366. Springer, 2011.

Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 253–260, 2002.

Matthias W Seeger, David Salinas, and Valentin Flunkert. Bayesian intermittent demand forecasting for large inventories. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Rujia Shen, Yi Guan, Liangliang Liu, Yang Yang, Boran Wang, Chao Zhao, and Jingchi Jiang. Causal discovery from time-series data with short-term invariance-based convolutional neural networks. *Mathematics*, 13(24):3979, 2025.

SS Sidharth, Keerthana AR, Gokul R, and Anas KP. Chebyshev polynomial-based kolmogorov-arnold networks: An efficient architecture for nonlinear function approximation. *arXiv preprint arXiv:2405.07200*, 2024.

Gideon Stein, Maha Shadaydeh, Jan Blunk, Niklas Penzel, and Joachim Denzler. Causalrivers–scaling up benchmarking of causal discovery for real-world time-series. In *International Conference on Learning Representations*, 2025.

Aris A Syntetos and John E Boylan. The accuracy of intermittent demand estimates. *International Journal of Forecasting*, 21(2):303–314, 2005.

Yunshi Wen, Tengfei Ma, Ronny Luss, Debarun Bhattacharjya, Achille Fokoue, and Anak Agung Julius. Shedding light on time series classification using interpretability gated networks. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yannick Wölker, Christian Beth, Matthias Renz, and Arne Biastoch. Suster: Sparse unstructured spatio temporal reconstruction on traffic prediction. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pp. 1–10, 2023.

Qiao Xiao, Boqian Wu, Yu Zhang, Shiwei Liu, Mykola Pechenizkiy, Elena Mocanu, and Decebal Constantin Mocanu. Dynamic sparse network for time series classification: Learning what to "see". In *Advances in Neural Information Processing Systems*, volume 35, pp. 16849–16862, 2022.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. Dag-gnn: Dag structure learning with graph neural networks. In *International conference on machine learning*, pp. 7154–7163. PMLR, 2019.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Quan Zhou, Changhua Pei, Fei Sun, Jing Han, Zhengwei Gao, Dan Pei, Haiming Zhang, Gaogang Xie, and Jianhui Li. KAN-AD: Time series anomaly detection with Kolmogorov-Arnold networks. In *International Conference on Machine Learning*, 2025.

A Related Work

A.1 Sparse Sequence Classification

Sparse temporal prediction arises naturally when activity is intermittent and exact zeros dominate the observed sequence. Classical statistical methods remain useful. Croston's method Croston (1972) and ZIP models Lambert (1992), for example, separate occurrence from magnitude and provide principled baselines for sparse observations. Their relatively restrictive distributional assumptions, however, limit their ability to capture rich nonlinear temporal dependence.

More recent sequence models target irregularity, missingness, and highly skewed observations with greater capacity. Sparse Unstructured Spatio Temporal Reconstruction Wölker et al. (2023) studies state reconstruction under severe sparsity, while time-series foundation models employ transformations designed for skewed inputs Ansari et al. (2024). These approaches do not necessarily distinguish an observed inactive state from the conditional magnitude process used for structural learning. CAST focuses on this distinction. It treats the exact-zero pattern as an observed hurdle, models activity probability separately from active magnitude, and learns temporally ordered structure under the resulting two-part objective. The objective is not to infer competing latent explanations for each zero, but to prevent inactive observations from being absorbed into the continuous response model as ordinary low-magnitude realizations.

A.2 Temporal Structure Learning and Structural Architectures

Learning directed structure from observational sequences requires both an appropriate graph parameterization and a well-specified observation model. Continuous optimization methods such as NOTEARS Zheng et al. (2018) made DAG estimation amenable to gradient-based learning, while recent temporal approaches such as Short-Term Invariance-based Convolutional networks Shen et al. (2025) and benchmarks such as CausalRivers Stein et al. (2025) emphasize the role of distributional and observational assumptions. In our setting, fitting zero-masked responses with a continuous Gaussian objective generally changes the population target. Proposition 4.1 characterizes this shift through the population-risk decomposition and the corresponding score. The learned graph is therefore treated as a temporally ordered structural dependency by default, with a causal interpretation reserved for the conditions stated in Assumption 3.1.

A parallel line of work incorporates structural inductive bias into deep architectures. KANs replace conventional MLP transformations with learnable basis-function representations. HG-KAN models heterogeneous spatio-temporal dependencies Bohra et al. (2026), while KAN-AD applies related basis-function machinery to temporal anomaly detection Zhou et al. (2025). Hyperbolic geometry provides a natural representation space for hierarchical and tree-like structure Nickel & Kiela (2017); Chami et al. (2019). CAST connects these directions by combining zero-inflation-aware structural learning, hyperbolic graph representation, and sample-adaptive KAN basis gating for sparse sequence classification. The gate modulates realized basis activity per sample; because KAN parameters are shared across temporal positions, it represents activation control rather than stored-parameter sparsification.

B Notation Recap and Supplementary Conventions

The main notations are summarized in Table 3. We distinguish the generic mathematical sample space from the nonnegative real-world datasets used in Section 5.

- Data: $\boldsymbol{X} \in \mathbb{R}^{N\times T}$, activity mask $m_{i,t} = \mathbb{1}[X_{i,t} \neq 0]$, and labels $y_i \in \{0, 1\}$. For the nonnegative real-world datasets, $m_{i,t} = \mathbb{1}[X_{i,t} > 0]$.
- Structural graphs: $\mathcal{G}^* = ([T], \boldsymbol{A}^*)$ denotes the data-generating temporally ordered graph, with $A^*_{s,t} \neq 0$ only if $s < t$; $\mathcal{G} = ([T], \boldsymbol{A})$ denotes its learned counterpart. Both are read as directed temporal dependency structures unless Assumption 3.1 is invoked.
- Structural predecessor set: $\mathrm{pa}(t) = \{s : A^*_{s,t} \neq 0\}$ for $\mathcal{G}^*$. The directed edge count is $s_{\mathcal{E}} = |\mathcal{E}|$.

- Designated support: $\mathcal{S} \subseteq [T]$, $r = |\mathcal{S}|$, and $q = r/T$. In the gate-energy and sensitivity analyses, $\mathcal{S}$ is the synthetic target-support set. When graph incidence is required, we use $\mathcal{S}_{\mathcal{G}} := \{t : \mathrm{pa}(t) \cup \mathrm{ch}(t) \neq \emptyset\}$.
- ZI-SEM observation model: Eq. (1) in the main text, with inactivity probability $\pi_t(\boldsymbol{z})$ and active-state density $f_t(\cdot \mid \boldsymbol{z})$.
- Hyperboloid: $\mathcal{H}^d$ with Lorentz metric $\boldsymbol{g}_l = \mathrm{diag}(-1, 1, \ldots, 1)$ and origin $\boldsymbol{p}_o = (1, \mathbf{0}_d)^\top$, with distance
$$d_l(\boldsymbol{p}_s, \boldsymbol{p}_t) = \mathrm{arcosh}\big(-\langle \boldsymbol{p}_s, \boldsymbol{p}_t \rangle_l\big).$$
- Poincaré ball: $\mathcal{B}^d = \{\boldsymbol{e} \in \mathbb{R}^d : \|\boldsymbol{e}\|_2 < 1\}$, with
$$d_p(\boldsymbol{e}_s, \boldsymbol{e}_t) = \mathrm{arcosh}\left(1 + 2\frac{\|\boldsymbol{e}_s - \boldsymbol{e}_t\|_2^2}{(1 - \|\boldsymbol{e}_s\|_2^2)(1 - \|\boldsymbol{e}_t\|_2^2)}\right).$$
- Diffeomorphism: $f_d : \mathcal{H}^d \to \mathcal{B}^d$,
$$f_d(\boldsymbol{p}) = \frac{(p^{(1)}, \ldots, p^{(d)})^\top}{p^{(0)} + 1}.$$
- CAST-KAN: L layers, K basis functions $\{\varphi_k\}_{k=1}^K$, hidden width C, branch-weight bound B_w, and basis Lipschitz constant L_φ.
- Layer-specific gate:
$$\boldsymbol{g}_t^{(i,\ell)} = \sigma\Big(\boldsymbol{W}_g^{(\ell)} \boldsymbol{\phi}_t^{(i)} + \boldsymbol{b}_g^{(\ell)}\Big) \in (0,1)^K.$$

Table 3: Summary of main notations.

Notation	Description
N	Number of samples
T	Number of temporal positions/features
$\boldsymbol{A}^* \in \mathbb{R}^{T \times T}$	Data-generating temporally ordered adjacency
$\boldsymbol{A} \in \mathbb{R}^{T \times T}$	Learned temporally ordered adjacency
$s_{\mathcal{E}} = \lvert\mathcal{E}\rvert$	Number of directed edges
$r = \lvert\mathcal{S}\rvert$	Size of the designated support
$q = r/T$	Fraction of supported temporal positions
$\eta_{i,\ell} \in [0,1]$	RMS gate level outside $\mathcal{S}$
d	Hyperbolic positional-embedding dimension
C	CAST-KAN hidden dimension
K	Number of basis functions per layer
L	Number of CAST-KAN layers
B_w	Spectral-norm bound for flattened branch weights
$\pi_t(\boldsymbol{z}) \in (0,1)$	Conditional inactivity probability
$m_{i,t} \in \{0,1\}$	Observed activity indicator
$\boldsymbol{p}_t \in \mathcal{H}^d$	Hyperboloid coordinate
$\boldsymbol{e}_t \in \mathcal{B}^d$	Poincaré-ball coordinate
$\boldsymbol{\phi}_t \in \mathbb{R}^d$	Structural positional encoding
$\boldsymbol{g}_t^{(i,\ell)} \in (0,1)^K$	Sample-, position-, and layer-specific gate
$\gamma_{i,t,\ell} \in [0,1]$	$\lVert\boldsymbol{g}_t^{(i,\ell)}\rVert_2/\sqrt{K}$
$\boldsymbol{\pi}^{\mathrm{PR}} \in (0,1)^T$	Reverse-edge PageRank vector
$\varphi_k(\cdot)$	k-th KAN basis function
L_φ	Basis Lipschitz constant on the analysis domain

All standing assumptions used by the main propositions are stated in Section 4, except the causal-interpretation requirement in Assumption 3.1, which is stated with the problem formulation before Stage I. This appendix intentionally does not restate them. The notation above is only a recap; the proof sections below add local algebraic definitions when needed and never strengthen a main-text claim silently.

C Proof Roadmap and Scope of the Guarantees

The theoretical results isolate distinct interfaces of CAST rather than forming a single end-to-end recovery theorem. Proposition 4.1 characterizes the population target of the hurdle observation

model and is separate from finite-sample structural recovery. Proposition 4.2 begins from a fixed graph metric and studies hyperbolic representability under the stated tree-approximation condition. Proposition 4.3 characterizes the radial effect of reverse-edge PageRank under matched non-PageRank forces. Propositions 4.4 and 4.5 concern the downstream predictor, quantifying realized gate energy and gate-conditioned sensitivity.

These distinctions also determine the role of the experiments. The controlled synthetic studies probe the mechanisms associated with the corresponding theoretical results, whereas the real-world experiments evaluate the predictive utility of the complete system. In the synthetic experiment associated with Proposition 4.1, predictor coordinates are additionally masked, so the reported structural-recovery behavior also reflects effects beyond the response-side population calculation.

The assumptions are local to the results they support. Misspecification of the active-state density replaces the data-generating parameter in Proposition 4.1 by the corresponding pseudo-true population target. Proposition 4.2 requires the stated tree-metric approximation, Proposition 4.3 requires the common radial-force condition, and Propositions 4.4 and 4.5 depend on the realized gate pattern. Assumption 3.1 is required only for a causal interpretation of the Stage I graph; without it, the analysis concerns a temporally ordered structural dependency model.

D Methodological Identities Deferred from the Main Text

This appendix records algebraic consequences of the definitions used in Section 3. They are collected here to keep the main methodological development compact.

Exact acyclicity under the temporal mask. Let

$$U_{s,t} = \mathbb{1}[s < t], \qquad \boldsymbol{A} = \boldsymbol{A} \odot \boldsymbol{U}, \qquad \boldsymbol{B} := \boldsymbol{A} \odot \boldsymbol{A}. \tag{33}$$

Both $\boldsymbol{A}$ and $\boldsymbol{B}$ are strictly upper triangular. Hence $\boldsymbol{B}$ is nilpotent with index at most T,

$$\boldsymbol{B}^T = \boldsymbol{0}, \qquad \operatorname{tr}(\boldsymbol{B}^k) = 0, \qquad k = 1, \dots, T-1. \tag{34}$$

The matrix exponential therefore reduces to the finite series

$$\begin{aligned} \exp(\boldsymbol{B}) &= \boldsymbol{I} + \boldsymbol{B} + \frac{\boldsymbol{B}^2}{2!} + \cdots + \frac{\boldsymbol{B}^{T-1}}{(T-1)!}, \\ \operatorname{tr}(\exp(\boldsymbol{B})) &= \operatorname{tr}(\boldsymbol{I}) + \sum_{k=1}^{T-1} \frac{\operatorname{tr}(\boldsymbol{B}^k)}{k!} = T. \end{aligned} \tag{35}$$

Consequently,

$$h(\boldsymbol{A}) = \operatorname{tr}(\exp(\boldsymbol{A} \odot \boldsymbol{A})) - T = 0. \tag{36}$$

Thus the temporal parameterization itself guarantees acyclicity. The NOTEARS term vanishes throughout the masked feasible set and is needed only for the optional unrestricted implementation.

Hurdle semantics and the mixed observation space. For one temporal coordinate, consider the sample space

$$\mathcal{X}_{\mathrm{H}} = \big(\{0\} \times \{0\}\big) \cup \big(\{1\} \times \mathcal{X}_{\mathrm{act}}\big), \tag{37}$$

equipped with reference measure

$$\nu(dm, dx) = \delta_0(dm)\delta_0(dx) + \delta_1(dm)\lambda(dx), \tag{38}$$

where λ is Lebesgue measure on the active-response space. Conditional on structural context $\boldsymbol{z}$, the hurdle law has density

$$q_{\psi,\theta}(m, x \mid \boldsymbol{z}) = \pi_\psi(\boldsymbol{z})\, \mathbb{1}[m = 0, x = 0] + \big(1 - \pi_\psi(\boldsymbol{z})\big) f_\theta(x \mid \boldsymbol{z})\, \mathbb{1}[m = 1]. \tag{39}$$

On the support of ν, this is equivalently written as

$$q_{\psi,\theta}(m, x \mid \boldsymbol{z}) = \Big[\big(1 - \pi_\psi(\boldsymbol{z})\big) f_\theta(x \mid \boldsymbol{z})\Big]^m \pi_\psi(\boldsymbol{z})^{1-m}. \tag{40}$$

Hence

$$\log q_{\psi,\theta}(m, x \mid \boldsymbol{z}) = m \log\big(1 - \pi_\psi(\boldsymbol{z})\big) + m \log f_\theta(x \mid \boldsymbol{z}) + (1 - m) \log \pi_\psi(\boldsymbol{z}). \tag{41}$$

The inactive state is therefore represented by probability mass, whereas the active state is represented by a conditional density. For the real-world nonnegative sequences, the Gaussian active model used in the implementation is a working likelihood; the population identities below apply to the correctly specified active density.

Interpretation of the gate penalty. The KAN coefficients $w_{j,c,k}^{(\ell)}$ are shared model parameters, whereas

$$g_{t,k}^{(i,\ell)} = \sigma\Big([\boldsymbol{W}_g^{(\ell)} \boldsymbol{\phi}_t^{(i)} + \boldsymbol{b}_g^{(\ell)}]_k\Big) \in (0, 1) \tag{42}$$

is sample-, position-, and layer-dependent. Therefore

$$\mathcal{L}_{\text{gate}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{k=1}^{K} \bar{g}_{t,k}^{(i)}, \qquad \bar{g}_{t,k}^{(i)} = \frac{1}{L} \sum_{\ell=1}^{L} g_{t,k}^{(i,\ell)}, \tag{43}$$

penalizes realized activation rather than the number of stored KAN coefficients. Propositions 4.4 and 4.5 quantify this distinction through gate energy and conditional sensitivity.

E Hyperboloid Optimization Details

We consider the d-dimensional hyperboloid

$$\mathcal{H}^d = \Big\{ \boldsymbol{p} \in \mathbb{R}^{d+1} : \langle \boldsymbol{p}, \boldsymbol{p} \rangle_l = -1,\ p^{(0)} > 0 \Big\}, \qquad \boldsymbol{g}_l = \operatorname{diag}(-1, 1, \ldots, 1), \tag{44}$$

where

$$\langle \boldsymbol{p}, \boldsymbol{p} \rangle_l = \boldsymbol{p}^\top \boldsymbol{g}_l \boldsymbol{p}. \tag{45}$$

The hyperboloid is equipped with the Riemannian metric induced by the Lorentzian bilinear form

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle_l = -u^{(0)} v^{(0)} + \sum_{j=1}^{d} u^{(j)} v^{(j)}. \tag{46}$$

Equivalently,

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle_l = \boldsymbol{u}^\top \boldsymbol{g}_l \boldsymbol{v}. \tag{47}$$

Since $\langle \boldsymbol{p}, \boldsymbol{p} \rangle_l = -1$, differentiating the defining constraint along a smooth curve through $\boldsymbol{p}$ gives the tangent space

$$T_{\boldsymbol{p}} \mathcal{H}^d = \big\{ \boldsymbol{v} \in \mathbb{R}^{d+1} : \langle \boldsymbol{p}, \boldsymbol{v} \rangle_l = 0 \big\}. \tag{48}$$

The restriction of $\langle \cdot, \cdot \rangle_l$ to $T_{\boldsymbol{p}} \mathcal{H}^d$ is positive definite, and therefore

$$\|\boldsymbol{v}\|_l = \sqrt{\langle \boldsymbol{v}, \boldsymbol{v} \rangle_l} \tag{49}$$

defines the Riemannian norm of every $\boldsymbol{v} \in T_{\boldsymbol{p}} \mathcal{H}^d$.

Lemma E.1 (Lorentz-orthogonal tangent projection)**.** *For any $\boldsymbol{p} \in \mathcal{H}^d$ and $\boldsymbol{u} \in \mathbb{R}^{d+1}$, the Lorentz-orthogonal projection of $\boldsymbol{u}$ onto $T_{\boldsymbol{p}} \mathcal{H}^d$ is*

$$\Pi_{\boldsymbol{p}}(\boldsymbol{u}) = \boldsymbol{u} + \langle \boldsymbol{p}, \boldsymbol{u} \rangle_l \boldsymbol{p}. \tag{50}$$

Equivalently,

$$\boldsymbol{u} = \Pi_{\boldsymbol{p}}(\boldsymbol{u}) - \langle \boldsymbol{p}, \boldsymbol{u} \rangle_l \boldsymbol{p}, \tag{51}$$

where $\Pi_{\boldsymbol{p}}(\boldsymbol{u}) \in T_{\boldsymbol{p}} \mathcal{H}^d$ and the second term belongs to the Lorentz-normal space $\operatorname{span}\{\boldsymbol{p}\}$.

