OGC-MoE: Online Granger-Causal Mixture-of-Experts for High-Dimensional Nonstationary Time Series

Anonymous authors
Paper under double-blind review

Abstract

Causal discovery in high-dimensional multivariate time series is challenging because the number of candidate relations grows rapidly, while nonstationarity may cause the underlying causal mechanisms to evolve over time. Existing methods often assume a single stable graph or estimate local graphs independently, limiting the reuse of historical knowledge and producing unstable estimates from short windows. We propose Online Granger-Causal Mixture-of-Experts (OGC-MoE), an online graph-level mixture-of-experts framework for causal discovery in high-dimensional nonstationary time series. OGC-MoE maintains a library of causal experts, each pairing a graph-conditioned temporal predictor with a sparse Granger-causal graph. For each incoming window, the model evaluates historical experts using prediction and causal-residual information, refines their routing weights, and combines the selected graphs to represent either a single historical mechanism or a recombination of multiple mechanisms. When the mixture is insufficient, sparse local adaptation is applied to moderate deviations, whereas a window identified as a genuinely unseen state is used directly to construct a new expert. Experiments on synthetic benchmarks and real-world financial data demonstrate the strong and robust causal discovery performance of OGC-MoE.

1 Introduction

High-dimensional multivariate time series arise in financial markets, industrial systems, transportation networks, healthcare monitoring, and complex dynamical systems. Discovering causal relations among their variables can reveal system dynamics, improve forecasting, and support reliable decision-making (Assaad et al., 2022). Granger causality characterizes temporal influence from a predictive perspective: one variable Granger-causes another if its historical information improves prediction beyond the information already available (Granger, 1969). Sparse statistical methods have extended this principle to large systems, but the rapidly growing number of candidate directed relations continues to make reliable high-dimensional causal estimation difficult (Shojaie & Michailidis, 2010; Basu et al., 2015).

Most time-series causal discovery methods assume a stable data-generating mechanism. In real systems, however, nonstationarity is not limited to changes in the mean, variance, or correlation structure; it may also alter the causal relations themselves (Huang et al., 2020; Gao & Yang, 2022). As the system changes state, causal edges may appear or disappear, their directions may change, and their strengths may vary. Such causal nonstationarity is common in financial markets, where intersector dependencies and risk-transmission pathways evolve across market regimes. A single fixed graph may therefore obscure state-specific relations and fail to represent emerging causal structures.

These difficulties become more severe in online settings. The causal graph of each incoming window must be inferred using only current and previously observed data, yet a short window may contain insufficient samples for estimating a reliable high-dimensional graph independently. Relearning a complete graph for every window is also computationally expensive and sensitive to transient noise. Existing approaches address temporal variation using sliding windows, time-varying models, latent contexts, or change-point detection (Huang et al., 2020; Gao et al., 2023). Amortized causal discovery shares information across samples with different graphs (Löwe et al., 2022), while mixture-based

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.

approaches represent heterogeneous causal graphs or temporal patterns (Varıcı et al., 2024; Liu et al., 2026). Although these methods capture different forms of variation, they do not jointly support online reuse, direct graph-level combination, local structural adaptation, and continual expansion of historical causal mechanisms. They also provide limited support for distinguishing a recombination of known mechanisms from a genuinely unseen causal state.

To address these limitations, we propose **Online Granger-Causal Mixture-of-Experts (OGC-MoE)** for high-dimensional nonstationary time series. OGC-MoE maintains predictor–graph experts learned from historical states, routes each incoming window using prediction and causal-residual compatibility, and combines selected expert graphs into a window-specific causal structure. A post-mixture assessment retains the mixture, applies a sparse current-window correction, or constructs a new expert directly from a window identified as an unseen state.

Our contributions are threefold: **1)** an online Granger-causal expert library for reusing historical mechanisms; **2)** compatibility-based routing and graph-level mixture for reconstructing window-specific causal structures; and **3)** local causal adaptation and continual expert expansion for variations not explained by the existing library.