Proof. Using $\langle \boldsymbol{p}, \boldsymbol{p}\rangle_l = -1$, we obtain

$$\begin{aligned}\langle \boldsymbol{p}, \Pi_{\boldsymbol{p}}(\boldsymbol{u})\rangle_l &= \langle \boldsymbol{p}, \boldsymbol{u} + \langle \boldsymbol{p}, \boldsymbol{u}\rangle_l \boldsymbol{p}\rangle_l, \\ &= \langle \boldsymbol{p}, \boldsymbol{u}\rangle_l + \langle \boldsymbol{p}, \boldsymbol{u}\rangle_l \langle \boldsymbol{p}, \boldsymbol{p}\rangle_l, \\ &= \langle \boldsymbol{p}, \boldsymbol{u}\rangle_l - \langle \boldsymbol{p}, \boldsymbol{u}\rangle_l, \\ &= 0. \end{aligned} \tag{52}$$

Hence $\Pi_{\boldsymbol{p}}(\boldsymbol{u}) \in T_{\boldsymbol{p}}\mathcal{H}^d$. Moreover,

$$\boldsymbol{u} - \Pi_{\boldsymbol{p}}(\boldsymbol{u}) = -\langle \boldsymbol{p}, \boldsymbol{u}\rangle_l \boldsymbol{p}, \tag{53}$$

and therefore

$$\boldsymbol{u} - \Pi_{\boldsymbol{p}}(\boldsymbol{u}) \in \operatorname{span}\{\boldsymbol{p}\}. \tag{54}$$

This establishes the tangent–normal decomposition and therefore the projection formula. □

Lemma E.2 (Exponential map on the hyperboloid)**.** *Let $\boldsymbol{p} \in \mathcal{H}^d$ and $\boldsymbol{v} \in T_{\boldsymbol{p}}\mathcal{H}^d$. Define*

$$r = \|\boldsymbol{v}\|_l . \tag{55}$$

Equivalently,

$$r = \sqrt{\langle \boldsymbol{v}, \boldsymbol{v}\rangle_l}. \tag{56}$$

Then the Riemannian exponential map is

$$\exp_{\boldsymbol{p}}(\boldsymbol{v}) = \cosh(r)\boldsymbol{p} + \sinh(r)\frac{\boldsymbol{v}}{r}, \qquad r > 0, \tag{57}$$

with the continuous extension

$$\exp_{\boldsymbol{p}}(\boldsymbol{0}) = \boldsymbol{p}. \tag{58}$$

Proof. For $\boldsymbol{v} \neq \boldsymbol{0}$, define

$$\boldsymbol{\gamma}(s) = \cosh(sr)\boldsymbol{p} + \sinh(sr)\frac{\boldsymbol{v}}{r}, \qquad s \in \mathbb{R}. \tag{59}$$

Since $\langle \boldsymbol{p}, \boldsymbol{p}\rangle_l = -1$, $\langle \boldsymbol{p}, \boldsymbol{v}\rangle_l = 0$, and $\langle \boldsymbol{v}, \boldsymbol{v}\rangle_l = r^2$, we have

$$\begin{aligned}\langle \boldsymbol{\gamma}(s), \boldsymbol{\gamma}(s)\rangle_l &= -\cosh^2(sr) + \sinh^2(sr), \\ &= -1. \end{aligned} \tag{60}$$

Here the cross term vanishes because $\langle \boldsymbol{p}, \boldsymbol{v}\rangle_l = 0$, while $\langle \boldsymbol{v}, \boldsymbol{v}\rangle_l / r^2 = 1$.

Furthermore,

$$\boldsymbol{\gamma}(0) = \boldsymbol{p}, \tag{61}$$

and

$$\dot{\boldsymbol{\gamma}}(s) = r\sinh(sr)\boldsymbol{p} + \cosh(sr)\boldsymbol{v}. \tag{62}$$

Thus,

$$\dot{\boldsymbol{\gamma}}(0) = \boldsymbol{v}. \tag{63}$$

Differentiating once more gives

$$\ddot{\boldsymbol{\gamma}}(s) = r^2\cosh(sr)\boldsymbol{p} + r\sinh(sr)\boldsymbol{v}. \tag{64}$$

Hence,

$$\ddot{\boldsymbol{\gamma}}(s) = r^2\boldsymbol{\gamma}(s). \tag{65}$$

Because the Lorentz-normal space at $\boldsymbol{\gamma}(s)$ is $\operatorname{span}\{\boldsymbol{\gamma}(s)\}$, the ambient acceleration $\ddot{\boldsymbol{\gamma}}(s) = r^2\boldsymbol{\gamma}(s)$ has zero tangent component. Therefore $\boldsymbol{\gamma}$ is the geodesic satisfying $\boldsymbol{\gamma}(0) = \boldsymbol{p}$ and $\dot{\boldsymbol{\gamma}}(0) = \boldsymbol{v}$. Consequently,

$$\exp_{\boldsymbol{p}}(\boldsymbol{v}) = \boldsymbol{\gamma}(1), \tag{66}$$

and therefore

$$\exp_{\boldsymbol{p}}(\boldsymbol{v}) = \cosh(r)\boldsymbol{p} + \sinh(r)\frac{\boldsymbol{v}}{r}. \tag{67}$$

The case $\boldsymbol{v} = \boldsymbol{0}$ follows by continuity. □

Riemannian gradient. Let $\mathcal{L} : \mathcal{H}^d \to \mathbb{R}$ be differentiable, and let $\overline{\mathcal{L}}$ be any smooth extension of $\mathcal{L}$ to a neighborhood of $\mathcal{H}^d$ in $\mathbb{R}^{d+1}$. Denote its Euclidean coordinate gradient by $\nabla^E_{\boldsymbol{p}}\overline{\mathcal{L}}$. The corresponding ambient Lorentz gradient is

$$\nabla^L_{\boldsymbol{p}}\overline{\mathcal{L}} = \boldsymbol{g}_l^{-1}\nabla^E_{\boldsymbol{p}}\overline{\mathcal{L}}. \tag{68}$$

Since

$$\boldsymbol{g}_l^{-1} = \boldsymbol{g}_l, \tag{69}$$

we obtain

$$\nabla^L_{\boldsymbol{p}}\overline{\mathcal{L}} = \boldsymbol{g}_l\nabla^E_{\boldsymbol{p}}\overline{\mathcal{L}}. \tag{70}$$

Moreover,

$$\boldsymbol{g}_l^2 = \boldsymbol{I}_{d+1}. \tag{71}$$

Indeed, for every ambient direction $\boldsymbol{u} \in \mathbb{R}^{d+1}$,

$$\begin{aligned} D\overline{\mathcal{L}}(\boldsymbol{p})[\boldsymbol{u}] &= \left(\nabla^E_{\boldsymbol{p}}\overline{\mathcal{L}}\right)^\top \boldsymbol{u}, \\ &= \left(\boldsymbol{g}_l\nabla^E_{\boldsymbol{p}}\overline{\mathcal{L}}\right)^\top \boldsymbol{g}_l\boldsymbol{u}, \\ &= \langle \boldsymbol{g}_l\nabla^E_{\boldsymbol{p}}\overline{\mathcal{L}},\, \boldsymbol{u}\rangle_l, \\ &= \langle \nabla^L_{\boldsymbol{p}}\overline{\mathcal{L}},\, \boldsymbol{u}\rangle_l. \end{aligned} \tag{72}$$

The Riemannian gradient on $\mathcal{H}^d$ is the tangent component of the ambient Lorentz gradient:

$$\begin{aligned} \operatorname{grad}_{\boldsymbol{p}}\mathcal{L} &= \Pi_{\boldsymbol{p}}\left(\nabla^L_{\boldsymbol{p}}\overline{\mathcal{L}}\right), \\ &= \nabla^L_{\boldsymbol{p}}\overline{\mathcal{L}} + \langle \boldsymbol{p},\, \nabla^L_{\boldsymbol{p}}\overline{\mathcal{L}}\rangle_l \boldsymbol{p}, \\ &= \boldsymbol{g}_l\nabla^E_{\boldsymbol{p}}\overline{\mathcal{L}} + \langle \boldsymbol{p},\, \boldsymbol{g}_l\nabla^E_{\boldsymbol{p}}\overline{\mathcal{L}}\rangle_l \boldsymbol{p}. \end{aligned} \tag{73}$$

To verify the defining property of the Riemannian gradient, let $\boldsymbol{v} \in T_{\boldsymbol{p}}\mathcal{H}^d$. Then

$$\begin{aligned} D\mathcal{L}(\boldsymbol{p})[\boldsymbol{v}] &= D\overline{\mathcal{L}}(\boldsymbol{p})[\boldsymbol{v}], \\ &= \langle \nabla^L_{\boldsymbol{p}}\overline{\mathcal{L}},\, \boldsymbol{v}\rangle_l, \\ &= \langle \operatorname{grad}_{\boldsymbol{p}}\mathcal{L},\, \boldsymbol{v}\rangle_l. \end{aligned} \tag{74}$$

The final equality follows from $\langle \boldsymbol{p},\, \boldsymbol{v}\rangle_l = 0$. Consequently, the tangent gradient does not depend on the choice of the ambient extension.

Riemannian gradient update. For a step size $\eta > 0$, define

$$\boldsymbol{v}_{\boldsymbol{p}} = -\eta \operatorname{grad}_{\boldsymbol{p}}\mathcal{L}. \tag{75}$$

Then

$$\boldsymbol{v}_{\boldsymbol{p}} \in T_{\boldsymbol{p}}\mathcal{H}^d. \tag{76}$$

Define also

$$r_{\boldsymbol{p}} = \|\boldsymbol{v}_{\boldsymbol{p}}\|_l. \tag{77}$$

The Riemannian gradient step is

$$\begin{aligned} \boldsymbol{p}^+ &= \exp_{\boldsymbol{p}}(\boldsymbol{v}_{\boldsymbol{p}}), \\ &= \cosh(r_{\boldsymbol{p}})\boldsymbol{p} + \sinh(r_{\boldsymbol{p}})\frac{\boldsymbol{v}_{\boldsymbol{p}}}{r_{\boldsymbol{p}}}. \end{aligned} \tag{78}$$

Since

$$r_{\boldsymbol{p}} = \eta \left\|\operatorname{grad}_{\boldsymbol{p}}\mathcal{L}\right\|_l, \tag{79}$$

we obtain

$$\begin{aligned} \boldsymbol{p}^+ = &\cosh\Big(\eta \left\|\operatorname{grad}_{\boldsymbol{p}}\mathcal{L}\right\|_l\Big)\boldsymbol{p} \\ &- \sinh\Big(\eta \left\|\operatorname{grad}_{\boldsymbol{p}}\mathcal{L}\right\|_l\Big)\frac{\operatorname{grad}_{\boldsymbol{p}}\mathcal{L}}{\left\|\operatorname{grad}_{\boldsymbol{p}}\mathcal{L}\right\|_l}. \end{aligned} \tag{80}$$

When $\mathrm{grad}_{\boldsymbol{p}} \mathcal{L} = \mathbf{0}$, the update is defined by continuity as

$$\boldsymbol{p}^{+} = \boldsymbol{p}. \tag{81}$$

It remains to verify that the update stays on the upper sheet. For $r_{\boldsymbol{p}} > 0$, using $\langle \boldsymbol{p}, \boldsymbol{p} \rangle_l = -1$, $\langle \boldsymbol{p}, \boldsymbol{v}_{\boldsymbol{p}} \rangle_l = 0$, and $\langle \boldsymbol{v}_{\boldsymbol{p}}, \boldsymbol{v}_{\boldsymbol{p}} \rangle_l = r_{\boldsymbol{p}}^2$, we obtain

$$\begin{aligned} \langle \boldsymbol{p}^{+}, \boldsymbol{p}^{+} \rangle_l &= -\cosh^2(r_{\boldsymbol{p}}) + \sinh^2(r_{\boldsymbol{p}}), \\ &= -1. \end{aligned} \tag{82}$$

To establish the upper-sheet condition, write

$$\boldsymbol{p} = \begin{pmatrix} p^{(0)} \\ \widetilde{\boldsymbol{p}} \end{pmatrix}, \qquad \boldsymbol{v}_{\boldsymbol{p}} = \begin{pmatrix} v_{\boldsymbol{p}}^{(0)} \\ \widetilde{\boldsymbol{v}}_{\boldsymbol{p}} \end{pmatrix}. \tag{83}$$

The hyperboloid constraint implies

$$\left(p^{(0)}\right)^2 - \|\widetilde{\boldsymbol{p}}\|_2^2 = 1. \tag{84}$$

Therefore,

$$p^{(0)} = \sqrt{1 + \|\widetilde{\boldsymbol{p}}\|_2^2}. \tag{85}$$

In particular,

$$p^{(0)} > \|\widetilde{\boldsymbol{p}}\|_2 . \tag{86}$$

The tangency constraint implies

$$p^{(0)} v_{\boldsymbol{p}}^{(0)} = \widetilde{\boldsymbol{p}}^{\top} \widetilde{\boldsymbol{v}}_{\boldsymbol{p}}. \tag{87}$$

By Cauchy–Schwarz,

$$\begin{aligned} \left(p^{(0)}\right)^2 \left(v_{\boldsymbol{p}}^{(0)}\right)^2 &\le \|\widetilde{\boldsymbol{p}}\|_2^2 \, \|\widetilde{\boldsymbol{v}}_{\boldsymbol{p}}\|_2^2 , \\ &= \|\widetilde{\boldsymbol{p}}\|_2^2 \left[r_{\boldsymbol{p}}^2 + \left(v_{\boldsymbol{p}}^{(0)}\right)^2 \right]. \end{aligned} \tag{88}$$

Since

$$\left(p^{(0)}\right)^2 = 1 + \|\widetilde{\boldsymbol{p}}\|_2^2 , \tag{89}$$

the preceding inequality gives

$$\begin{aligned} \left(v_{\boldsymbol{p}}^{(0)}\right)^2 &\le \|\widetilde{\boldsymbol{p}}\|_2^2 \, r_{\boldsymbol{p}}^2, \\ \frac{\left|v_{\boldsymbol{p}}^{(0)}\right|}{r_{\boldsymbol{p}}} &\le \|\widetilde{\boldsymbol{p}}\|_2 . \end{aligned} \tag{90}$$

Therefore,

$$\begin{aligned} p^{+(0)} &= p^{(0)} \cosh(r_{\boldsymbol{p}}) + \frac{v_{\boldsymbol{p}}^{(0)}}{r_{\boldsymbol{p}}} \sinh(r_{\boldsymbol{p}}), \\ &\ge p^{(0)} \cosh(r_{\boldsymbol{p}}) - \|\widetilde{\boldsymbol{p}}\|_2 \sinh(r_{\boldsymbol{p}}), \\ &= \left(p^{(0)} - \|\widetilde{\boldsymbol{p}}\|_2\right) \cosh(r_{\boldsymbol{p}}) + \|\widetilde{\boldsymbol{p}}\|_2 \, e^{-r_{\boldsymbol{p}}}, \\ &> 0. \end{aligned} \tag{91}$$

For $r_{\boldsymbol{p}} = 0$, we have

$$\boldsymbol{p}^{+} = \boldsymbol{p}, \tag{92}$$

and hence

$$p^{+(0)} = p^{(0)} > 0. \tag{93}$$

Thus every Riemannian gradient update satisfies

$$\langle \boldsymbol{p}^+, \boldsymbol{p}^+ \rangle_l = -1, \tag{94}$$

together with

$$p^{+(0)} > 0. \tag{95}$$

Therefore,

$$\boldsymbol{p}^+ \in \mathcal{H}^d. \tag{96}$$

F Supplementary Hyperboloid Identities

We consider the d-dimensional hyperboloid

$$\mathcal{H}^d = \left\{ \boldsymbol{p} \in \mathbb{R}^{d+1} : \boldsymbol{p}^\top \boldsymbol{g}_l \boldsymbol{p} = -1,\ p^{(0)} > 0 \right\},$$
$$\boldsymbol{g}_l = \mathrm{diag}(-1, 1, \ldots, 1). \tag{97}$$

Equivalently,

$$\langle \boldsymbol{p}, \boldsymbol{p} \rangle_l = \boldsymbol{p}^\top \boldsymbol{g}_l \boldsymbol{p}. \tag{98}$$

The hyperboloid is equipped with the Riemannian metric induced by the Lorentzian bilinear form

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle_l = -u^{(0)} v^{(0)} + \sum_{j=1}^{d} u^{(j)} v^{(j)}. \tag{99}$$

Since $\langle \boldsymbol{p}, \boldsymbol{p} \rangle_l = -1$, differentiating the defining constraint along a smooth curve through $\boldsymbol{p}$ gives the tangent space

$$T_{\boldsymbol{p}} \mathcal{H}^d = \left\{ \boldsymbol{v} \in \mathbb{R}^{d+1} : \langle \boldsymbol{p}, \boldsymbol{v} \rangle_l = 0 \right\}. \tag{100}$$

The restriction of $\langle \cdot, \cdot \rangle_l$ to $T_{\boldsymbol{p}} \mathcal{H}^d$ is positive definite, and therefore

$$\|\boldsymbol{v}\|_l = \sqrt{\langle \boldsymbol{v}, \boldsymbol{v} \rangle_l} \tag{101}$$

defines the Riemannian norm of every $\boldsymbol{v} \in T_{\boldsymbol{p}} \mathcal{H}^d$.

Lemma F.1 (Lorentz-orthogonal tangent projection)**.** *For any $\boldsymbol{p} \in \mathcal{H}^d$ and $\boldsymbol{u} \in \mathbb{R}^{d+1}$, the Lorentz-orthogonal projection of $\boldsymbol{u}$ onto $T_{\boldsymbol{p}} \mathcal{H}^d$ is*

$$\Pi_{\boldsymbol{p}}(\boldsymbol{u}) = \boldsymbol{u} + \langle \boldsymbol{p}, \boldsymbol{u} \rangle_l \boldsymbol{p}. \tag{102}$$

Equivalently,

$$\boldsymbol{u} = \Pi_{\boldsymbol{p}}(\boldsymbol{u}) - \langle \boldsymbol{p}, \boldsymbol{u} \rangle_l \boldsymbol{p}, \tag{103}$$

where $\Pi_{\boldsymbol{p}}(\boldsymbol{u}) \in T_{\boldsymbol{p}} \mathcal{H}^d$ and the second term belongs to the Lorentz-normal space $\mathrm{span}\{\boldsymbol{p}\}$.

Proof. Using $\langle \boldsymbol{p}, \boldsymbol{p} \rangle_l = -1$, we obtain

$$\begin{aligned} \langle \boldsymbol{p}, \Pi_{\boldsymbol{p}}(\boldsymbol{u}) \rangle_l &= \langle \boldsymbol{p}, \boldsymbol{u} \rangle_l + \langle \boldsymbol{p}, \boldsymbol{u} \rangle_l \langle \boldsymbol{p}, \boldsymbol{p} \rangle_l, \\ &= \langle \boldsymbol{p}, \boldsymbol{u} \rangle_l - \langle \boldsymbol{p}, \boldsymbol{u} \rangle_l, \\ &= 0. \end{aligned} \tag{104}$$

Hence $\Pi_{\boldsymbol{p}}(\boldsymbol{u}) \in T_{\boldsymbol{p}} \mathcal{H}^d$. Moreover,

$$\boldsymbol{u} - \Pi_{\boldsymbol{p}}(\boldsymbol{u}) = -\langle \boldsymbol{p}, \boldsymbol{u} \rangle_l \boldsymbol{p} \in \mathrm{span}\{\boldsymbol{p}\}, \tag{105}$$

which establishes the tangent–normal decomposition and therefore the projection formula. □

Lemma F.2 (Exponential map on the hyperboloid)**.** *Let $\boldsymbol{p} \in \mathcal{H}^d$ and $\boldsymbol{v} \in T_{\boldsymbol{p}} \mathcal{H}^d$. For $r = \|\boldsymbol{v}\|_l$, the Riemannian exponential map is*

$$\exp_{\boldsymbol{p}}(\boldsymbol{v}) = \cosh(r) \boldsymbol{p} + \sinh(r) \frac{\boldsymbol{v}}{r}, \qquad r = \sqrt{\langle \boldsymbol{v}, \boldsymbol{v} \rangle_l}, \tag{106}$$

with the continuous extension

$$\exp_{\boldsymbol{p}}(\boldsymbol{0}) = \boldsymbol{p}. \tag{107}$$

Proof. For $\boldsymbol{v} \neq \mathbf{0}$, define

$$\boldsymbol{\gamma}(s) = \cosh(sr)\boldsymbol{p} + \sinh(sr)\frac{\boldsymbol{v}}{r}, \qquad s \in \mathbb{R}. \tag{108}$$

Since $\langle \boldsymbol{p},\, \boldsymbol{p}\rangle_l = -1$, $\langle \boldsymbol{p},\, \boldsymbol{v}\rangle_l = 0$, and $\langle \boldsymbol{v},\, \boldsymbol{v}\rangle_l = r^2$, we obtain

$$\begin{aligned}\langle \boldsymbol{\gamma}(s),\, \boldsymbol{\gamma}(s)\rangle_l &= -\cosh^2(sr) + \sinh^2(sr),\\ &= -1.\end{aligned} \tag{109}$$