2 Related Work

2.1 Time-Series Causal Discovery under Nonstationarity

Time-series causal discovery aims to identify directed temporal dependencies from observational sequences. Classical approaches include structural-equation models, sparse vector autoregression, and conditional-independence tests such as PCMCI and PCMCI+ (Granger, 1969; Haufe et al., 2010; Peters et al., 2013; Runge et al., 2019; Runge, 2020). Neural approaches use structured predictors to recover nonlinear Granger-causal relations (Tank et al., 2021; Nauta et al., 2019; Bussmann et al., 2021; Marcinkevičs & Vogt, 2021), while deep temporal models further accommodate flexible dynamics and history-dependent noise (Pamfil et al., 2020; Gong et al., 2022).

To address nonstationarity, existing methods use state-space models, sliding windows, time-varying parameters, change-point detection, or contextual variables to estimate causal structures across different periods and environments (Huang et al., 2019; 2020; Günther et al., 2023). PCMCI$_\Omega$ models alternating and recurring causal mechanisms in semi-stationary time series (Gao et al., 2023), while TS-CausalNN learns nonlinear causal relations from nonstationary observations (Faruque et al., 2025). DCD decomposes nonstationary time series into trend, seasonal, and residual components and integrates their causal relations into a unified graph (Ferdous & Gani, 2026). These methods capture different forms of temporal variation, but do not perform sequential causal-graph adaptation during online inference.

2.2 Mixture-of-Experts for Causal Discovery

Mixture-of-experts models combine specialized predictors through input-dependent routing (Jacobs et al., 1991). In causal discovery, mixture-based methods represent heterogeneous observations using multiple latent causal graphs, but mainly focus on static observational or interventional settings (Varıcı et al., 2024; Björkman et al., 2025). Only a few studies have explored mixture-of-experts architectures for causal discovery. Among them, CausalMoE dynamically routes time-series patches to heterogeneous representation experts and aggregates their outputs before recovering a Granger-causal graph through causality-aware attention (Liu et al., 2026). Other recent work uses expert architectures to identify local causal patterns (Kafantaris, 2026).

OGC-MoE differs from CausalMoE in both the expert representation and inference process. CausalMoE performs dynamic routing and mixture at the representation level, whereas OGC-MoE treats each expert as a predictor–graph pair and directly combines selected expert graphs for each incoming window. OGC-MoE further supports online adaptation of the current causal structure and direct expert construction for unseen states. Table 1 compares representative methods in terms of nonstationarity, high-dimensional evaluation, and online adaptation, where online adaptation denotes sequentially adapting the causal structure using only observations available up to the current window.

Table 1: Comparison of representative time-series causal discovery methods. HD indicates reported evaluation with at least 100 variables.

Method	Nonstationarity	High dimensions (HD)	Online adaptation
PCMCI$_\Omega$ (2023)	✓	–	–
CUTS+ (2024)	–	✓	–
TS-CausalNN (2025)	✓	–	–
DCD (2026)	✓	–	–
CausalMoE (2026)	✓	–	–
OGC-MoE (Ours)	✓	✓	✓

3 METHODOLOGY

3.1 PROBLEM SETUP AND FRAMEWORK OVERVIEW

Let $X = \{\mathbf{x}_s\}_{s=1}^{T}$, where $\mathbf{x}_s \in \mathbb{R}^N$, denote a high-dimensional multivariate time series. Historical observations are provided as K temporal domains $\{\mathcal{D}_k\}_{k=1}^{K}$, each representing a relatively stable causal state. From domain $\mathcal{D}_k$, we learn a causal expert $\mathcal{E}_k = (f_{\theta_k}, \mathbf{A}_k)$, where f_{θ_k} is a graph-conditioned temporal predictor and $\mathbf{A}_k \in [0,1]^{N\times N}$ is the corresponding soft Granger causal-strength matrix. These experts form a historical causal expert library:

$$\mathcal{E} = \{\mathcal{E}_k = (f_{\theta_k}, \mathbf{A}_k)\}_{k=1}^{K}. \tag{1}$$

We formulate causal nonstationarity at the window level. Let $\mathbf{A}_t^\star$ denote the underlying causal-strength matrix of the current window. We model it as either a single stable historical mechanism or a weighted combination of multiple historical mechanisms, supplemented when necessary by a sparse local structural variation:

$$\mathbf{A}_t^\star \approx \mathrm{clip}\left(\sum_{k=1}^{K} \alpha_{t,k}\mathbf{A}_k + \Delta\mathbf{A}_t, 0, 1\right), \qquad \boldsymbol{\alpha}_t \in \Delta^{K-1}, \tag{2}$$

where $\boldsymbol{\alpha}_t = (\alpha_{t,1}, \ldots, \alpha_{t,K})$ contains nonnegative mixture weights satisfying $\sum_{k=1}^{K}\alpha_{t,k} = 1$, and $\Delta\mathbf{A}_t$ denotes a sparse local causal-strength correction. Concentrated weights represent a single historical mechanism, while distributed weights recombine several mechanisms. The correction captures relations not adequately represented by either case.