Furthermore,

$$\boldsymbol{\gamma}(0) = \boldsymbol{p}, \tag{110}$$

$$\dot{\boldsymbol{\gamma}}(s) = r\sinh(sr)\boldsymbol{p} + \cosh(sr)\boldsymbol{v}, \tag{111}$$

$$\dot{\boldsymbol{\gamma}}(0) = \boldsymbol{v}, \tag{112}$$

$$\begin{aligned}\ddot{\boldsymbol{\gamma}}(s) &= r^2\cosh(sr)\boldsymbol{p} + r\sinh(sr)\boldsymbol{v},\\ &= r^2\boldsymbol{\gamma}(s).\end{aligned} \tag{113}$$

Because the Lorentz-normal space at $\boldsymbol{\gamma}(s)$ is $\operatorname{span}\{\boldsymbol{\gamma}(s)\}$, the ambient acceleration $\ddot{\boldsymbol{\gamma}}(s) = r^2\boldsymbol{\gamma}(s)$ has zero tangent component. Therefore $\boldsymbol{\gamma}$ is the geodesic satisfying $\boldsymbol{\gamma}(0) = \boldsymbol{p}$ and $\dot{\boldsymbol{\gamma}}(0) = \boldsymbol{v}$. Hence

$$\begin{aligned}\exp_{\boldsymbol{p}}(\boldsymbol{v}) &= \boldsymbol{\gamma}(1),\\ &= \cosh(r)\boldsymbol{p} + \sinh(r)\frac{\boldsymbol{v}}{r}.\end{aligned} \tag{114}$$

The case $\boldsymbol{v} = \mathbf{0}$ follows by continuity. □

Riemannian gradient. Let $\mathcal{L} : \mathcal{H}^d \to \mathbb{R}$ be differentiable, and let $\overline{\mathcal{L}}$ be any smooth extension of $\mathcal{L}$ to a neighborhood of $\mathcal{H}^d$ in $\mathbb{R}^{d+1}$. Denote its Euclidean coordinate gradient by $\nabla^E_{\boldsymbol{p}}\overline{\mathcal{L}}$. The corresponding ambient Lorentz gradient is

$$\begin{aligned}\nabla^L_{\boldsymbol{p}}\overline{\mathcal{L}} &= \boldsymbol{g}_l^{-1}\nabla^E_{\boldsymbol{p}}\overline{\mathcal{L}},\\ &= \boldsymbol{g}_l\nabla^E_{\boldsymbol{p}}\overline{\mathcal{L}},\end{aligned} \tag{115}$$

where the second equality follows from

$$\boldsymbol{g}_l^{-1} = \boldsymbol{g}_l, \qquad \boldsymbol{g}_l^2 = \boldsymbol{I}_{d+1}. \tag{116}$$

Indeed, for every ambient direction $\boldsymbol{u} \in \mathbb{R}^{d+1}$,

$$\begin{aligned}D\overline{\mathcal{L}}(\boldsymbol{p})[\boldsymbol{u}] &= \left(\nabla^E_{\boldsymbol{p}}\overline{\mathcal{L}}\right)^\top \boldsymbol{u},\\ &= \left(\boldsymbol{g}_l\nabla^E_{\boldsymbol{p}}\overline{\mathcal{L}}\right)^\top \boldsymbol{g}_l\boldsymbol{u},\\ &= \langle \boldsymbol{g}_l\nabla^E_{\boldsymbol{p}}\overline{\mathcal{L}},\, \boldsymbol{u}\rangle_l,\\ &= \langle \nabla^L_{\boldsymbol{p}}\overline{\mathcal{L}},\, \boldsymbol{u}\rangle_l.\end{aligned} \tag{117}$$

The Riemannian gradient on $\mathcal{H}^d$ is the tangent component of the ambient Lorentz gradient:

$$\begin{aligned}\operatorname{grad}_{\boldsymbol{p}}\mathcal{L} &= \Pi_{\boldsymbol{p}}\left(\nabla^L_{\boldsymbol{p}}\overline{\mathcal{L}}\right),\\ &= \nabla^L_{\boldsymbol{p}}\overline{\mathcal{L}} + \langle \boldsymbol{p},\, \nabla^L_{\boldsymbol{p}}\overline{\mathcal{L}}\rangle_l\boldsymbol{p},\\ &= \boldsymbol{g}_l\nabla^E_{\boldsymbol{p}}\overline{\mathcal{L}} + \langle \boldsymbol{p},\, \boldsymbol{g}_l\nabla^E_{\boldsymbol{p}}\overline{\mathcal{L}}\rangle_l\boldsymbol{p}.\end{aligned} \tag{118}$$

To verify the defining property of the Riemannian gradient, for every $\boldsymbol{v} \in T_{\boldsymbol{p}}\mathcal{H}^d$,

$$\begin{aligned}D\mathcal{L}(\boldsymbol{p})[\boldsymbol{v}] &= D\overline{\mathcal{L}}(\boldsymbol{p})[\boldsymbol{v}],\\ &= \langle \nabla^L_{\boldsymbol{p}}\overline{\mathcal{L}},\, \boldsymbol{v}\rangle_l,\\ &= \langle \operatorname{grad}_{\boldsymbol{p}}\mathcal{L},\, \boldsymbol{v}\rangle_l.\end{aligned} \tag{119}$$

The last equality follows from $\langle \boldsymbol{p},\, \boldsymbol{v}\rangle_l = 0$. Consequently, the tangent gradient does not depend on the choice of the ambient extension.

Riemannian gradient update. For a step size $\eta > 0$, define

$$\boldsymbol{v}_{\boldsymbol{p}} = -\eta \operatorname{grad}_{\boldsymbol{p}} \mathcal{L} \in T_{\boldsymbol{p}}\mathcal{H}^d, \qquad r_{\boldsymbol{p}} = \|\boldsymbol{v}_{\boldsymbol{p}}\|_l \,. \tag{120}$$

The Riemannian gradient step is

$$\begin{aligned} \boldsymbol{p}^+ &= \exp_{\boldsymbol{p}}(\boldsymbol{v}_{\boldsymbol{p}}), \\ &= \cosh(r_{\boldsymbol{p}})\boldsymbol{p} + \sinh(r_{\boldsymbol{p}})\frac{\boldsymbol{v}_{\boldsymbol{p}}}{r_{\boldsymbol{p}}}, \\ &= \cosh\Big(\eta \left\|\operatorname{grad}_{\boldsymbol{p}} \mathcal{L}\right\|_l\Big)\, \boldsymbol{p} - \sinh\Big(\eta \left\|\operatorname{grad}_{\boldsymbol{p}} \mathcal{L}\right\|_l\Big)\, \frac{\operatorname{grad}_{\boldsymbol{p}} \mathcal{L}}{\left\|\operatorname{grad}_{\boldsymbol{p}} \mathcal{L}\right\|_l}. \end{aligned} \tag{121}$$

Here the line breaks in the source code inside the last expression do not create additional displayed formula lines. When $\operatorname{grad}_{\boldsymbol{p}} \mathcal{L} = \mathbf{0}$, the update is understood by continuity as $\boldsymbol{p}^+ = \boldsymbol{p}$.

It remains to verify that the update stays on the upper sheet. For $r_{\boldsymbol{p}} > 0$, using $\langle \boldsymbol{p},\, \boldsymbol{p}\rangle_l = -1$, $\langle \boldsymbol{p},\, \boldsymbol{v}_{\boldsymbol{p}}\rangle_l = 0$, and $\langle \boldsymbol{v}_{\boldsymbol{p}},\, \boldsymbol{v}_{\boldsymbol{p}}\rangle_l = r_{\boldsymbol{p}}^2$, we obtain

$$\begin{aligned} \langle \boldsymbol{p}^+,\, \boldsymbol{p}^+\rangle_l &= -\cosh^2(r_{\boldsymbol{p}}) + \sinh^2(r_{\boldsymbol{p}}), \\ &= -1. \end{aligned} \tag{122}$$

To establish the upper-sheet condition, write

$$\boldsymbol{p} = \begin{pmatrix} p^{(0)} \\ \widetilde{\boldsymbol{p}} \end{pmatrix}, \qquad \boldsymbol{v}_{\boldsymbol{p}} = \begin{pmatrix} v_{\boldsymbol{p}}^{(0)} \\ \widetilde{\boldsymbol{v}}_{\boldsymbol{p}} \end{pmatrix}. \tag{123}$$

The hyperboloid and tangency constraints imply

$$\Big(p^{(0)}\Big)^2 - \|\widetilde{\boldsymbol{p}}\|_2^2 = 1, \tag{124}$$

$$\begin{aligned} p^{(0)} &= \sqrt{1 + \|\widetilde{\boldsymbol{p}}\|_2^2}, \\ p^{(0)} &> \|\widetilde{\boldsymbol{p}}\|_2 \,, \\ p^{(0)} v_{\boldsymbol{p}}^{(0)} &= \widetilde{\boldsymbol{p}}^\top \widetilde{\boldsymbol{v}}_{\boldsymbol{p}}. \end{aligned} \tag{125}$$

By Cauchy–Schwarz,

$$\begin{aligned} \Big(p^{(0)}\Big)^2 \Big(v_{\boldsymbol{p}}^{(0)}\Big)^2 &\le \|\widetilde{\boldsymbol{p}}\|_2^2 \, \|\widetilde{\boldsymbol{v}}_{\boldsymbol{p}}\|_2^2 \,, \\ &= \|\widetilde{\boldsymbol{p}}\|_2^2 \left[r_{\boldsymbol{p}}^2 + \Big(v_{\boldsymbol{p}}^{(0)}\Big)^2 \right]. \end{aligned} \tag{126}$$

Since $\big(p^{(0)}\big)^2 = 1 + \|\widetilde{\boldsymbol{p}}\|_2^2$, it follows that

$$\begin{aligned} \Big(v_{\boldsymbol{p}}^{(0)}\Big)^2 &\le \|\widetilde{\boldsymbol{p}}\|_2^2 \, r_{\boldsymbol{p}}^2, \\ \frac{\left|v_{\boldsymbol{p}}^{(0)}\right|}{r_{\boldsymbol{p}}} &\le \|\widetilde{\boldsymbol{p}}\|_2 \,. \end{aligned} \tag{127}$$

Therefore,

$$\begin{aligned} p^{+(0)} &= p^{(0)} \cosh(r_{\boldsymbol{p}}) + \frac{v_{\boldsymbol{p}}^{(0)}}{r_{\boldsymbol{p}}} \sinh(r_{\boldsymbol{p}}), \\ &\ge p^{(0)} \cosh(r_{\boldsymbol{p}}) - \|\widetilde{\boldsymbol{p}}\|_2 \sinh(r_{\boldsymbol{p}}), \\ &= \Big(p^{(0)} - \|\widetilde{\boldsymbol{p}}\|_2\Big) \cosh(r_{\boldsymbol{p}}) + \|\widetilde{\boldsymbol{p}}\|_2 \, e^{-r_{\boldsymbol{p}}}, \\ &> 0. \end{aligned} \tag{128}$$

For $r_{\boldsymbol{p}} = 0$, we have $\boldsymbol{p}^+ = \boldsymbol{p}$ and hence $p^{+(0)} = p^{(0)} > 0$. Thus every Riemannian gradient update satisfies

$$\langle \boldsymbol{p}^+,\, \boldsymbol{p}^+\rangle_l = -1, \qquad p^{+(0)} > 0, \qquad \boldsymbol{p}^+ \in \mathcal{H}^d. \tag{129}$$

G Proof of the Tree-Approximation Transfer Proposition

Let $\widetilde{\mathcal{G}} = (\mathcal{V}, \widetilde{\mathcal{E}})$ denote the connected symmetrized skeleton of the directed structural graph, where $|\mathcal{V}| = T$, and let $d_{\mathcal{G}}$ denote its unweighted shortest-path metric. Let $\mathcal{T} = (\mathcal{V}, \mathcal{E}_{\mathcal{T}})$ be a tree on the same vertex set with tree metric $d_{\mathcal{T}}$. Throughout this result, the embedding dimension satisfies $d \geq 2$.

Scale-free multiplicative distortion. For a metric space $(\mathcal{M}, d_{\mathcal{M}})$, define

$$D_{\mathcal{M}}^{\max}(\mathcal{G}) := \inf_{\substack{\phi:\mathcal{V}\hookrightarrow\mathcal{M} \\ \alpha>0}} \max_{\substack{u,v\in\mathcal{V} \\ u\neq v}} \max\left\{ \frac{d_{\mathcal{M}}(\phi(u),\phi(v))}{\alpha d_{\mathcal{G}}(u,v)},\ \frac{\alpha d_{\mathcal{G}}(u,v)}{d_{\mathcal{M}}(\phi(u),\phi(v))} \right\}. \tag{130}$$

Equivalently, for a fixed embedding ϕ and scale α, a number $D \geq 1$ is admissible if and only if

$$\frac{\alpha d_{\mathcal{G}}(u,v)}{D} \leq d_{\mathcal{M}}(\phi(u),\phi(v)) \leq D\alpha d_{\mathcal{G}}(u,v) \qquad \text{for all } u \neq v. \tag{131}$$

Graph-growth condition. For $v_0 \in \mathcal{V}$ and $r \geq 0$, write

$$B_{\widetilde{\mathcal{G}}}(v_0, r) := \{v \in \mathcal{V} : d_{\mathcal{G}}(v_0, v) \leq r\}. \tag{132}$$

The Euclidean comparison assumes that there exist constants $b > 0$ and $c_b \in (0, 1]$, independent of T, together with a vertex $v_0 \in \mathcal{V}$ and a radius $r_T \geq 1$, such that

$$r_T \leq b \log T, \qquad \left|B_{\widetilde{\mathcal{G}}}(v_0, r_T)\right| \geq c_b T, \qquad c_b T > 1. \tag{133}$$

Assume further that the tree approximation satisfies

$$d_{\mathcal{G}}(u,v) \leq d_{\mathcal{T}}(u,v) \leq d_{\mathcal{G}}(u,v) + \Delta_T, \qquad \Delta_T \geq 0, \qquad u, v \in \mathcal{V}. \tag{134}$$

For every finite tree and every $\varepsilon > 0$, the low-distortion tree-to-hyperbolic embedding result gives a scale $\alpha_\varepsilon > 0$ and an embedding $\phi_\varepsilon : \mathcal{V} \to \mathbb{H}^2$ such that Sarkar (2011)

$$\alpha_\varepsilon d_{\mathcal{T}}(u,v) \leq d_{\mathbb{H}}\left(\phi_\varepsilon(u), \phi_\varepsilon(v)\right) \leq (1+\varepsilon)\alpha_\varepsilon d_{\mathcal{T}}(u,v) \qquad \text{for all } u, v \in \mathcal{V}. \tag{135}$$

Since $\mathbb{H}^2$ embeds isometrically as a totally geodesic subspace of $\mathbb{H}^d$ for every $d \geq 2$, the same inequalities hold for an embedding into $\mathbb{H}^d$.

Proof. We first establish the hyperbolic upper bound. Fix $\varepsilon > 0$ and let $(\phi_\varepsilon, \alpha_\varepsilon)$ satisfy Eq. (135). For any distinct $u, v \in \mathcal{V}$, Eqs. (134) and (135) give

$$\begin{aligned} \alpha_\varepsilon d_{\mathcal{G}}(u,v) &\leq \alpha_\varepsilon d_{\mathcal{T}}(u,v) \\ &\leq d_{\mathbb{H}}\left(\phi_\varepsilon(u), \phi_\varepsilon(v)\right) \\ &\leq (1+\varepsilon)\alpha_\varepsilon d_{\mathcal{T}}(u,v) \\ &\leq (1+\varepsilon)\alpha_\varepsilon \left[d_{\mathcal{G}}(u,v) + \Delta_T\right]. \end{aligned} \tag{136}$$

Because $\widetilde{\mathcal{G}}$ is an unweighted connected graph and $u \neq v$,

$$d_{\mathcal{G}}(u,v) \geq 1, \tag{137}$$

and therefore

$$\begin{aligned} d_{\mathcal{G}}(u,v) + \Delta_T &= d_{\mathcal{G}}(u,v)\left[1 + \frac{\Delta_T}{d_{\mathcal{G}}(u,v)}\right] \\ &\leq d_{\mathcal{G}}(u,v)\left(1 + \Delta_T\right). \end{aligned} \tag{138}$$

Substituting Eq. (138) into Eq. (136) yields

$$\begin{aligned} \alpha_\varepsilon d_{\mathcal{G}}(u,v) &\leq d_{\mathbb{H}}\left(\phi_\varepsilon(u), \phi_\varepsilon(v)\right) \\ &\leq (1+\varepsilon)(1+\Delta_T)\alpha_\varepsilon d_{\mathcal{G}}(u,v). \end{aligned} \tag{139}$$

Equivalently,

$$\begin{aligned} 1 &\le \frac{d_{\mathbb{H}}\left(\phi_\varepsilon(u), \phi_\varepsilon(v)\right)}{\alpha_\varepsilon d_{\mathcal{G}}(u,v)} \\ &\le (1+\varepsilon)\left[1 + \frac{\Delta_T}{d_{\mathcal{G}}(u,v)}\right] \\ &\le (1+\varepsilon)(1+\Delta_T), \end{aligned} \tag{140}$$

while the first inequality in Eq. (139) also implies

$$0 < \frac{\alpha_\varepsilon d_{\mathcal{G}}(u,v)}{d_{\mathbb{H}}\left(\phi_\varepsilon(u), \phi_\varepsilon(v)\right)} \le 1 \le (1+\varepsilon)(1+\Delta_T). \tag{141}$$

Hence, for the particular pair $(\phi_\varepsilon, \alpha_\varepsilon)$,

$$\begin{aligned} \max_{\substack{u,v\in\mathcal{V}\\ u\neq v}} \max\left\{ \frac{d_{\mathbb{H}}\left(\phi_\varepsilon(u), \phi_\varepsilon(v)\right)}{\alpha_\varepsilon d_{\mathcal{G}}(u,v)},\ \frac{\alpha_\varepsilon d_{\mathcal{G}}(u,v)}{d_{\mathbb{H}}\left(\phi_\varepsilon(u), \phi_\varepsilon(v)\right)} \right\} \\ \le (1+\varepsilon)(1+\Delta_T). \end{aligned} \tag{142}$$

Taking the infimum over all admissible embeddings and scales in Eq. (130) gives

$$D_{\mathbb{H}}^{\max}(\mathcal{G}) \le (1+\varepsilon)(1+\Delta_T). \tag{143}$$

Since Eq. (143) holds for every $\varepsilon > 0$, we also have

$$\begin{aligned} D_{\mathbb{H}}^{\max}(\mathcal{G}) &\le \inf_{\varepsilon>0}(1+\varepsilon)(1+\Delta_T) \\ &= 1 + \Delta_T. \end{aligned} \tag{144}$$

If $\widetilde{\mathcal{G}}$ is itself a tree, we may choose $d_{\mathcal{T}} = d_{\mathcal{G}}$ and $\Delta_T = 0$. Equation (144) then gives

$$D_{\mathbb{H}}^{\max}(\mathcal{G}) \le 1. \tag{145}$$

On the other hand, for every positive number a,

$$\max\{a, a^{-1}\} \ge 1, \tag{146}$$

so Eq. (130) implies

$$D_{\mathbb{H}}^{\max}(\mathcal{G}) \ge 1. \tag{147}$$

Consequently,

$$D_{\mathbb{H}}^{\max}(\mathcal{G}) = 1. \tag{148}$$

Here the equality refers to the infimal scale-free distortion; it does not require the infimum to be attained by a finite-scale embedding.