During online inference, OGC-MoE processes the test stream using windows $\mathcal{W}_t = \{\mathbf{x}_s\}_{s=t-W+1}^{t}$, where W is the online-window length. The temporal predictor uses a separate lookback length H for each one-step prediction. For each window, OGC-MoE evaluates the historical experts, refines the routing weights, and combines the selected causal-strength matrices. Post-mixture assessment then retains the mixture, performs a current-window local correction, or uses an unseen-state window to construct a new expert. Figure 1 summarizes the overall framework.

3.2 OFFLINE CAUSAL EXPERT LEARNING

For domain $\mathcal{D}_k$, the causal-strength matrix is parameterized by a learnable logit matrix $\mathbf{Z}_k$. We use the convention that $A_{k,ij}$ represents the directed relation $i \to j$:

$$\mathbf{A}_k = \sigma(\mathbf{Z}_k), \qquad A_{k,ii} = 0, \qquad \hat{\mathbf{x}}_s^{(k)} = f_{\theta_k}(\mathbf{x}_{s-H:s-1}, \mathbf{A}_k). \tag{3}$$

Equation (3) gives the deterministic soft matrix used for inference. During graph optimization, we use the standard fixed-temperature Gumbel–Sigmoid relaxation and mask its diagonal to zero.

Each off-diagonal entry $A_{k,ij} = \sigma(Z_{k,ij}) \in (0,1)$ gates the contribution of variable i's history to the prediction of variable j. All domains use the same training-set standardization, and all experts share the predictor architecture, graph parameterization, objective, and sparsity regularization to support a common model-internal scale. We therefore interpret $A_{k,ij}$ as the normalized predictive Granger-causal strength of the directed relation $i \to j$.

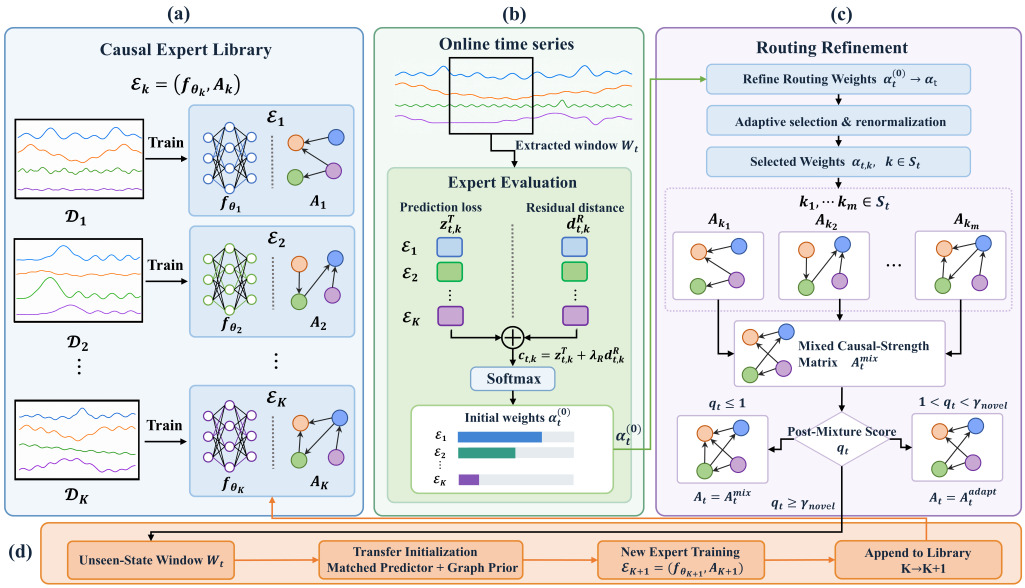

Figure 1: Overview of OGC-MoE: (a) offline causal expert learning, (b) online expert evaluation and initial routing, (c) causal graph construction and adaptation, and (d) continual expert expansion.

The predictor uses graph-conditioned message passing followed by recurrent state updates. Reciprocal lagged effects are allowed, so no acyclicity constraint is imposed on the time-collapsed Granger graph. The zero diagonal excludes reported self-edges, while each variable's own history remains available through the recurrent pathway.

Let $\mathcal{I}_k$ denote the valid prediction indices in $\mathcal{D}_k$ after constructing the length-H lagged inputs. The expert objective combines mean squared prediction error with graph sparsity:

$$\mathcal{L}_{\text{pred}}^{(k)} = \frac{1}{|\mathcal{I}_k|N} \sum_{s \in \mathcal{I}_k} \sum_{i=1}^{N} \left(x_{s,i} - \hat{x}_{s,i}^{(k)}\right)^2,$$
$$\mathcal{L}_{\text{expert}}^{(k)} = \mathcal{L}_{\text{pred}}^{(k)} + \lambda_{\text{sp}} \|\mathbf{A}_k\|_1. \tag{4}$$

We alternately update θ_k with the causal-strength matrix fixed and $\mathbf{Z}_k$ with the predictor fixed. For high-dimensional data, the sparsity penalty suppresses irrelevant relations among the $O(N^2)$ candidates, and a validation-selected threshold retains a sparse active edge set for message passing. During online inference, the expert graphs remain frozen and only the K routing weights are routinely optimized; sparse graph parameters are introduced only when local adaptation is triggered.

3.3 Online Causal Graph Construction and Adaptation

Expert Evaluation and Initial Routing. Before online inference, each causal expert is evaluated on held-out historical windows to construct its reference prediction and residual profiles. For an incoming window $\mathcal{W}_t$, expert k produces the prediction $\widehat{\mathbf{Y}}_t^{(k)}$ using its predictor f_{θ_k} and causal-strength matrix $\mathbf{A}_k$, where $\mathbf{Y}_t$ denotes the observed prediction targets associated with the window. These expert-specific predictions, each evaluated with its own matrix $\mathbf{A}_k$, are used for expert matching and routing. Let $\ell_{t,k} = \mathcal{L}_{\text{pred}}(\mathbf{Y}_t, \widehat{\mathbf{Y}}_t^{(k)})$ and let $e_{s,k,i} = M_{s,i}|\hat{x}_{s,i}^{(k)} - x_{s,i}|$ be its absolute residual for variable i, where $M_{s,i}$ is the observation indicator (and equals one for fully observed data); masked entries are excluded from the window statistics. We smooth and scale the residual vector component-wise as $\widetilde{\mathbf{e}}_{s,k} = \text{MA}(\mathbf{e}_{s,k})/(\text{RStd}(\text{MA}(\mathbf{e}_{s,k})) + \epsilon)$, and obtain the graph-weighted residual $\mathbf{g}_{s,k} = \widetilde{\mathbf{e}}_{s,k}\mathbf{A}_k^\top$, where MA and RStd denote the moving average and rolling standard deviation.

The window profile concatenates three variable-wise statistics,

$$\mathbf{h}_{t,k} = \left[\mathrm{Mean}_{s\in\mathcal{W}_t}\,|\mathbf{e}_{s,k}|\,;\mathrm{Std}_{s\in\mathcal{W}_t}(\mathbf{e}_{s,k})\,;\mathrm{Mean}_{s\in\mathcal{W}_t}\,|\mathbf{g}_{s,k}|\right] \in \mathbb{R}^{3N}. \tag{5}$$

Using held-out historical windows, we estimate the reference mean $\boldsymbol{\mu}_k^R$ and scale $\boldsymbol{\sigma}_k^R$ of this profile, together with the prediction-loss statistics μ_k^T and σ_k^T. The two normalized compatibility scores are

$$z_{t,k}^T = \frac{\ell_{t,k} - \mu_k^T}{\sigma_k^T + \epsilon}, \qquad d_{t,k}^R = \frac{1}{3N}\sum_{j=1}^{3N} \mathrm{clip}\left(\frac{|h_{t,k,j} - \mu_{k,j}^R|}{\max(\sigma_{k,j}^R, \epsilon)}, 0, 20\right), \tag{6}$$

where the fixed clipping bound limits the influence of extreme coordinates. Here, T and R denote the temporal-prediction and residual-profile scores, respectively; lower values indicate greater compatibility with expert k.