Euclidean lower bound. Consider an arbitrary injective embedding

$$\phi : \mathcal{V} \to \mathbb{R}^d \tag{149}$$

and an arbitrary scale $\alpha > 0$. Let

$$\begin{aligned} D := \max_{\substack{u,v\in\mathcal{V}\\ u\neq v}} \max\left\{ \frac{\|\phi(u)-\phi(v)\|_2}{\alpha d_{\mathcal{G}}(u,v)},\ \frac{\alpha d_{\mathcal{G}}(u,v)}{\|\phi(u)-\phi(v)\|_2} \right\} \\ \ge 1. \end{aligned} \tag{150}$$

By definition of D,

$$\frac{\alpha d_{\mathcal{G}}(u,v)}{D} \le \|\phi(u)-\phi(v)\|_2 \le D\alpha d_{\mathcal{G}}(u,v) \qquad \text{for all } u \neq v. \tag{151}$$

Let

$$B_T := B_{\widetilde{\mathcal{G}}}(v_0, r_T), \qquad n_T := |B_T| \ge c_b T. \tag{152}$$

For every $v \in B_T$, Eq. (151) gives

$$\begin{aligned}\|\phi(v) - \phi(v_0)\|_2 &\le D\alpha d_{\mathcal{G}}(v, v_0) \\ &\le D\alpha r_T. \end{aligned} \tag{153}$$

For any two distinct $u, v \in B_T$, the unweighted graph metric satisfies $d_{\mathcal{G}}(u, v) \ge 1$, and hence

$$\begin{aligned}\|\phi(u) - \phi(v)\|_2 &\ge \frac{\alpha d_{\mathcal{G}}(u, v)}{D} \\ &\ge \frac{\alpha}{D}. \end{aligned} \tag{154}$$

Define

$$\rho := \frac{\alpha}{2D}, \qquad R := D\alpha r_T + \rho = D\alpha r_T + \frac{\alpha}{2D}. \tag{155}$$

Equation (154) implies

$$\|\phi(u) - \phi(v)\|_2 \ge 2\rho \qquad \text{for all distinct } u, v \in B_T, \tag{156}$$

so the open Euclidean balls

$$\{B_{\mathbb{R}^d}(\phi(v), \rho) : v \in B_T\} \tag{157}$$

are pairwise disjoint. Moreover, if $y \in B_{\mathbb{R}^d}(\phi(v), \rho)$ for some $v \in B_T$, then by the triangle inequality and Eq. (153),

$$\begin{aligned}\|y - \phi(v_0)\|_2 &\le \|y - \phi(v)\|_2 + \|\phi(v) - \phi(v_0)\|_2 \\ &< \rho + D\alpha r_T \\ &= R. \end{aligned} \tag{158}$$

Therefore,

$$\bigcup_{v \in B_T} B_{\mathbb{R}^d}(\phi(v), \rho) \subseteq B_{\mathbb{R}^d}(\phi(v_0), R)\,. \tag{159}$$

Let

$$C_d := \mathrm{Vol}\left(B_{\mathbb{R}^d}(\mathbf{0}, 1)\right) \tag{160}$$

denote the volume of the unit ball in $\mathbb{R}^d$. Using pairwise disjointness, containment, and the scaling law for Euclidean volume,

$$\begin{aligned} n_T C_d \rho^d &= \sum_{v \in B_T} \mathrm{Vol}\left(B_{\mathbb{R}^d}(\phi(v), \rho)\right) \\ &= \mathrm{Vol}\left(\bigcup_{v \in B_T} B_{\mathbb{R}^d}(\phi(v), \rho)\right) \\ &\le \mathrm{Vol}\left(B_{\mathbb{R}^d}(\phi(v_0), R)\right) \\ &= C_d R^d. \end{aligned} \tag{161}$$

Since $C_d > 0$ and $\rho > 0$, Eq. (161) yields

$$\begin{aligned} n_T &\le \left(\frac{R}{\rho}\right)^d \\ &= \left(\frac{D\alpha r_T + \alpha/(2D)}{\alpha/(2D)}\right)^d \\ &= \left(\frac{2D^2\alpha r_T + \alpha}{\alpha}\right)^d \\ &= \left(2D^2 r_T + 1\right)^d. \end{aligned} \tag{162}$$

Taking the d-th root of both sides gives

$$n_T^{1/d} \le 2D^2 r_T + 1, \tag{163}$$

$$n_T^{1/d} - 1 \le 2D^2 r_T, \tag{164}$$

$$D^2 \ge \frac{n_T^{1/d} - 1}{2r_T}, \tag{165}$$

$$D \ge \left[\frac{n_T^{1/d} - 1}{2r_T}\right]^{1/2}. \tag{166}$$

Using $n_T \ge c_b T$ and monotonicity of $x \mapsto x^{1/d}$, we further obtain

$$\begin{aligned} D &\ge \left[\frac{n_T^{1/d} - 1}{2r_T}\right]^{1/2} \\ &\ge \left[\frac{(c_b T)^{1/d} - 1}{2r_T}\right]^{1/2}. \end{aligned} \tag{167}$$

The embedding ϕ and scale α were arbitrary. Taking the infimum over all admissible (ϕ, α) therefore preserves the lower bound:

$$D_{\mathbb{E}}^{\max}(\mathcal{G}) \ge \left[\frac{(c_b T)^{1/d} - 1}{2r_T}\right]^{1/2}. \tag{168}$$

Finally, under $r_T \le b \log T$, for all sufficiently large T such that

$$(c_b T)^{1/d} \ge 2, \tag{169}$$

we have

$$(c_b T)^{1/d} - 1 \ge \frac{1}{2}(c_b T)^{1/d}. \tag{170}$$

Substituting these two inequalities into Eq. (168) gives

$$\begin{aligned} D_{\mathbb{E}}^{\max}(\mathcal{G}) &\ge \left[\frac{(c_b T)^{1/d} - 1}{2r_T}\right]^{1/2} \\ &\ge \left[\frac{(c_b T)^{1/d}}{4b \log T}\right]^{1/2} \\ &= \frac{c_b^{1/(2d)}}{2\sqrt{b}} \frac{T^{1/(2d)}}{\sqrt{\log T}} \\ &= \Omega_{b,c_b,d}\left(\frac{T^{1/(2d)}}{\sqrt{\log T}}\right). \end{aligned} \tag{171}$$

This proves the stated comparison. □

Remark G.1 (Scope)**.** *The hyperbolic conclusion is a representability statement conditional on the additive tree approximation in Eq.* (134)*. The Euclidean lower bound additionally uses the graph-growth condition in Eq.* (133)*. Neither conclusion asserts that the Stage II training objective reaches an embedding attaining $D_{\mathbb{H}}^{\max}(\mathcal{G})$ or $D_{\mathbb{E}}^{\max}(\mathcal{G})$. In particular, $D_{\mathbb{H}}^{\max}(\mathcal{G}) = 1$ for an exact tree metric is an equality of infimal distortions and need not correspond to an attained finite-scale isometry.*

H Proof of the Conditional PageRank–Radius Proposition

We consider the nondegenerate PageRank-regularized regime

$$\lambda_{\mathrm{hyp}} \lambda_g > 0. \tag{172}$$

The radial-force condition in the proposition is understood for stationary nodes with positive radii. This is the regime in which the outward unit radial direction and its associated one-dimensional force law are well defined.

H.1 RADIAL GEOMETRY

For node u, let

$$r_u := d_l(\boldsymbol{p}_u, \boldsymbol{p}_o), \qquad \boldsymbol{p}_o = (1, \mathbf{0}_d)^\top. \tag{173}$$

Since the hyperbolic distance on the hyperboloid satisfies

$$r_u = \operatorname{arcosh}(-\langle \boldsymbol{p}_u, \boldsymbol{p}_o \rangle_l) = \operatorname{arcosh}\Big(p_u^{(0)}\Big), \tag{174}$$

$$-\langle \boldsymbol{p}_u, \boldsymbol{p}_o \rangle_l = p_u^{(0)} = \cosh r_u, \tag{175}$$

the hyperboloid constraint $\langle \boldsymbol{p}_u, \boldsymbol{p}_u \rangle_l = -1$ gives

$$\begin{aligned} -\big(p_u^{(0)}\big)^2 + \|\widetilde{\boldsymbol{p}}_u\|_2^2 &= -1, \\ \|\widetilde{\boldsymbol{p}}_u\|_2^2 &= \big(p_u^{(0)}\big)^2 - 1 = \cosh^2 r_u - 1 = \sinh^2 r_u, \\ \|\widetilde{\boldsymbol{p}}_u\|_2 &= \sinh r_u, \end{aligned} \tag{176}$$

where the final equality follows from $r_u \geq 0$.

Under the hyperboloid-to-Poincaré map

$$\boldsymbol{e}_u = \frac{\widetilde{\boldsymbol{p}}_u}{p_u^{(0)} + 1}, \tag{177}$$

we therefore obtain

$$\begin{aligned} \|\boldsymbol{e}_u\|_2 &= \frac{\|\widetilde{\boldsymbol{p}}_u\|_2}{p_u^{(0)} + 1} = \frac{\sinh r_u}{\cosh r_u + 1} \\ &= \frac{2\sinh(r_u/2)\cosh(r_u/2)}{2\cosh^2(r_u/2)} = \frac{\sinh(r_u/2)}{\cosh(r_u/2)} = \tanh\Big(\frac{r_u}{2}\Big). \end{aligned} \tag{178}$$

Moreover,

$$\frac{d}{dr}\tanh\Big(\frac{r}{2}\Big) = \frac{1}{2}\operatorname{sech}^2\Big(\frac{r}{2}\Big) > 0, \qquad r \geq 0. \tag{179}$$

Hence

$$r_s < r_t \quad \Longrightarrow \quad \|\boldsymbol{e}_s\|_2 < \|\boldsymbol{e}_t\|_2. \tag{180}$$

For $r_u > 0$, define the unit spatial direction

$$\boldsymbol{\xi}_u := \frac{\widetilde{\boldsymbol{p}}_u}{\|\widetilde{\boldsymbol{p}}_u\|_2} \in \mathbb{R}^d, \qquad \|\boldsymbol{\xi}_u\|_2 = 1. \tag{181}$$

Then

$$\boldsymbol{p}_u = \begin{pmatrix} \cosh r_u \\ \sinh r_u\, \boldsymbol{\xi}_u \end{pmatrix}. \tag{182}$$

The geodesic ray from $\boldsymbol{p}_o$ through $\boldsymbol{p}_u$ is

$$\boldsymbol{\gamma}_u(r) = \begin{pmatrix} \cosh r \\ \sinh r\, \boldsymbol{\xi}_u \end{pmatrix}, \qquad r \geq 0, \tag{183}$$

and its outward tangent vector is

$$\boldsymbol{v}_u(r) := \dot{\boldsymbol{\gamma}}_u(r) = \begin{pmatrix} \sinh r \\ \cosh r\, \boldsymbol{\xi}_u \end{pmatrix}. \tag{184}$$

Using $\|\boldsymbol{\xi}_u\|_2 = 1$, we have

$$\langle \boldsymbol{\gamma}_u(r), \boldsymbol{v}_u(r) \rangle_l = -\cosh r\, \sinh r + \sinh r\, \cosh r\, \|\boldsymbol{\xi}_u\|_2^2 = 0, \tag{185}$$

$$\langle \boldsymbol{v}_u(r), \boldsymbol{v}_u(r) \rangle_l = -\sinh^2 r + \cosh^2 r\, \|\boldsymbol{\xi}_u\|_2^2 = 1. \tag{186}$$

Thus

$$\boldsymbol{v}_u(r) \in T_{\boldsymbol{\gamma}_u(r)}\mathcal{H}^d, \qquad \|\boldsymbol{v}_u(r)\|_l = 1. \tag{187}$$

At $r = r_u$, write $\boldsymbol{v}_u := \boldsymbol{v}_u(r_u)$. Since

$$d_l(\boldsymbol{\gamma}_u(r), \boldsymbol{p}_o) = r, \tag{188}$$

the directional derivative of the radius along the outward radial direction is

$$D_{\boldsymbol{v}_u} r_u = \frac{d}{dr} d_l(\boldsymbol{\gamma}_u(r), \boldsymbol{p}_o)\Big|_{r=r_u} = 1. \tag{189}$$

H.2 RADIAL FORCE BALANCE

For the frozen Stage I graph, π_u^{PR} depends only on the learned graph and is independent of the hyperbolic coordinate $\boldsymbol{p}_u$. The PageRank radial term associated with node u is therefore

$$\Omega_u(\boldsymbol{p}_u) = \pi_u^{\mathrm{PR}} r_u = \pi_u^{\mathrm{PR}} d_l(\boldsymbol{p}_u, \boldsymbol{p}_o). \tag{190}$$

For $r_u > 0$, the defining property of the Riemannian gradient gives

$$\begin{aligned}\langle \nabla^T_{\boldsymbol{p}_u} \Omega_u,\, \boldsymbol{v}_u \rangle_l &= D_{\boldsymbol{v}_u} \Omega_u = D_{\boldsymbol{v}_u} \left(\pi_u^{\mathrm{PR}} r_u\right) \\ &= \pi_u^{\mathrm{PR}} D_{\boldsymbol{v}_u} r_u = \pi_u^{\mathrm{PR}} \cdot 1 = \pi_u^{\mathrm{PR}}.\end{aligned} \tag{191}$$

To match the Stage II–III objective, define

$$\mathcal{L}_{\mathrm{gate}} := \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{k=1}^{K} \bar{g}_{t,k}^{(i)}. \tag{192}$$

Using

$$\mathcal{L}_{\mathcal{H}} = \frac{1}{T} \sum_{v=1}^{T} \left[\mathcal{L}_{\mathrm{con}}(\boldsymbol{p}_v) + \lambda_g \pi_v^{\mathrm{PR}} r_v \right], \tag{193}$$

the fine-tuning objective can be decomposed as

$$\mathcal{L}_{\mathrm{FT}} = \mathcal{L}_{\mathrm{rest}} + \frac{\lambda_{\mathrm{hyp}} \lambda_g}{T} \sum_{v=1}^{T} \pi_v^{\mathrm{PR}} r_v, \tag{194}$$

$$\mathcal{L}_{\mathrm{rest}} = \mathcal{L}_{\mathrm{task}} + \lambda_{\mathrm{sp}} \mathcal{L}_{\mathrm{gate}} + \frac{\lambda_{\mathrm{hyp}}}{T} \sum_{v=1}^{T} \mathcal{L}_{\mathrm{con}}(\boldsymbol{p}_v). \tag{195}$$

All dependence of the task loss, gate penalty, and contrastive loss on the hyperbolic coordinates is retained in $\mathcal{L}_{\mathrm{rest}}$. In particular, although $\mathcal{L}_{\mathrm{con}}(\boldsymbol{p}_v)$ is indexed by its anchor node v, its pairwise-distance terms may also depend on other node embeddings, and these contributions are included when differentiating $\mathcal{L}_{\mathrm{rest}}$ with respect to $\boldsymbol{p}_u$.

Let Θ^* denote a stationary configuration of all Stage II–III trainable quantities, and let $\boldsymbol{p}_u^*$ be the stationary embedding of node u, with

$$r_u^* := d_l(\boldsymbol{p}_u^*, \boldsymbol{p}_o) > 0. \tag{196}$$

Write

$$\boldsymbol{\xi}_u^* := \frac{\widetilde{\boldsymbol{p}}_u^*}{\|\widetilde{\boldsymbol{p}}_u^*\|_2}, \qquad \boldsymbol{\gamma}_u^*(r) := \begin{pmatrix} \cosh r \\ \sinh r\, \boldsymbol{\xi}_u^* \end{pmatrix}, \tag{197}$$

so that

$$\boldsymbol{\gamma}_u^*(r_u^*) = \boldsymbol{p}_u^*. \tag{198}$$

Let Θ^*_{-u} collect all components of Θ^* other than $\boldsymbol{p}_u^*$, and define the restriction of the non-PageRank objective to the stationary radial ray by

$$\ell_u^{\mathrm{rad}}(r) := \mathcal{L}_{\mathrm{rest}} \left(\Theta^*_{-u}, \boldsymbol{p}_u = \boldsymbol{\gamma}_u^*(r)\right). \tag{199}$$

The outward radial force profile along this ray is

$$\begin{aligned}\mathcal{F}_u(r) &:= -\frac{d}{dr} \ell_u^{\mathrm{rad}}(r) \\ &= -\langle \nabla^T_{\boldsymbol{p}_u} \mathcal{L}_{\mathrm{rest}} \left(\Theta^*_{-u}, \boldsymbol{p}_u = \boldsymbol{\gamma}_u^*(r)\right),\, \dot{\boldsymbol{\gamma}}_u^*(r) \rangle_l.\end{aligned} \tag{200}$$

Thus $\mathcal{F}_u(r)$ is a one-dimensional radial force law obtained by varying only the radius of node u along its stationary geodesic ray while all other trainable quantities remain fixed at their stationary values.

For the pair of nodes s and t in the proposition, the matched-force condition means that there exists an interval $I \subset (0, \infty)$ containing both r_s^* and r_t^* and a common function

$$\mathcal{F}_{\mathrm{out}} : I \to \mathbb{R} \tag{201}$$

such that

$$\mathcal{F}_s(r) = \mathcal{F}_t(r) = \mathcal{F}_{\mathrm{out}}(r), \qquad r \in I, \tag{202}$$

with

$$r_1 < r_2 \quad \Longrightarrow \quad \mathcal{F}_{\mathrm{out}}(r_1) > \mathcal{F}_{\mathrm{out}}(r_2). \tag{203}$$

Proof. At the stationary configuration Θ^*, the Riemannian gradient with respect to $\boldsymbol{p}_u$ vanishes:

$$\begin{aligned}\mathbf{0} &= \nabla^T_{\boldsymbol{p}_u}\mathcal{L}_{\mathrm{FT}}\big|_{\Theta=\Theta^*} \\ &= \nabla^T_{\boldsymbol{p}_u}\mathcal{L}_{\mathrm{rest}}\big|_{\Theta=\Theta^*} + \frac{\lambda_{\mathrm{hyp}}\lambda_g}{T}\nabla^T_{\boldsymbol{p}_u}\left(\sum_{v=1}^{T}\pi_v^{\mathrm{PR}} r_v\right)\Bigg|_{\Theta=\Theta^*} \\ &= \nabla^T_{\boldsymbol{p}_u}\mathcal{L}_{\mathrm{rest}}\big|_{\Theta=\Theta^*} + \frac{\lambda_{\mathrm{hyp}}\lambda_g}{T}\pi_u^{\mathrm{PR}}\nabla^T_{\boldsymbol{p}_u} r_u\big|_{\boldsymbol{p}_u=\boldsymbol{p}_u^*},\end{aligned} \tag{204}$$

because r_v does not depend on $\boldsymbol{p}_u$ for $v \neq u$, and π_v^{PR} is fixed during Stage II–III optimization.

Let

$$\boldsymbol{v}_u^* := \dot{\boldsymbol{\gamma}}_u^*(r_u^*) \in T_{\boldsymbol{p}_u^*}\mathcal{H}^d \tag{205}$$

be the outward unit radial tangent at $\boldsymbol{p}_u^*$. Taking the Lorentzian inner product of the stationarity equation with $\boldsymbol{v}_u^*$ and using Eq. (191), we obtain

$$\begin{aligned}0 &= \langle \nabla^T_{\boldsymbol{p}_u}\mathcal{L}_{\mathrm{rest}}\big|_{\Theta=\Theta^*}, \boldsymbol{v}_u^*\rangle_l + \frac{\lambda_{\mathrm{hyp}}\lambda_g}{T}\langle \nabla^T_{\boldsymbol{p}_u}\Omega_u\big|_{\boldsymbol{p}_u=\boldsymbol{p}_u^*}, \boldsymbol{v}_u^*\rangle_l \\ &= -\mathcal{F}_u(r_u^*) + \frac{\lambda_{\mathrm{hyp}}\lambda_g}{T}\pi_u^{\mathrm{PR}}.\end{aligned} \tag{206}$$

Consequently,

$$\mathcal{F}_u(r_u^*) = \frac{\lambda_{\mathrm{hyp}}\lambda_g}{T}\pi_u^{\mathrm{PR}}. \tag{207}$$

Applying Eq. (207) to nodes s and t and using their common radial force law gives

$$\mathcal{F}_{\mathrm{out}}(r_s^*) = \mathcal{F}_s(r_s^*) = \frac{\lambda_{\mathrm{hyp}}\lambda_g}{T}\pi_s^{\mathrm{PR}}, \tag{208}$$

$$\mathcal{F}_{\mathrm{out}}(r_t^*) = \mathcal{F}_t(r_t^*) = \frac{\lambda_{\mathrm{hyp}}\lambda_g}{T}\pi_t^{\mathrm{PR}}. \tag{209}$$

Since

$$\frac{\lambda_{\mathrm{hyp}}\lambda_g}{T} > 0, \tag{210}$$

we have

$$\begin{aligned}\pi_s^{\mathrm{PR}} > \pi_t^{\mathrm{PR}} &\Longrightarrow \frac{\lambda_{\mathrm{hyp}}\lambda_g}{T}\pi_s^{\mathrm{PR}} > \frac{\lambda_{\mathrm{hyp}}\lambda_g}{T}\pi_t^{\mathrm{PR}} \\ &\Longrightarrow \mathcal{F}_{\mathrm{out}}(r_s^*) > \mathcal{F}_{\mathrm{out}}(r_t^*) \\ &\Longrightarrow r_s^* < r_t^*,\end{aligned} \tag{211}$$

where the final implication follows from strict monotonic decrease of $\mathcal{F}_{\mathrm{out}}$.