The two scores are combined into an expert-matching cost and converted into the initial routing weights:

$$c_{t,k} = z_{t,k}^T + \lambda_R d_{t,k}^R, \qquad \alpha_{t,k}^{(0)} = \frac{\exp(-c_{t,k}/\tau)}{\sum_{j=1}^K \exp(-c_{t,j}/\tau)}, \tag{7}$$

where $\lambda_R \geq 0$ controls the relative contribution of the residual signal and $\tau > 0$ is the routing temperature. The initial weight vector $\boldsymbol{\alpha}_t^{(0)}$ is subsequently refined using the routing-prediction objective.

Routing Refinement and Causal-Strength Mixture. Throughout online inference, historical predictor parameters $\{\theta_k\}$ and graph logits $\{\mathbf{Z}_k\}$ remain frozen. For a candidate routing vector $\boldsymbol{\alpha}$, define the routing prediction as $\widehat{\mathbf{Y}}_t^{\mathrm{route}}(\boldsymbol{\alpha}) = \sum_{k=1}^K \alpha_k f_{\theta_k}(\mathcal{W}_t, \mathbf{A}_k)$, where each expert retains its own causal-strength matrix. Starting from $\boldsymbol{\alpha}_t^{(0)}$, only routing weights are optimized on the probability simplex:

$$\boldsymbol{\alpha}_t = \arg\min_{\boldsymbol{\alpha}\in\Delta^{K-1}} \left[\mathcal{L}_{\mathrm{pred}}\left(\mathbf{Y}_t, \widehat{\mathbf{Y}}_t^{\mathrm{route}}(\boldsymbol{\alpha})\right) + \beta_h \mathcal{H}(\boldsymbol{\alpha}) + \beta_s \left\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_{t-1}\right\|_2^2\right], \tag{8}$$

where $\mathcal{H}(\boldsymbol{\alpha}) = -\sum_{k=1}^K \alpha_k \log(\alpha_k + \epsilon)$ is the routing entropy, with $\beta_h, \beta_s \geq 0$ and $\epsilon > 0$. We optimize the simplex-constrained objective through a softmax parameterization initialized by $\boldsymbol{\alpha}_t^{(0)}$. The three terms encourage accurate prediction, concentrated routing, and temporal smoothness, respectively. Optimization and boundary-case details are given in Appendix B.3.

OGC-MoE selects the two experts with the largest refined routing weights, denoted by $\mathcal{S}_t$, and renormalizes their weights over the selected set. For simplicity, the renormalized weights are still denoted by $\alpha_{t,k}$. The same Top-2 rule is used for every dataset and online window. After routing, OGC-MoE constructs a shared window-level causal-strength matrix and evaluates a candidate matrix through the selected frozen predictors:

$$\mathbf{A}_t^{\mathrm{mix}} = \sum_{k\in\mathcal{S}_t} \alpha_{t,k}\mathbf{A}_k, \qquad \widehat{\mathbf{Y}}_t^{\mathrm{graph}}(\mathbf{A}) = \sum_{k\in\mathcal{S}_t} \alpha_{t,k} f_{\theta_k}(\mathcal{W}_t, \mathbf{A}). \tag{9}$$

Routing evaluates experts under their own graphs, whereas graph evaluation supplies the shared candidate matrix to all selected predictors, with predictor parameters and routing weights fixed. The convex mixture has entries in $[0, 1]$ and zero diagonal. Binary graphs, when required, use a threshold selected on held-out data.

Post-Mixture Decision and Local Causal Adaptation. We assess the mixed graph using its normalized graph-conditioned prediction loss z_t^{mix}, together with the routing-weighted residual discrepancy $d_t^{\mathrm{mix}} = \sum_{k\in\mathcal{S}_t} \alpha_{t,k} d_{t,k}^R$. Each $d_{t,k}^R$ is computed from expert k's prediction under its own graph $\mathbf{A}_k$ and its own reference profile. Reference statistics and normalization are detailed in Appendix B.4. With positive reference thresholds ρ_z, ρ_d calibrated on held-out historical windows, the post-mixture score is

$$q_t = \max\left(\frac{z_t^{\mathrm{mix}}}{\rho_z}, \frac{d_t^{\mathrm{mix}}}{\rho_d}\right). \tag{10}$$

Here, ρ_z and ρ_d are calibration scales rather than learned model parameters. The fixed boundary $q_t = 1$ separates adequately represented windows from those requiring adaptation, while the validation-calibrated threshold $\gamma_{\text{novel}} > 1$ identifies novel-state candidates. Its calibration is described in Appendix B.5.

When $q_t \leq 1$, the historical mixture adequately explains the current observations, and $\mathbf{A}_t = \mathbf{A}_t^{\text{mix}}$ is directly used as the causal-strength estimate of the current window. When $1 < q_t < \gamma_{\text{novel}}$, the mixed matrix provides a useful initialization but requires local refinement. Keeping all historical experts fixed, OGC-MoE directly optimizes a current-window causal-strength matrix. Let $\mathcal{A} = \{\mathbf{A} \in [0,1]^{N\times N} : \text{diag}(\mathbf{A}) = \mathbf{0}\}$. Here, $f_{\theta_k}(\mathcal{W}_t, \mathbf{A})$ denotes predictions formed only from the lagged inputs associated with $\mathcal{W}_t$. The selected predictor parameters and routing weights remain fixed while only the current-window matrix is optimized:

$$\begin{aligned}\mathbf{A}_t^{\text{adapt}} = \arg\min_{\mathbf{A}\in\mathcal{A}} \; & \mathcal{L}_{\text{pred}}\left(\mathbf{Y}_t, \widehat{\mathbf{Y}}_t^{\text{graph}}(\mathbf{A})\right) \\ & + \lambda_{\text{sp}} \|\mathbf{A}\|_1 + \lambda_{\text{anc}} \left\|\mathbf{A} - \mathbf{A}_t^{\text{mix}}\right\|_1 .\end{aligned} \tag{11}$$

The regularizers promote graph sparsity and anchor adaptation to the historical mixture. Implementation uses sparse-correction initialization followed by graph refinement, with stage-specific anchor coefficients reported in Appendix B.3. The adapted matrix is directly adopted as $\mathbf{A}_t = \mathbf{A}_t^{\text{adapt}}$.

When $q_t \geq \gamma_{\text{novel}}$, the current state is considered insufficiently represented by the existing expert library. Local adaptation is not performed; instead, the current window is used directly to construct a new expert as described in Section 3.4. The mixed matrix serves as the provisional causal estimate for the current window, and the new expert becomes available only to subsequent windows. Local corrections are not propagated; each subsequent window repeats the routing, mixture, and decision procedure, with temporal coupling only through routing smoothness and the updated expert library.

3.4 Continual Expert Expansion

Immediate Expert Construction. For a window satisfying $q_t \geq \gamma_{\text{novel}}$, OGC-MoE constructs a new expert from the observations in $\mathcal{W}_t$. Let $\mathbf{A}^{\text{cons}} = K^{-1}\sum_{k=1}^{K} \mathbf{A}_k$ denote the library consensus. The new predictor is initialized from the historical expert with the lowest current-window matching cost, while its graph combines the current mixed matrix and the library consensus:

$$\begin{aligned} k^\star = \arg\min_{1\leq k\leq K} c_{t,k}, \qquad \theta_{K+1}^{(0)} = \theta_{k^\star}, \\ \mathbf{A}_{K+1}^{(0)} = (1-\lambda_c)\mathbf{A}_t^{\text{mix}} + \lambda_c \mathbf{A}^{\text{cons}}, \end{aligned} \tag{12}$$

where $\lambda_c \in [0,1]$ is a fixed transfer coefficient shared across datasets; its value is reported in Appendix B.6. This transfer initialization provides a warm start, while all off-diagonal graph parameters of the new expert remain trainable. Implementation details are given in Appendix B.3.

New Expert Training and Library Update. The new expert $\mathcal{E}_{K+1} = (f_{\theta_{K+1}}, \mathbf{A}_{K+1})$ is trained on the current window $\mathcal{W}_t$ using the objective in Section 3.2, while all existing experts remain fixed. After training, it is appended to the library and K increases to $K+1$. To preserve causal online evaluation, the new expert is used only from window $t+1$ onward.