Finally, by Eq. (178) and the strict monotonicity of $r \mapsto \tanh(r/2)$,

$$\begin{aligned}r_s^* < r_t^* &\Longrightarrow \frac{r_s^*}{2} < \frac{r_t^*}{2} \\ &\Longrightarrow \tanh\left(\frac{r_s^*}{2}\right) < \tanh\left(\frac{r_t^*}{2}\right) \\ &\Longrightarrow \|\boldsymbol{e}_s\|_2 = \tanh\left(\frac{r_s^*}{2}\right) < \tanh\left(\frac{r_t^*}{2}\right) = \|\boldsymbol{e}_t\|_2 .\end{aligned} \tag{212}$$

Therefore,

$$\pi_s^{\mathrm{PR}} > \pi_t^{\mathrm{PR}} \quad\Longrightarrow\quad r_s^* < r_t^* \quad\Longrightarrow\quad \|\boldsymbol{e}_s\|_2 < \|\boldsymbol{e}_t\|_2 , \tag{213}$$

which proves the claimed pairwise ordering. □

Remark H.1 (Scope)**.** *The ordering is conditional on equality of the two radial force laws along the stationary geodesic rays and on their strict decrease over an interval containing the two stationary radii. Similarity of local graph neighborhoods alone does not guarantee this condition, because the task loss, gate penalty, contrastive interactions with other nodes, and the remaining trainable variables may break the required radial symmetry. The argument applies to positive stationary radii; at the hyperbolic origin the distance-to-origin function does not admit a unique outward radial direction, and a separate subgradient analysis would be required.*

I PROOF OF THE SUPPORT–LEAKAGE PROPOSITION

The proof uses exactly the conditions stated in the main-text proposition. Fix a sample i and a CAST-KAN layer ℓ. Let $\mathcal{S} \subseteq [T]$ be the designated support, $r := |\mathcal{S}|$, and $q := r/T$. Define the total realized gate energy by

$$E_{i,\ell} := \frac{1}{TK} \sum_{t=1}^{T} \sum_{k=1}^{K} \big(g_{t,k}^{(i,\ell)}\big)^2.$$

To make the definitions valid also at the boundary cases, define

$$E_{i,\ell}^{\text{in}} := \begin{cases} \dfrac{1}{Kr} \displaystyle\sum_{t\in\mathcal{S}} \sum_{k=1}^{K} \big(g_{t,k}^{(i,\ell)}\big)^2, & r > 0, \\ 0, & r = 0, \end{cases} \qquad \eta_{i,\ell}^2 := \begin{cases} \dfrac{1}{K(T-r)} \displaystyle\sum_{t\notin\mathcal{S}} \sum_{k=1}^{K} \big(g_{t,k}^{(i,\ell)}\big)^2, & r < T, \\ 0, & r = T. \end{cases}$$

With these conventions, the identities

$$KrE_{i,\ell}^{\text{in}} = \sum_{t\in\mathcal{S}} \sum_{k=1}^{K} \big(g_{t,k}^{(i,\ell)}\big)^2, \qquad K(T-r)\eta_{i,\ell}^2 = \sum_{t\notin\mathcal{S}} \sum_{k=1}^{K} \big(g_{t,k}^{(i,\ell)}\big)^2$$

hold for every $0 \le r \le T$.

Proof. For every $0 \le r \le T$, the total gate energy admits the exact decomposition

$$\begin{aligned} E_{i,\ell} &= \frac{1}{TK} \left[\sum_{t\in\mathcal{S}} \sum_{k=1}^{K} \big(g_{t,k}^{(i,\ell)}\big)^2 + \sum_{t\notin\mathcal{S}} \sum_{k=1}^{K} \big(g_{t,k}^{(i,\ell)}\big)^2 \right] \\ &= \frac{1}{TK} \left[KrE_{i,\ell}^{\text{in}} + K(T-r)\eta_{i,\ell}^2 \right] = \frac{r}{T} E_{i,\ell}^{\text{in}} + \frac{T-r}{T} \eta_{i,\ell}^2 = qE_{i,\ell}^{\text{in}} + (1-q)\eta_{i,\ell}^2. \end{aligned} \tag{214}$$

Since each gate is obtained through the sigmoid map,

$$0 < g_{t,k}^{(i,\ell)} < 1, \qquad 1 \le t \le T, \qquad 1 \le k \le K. \tag{215}$$

Together with the boundary conventions above, this implies

$$0 \le E_{i,\ell}^{\text{in}} \le 1, \qquad 0 \le \eta_{i,\ell}^2 \le 1, \qquad 0 \le E_{i,\ell} \le 1. \tag{216}$$

Using Eq. (214) and $E_{i,\ell}^{\text{in}} \le 1$, we obtain

$$E_{i,\ell} = qE_{i,\ell}^{\text{in}} + (1-q)\eta_{i,\ell}^2 \le q + (1-q)\eta_{i,\ell}^2. \tag{217}$$

Moreover, the complement of the realized gate energy satisfies

$$\begin{aligned} 1 - E_{i,\ell} &= 1 - qE_{i,\ell}^{\text{in}} - (1-q)\eta_{i,\ell}^2 \\ &= q\big(1 - E_{i,\ell}^{\text{in}}\big) + (1-q)\big(1 - \eta_{i,\ell}^2\big) \ge (1-q)\big(1 - \eta_{i,\ell}^2\big), \end{aligned} \tag{218}$$

where the final inequality follows from $q \ge 0$ and $E_{i,\ell}^{\text{in}} \le 1$.

For a threshold $\tau \in (0, 1]$, define the number of off-support gates whose realized values exceed τ by

$$k_{\text{out}}^{(i,\ell)}(\tau) = \sum_{t\notin\mathcal{S}} \sum_{k=1}^{K} \mathbb{1}\left[g_{t,k}^{(i,\ell)} > \tau \right]. \tag{219}$$

For every $a \ge 0$ and every $\tau > 0$,

$$\mathbb{1}[a > \tau] \le \frac{a^2}{\tau^2}. \tag{220}$$

Applying this inequality pointwise to the off-support gate values gives

$$\begin{aligned} k_{\text{out}}^{(i,\ell)}(\tau) &= \sum_{t\notin\mathcal{S}} \sum_{k=1}^{K} \mathbb{1}\left[g_{t,k}^{(i,\ell)} > \tau \right] \\ &\le \frac{1}{\tau^2} \sum_{t\notin\mathcal{S}} \sum_{k=1}^{K} \big(g_{t,k}^{(i,\ell)}\big)^2 = \frac{K(T-r)\eta_{i,\ell}^2}{\tau^2}. \end{aligned} \tag{221}$$

When $r = T$, both sides are zero because the off-support index set is empty and $\eta_{i,\ell}^2 = 0$ by definition. When $r = 0$, the same calculation applies to all T temporal positions. Hence the bound holds for every $0 \le r \le T$. □

A layer-averaged form follows by defining

$$\overline{E}_i = \frac{1}{L}\sum_{\ell=1}^{L} E_{i,\ell}, \qquad \overline{E}_i^{\text{in}} = \frac{1}{L}\sum_{\ell=1}^{L} E_{i,\ell}^{\text{in}}, \qquad \overline{\eta}_i^{\,2} = \frac{1}{L}\sum_{\ell=1}^{L} \eta_{i,\ell}^2. \tag{222}$$

Averaging Eq. (214) over layers and using the preceding bounds gives

$$\overline{E}_i = \frac{1}{L}\sum_{\ell=1}^{L}\left[qE_{i,\ell}^{\text{in}} + (1-q)\eta_{i,\ell}^2\right] = q\overline{E}_i^{\text{in}} + (1-q)\overline{\eta}_i^{\,2} \le q + (1-q)\overline{\eta}_i^{\,2}, \qquad 1-\overline{E}_i \ge (1-q)\big(1-\overline{\eta}_i^{\,2}\big). \tag{223}$$

Remark I.1 (Scope)**.** *The decomposition characterizes the realized gate pattern rather than the number of stored model parameters. In the gate-energy and sensitivity experiments, $\mathcal{S}$ denotes the synthetic target-support set rather than the edge-incidence set of the learned structural graph. Consequently, the proposition quantifies on-support realized gate energy and off-support leakage with respect to that designated support.*

J Proof of the Gate-Conditioned Sensitivity Proposition

The proof uses exactly the conditions stated in the main-text proposition. Throughout the argument, the gate vectors and structure-conditioned rotations are held fixed when evaluating the corresponding Lipschitz constants.

Proof. Fix a sample i, temporal position t, and layer ℓ, and temporarily suppress the indices (i, t, ℓ). Define the gated basis map

$$\Psi_{\boldsymbol{g}}(\boldsymbol{h}) = \left(g_k\varphi_k(h_c)\right)_{c\in[C],\,k\in[K]} \in \mathbb{R}^{CK}. \tag{224}$$

For any $\boldsymbol{h}, \boldsymbol{h}' \in \mathbb{R}^C$ whose coordinates lie in the analysis domain of the basis functions, the uniform L_φ-Lipschitz condition gives

$$\begin{aligned}\|\Psi_{\boldsymbol{g}}(\boldsymbol{h}) - \Psi_{\boldsymbol{g}}(\boldsymbol{h}')\|_2^2 &= \sum_{c=1}^{C}\sum_{k=1}^{K} g_k^2\left[\varphi_k(h_c) - \varphi_k(h_c')\right]^2 \\ &\le \sum_{c=1}^{C}\sum_{k=1}^{K} g_k^2 L_\varphi^2 (h_c - h_c')^2 = L_\varphi^2\left(\sum_{k=1}^{K} g_k^2\right)\left(\sum_{c=1}^{C}(h_c - h_c')^2\right) \\ &= L_\varphi^2\,\|\boldsymbol{g}\|_2^2\,\|\boldsymbol{h} - \boldsymbol{h}'\|_2^2\,.\end{aligned} \tag{225}$$

Taking square roots on both sides yields

$$\|\Psi_{\boldsymbol{g}}(\boldsymbol{h}) - \Psi_{\boldsymbol{g}}(\boldsymbol{h}')\|_2 \le L_\varphi\,\|\boldsymbol{g}\|_2\,\|\boldsymbol{h} - \boldsymbol{h}'\|_2\,. \tag{226}$$

Hence

$$\mathrm{Lip}(\Psi_{\boldsymbol{g}}) \le L_\varphi\,\|\boldsymbol{g}\|_2\,. \tag{227}$$

For the structure-conditioned rotation, let $M = \min\{d, \lfloor C/2 \rfloor\}$. At fixed (i, t, ℓ), the channel transformation in Eq. (17) is represented by the block-diagonal matrix

$$\mathcal{R}_{i,t}^{(\ell)} = \mathrm{diag}\left(R(\theta_{t,0}^{(i)}), \ldots, R(\theta_{t,M-1}^{(i)}), \boldsymbol{I}_{C-2M}\right), \qquad R(\theta) = \begin{pmatrix}\cos\theta & -\sin\theta \\ \sin\theta & \cos\theta\end{pmatrix}. \tag{228}$$

Each 2×2 block satisfies $R(\theta)^\top R(\theta) = \boldsymbol{I}_2$, while the unmatched channels are unchanged. Therefore

$$\left(\mathcal{R}_{i,t}^{(\ell)}\right)^\top \mathcal{R}_{i,t}^{(\ell)} = \boldsymbol{I}_C, \qquad \left\|\mathcal{R}_{i,t}^{(\ell)}\right\|_2 = 1, \qquad \left\|\mathcal{R}_{i,t}^{(\ell)}\boldsymbol{h} - \mathcal{R}_{i,t}^{(\ell)}\boldsymbol{h}'\right\|_2 = \|\boldsymbol{h} - \boldsymbol{h}'\|_2\,. \tag{229}$$

Let $\boldsymbol{W}^{(\ell)} \in \mathbb{R}^{C\times CK}$ denote the flattened linear operator whose entries are the CAST-KAN weights $w_{j,c,k}^{(\ell)}$, and assume $\left\|\boldsymbol{W}^{(\ell)}\right\|_2 \le B_w$ as in the standing conditions. The gated nonlinear branch can then be written as

$$G_{i,t}^{(\ell)}(\boldsymbol{h}) = \boldsymbol{W}^{(\ell)}\Psi_{\boldsymbol{g}_t^{(i,\ell)}}\left(\mathcal{R}_{i,t}^{(\ell)}\boldsymbol{h}\right) + \boldsymbol{b}^{(\ell)}, \qquad \left\|\boldsymbol{W}^{(\ell)}\right\|_2 \le B_w. \tag{230}$$

For any $\boldsymbol{h}, \boldsymbol{h}' \in \mathbb{R}^C$, the bias cancels and

$$\begin{aligned}\left\|G_{i,t}^{(\ell)}(\boldsymbol{h}) - G_{i,t}^{(\ell)}(\boldsymbol{h}')\right\|_2 &= \left\|\boldsymbol{W}^{(\ell)}\left[\Psi_{\boldsymbol{g}_t^{(i,\ell)}}\left(\mathcal{R}_{i,t}^{(\ell)}\boldsymbol{h}\right) - \Psi_{\boldsymbol{g}_t^{(i,\ell)}}\left(\mathcal{R}_{i,t}^{(\ell)}\boldsymbol{h}'\right)\right]\right\|_2 \\ &\le \left\|\boldsymbol{W}^{(\ell)}\right\|_2 \left\|\Psi_{\boldsymbol{g}_t^{(i,\ell)}}\left(\mathcal{R}_{i,t}^{(\ell)}\boldsymbol{h}\right) - \Psi_{\boldsymbol{g}_t^{(i,\ell)}}\left(\mathcal{R}_{i,t}^{(\ell)}\boldsymbol{h}'\right)\right\|_2 \\ &\le B_w L_\varphi \left\|\boldsymbol{g}_t^{(i,\ell)}\right\|_2 \left\|\mathcal{R}_{i,t}^{(\ell)}\boldsymbol{h} - \mathcal{R}_{i,t}^{(\ell)}\boldsymbol{h}'\right\|_2 \\ &= B_w L_\varphi \left\|\boldsymbol{g}_t^{(i,\ell)}\right\|_2 \|\boldsymbol{h}-\boldsymbol{h}'\|_2 = B_w L_\varphi \sqrt{K}\,\gamma_{i,t,\ell}\,\|\boldsymbol{h}-\boldsymbol{h}'\|_2 .\end{aligned} \tag{231}$$

Consequently,

$$\mathrm{Lip}\Big(G_{i,t}^{(\ell)} \mid g\Big) \le B_w L_\varphi \left\|\boldsymbol{g}_t^{(i,\ell)}\right\|_2 = B_w L_\varphi \sqrt{K}\,\gamma_{i,t,\ell},$$

which proves the first inequality in Eq. (32).

For completeness, if the nonlinear branches are composed without residual connections and

$$G_{i,t}^{(1:L)} := G_{i,t}^{(L)} \circ G_{i,t}^{(L-1)} \circ \cdots \circ G_{i,t}^{(1)},$$

then submultiplicativity of Lipschitz constants gives

$$\mathrm{Lip}\Big(G_{i,t}^{(1:L)} \mid g\Big) \le \prod_{\ell=1}^{L} \mathrm{Lip}\Big(G_{i,t}^{(\ell)} \mid g\Big) \le \prod_{\ell=1}^{L}\Big(B_w L_\varphi \sqrt{K}\,\gamma_{i,t,\ell}\Big) = \Big(B_w L_\varphi \sqrt{K}\Big)^L \prod_{\ell=1}^{L}\gamma_{i,t,\ell}. \tag{232}$$

Now consider the residualized layer

$$H_{i,t}^{(\ell)}(\boldsymbol{h}) := \boldsymbol{h} + G_{i,t}^{(\ell)}(\boldsymbol{h}).$$

For any $\boldsymbol{h}, \boldsymbol{h}' \in \mathbb{R}^C$,

$$\begin{aligned}\left\|H_{i,t}^{(\ell)}(\boldsymbol{h}) - H_{i,t}^{(\ell)}(\boldsymbol{h}')\right\|_2 &= \left\|(\boldsymbol{h}-\boldsymbol{h}') + G_{i,t}^{(\ell)}(\boldsymbol{h}) - G_{i,t}^{(\ell)}(\boldsymbol{h}')\right\|_2 \\ &\le \|\boldsymbol{h}-\boldsymbol{h}'\|_2 + \left\|G_{i,t}^{(\ell)}(\boldsymbol{h}) - G_{i,t}^{(\ell)}(\boldsymbol{h}')\right\|_2 \\ &\le \left(1 + B_w L_\varphi \sqrt{K}\,\gamma_{i,t,\ell}\right)\|\boldsymbol{h}-\boldsymbol{h}'\|_2 .\end{aligned} \tag{233}$$

Therefore,

$$\mathrm{Lip}\Big(H_{i,t}^{(\ell)} \mid g\Big) \le 1 + B_w L_\varphi \sqrt{K}\,\gamma_{i,t,\ell}.$$

Defining

$$H_{i,t}^{(1:L)} := H_{i,t}^{(L)} \circ H_{i,t}^{(L-1)} \circ \cdots \circ H_{i,t}^{(1)},$$

we obtain

$$\mathrm{Lip}\Big(H_{i,t}^{(1:L)} \mid g\Big) \le \prod_{\ell=1}^{L} \mathrm{Lip}\Big(H_{i,t}^{(\ell)} \mid g\Big) \le \prod_{\ell=1}^{L}\Big(1 + B_w L_\varphi \sqrt{K}\,\gamma_{i,t,\ell}\Big), \tag{234}$$

which proves the second inequality in Eq. (32). □

Corollary J.1 (Average branch sensitivity)**.** *For a fixed sample i and layer ℓ, with the gate vectors and structure-conditioned rotations held fixed,*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathrm{Lip}\Big(G_{i,t}^{(\ell)} \mid g\Big)^2 &\le \frac{B_w^2 L_\varphi^2}{T}\sum_{t=1}^{T}\left\|\boldsymbol{g}_t^{(i,\ell)}\right\|_2^2 = \frac{B_w^2 L_\varphi^2}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\big(g_{t,k}^{(i,\ell)}\big)^2 \\ &= B_w^2 L_\varphi^2 K\left[\frac{1}{TK}\sum_{t=1}^{T}\sum_{k=1}^{K}\big(g_{t,k}^{(i,\ell)}\big)^2\right] = B_w^2 L_\varphi^2 K E_{i,\ell} \\ &\le B_w^2 L_\varphi^2 K\left[q + (1-q)\eta_{i,\ell}^2\right].\end{aligned} \tag{235}$$

Corollary J.2 (Within-mask local stability)**.** *Fix a trained CAST model and two inputs $\boldsymbol{X}, \boldsymbol{X}'$ with the same activity mask,*

$$\boldsymbol{M}(\boldsymbol{X}) = \boldsymbol{M}(\boldsymbol{X}') = \boldsymbol{M}. \tag{236}$$

Because the sample-adaptive structural representation in Eq. (16) *depends on the input through the activity mask, the two inputs induce the same structure-conditioned quantities $\boldsymbol{\phi}_t^{(\boldsymbol{M})}$, $\mathcal{R}_t^{(\boldsymbol{M},\ell)}$, and $\boldsymbol{g}_t^{(\boldsymbol{M},\ell)}$ at every position and layer. Define*

$$\gamma_\ell(\boldsymbol{M}) := \max_{1\le t\le T} \frac{\left\|\boldsymbol{g}_t^{(\boldsymbol{M},\ell)}\right\|_2}{\sqrt{K}}. \tag{237}$$

If the CAST-KAN backbone uses residualized layers and the input-side transformation covered by the standing conditions is non-expansive, then

$$\|F_{\mathrm{back}}(\boldsymbol{X}) - F_{\mathrm{back}}(\boldsymbol{X}')\|_F \le \prod_{\ell=1}^{L} \left(1 + B_w L_\varphi \sqrt{K}\, \gamma_\ell(\boldsymbol{M})\right) \|\boldsymbol{X} - \boldsymbol{X}'\|_F . \tag{238}$$

If the subsequent pooling and prediction map is L_{head}-Lipschitz on the analysis domain, then

$$\begin{aligned} |f_{\mathrm{CAST}}(\boldsymbol{X}) - f_{\mathrm{CAST}}(\boldsymbol{X}')| &\le L_{\mathrm{head}} \|F_{\mathrm{back}}(\boldsymbol{X}) - F_{\mathrm{back}}(\boldsymbol{X}')\|_F \\ &\le L_{\mathrm{head}} \prod_{\ell=1}^{L} \left(1 + B_w L_\varphi \sqrt{K}\, \gamma_\ell(\boldsymbol{M})\right) \|\boldsymbol{X} - \boldsymbol{X}'\|_F . \end{aligned} \tag{239}$$

Proof. Let $\boldsymbol{h}_t^{(0)}$ and $\boldsymbol{h}_t'^{(0)}$ denote the layer-0 representations induced by $\boldsymbol{X}$ and $\boldsymbol{X}'$, respectively. The common activity mask implies

$$\boldsymbol{\phi}_t^{(\boldsymbol{X})} = \boldsymbol{\phi}_t^{(\boldsymbol{X}')} = \boldsymbol{\phi}_t^{(\boldsymbol{M})}, \qquad \boldsymbol{g}_t^{(\boldsymbol{X},\ell)} = \boldsymbol{g}_t^{(\boldsymbol{X}',\ell)} = \boldsymbol{g}_t^{(\boldsymbol{M},\ell)}, \qquad \mathcal{R}_t^{(\boldsymbol{X},\ell)} = \mathcal{R}_t^{(\boldsymbol{X}',\ell)} = \mathcal{R}_t^{(\boldsymbol{M},\ell)}. \tag{240}$$

For each layer ℓ and temporal position t, the residual bound proved above therefore applies with the same gate and rotation to both trajectories:

$$\begin{aligned} \left\|\boldsymbol{h}_t^{(\ell+1)} - \boldsymbol{h}_t'^{(\ell+1)}\right\|_2 &\le \left(1 + B_w L_\varphi \sqrt{K}\, \gamma_{i,t,\ell}\right) \left\|\boldsymbol{h}_t^{(\ell)} - \boldsymbol{h}_t'^{(\ell)}\right\|_2 \\ &\le \left(1 + B_w L_\varphi \sqrt{K}\, \gamma_\ell(\boldsymbol{M})\right) \left\|\boldsymbol{h}_t^{(\ell)} - \boldsymbol{h}_t'^{(\ell)}\right\|_2 . \end{aligned} \tag{241}$$

Squaring, summing over temporal positions, and using that the multiplicative factor is independent of t gives

$$\sum_{t=1}^{T} \left\|\boldsymbol{h}_t^{(\ell+1)} - \boldsymbol{h}_t'^{(\ell+1)}\right\|_2^2 \le \left(1 + B_w L_\varphi \sqrt{K}\, \gamma_\ell(\boldsymbol{M})\right)^2 \sum_{t=1}^{T} \left\|\boldsymbol{h}_t^{(\ell)} - \boldsymbol{h}_t'^{(\ell)}\right\|_2^2 . \tag{242}$$

Taking square roots yields

$$\left[\sum_{t=1}^{T} \left\|\boldsymbol{h}_t^{(\ell+1)} - \boldsymbol{h}_t'^{(\ell+1)}\right\|_2^2\right]^{1/2} \le \left(1 + B_w L_\varphi \sqrt{K}\, \gamma_\ell(\boldsymbol{M})\right) \left[\sum_{t=1}^{T} \left\|\boldsymbol{h}_t^{(\ell)} - \boldsymbol{h}_t'^{(\ell)}\right\|_2^2\right]^{1/2} .$$

Iterating this inequality over $\ell = 1, \dots, L$ gives

$$\begin{aligned} \|F_{\mathrm{back}}(\boldsymbol{X}) - F_{\mathrm{back}}(\boldsymbol{X}')\|_F &\le \prod_{\ell=1}^{L} \left(1 + B_w L_\varphi \sqrt{K}\, \gamma_\ell(\boldsymbol{M})\right) \left[\sum_{t=1}^{T} \left\|\boldsymbol{h}_t^{(0)} - \boldsymbol{h}_t'^{(0)}\right\|_2^2\right]^{1/2} \\ &\le \prod_{\ell=1}^{L} \left(1 + B_w L_\varphi \sqrt{K}\, \gamma_\ell(\boldsymbol{M})\right) \|\boldsymbol{X} - \boldsymbol{X}'\|_F , \end{aligned} \tag{243}$$

where the final inequality follows from the non-expansiveness of the input-side transformation assumed in the standing conditions. Applying the L_{head}-Lipschitz bound to the subsequent pooling and prediction map gives Eq. (239). □

Remark J.3 (Scope)**.** *The sensitivity bounds condition on the realized gate vectors and structure-conditioned rotations. The within-mask result converts this conditional statement into a local input-stability statement because inputs sharing the same activity mask induce the same sample-conditioned structural representation, rotations, and gates. If an input perturbation changes the activity mask, these quantities may change discontinuously at an exact-zero boundary, and the within-mask argument no longer applies. The result therefore characterizes stability within a common observation-state stratum rather than across mask transitions.*

K Summary of Theoretical Contributions

For reference, the principal statements are Propositions 4.1–4.5. They concern, respectively, population targeting, graph-metric representation, conditional radial ordering, realized gate energy, and gate-conditioned sensitivity. The main text states the propositions and standing assumptions; the preceding appendix sections provide supporting definitions, derivations, corollaries, and local technical conditions.

L Empirical Evaluation of Theoretical Mechanisms with Synthetic Data

To evaluate the mechanisms characterized by the propositions in Section 4, we conduct controlled synthetic experiments. The experiments are not used as substitutes for the proofs; instead, each scenario is chosen to instantiate the corresponding assumptions and to test whether the predicted qualitative behavior appears at finite sample sizes.

L.1 Synthetic Data Generation Protocols

The synthetic datasets use a linear SEM over a temporally ordered DAG. To avoid a matrix-orientation ambiguity, we distinguish the simulator coefficient matrix from the source-by-target adjacency used in the main text. Let $\boldsymbol{B}$ use the target-by-source convention, so $B_{t,s}$ is the coefficient multiplying source s in the structural equation for target t; the paper-level adjacency is $\boldsymbol{A} = \boldsymbol{B}^\top$, hence $A_{s,t} = B_{t,s}$. With row-vector samples, the reported generator is

$$\boldsymbol{X} = \boldsymbol{E}(\boldsymbol{I} - \boldsymbol{B}^\top)^{-1}, \qquad E_{i,t} \sim \mathcal{N}(0, \sigma^2). \tag{244}$$

Because the temporally ordered coefficient matrix is strictly triangular, $\boldsymbol{I} - \boldsymbol{B}^\top$ is invertible exactly; no spectral-radius condition is required for acyclicity or invertibility. The implementation retains the same edge-weight scaling used to generate the reported data, but that numerical scaling is not used as a theoretical stability premise.

To simulate zero inflation, we introduce a stochastic mask $\boldsymbol{M} \in \{0,1\}^{N\times T}$ with $m_{i,t} \sim \mathrm{Ber}(1-\pi)$ and form $\widetilde{\boldsymbol{X}} = \boldsymbol{M} \odot \boldsymbol{X}$. In the hurdle stress test the mask is generated independently of the SEM coefficients and active residuals, so the mask itself carries no additional edge information. This is intentionally an *observation-model stress test*, not an exact simulator of the full hurdle SEM in Eq. (1): predictor coordinates are also masked before structural fitting. Proposition 4.1 isolates the response-side population target shift caused by treating a masked zero as an ordinary Gaussian response; the complete hurdle stress-test experiment additionally contains an errors-in-variables effect from masked predictors. We therefore interpret its SHD/TPR curves as complementary finite-sample evidence rather than as a direct numerical verification of the score formula.

For classification and sequence modeling tasks, the binary target y is synthesized via a nonlinear decision boundary spanning a designated sparse target-support set $\mathcal{S}$:

$$y = \mathbb{1}\left[\sigma\left(\sum_{j\in\mathcal{S}} f(X_j) + \varepsilon\right) > 0.5\right], \tag{245}$$

where $f(\cdot)$ applies non-linear transformations (e.g., sine, quadratic) and ε is Gaussian noise.

The geometric hierarchy is systematically organized by mapping nodes into the continuous space shaped by the PageRank regularizer:

$$\Omega(\mathcal{H}) = \lambda_g \sum_{t=1}^{T} \pi_t^{\mathrm{PR}} \cdot d_l(\boldsymbol{p}_t, \boldsymbol{p}_o). \tag{246}$$

To avoid redundant text descriptions, the specific topological configurations and distributional hyperparameters across all five evaluation scenarios are systematically detailed in Table 4.

Table 4: **Synthetic data generation and experimental parameters.** Summary of the topological and distributional settings used to evaluate the five theoretical mechanisms.

Evaluation Scenario	Hyperparameter	Values / Range
ZI-SEM Recovery	Number of nodes (T)	20
	Base edge probability (p_e)	0.15
	Zero-inflation rate (π)	$\{0.0, 0.1, 0.3, 0.5, 0.7\}$
Hyperbolic Distortion	Tree sizes (T)	$\{15, 31, 63, 127\}$
	Manifold dimensions (d)	$\{2, 3, 5, 10\}$
Hierarchical Reg.	Total nodes (T)	21 (Ratio 3 : 6 : 12)
	Skip connection probability	0.15
	PageRank damping factor	0.15
Gating Activation	Sequence length (T)	30
	Target-support fraction ($\rho_s = \lvert\mathcal{S}\rvert/T$)	$\{0.05, 0.1, 0.2, 0.3, 0.5, 0.8\}$
Low-Sample Predictive Behavior	Total features (T)	50
	Target-support size ($r = \lvert\mathcal{S}\rvert$)	$\{3, 5, 10, 25\}$
	Training sample size (N)	$[100, 5000]$ (Log-scaled)

L.2 Empirical Evaluation Results

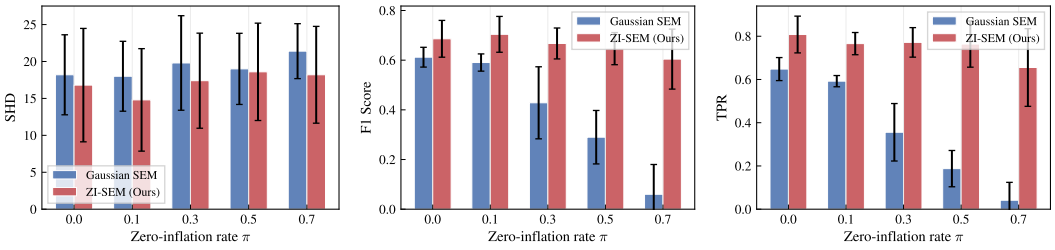
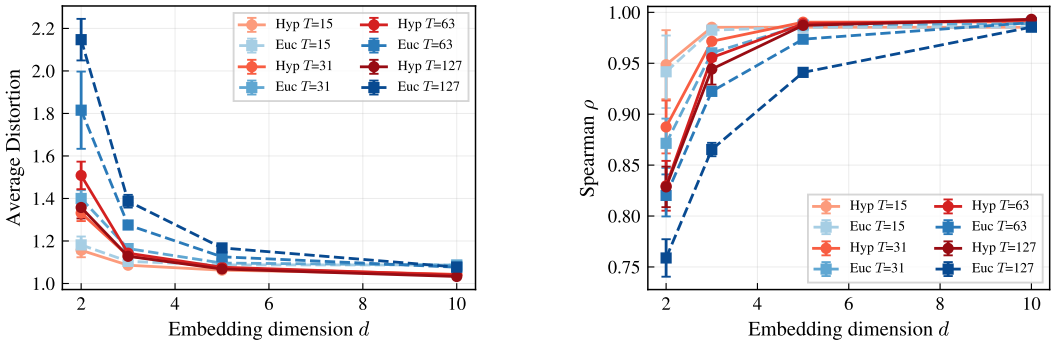


Figure 4: **Empirical evaluation of two-part structural learning under masking.** Performance comparison between the Gaussian baseline and CAST across varying sparsity levels. From left to right: structural distance, edge score, and true positive rate.

Evaluation of Two-Part Structural Learning under Masking. Figure 4 evaluates the complete finite-sample masking stress test associated with Proposition 4.1. Because the Bernoulli mask is independent of the SEM coefficients and active residuals, inactive indicators do not provide extra edge information. The experiment nevertheless differs from the proposition in one important respect: predictor coordinates are also masked, so structural fitting contains an errors-in-variables component in addition to the response-side target shift isolated by Eq. (27). At zero inflation 0.7, the standard baseline attains a true positive rate of 0.04 and an edge score of 0.06, whereas CAST retains an edge score of 0.60 and a true positive rate of 0.65; structural distance is also lower across the evaluated sparsity levels. These curves are therefore consistent with the motivation for the hurdle objective, but they are not claimed as a finite-sample consequence of the KL/score calculation alone.

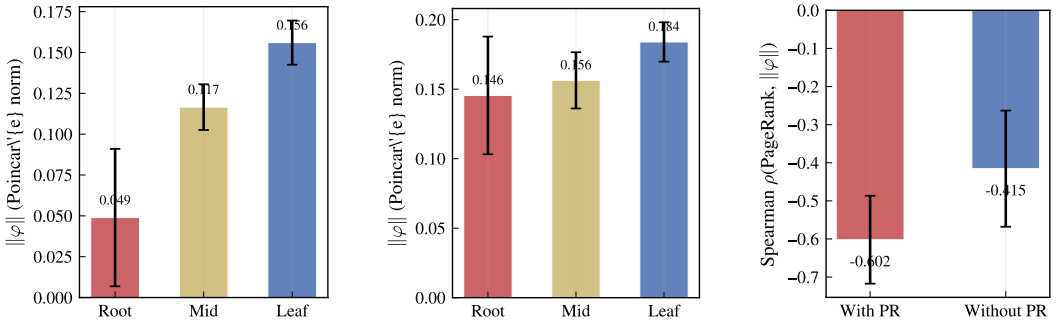
Evaluation of Hyperbolic Metric Representation. The tree-approximation graphs are deliberately tree-like, matching the regime covered by Proposition 4.2. Figure 5 shows that the measured average



(a) Space Distortion Comparison (b) Hierarchical Rank Correlation

Figure 5: **Empirical evaluation of hyperbolic representation.** Performance comparison between hyperbolic and Euclidean embeddings across varying dimensions and sequence lengths. From left to right: average geometric distortion and hierarchical rank correlation.

distortion of the hyperbolic embeddings is lower than the Euclidean alternative across all tested dimensions and sequence lengths. For $T = 127$ in two dimensions, the Euclidean distortion exceeds 2.15, while the hyperbolic distortion is near 1.30. The corresponding hierarchical rank correlation remains above 0.83 for the hyperbolic representation and drops to 0.76 for the Euclidean baseline. These finite experiments are consistent with the low-distortion behavior predicted for tree-like metrics; we do not extrapolate the asymptotic Euclidean lower bound to arbitrary real-world temporal graphs.



(a) Norms with Regularization (b) Norms without Regularization (c) Rank Correlation

Figure 6: **Empirical evaluation of hierarchical regularization.** The panels compare the spatial distribution of nodes based on their topological roles. From left to right: average distances to the origin with the regularization mechanism, distances without the mechanism, and the resulting hierarchical rank correlation.

Evaluation of Hierarchical Regularization. Proposition 4.3 proves a pairwise radial ordering result under a common monotone force law; Figure 6 asks whether the same tendency is visible globally. With PageRank regularization, root nodes have average radius 0.05 and leaf nodes average radius 0.16, producing a PageRank–radius rank correlation of -0.60. Without the regularizer, the root-node radius increases to 0.15 and the correlation weakens to -0.42. The experiment therefore supports the intended global ordering effect while remaining logically distinct from the conditional pairwise proposition.

Evaluation of Realized Gate Energy. Figure 7 varies the target-support fraction $\rho_s = |\mathcal{S}|/T$, the same quantity denoted by q in Proposition 4.4. The ungated baseline remains at 8 active basis functions across support fractions, whereas the learned gates reduce realized effective activation as the designated support becomes smaller. At target-support fraction 0.8, the measured effective activation is 2.5 basis functions, and it approaches zero at the lowest fractions. Because CAST shares KAN weights across time, we interpret this as reduced *realized basis activation*, not literal deletion of stored parameters. Proposition 4.4 should be read as a post-training certificate: the empirical

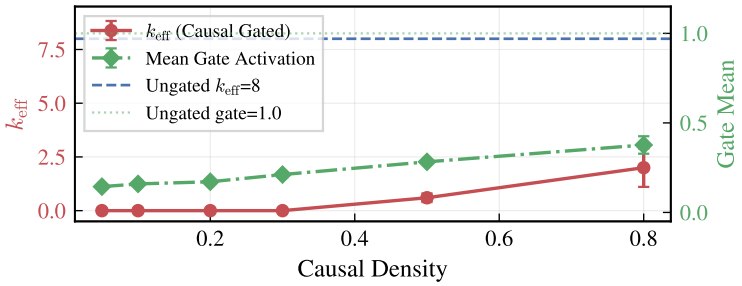
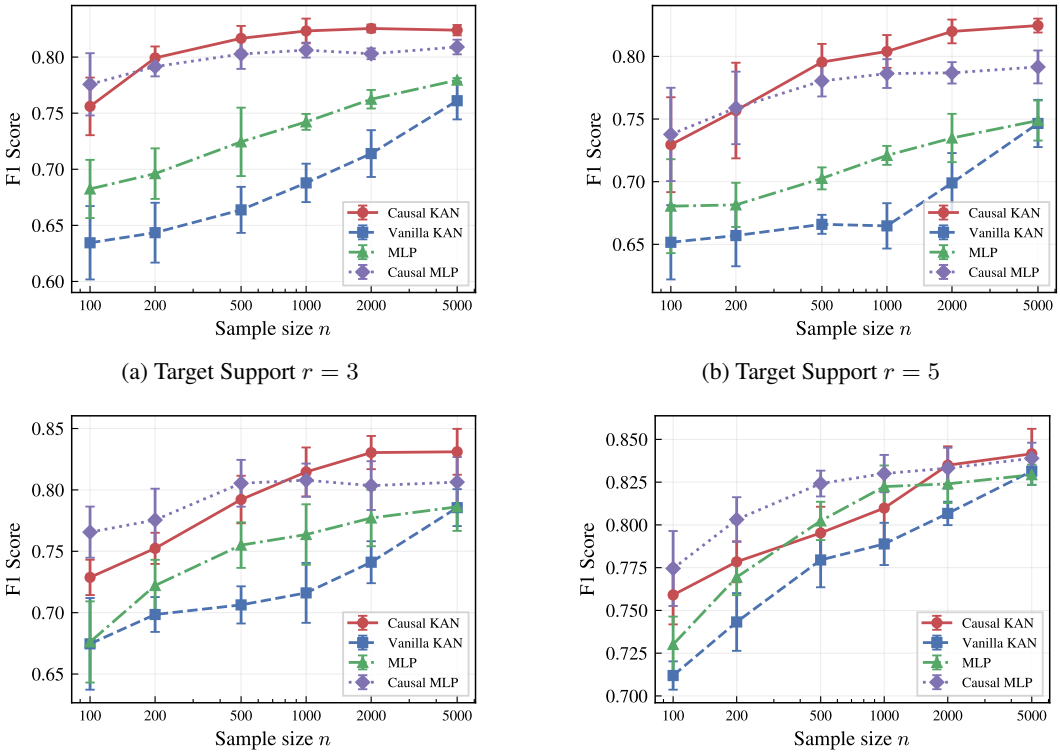


Figure 7: **Empirical evaluation of realized gate energy.** The plot reports the realized effective basis activation while varying the designated target-support fraction, interpreted as activation-level sparsification rather than stored-parameter pruning.

curve supplies the nontrivial observation that the learned leakage η is small enough for the bound in Eq. (31) to be informative.



(a) Target Support $r = 3$

(b) Target Support $r = 5$

(c) Target Support $r = 10$

(d) Target Support $r = 25$

Figure 8: **Empirical evaluation of gate-conditioned predictive behavior.** The panels display held-out predictive performance across training sample sizes while increasing the designated target-support size from $r = 3$ to $r = 25$. The experiment varies predictive support size, not graph density.