4 Experiments

4.1 Experimental setup

Unless otherwise stated, tabular results are reported as mean $\pm$ standard deviation over five independent runs. Ablation results are reported as five-run means, with standard deviations omitted for compactness. The implementation code is provided in the supplementary material.

Datasets. We evaluate OGC-MoE on two synthetic datasets generated from linear VAR and nonlinear Lorenz-96 systems, together with a real-world THS concept-index dataset. Both synthetic streams exhibit causal nonstationarity, with their directed causal structures and relation strengths

varying across temporal regimes. Table 2 summarizes the dataset sizes and online-window configurations used in the main comparison; the separate diagnostic stream for evaluating expert expansion is described in Section 4.2.3. Detailed generation procedures and dataset descriptions are provided in Appendix A and Appendix B.

Table 2: Dataset statistics and online-window configurations. Parentheses indicate the additional 256-variable experiments for OGC-MoE.

Dataset	Variables	Training	Testing	Window	Stride
VAR	128 (256)	3,000	2,000	400	200
Lorenz-96	128 (256)	3,000	2,000	400	200
THS	102	1,440	500	120	60

Evaluation Metrics. For synthetic data, we report window-averaged AUROC and AUPRC for causal-graph recovery, excluding self-loops. For financial data without a ground-truth graph, let $\mathcal{G}_{t,K}$ be the Top-K directed edges in window t, and $\mathcal{V}_{t,K}$ the subset with valid forward and reverse lifts. Using the cumulative 1–5-day lagged Lift defined in Appendix B.2, we compute

$$\begin{aligned}
\mathrm{LiftRate}_{t,K} &= \frac{1}{|\mathcal{G}_{t,K}|} \sum_{(i,j)\in\mathcal{G}_{t,K}} \mathbb{I}[\mathrm{Lift}_{1:5}(i \to j) > 1]\,, \\
\mathrm{MEL}_{t,K} &= \frac{1}{|\mathcal{G}_{t,K}|} \sum_{(i,j)\in\mathcal{G}_{t,K}} \mathrm{Lift}_{1:5}(i \to j), \\
\mathrm{DS}_{t,K} &= \frac{1}{|\mathcal{V}_{t,K}|} \sum_{(i,j)\in\mathcal{V}_{t,K}} \mathbb{I}[\mathrm{Lift}_{1:5}(i \to j) > \mathrm{Lift}_{1:5}(j \to i)]\,.
\end{aligned} \tag{13}$$

These metrics measure the fraction of above-baseline associations, average lift, and forward-over-reverse support, respectively. We evaluate $K \in \{50, 100, 200\}$ and average over test windows. They provide observational support rather than ground-truth causal accuracy; further details appear in Appendix B.2.

Baselines and Online Protocol. We compare OGC-MoE with eSRU (Khanna & Tan, 2019), NGM (Bellot et al., 2021), CUTS+ (Cheng et al., 2024), PCMCI$_\Omega$ (Gao et al., 2023), TS-CausalNN (Faruque et al., 2025), DCD (Ferdous & Gani, 2026), and CausalMoE (Liu et al., 2026). The test stream is processed chronologically. At window t, static-graph baselines are refitted using the training data and test observations available up to t, whereas CausalMoE and OGC-MoE follow their native inference procedures. No method accesses future windows or test ground-truth graphs. Further baseline descriptions and protocol details are given in Appendix B.8.

4.2 Experimental Results

4.2.1 Results on Synthetic Datasets

At 128 variables, OGC-MoE achieves the best AUROC and AUPRC on both synthetic datasets in Table 3, outperforming CausalMoE by particularly clear margins on VAR. The consistent gains under linear and nonlinear dynamics demonstrate its ability to recover evolving causal structures. With 256 variables and the same evaluation protocol, OGC-MoE achieves AUROCs of 0.929 and 0.951 on VAR and Lorenz-96, respectively, providing further evidence of graph-recovery performance in higher-dimensional settings.

4.2.2 Results on the Real-World Financial Dataset

Table 4 shows that OGC-MoE performs best across all metrics and Top-K settings, improving the Top-50 Lift> 1 Rate from CausalMoE's 0.603 to 0.665 with consistent gains at larger K.