Evaluation of Gate-Conditioned Predictive Behavior. Figure 8 varies the designated target-support size $r \in \{3, 5, 10, 25\}$ while changing the number of training examples. With only 100 training samples and support size $r = 3$, CAST achieves a predictive score of 0.75 compared with 0.63 for the unconstrained baseline. The regularized model also maintains a performance margin as the support increases to 25. Proposition 4.5 does not assert an exact $(r/T)^L$ sample-complexity law; instead, it shows that, conditional on the realized structural modulation, smaller gate norms contract the local KAN sensitivity by the multiplicative factor $\prod_\ell \gamma_{i,t,\ell}$. The synthetic curves provide the complementary end-to-end observation that smaller target support is translated by training into

Table 6: Fixed historical observation and future target windows.

Dataset	Bin	T_d	H_{pri}	H_{aux}
CDNOW	month	12	3	3
TAFENG	week	12	4	4
RETAIL	month	12	3	3
INSTACART	week	12	4	4
SALES-WEEKLY	week	52	13	52

better held-out performance in this controlled setting. This predictive pattern is not itself a causal-identification claim.

M More Details of the Real-World Datasets and Modeling Tasks

The datasets utilized in this study represent diverse behavioral patterns across the financial, retail, and electronic commerce sectors. All public benchmarks are preprocessed into consistent temporal aggregates to ensure alignment across modalities. To provide a concise overview, the fundamental properties of these benchmarks are summarized in Table 5.

Each benchmark defines a prediction problem at a fixed cutoff: the model sees only a historical observation window, whereas the binary outcome is constructed only from the subsequent target window. We next specify the windows, label functions, split protocol, and leakage controls before summarizing the dataset-specific tasks in Table 7.

Table 5: Summary statistics of evaluated datasets.

Dataset	Samples	Pos. Ratio	Sparsity
CDNOW	23,502	35.15%	84.53%
TAFENG	32,266	35.00%	82.83%
RETAIL	4,338	34.99%	76.85%
INSTACART	102,589	35.00%	47.30%
SALES WEEKLY	811	35.02%	28.22%

M.1 Prediction Windows, Label Construction, and Leakage Control

Observation and target windows. Let $\mathcal{R}_i$ denote the timestamped records of entity i and let τ be its prediction cutoff. For dataset d and task k, we use non-overlapping historical and future windows

$$\mathcal{W}_{i,\tau}^{\text{in}} = [\tau - T_d + 1, \tau], \qquad \mathcal{W}_{i,\tau}^{\text{out},k} = [\tau + 1, \tau + H_{d,k}], \tag{247}$$

with the fixed aggregation and horizon choices in Table 6. An entity is retained only when the complete historical window and the corresponding future window are observable. We require at least one qualifying event in the historical window, but never condition eligibility on future activity or on the future label. For retail transactions, qualifying purchases have positive quantity and positive unit price; cancelled transactions, returns, and administrative records are excluded from purchase counts and monetary totals. Instacart uses the released order and product records, and Sales-Weekly uses the published weekly sales series without outcome-dependent filtering.

All model inputs, including the raw aggregate sequence $\boldsymbol{x}_{i,\tau} = \text{Aggregate}_d(\mathcal{R}_i|_{\mathcal{W}_{i,\tau}^{\text{in}}})$ and its activity mask, are functions of $\mathcal{W}_{i,\tau}^{\text{in}}$ only. Records in $\mathcal{W}_{i,\tau}^{\text{out},k}$ are inaccessible to CAST and are used solely to construct the target defined below.

Training-only thresholding. Within a future target window, let A_h be positive monetary amount, C_h the number of qualifying transactions or orders, and U_h the number of purchased units in bin h. For Ta-Feng, let B_j be the number of distinct future baskets containing product j; for Sales-Weekly, let S_h denote weekly sales. Empty sums are zero, $(z)_+ = \max(z, 0)$, and $\epsilon = 10^{-8}$. Every empirical distribution and every non-churn label threshold is estimated after the split from training entities only. For a scalar training quantity v, define

$$\widehat{F}_v^{\text{tr}}(z) = \frac{1}{N_{\text{tr}}} \sum_{n=1}^{N_{\text{tr}}} \mathbb{1}[v_n \le z], \qquad Q_{0.65}^{\text{tr}}(v) = \inf\{z : \widehat{F}_v^{\text{tr}}(z) \ge 0.65\}. \tag{248}$$

For each non-churn task, a larger score denotes the named positive class and

$$y_{i,\tau}^{(k)} = \mathbb{1}\left[s_{i,\tau}^{(k)} \geq Q_{0.65}^{\mathrm{tr}}\left(s^{(k)}\right)\right]. \tag{249}$$

The threshold is frozen for validation and test data. Ties are neither randomly broken nor rebalanced, so the realized positive fraction can differ slightly from 35%.

CDNOW and RETAIL: future value and churn. Future value is the log-transformed total positive purchase amount in the next three months,

$$s_{\mathrm{value}} = \log\left(1 + \sum_{h=1}^{3} A_h\right), \qquad y_{\mathrm{churn}} = \mathbb{1}\left[\sum_{h=1}^{3} C_h = 0\right]. \tag{250}$$

The high-value target follows Eq. (249); churn requires no fitted threshold. Because purchase counts use qualifying positive purchases only, returns do not count as continuing activity.

INSTACART: future activity and churn. Let

$$O = \log\left(1 + \sum_{h=1}^{4} C_h\right), \qquad V = \log\left(1 + \sum_{h=1}^{4} U_h\right). \tag{251}$$

Future activity combines order frequency and basket volume through training-set empirical ranks rather than unit-dependent hand-tuned coefficients:

$$s_{\mathrm{activity}} = \frac{1}{2}\left[\widehat{F}_O^{\mathrm{tr}}(O) + \widehat{F}_V^{\mathrm{tr}}(V)\right], \qquad y_{\mathrm{churn}} = \mathbb{1}\left[\sum_{h=1}^{4} C_h = 0\right]. \tag{252}$$

The high-activity label follows Eq. (249). In particular, future basket-size statistics are target-side quantities and never enter the model input.

TAFENG: future risk and repurchase. For the four-week target window, define future inactivity, amount intensity, and transaction intensity by

$$R_0 = \frac{1}{4}\sum_{h=1}^{4} \mathbb{1}[C_h = 0], \qquad M = \log\left(1 + \sum_{h=1}^{4} A_h\right), \qquad F = \log\left(1 + \sum_{h=1}^{4} C_h\right). \tag{253}$$

The risk score assigns equal rank weight to high inactivity, low expenditure, and low transaction frequency:

$$s_{\mathrm{risk}} = \frac{1}{3}\left[\widehat{F}_{R_0}^{\mathrm{tr}}(R_0) + 1 - \widehat{F}_M^{\mathrm{tr}}(M) + 1 - \widehat{F}_F^{\mathrm{tr}}(F)\right]. \tag{254}$$

The coefficients are fixed by definition and are not tuned on validation performance. Repurchase is measured across distinct future baskets,

$$s_{\mathrm{repurchase}} = \begin{cases} \dfrac{\sum_j (B_j - 1)_+}{\sum_j B_j}, & \sum_j B_j > 0, \\ 0, & \sum_j B_j = 0. \end{cases} \tag{255}$$

Thus, multiple units of an item in one basket do not constitute repurchase; an item must appear in at least two distinct future baskets. Both binary targets follow Eq. (249).

SALES-WEEKLY: future decline risk and seasonality. For the 13-week decline target, fit an ordinary least-squares line only to future weekly sales. With $\bar{h} = 7$ and $\bar{S} = 13^{-1}\sum_{h=1}^{13} S_h$, define

$$\widehat{\beta} = \frac{\sum_{h=1}^{13}(h - \bar{h})(S_h - \bar{S})}{\sum_{h=1}^{13}(h - \bar{h})^2}, \qquad s_{\mathrm{decline}} = -\frac{\widehat{\beta}}{\max(|\bar{S}|, \epsilon)}. \tag{256}$$

Positive scores indicate a decreasing future sales trajectory. For the 52-week seasonality target, remove the future-window least-squares intercept and linear trend, $r_h = S_h - (\widehat{a} + \widehat{b}h)$, and define the one-sided periodogram

$$P_k = \left|\sum_{h=1}^{52} r_h \exp\left(-\frac{2\pi \mathrm{i} k h}{52}\right)\right|^2, \qquad k = 1, \ldots, 26. \tag{257}$$

The seasonality score is

$$s_{\text{seasonality}} = \begin{cases} \dfrac{P_1 + P_2}{\sum_{k=1}^{26} P_k}, & \sum_{k=1}^{26} P_k > \epsilon, \\ 0, & \text{otherwise.} \end{cases} \tag{258}$$

The annual and semiannual energy is therefore assessed after detrending, so a monotone future trend alone is not labeled seasonal. Both decline-risk and seasonality labels follow Eq. (249).

Entity-disjoint, temporally ordered splitting. Splitting is performed before any training-dependent label quantity or fitted preprocessing operation. For each dataset-task pair, at most one eligible prediction example is retained per entity, using the latest cutoff with complete input and target windows; this cutoff rule is independent of the future label. When entities have different cutoffs, examples are ordered by cutoff and assigned in a 70:15:15 train/validation/test ratio. Boundaries are placed between distinct timestamps whenever possible, with a deterministic hash of the entity identifier used only to break cutoff ties. If all entities share the same cutoff, temporal ordering cannot separate them, so the deterministic hash defines an entity-disjoint 70:15:15 split; we do not describe that fallback as a strict out-of-time evaluation. Consequently,

$$\mathcal{E}_{\text{tr}} \cap \mathcal{E}_{\text{val}} = \mathcal{E}_{\text{tr}} \cap \mathcal{E}_{\text{test}} = \mathcal{E}_{\text{val}} \cap \mathcal{E}_{\text{test}} = \varnothing. \tag{259}$$

Random seeds affect optimization only, not cutoff selection, label construction, or partition membership. All reported results are averaged over three random seeds: $\{42, 123, 456\}$.

Training-only preprocessing and leakage audit. All fitted preprocessing objects, including empirical CDFs, label quantiles, scaling or clipping constants when used, and class-frequency statistics, are estimated from the training partition and then frozen for validation and test. Sample-wise deterministic operations such as activity-mask extraction are applied independently after splitting. The construction satisfies the following operational invariants:

$$\begin{aligned} &\text{IDs}(\mathcal{D}_{\text{tr}}),\ \text{IDs}(\mathcal{D}_{\text{val}}),\ \text{IDs}(\mathcal{D}_{\text{test}}) \text{ are pairwise disjoint}, \\ &\max \mathcal{W}_{i,\tau}^{\text{in}} < \min \mathcal{W}_{i,\tau}^{\text{out},k}, \\ &\Theta_{\text{prep}} = \text{Fit}(\mathcal{D}_{\text{tr}}), \\ &\text{RecordsUsed}(\boldsymbol{x}_{i,\tau}, \boldsymbol{m}_{i,\tau}) \subseteq \mathcal{R}_i|_{\mathcal{W}_{i,\tau}^{\text{in}}}, \\ &y_{i,\tau}^{(k)} = \ell_{d,k}\left(\mathcal{R}_i|_{\mathcal{W}_{i,\tau}^{\text{out},k}}; \Theta_{\text{prep}}\right). \end{aligned} \tag{260}$$

These conditions make the distinction between predictors and targets operational: the future window is used to define the supervised outcome, never to construct CAST's input, mask, graph, gate, or fitted preprocessing parameters.

N More Details of the Baseline Methods

To evaluate CAST across distinct modeling assumptions, we selected 12 representative baselines spanning three methodological paradigms. The comparison is organized to separate the contribution of dedicated intermittent-data modeling, generic deep temporal representation, and the full structure-conditioned architecture under sparse observations.

1. Traditional Statistical and Zero-Inflated Models. We include **Croston** Croston (1972), **SBA** Syntetos & Boylan (2005), **ZIP** Lambert (1992), and **Hurdle** Mullahy (1986). *Rationale:* These are canonical statistical baselines for intermittent and zero-inflated data. They distinctly separate the occurrence of an event from its magnitude and therefore provide a direct control for the observation-modeling aspect of CAST. Their comparison with CAST helps distinguish the value of zero-state

[1] Available at: `http://www.brucehardie.com/datasets/`

[2] Available at: `https://www.kaggle.com/datasets/chiranjivdas09/ta-feng-grocery-dataset`

[3] Available at: `https://archive.ics.uci.edu/dataset/502/online+retail+ii`

[4] Available at: Kaggle Instacart Market Basket page.

[5] Available at: Kaggle Walmart Store Sales Forecasting page.

Table 7: Detailed task definitions, class labels, and source access for the evaluated public benchmarks.

Dataset	Original Basis and Access	Primary Task Description	Auxiliary Task Description
CDNOW	CD purchasing records from Kaggle Repository[1]	**Future Value:** Predicts whether next-three-month positive purchase value exceeds the training-only 65th-percentile target threshold. Labels: 'Low Value', 'High Value'.	**Future Churn:** Predicts whether no qualifying purchase occurs in the next three months. Labels: 'No Churn', 'Churn'.
TAFENG	TaFeng Grocery from UCI Machine Learning Repository[2]	**Future Risk:** Predicts a four-week risk score combining inactivity, low expenditure, and low transaction frequency, thresholded from training entities only. Labels: 'Low Risk', 'High Risk'.	**Future Repurchase:** Predicts repeated-product appearances across distinct baskets in the next four weeks. Labels: 'Non Repurchase', 'Repurchase'.
RETAIL	Online RETAIL II from UCI Machine Learning Repository[3]	**Future Value:** Predicts whether next-three-month positive purchase value exceeds the training-only 65th-percentile target threshold. Labels: 'Low Value', 'High Value'.	**Future Churn:** Predicts whether no qualifying purchase occurs in the next three months. Labels: 'No Churn', 'Churn'.
INSTACART	Instacart Market from Kaggle Repository[4]	**Future Activity:** Predicts four-week activity from a training-rank combination of future order count and purchased units. Labels: 'Low Activity', 'High Activity'.	**Future Churn:** Predicts whether no order occurs in the next four weeks. Labels: 'No Churn', 'Churn'.
SALES WEEKLY	Walmart Sales from Kaggle Repository[5]	**Future Decline Risk:** Predicts a normalized negative slope fitted to the next 13 weeks. Labels: 'Low Risk', 'High Risk'.	**Future Seasonality:** Predicts detrended annual/semiannual spectral concentration over the next 52 weeks. Labels: 'Non Seasonal', 'Seasonal'.

separation from the additional contribution of nonlinear temporal representation and learned structural conditioning.

2. Neural-Oriented Deep Sequence Models. This category includes **DSN** Xiao et al. (2022), **SoftShape** Liu et al. (2025a), **TimeMIL** Chen et al. (2024), and **InterpGN** Wen et al. (2025). *Rationale:* These models represent modern deep-learning approaches to complex sequences, including survival, shape-based, and graph-oriented components. Their inclusion tests whether high nonlinear capacity alone is sufficient in the zero-inflated setting, relative to the dedicated activity-state and structural inductive biases used by CAST.

3. Advanced State-Space and Foundation Models. This cutting-edge category comprises **UniTS** Gao et al. (2024), **Mamba** Gu & Dao (2024), **MambaSL** Jung & Kim (2026), and **GRU-D** Che et al. (2018). *Rationale:* GRU-D is designed for irregular sampling and missingness through temporal decay, while Mamba and MambaSL provide strong state-space sequence baselines and UniTS represents a general-purpose time-series foundation-model paradigm. Their inclusion tests whether generic sequence capacity is sufficient in the present zero-inflated setting. The reported tables show that these models remain below CAST on the aggregate metrics, motivating the use of a dedicated activity-state model and structure-conditioned gating rather than supporting a stronger causal-identification claim.

O Additional Evaluation Metrics

As discussed in Section 5.2, the evaluated sequence datasets exhibit extreme sparsity, with an average positive class ratio of approximately 35.0%. While the F1-Score serves as the primary metric to robustly evaluate the models' capacity to capture rare but critical state transitions under such severe class imbalance, Test Accuracy provides a necessary complementary perspective. It ensures that models do not compromise overall prediction stability or heavily skew their outputs merely to inflate minority class recall.

Table 8 details the comparative Test Accuracy across all ten configurations. Traditional statistical baselines, such as ZIP and Hurdle, exhibit relatively stable accuracy by naturally defaulting their predictions to the predominant inactive states. However, as evidenced by their F1-Scores in the main text, this stability comes at the cost of failing to capture complex active dynamics. Conversely, advanced generic sequence architectures like UniTS experience substantial degradation under extreme sparsity, providing empirical evidence that high-capacity temporal representation without appropriate structural priors can lead to severe overfitting.

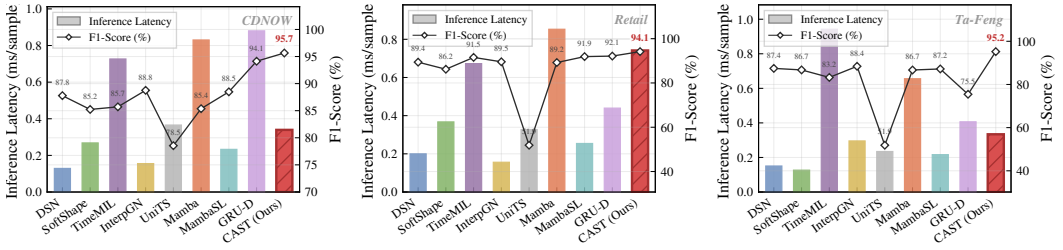
CAST achieves the highest overall average accuracy in Table 8, with 88.28%, while also retaining the F1 improvements reported in the main text. The agreement between the two metrics indicates that the gain is not obtained merely by shifting predictions toward the minority class. Together with the ablations, these results support the practical value of separating the activity state and using structure-conditioned basis gating. We interpret this as empirical predictive evidence; it does not by itself identify the learned temporal edges with the data-generating causal graph.

Table 8: **Main results: Test Accuracy (↑) on different tasks across datasets.** All results are reported as percentages (%). The best results are in **bold**, and the second-best results are underlined.

Model	*Retail*		*CDNOW*		*Instacart*		*Sales-Weekly*		*Ta-Feng*		*Overall*	
	Risk	Churn	Risk	Churn	Risk	Churn	Risk	Seasonality	Risk	Repurchase	Avg (%)	Δ
Croston	92.01±0.43	70.30±1.38	83.58±0.93	85.05±0.35	85.72±0.38	65.85±0.56	97.54±1.16	96.45±1.02	78.17±0.20	68.81±0.32	82.35	-6.11
SBA	92.01±0.43	70.30±1.38	83.57±0.94	85.05±0.35	85.71±0.39	65.85±0.56	97.54±1.16	96.45±1.02	78.17±0.20	68.80±0.32	82.35	-6.11
ZIP	92.47±0.45	70.30±1.62	81.06±1.96	85.04±0.24	89.76±0.33	66.26±0.88	97.54±1.16	98.36±0.67	75.50±0.13	67.32±0.20	82.36	-6.10
Hurdle	93.34±0.72	71.02±1.89	83.93±0.46	85.09±0.30	90.81±0.12	66.41±1.37	97.27±1.39	98.63±1.02	74.19±0.14	66.33±0.16	82.70	-5.76
DSN	92.17±0.33	70.15±1.62	91.99±0.53	84.69±0.24	97.30±0.10	65.56±1.20	93.72±0.77	97.54±1.16	91.15±0.48	72.18±0.44	85.64	-2.82
SoftShape	89.86±1.54	67.90±1.45	90.39±0.25	84.41±0.27	91.77±4.83	65.03±0.30	91.80±0.67	94.81±2.79	90.70±0.57	72.12±0.38	83.88	-4.58
TimeMIL	94.01±0.98	70.61±2.60	90.32±0.25	84.87±0.27	96.68±0.30	66.60±1.30	96.45±2.04	98.36±0.00	87.96±0.50	71.57±0.96	85.74	-2.72
InterpGN	92.06±0.14	71.33±1.51	92.50±0.19	84.85±0.38	97.79±0.10	67.15±0.67	95.36±1.02	99.45±0.39	91.93±0.73	72.18±0.43	86.46	-2.00
UniTS	93.74±0.25	41.01±0.00	86.06±0.50	82.84±0.00	35.00±0.00	36.64±0.00	93.92±0.72	96.72±0.67	84.86±1.10	52.42±0.00	70.32	-17.96
Mamba	92.11±0.40	70.97±0.45	90.25±0.19	84.93±0.26	96.74±0.27	65.11±1.19	92.35±0.39	97.54±0.67	90.63±0.44	72.04±0.34	85.27	-3.19
MambaSL	94.21±0.26	70.51±2.07	92.28±0.07	84.82±0.23	97.98±0.06	65.76±0.41	95.63±1.68	98.63±0.39	90.96±0.30	72.09±0.43	86.29	-2.17
GRU-D	94.42±0.19	72.20±2.39	95.93±0.28	85.31±0.29	98.06±0.08	66.65±1.20	97.54±1.77	97.54±1.16	80.39±0.60	72.16±0.45	86.02	-2.44
CAST (Ours)	**95.80**±0.14	**72.56**±1.17	**96.96**±0.17	**85.39**±0.33	**98.26**±0.13	**67.38**±0.34	**97.81**±1.55	**99.73**±0.39	**96.63**±0.30	**72.24**±0.31	**88.28**	–

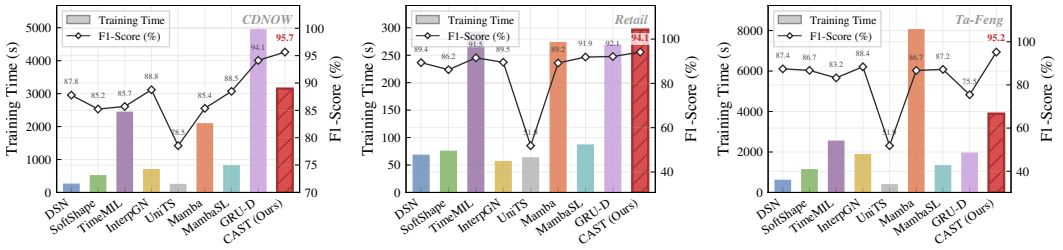
P Computational Efficiency Analysis

Comprehensive evaluation of the trade-off between predictive performance and computational cost is critical for sparse sequence modeling. We present detailed visualizations of inference latency and training time against the F1-Score across the CDNOW, Retail, and Ta-Feng datasets.



(a) CDNOW (b) Retail (c) Ta-Feng

Figure 9: Inference latency versus F1-Score across datasets. The bar chart represents inference latency in milliseconds per sample, and the line plot represents the F1-Score in percentage.



(a) CDNOW (b) Retail (c) Ta-Feng

Figure 10: Training time versus F1-Score across datasets. The bar chart represents total training time in seconds, and the line plot represents the F1-Score in percentage.

Figure 9 illustrates the inference latency per sample evaluated alongside the test F1-Score. CAST achieves strong predictive accuracy, recording F1-Scores of 95.7% on CDNOW, 94.1% on Retail, and 95.2% on Ta-Feng. Concurrently, its measured inference latency remains competitive with the evaluated recurrent and state-space baselines. This comparison indicates that CAST attains a favorable accuracy–latency trade-off among the evaluated models.

Figure 10 depicts the total training time relative to the F1-Score. The measured training duration of CAST is lower than that of several heavier sequence baselines in this implementation. While UniTS exhibits low training time and inference latency in this comparison, its reported predictive scores are 78.5% on CDNOW and 51.9% on both Retail and Ta-Feng. CAST therefore occupies a different point on the measured accuracy–efficiency frontier, with stronger predictive scores at a still practical computational cost.

Q Parameter Sensitivity Analysis

To assess the robustness of the proposed framework, we evaluate the impact of six key hyperparameters on the predictive performance. Figure 11 illustrates the deviation of the F1 score from the median performance across all datasets.

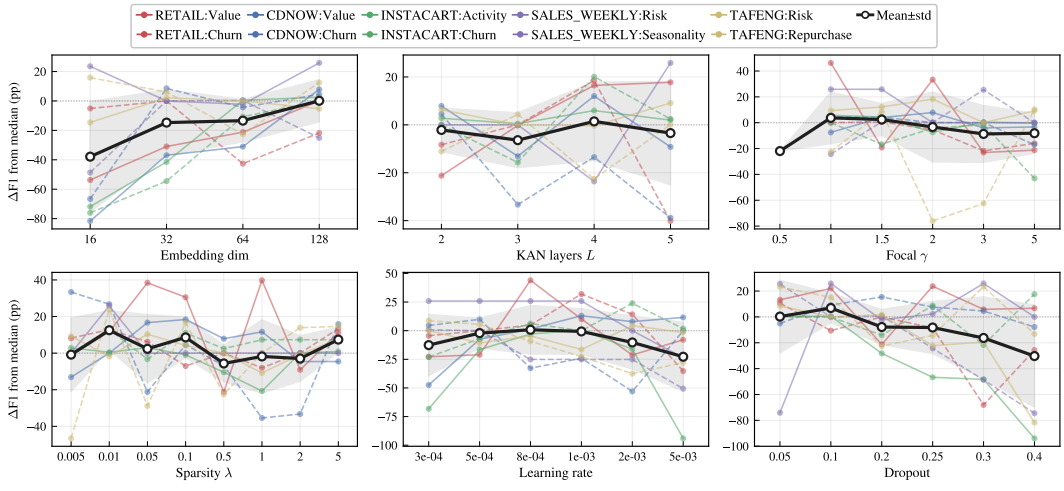


Figure 11: **Hyperparameter sensitivity analysis.** The plots show the variation in F1 score from the median across ten task configurations. The thick black line represents the average trend, with the shaded region denoting the standard deviation.

Representation Capacity. Across the tested grid, performance improves consistently as the embedding dimension increases from 16 to 128. Restricting the dimension to 16 causes a drop of up to 80 points across multiple datasets. Regarding network depth, 3 layers yield the best score among the evaluated settings, providing an average improvement of nearly 10 points above the

Table 9: **KAN basis function comparison: Test F1-Score (↑) on primary and auxiliary tasks across datasets.** All results are reported as percentages (%). Taylor is the default basis function used in **CAST**. The best results are in **bold**, and the second-best results are underlined.

Basis Function	*Retail*		*CDNOW*		*Instacart*		*Sales-Weekly*		*Ta-Feng*		*Overall*	
	Risk	Churn	Risk	Churn	Risk	Churn	Risk	Seasonality	Risk	Repurchase	Avg (%)	Δ
B-Spline Liu et al. (2025b)	93.68±0.47	72.83±0.88	95.89±0.29	91.69±0.12	97.12±0.09	64.40±0.30	96.26±1.89	99.47±0.75	94.57±0.42	73.96±0.43	87.99	-0.27
Chebyshev Sidharth et al. (2024)	93.85±0.56	72.57±0.88	95.69±0.37	91.69±0.10	97.03±0.18	64.38±0.29	96.26±1.89	99.22±0.64	94.09±0.48	73.97±0.44	87.88	-0.38
Fourier Imran & Ishmam (2024)	93.15±0.60	72.61±0.86	**95.97**±0.24	91.69±0.19	97.14±0.13	64.42±0.33	96.65±2.42	98.96±0.97	94.38±0.03	73.93±0.46	87.89	-0.36
Jacobi El Afia et al. (2025)	93.87±0.77	72.81±0.81	95.49±0.21	91.66±0.13	97.18±0.14	64.42±0.27	96.26±1.89	99.22±0.64	**95.22**±0.24	73.91±0.47	88.00	-0.25
RBF Chiu et al. (2026)	93.87±0.46	72.58±0.78	95.83±0.38	91.67±0.14	96.98±0.07	64.39±0.20	96.61±2.44	99.47±0.37	93.74±0.17	74.05±0.37	87.92	-0.34
Wavelet Bozorgasl & Chen (2024)	93.95±0.82	72.18±1.24	95.77±0.19	91.57±0.17	97.06±0.20	64.37±0.31	96.61±2.44	98.95±0.74	94.79±0.14	74.00±0.38	87.93	-0.33
Legendre Chen et al. (2025)	93.22±0.36	**72.86**±0.63	95.56±0.23	91.68±0.17	97.01±0.14	64.48±0.35	96.65±2.42	99.47±0.75	94.25±0.38	73.92±0.43	87.91	-0.34
CAST (Taylor)	**94.14**±0.13	72.84±1.17	95.65±0.26	**91.74**±0.18	**97.52**±0.18	**64.55**±0.34	**97.00**±2.12	**99.74**±0.37	95.20±0.44	**74.17**±0.45	**88.25**	–

median. Increasing the depth beyond 3 layers degrades held-out performance, which is consistent with overfitting at greater depth in this experimental regime.

Structural Regularization. Within the evaluated grid, the sparsity penalty performs best at 0.01; values exceeding 0.5 over-constrain the model and reduce predictive accuracy. A dropout rate of 0.1 is likewise the best tested setting. Increasing dropout to 0.4 produces a decline of up to 25 points, suggesting that strong random dropping is unnecessary when the learned structural gate already supplies substantial activation regularization.

Optimization Dynamics. The framework requires a carefully tuned learning rate, achieving peak stability at 0.001. Larger rates such as 0.005 lead to optimization instability, dropping performance by 100 points on certain datasets. The focal parameter remains generally stable across the tested range, but a value of 2.0 achieves the highest mean score, successfully balancing the learning focus on difficult samples.

Basis Function Variants. To evaluate the impact of the nonlinear basis choice, we compare the default Taylor basis with seven alternatives in Table 9. Taylor obtains the highest overall average F1-score of 88.25% and the best result in seven of the ten reported task configurations. Jacobi and B-Spline remain competitive but have slightly lower aggregate scores. These results support Taylor as the strongest basis choice among the alternatives evaluated here; they do not imply a universal optimality result for KAN bases outside the tested settings.

R Additional Structural Analysis for Interpretability

R.1 Panel-by-Panel Reading of the Main-Text Structural Illustration

The three panels in Figure 3 should be read as successive representations of the same learned structural signal, not as three independent claims of causal identification. We therefore separate what is imposed by construction, what is learned from the data, and what is induced downstream by the gating mechanism.

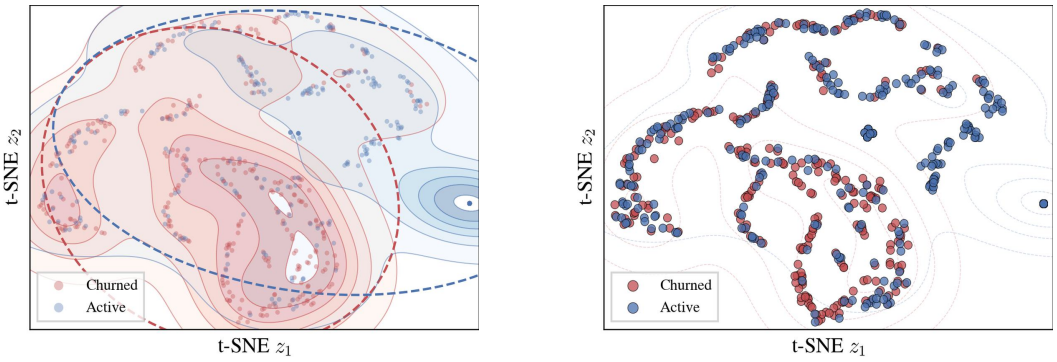
Temporally Ordered Structural Learning. Figure 3a illustrates the learned adjacency matrix of the ZI-SEM component. The zero lower-triangular region is imposed by the temporal mask and therefore verifies implementation of temporal precedence rather than serving as empirical evidence for acyclicity. Strong learned dependencies appear near the superdiagonal, with selected weights reaching 0.6, and several early positions have comparatively large outgoing influence. We interpret these nodes as structurally influential ancestors in the learned temporal graph; the visualization alone does not establish unobserved-confounder control or prove that every learned edge is a data-generating causal effect.

Hyperbolic Hierarchical Representation. Figure 3b projects the learned temporal graph into a continuous hyperbolic space. Node sizes scale with structural out-degree; the displayed configuration places several influential early nodes closer to the origin and more peripheral nodes farther outward, a pattern consistent with the PageRank–radius tendency analyzed in Proposition 4.3. For this real dataset the panel is a representation diagnostic, not a certificate of the tree-approximation premise required by Proposition 4.2. Its operational role is to provide a continuous geometric prior to the downstream sequence layers.

Structure-Conditioned Gate Modulation. Figure 3c compares the KAN basis responses before and after structure-conditioned modulation. The dashed polynomial responses reach larger magnitudes near the input boundaries, whereas the solid curves show the realized attenuation induced by the learned gates. This multiplicative modulation acts as an activation-level structural regularizer. It should not be interpreted as stored-parameter pruning or, on real data, as proof that a suppressed position is causally irrelevant. Further visualizations of the latent space, structural influence scores, and gate distributions are provided below.

The appropriate evidential reading is consequently layered. The adjacency panel shows the structure returned by the ZI-SEM under the temporal mask; the Poincaré panel shows how that structure is organized geometrically; and the gated-basis panel shows how the geometric representation changes realized nonlinear activity. None of the three panels removes the causal-sufficiency requirement in Assumption 3.1.

To provide deeper insights into the internal mechanisms of the proposed framework, we present supplementary visualizations of the latent space, feature importance, and gating dynamics evaluated on the retail churn dataset.



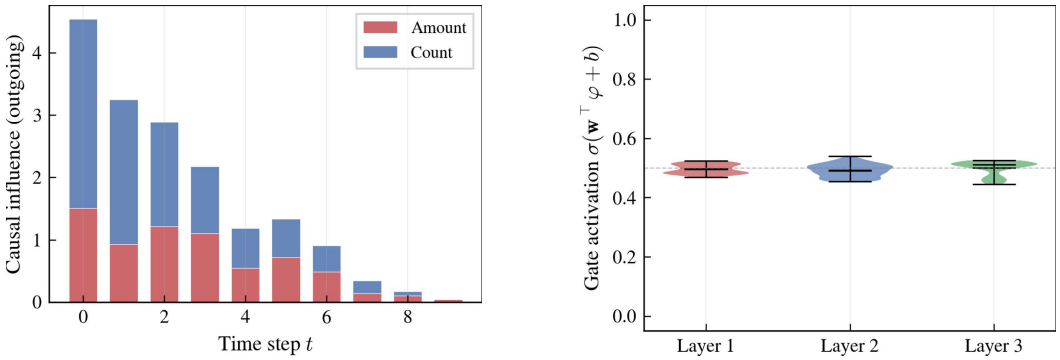
(a) Training Set (b) Test Set

Figure 12: **Latent space separation.** Projection of the high-dimensional hidden representations demonstrating distinct structural clustering between the churned and active customer profiles across both training and test distributions.

Latent Space Separation. Figure 12 projects the learned hidden representations into two dimensions for qualitative inspection. The training and test projections in Figures 12a and 12b show visible separation between the two customer groups. Because t-SNE is a nonlinear visualization method whose geometry depends on the projection procedure, we use these plots only as qualitative evidence of class-structured representations; predictive generalization is assessed by the held-out metrics rather than inferred from the projection alone.

Temporal Structural Influence. Figure 13a aggregates outgoing weights of the learned temporally ordered graph to form a model-internal structural influence score. The initial sequence step has the largest outgoing structural mass, exceeding 4.0, and the displayed score decreases toward the eighth step. The channel decomposition attributes more of this learned outgoing mass to transaction count at the earliest positions, while transaction amount contributes throughout the sequence. These quantities are neither interventional effects nor direct attributions of the final classifier; a causal interpretation requires Assumption 3.1.

Gating Mechanism Stability. Figure 13b shows the distribution of gate values across the three sequence layers. The medians are concentrated near 0.5 and the displayed distributions do not place most mass at the sigmoid boundaries. This is consistent with a largely non-saturated gating regime on the evaluated data. We do not infer gradient stability or absence of dead units from the violin plot alone; its role is to report the realized gate distribution used by the activation-level interpretation in Proposition 4.4.



(a) Temporal Structural Influence

(b) Gate Activation Distribution

Figure 13: **Temporal structural influence and gating dynamics.** The left panel displays aggregated outgoing weights of the learned structural graph by time step and feature channel. The right panel shows the realized gate distributions across three network layers. These are model-internal diagnostics, not interventional effect estimates.

S Computation Infrastructure and Hyperparameters

All experiments are conducted on a cloud server instance equipped with a 16-core CPU and a single NVIDIA GeForce RTX 4090 GPU with 48GB of VRAM. The system is supported by 48GB of RAM. All models are implemented in Python 3.11. The software stack includes deep learning libraries and Anaconda, with corresponding GPU drivers and compilation support ensuring hardware acceleration.

The codes and datasets utilized during our empirical experiments are included in the following Anonymous GitHub: `https://anonymous.4open.science/r/Nips-B305`. For compliance with the double-blind review policy, the authors will publicly release the full experimental code on GitHub upon acceptance.

This section reports the hyperparameters used for CAST and states the stage-wise optimization protocol clearly. The numerical settings in Table 10 are unchanged across the reported experiments.

Stage-wise optimization protocol. Training has two optimization phases. First, Stage I optimizes $(\Theta_{\mathrm{ZI}}, \boldsymbol{A})$ with $\mathcal{L}_{\mathrm{DAG}}$ for the listed pre-training epochs. The graph is then thresholded by τ and frozen, together with the Stage I likelihood parameters. Second, Stages II–III are optimized jointly for the listed fine-tuning epochs with $\mathcal{L}_{\mathrm{FT}}$ in Eq. (25). The hyperbolic coordinates $\{\boldsymbol{p}_t\}$ remain trainable and therefore receive both predictive gradients and the regularizing gradient from $\lambda_{\mathrm{hyp}}\mathcal{L}_{\mathcal{H}}$, whereas $\boldsymbol{A}$ and all Stage I likelihood parameters remain frozen. Thus $\mathcal{L}_{\mathrm{ZI}}$ and $\lambda_s\|\boldsymbol{A}\|_1$ are Stage I terms only. The reported "Pre-train Epochs" refer to Stage I, and "Fine-tune Epochs" to the joint Stage II–III phase. In the optional unmasked mode, λ_{dag} scales the augmented-Lagrangian block, ρ is its quadratic penalty, and α_{dag} is the associated dual multiplier. Under the exact temporal mask used in the reported CAST experiments, $h(\boldsymbol{A}) = 0$ identically, so this entire block is inactive.

Table 10: **Hyperparameter configurations across different datasets and tasks.** The table details the dataset-specific hyperparameters utilized for training the CAST framework.

Hyperparameter	***Retail***		***CDNOW***		***Instacart***		***Sales-Weekly***		***Ta-Feng***	
	Risk	**Churn**	**Risk**	**Churn**	**Risk**	**Churn**	**Risk**	**Seasonality**	**Risk**	**Repurchase**
Embedding Dim	64	32	128	128	64	128	32	64	64	64
KAN Hidden Dim	64	256	128	64	256	256	32	128	128	128
KAN Layers	2	3	2	2	3	2	2	3	3	4
KAN Order	8	8	8	8	8	8	8	8	8	8
KAN Basis Type	Taylor	Taylor	Taylor	Taylor	Taylor	Taylor	Taylor	Taylor	Taylor	Taylor
k-hop Neighbors	2	2	2	2	2	2	2	2	2	2
Learning Rate	2e-3	5e-4	2e-3	3e-4	1e-3	1e-3	3e-4	3e-4	3e-4	1e-3
Batch Size	128	128	128	128	512	256	128	128	128	128
Pre-train Epochs	100	100	100	100	100	100	100	100	100	100
Fine-tune Epochs	100	100	100	100	100	100	500	100	100	100
Dropout	0.2	0.2	0.1	0.3	0.05	0.05	0.05	0.1	0.1	0.2
Weight Decay	1e-4	5e-4	5e-4	1e-3	1e-4	1e-4	1e-3	1e-4	1e-3	1e-4
Sparsity Penalty (λ_s)	0.5	2.0	0.5	0.1	0.01	0.005	2.0	2.0	0.5	0.5
Focal Gamma (γ)	5.0	1.0	2.0	1.0	3.0	1.0	5.0	2.0	3.0	1.0
Focal Alpha (α)	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7
Unmasked DAG Penalty (λ_{dag})	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Hyp. Loss Weight (λ_{hyp})	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Gate Penalty (λ_{sp})	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
DAG Threshold (τ)	0.05	0.05	0.01	0.05	0.05	0.01	0.01	0.03	0.3	0.03
DAG ρ Init	1.0	1.0	1.0	1.0	0.1	0.1	1.0	1.0	1.0	1.0
DAG Outer Iter	10	10	10	10	3	10	10	10	10	10
DAG Inner Iter	300	300	300	300	100	300	300	300	300	300