Table 3: Causal-graph recovery on the synthetic datasets (higher is better). All rows use 128 variables except the final row; bold/underlined values mark the best/second-best results at 128 variables.

	VAR		Lorenz-96	
Model	AUROC	AUPRC	AUROC	AUPRC
eSRU	0.589 ± 0.018	0.155 ± 0.012	0.619 ± 0.020	0.097 ± 0.010
NGM	0.506 ± 0.014	0.191 ± 0.016	0.726 ± 0.018	0.362 ± 0.022
PCMCI$_\Omega$	0.681 ± 0.021	0.317 ± 0.024	0.731 ± 0.017	0.323 ± 0.020
CUTS+	0.703 ± 0.019	0.330 ± 0.021	0.836 ± 0.015	0.481 ± 0.021
TS-CausalNN	0.664 ± 0.017	0.267 ± 0.019	0.850 ± 0.013	0.495 ± 0.018
DCD	0.763 ± 0.016	0.474 ± 0.022	0.842 ± 0.014	0.623 ± 0.019
CausalMoE	$\underline{0.851 \pm 0.009}$	$\underline{0.622 \pm 0.012}$	$\underline{0.904 \pm 0.015}$	$\underline{0.728 \pm 0.010}$
OGC-MoE	$\mathbf{0.944 \pm 0.015}$	$\mathbf{0.802 \pm 0.018}$	$\mathbf{0.978 \pm 0.016}$	$\mathbf{0.815 \pm 0.014}$
OGC-MoE ($N = 256$)	0.929 ± 0.012	0.730 ± 0.023	0.951 ± 0.016	0.715 ± 0.025

Table 4: Results on the THS dataset across Top-K settings (higher is better).

Model	Lift>1 Rate	MEL	DS
Top-50			
eSRU	0.541 ± 0.018	1.015 ± 0.021	0.498 ± 0.024
NGM	0.572 ± 0.022	1.021 ± 0.017	0.458 ± 0.025
PCMCI$_\Omega$	0.554 ± 0.016	1.045 ± 0.025	0.515 ± 0.021
CUTS+	0.576 ± 0.024	1.032 ± 0.020	0.504 ± 0.027
TS-CausalNN	0.523 ± 0.026	1.011 ± 0.023	0.474 ± 0.029
DCD	0.584 ± 0.019	1.055 ± 0.026	0.524 ± 0.022
CausalMoE	0.603 ± 0.015	1.078 ± 0.020	0.546 ± 0.018
OGC-MoE	$\mathbf{0.665 \pm 0.020}$	$\mathbf{1.103 \pm 0.031}$	$\mathbf{0.561 \pm 0.025}$
Top-100			
eSRU	0.529 ± 0.020	1.012 ± 0.019	0.479 ± 0.026
NGM	0.564 ± 0.019	1.019 ± 0.016	0.454 ± 0.023
PCMCI$_\Omega$	0.564 ± 0.018	1.040 ± 0.022	0.503 ± 0.020
CUTS+	0.556 ± 0.023	1.024 ± 0.018	0.489 ± 0.025
TS-CausalNN	0.543 ± 0.022	1.013 ± 0.020	0.485 ± 0.027
DCD	0.558 ± 0.021	1.047 ± 0.023	0.515 ± 0.021
CausalMoE	0.591 ± 0.025	1.055 ± 0.021	0.524 ± 0.017
OGC-MoE	$\mathbf{0.635 \pm 0.018}$	$\mathbf{1.085 \pm 0.028}$	$\mathbf{0.549 \pm 0.023}$
Top-200			
eSRU	0.535 ± 0.017	1.012 ± 0.018	0.490 ± 0.022
NGM	0.561 ± 0.021	1.019 ± 0.015	0.472 ± 0.024
PCMCI$_\Omega$	0.568 ± 0.017	1.039 ± 0.021	0.504 ± 0.019
CUTS+	0.561 ± 0.020	1.025 ± 0.017	0.482 ± 0.024
TS-CausalNN	0.555 ± 0.024	1.014 ± 0.019	0.485 ± 0.026
DCD	0.552 ± 0.018	$\underline{1.045 \pm 0.022}$	0.502 ± 0.020
CausalMoE	$\underline{0.575 \pm 0.021}$	1.043 ± 0.011	$\underline{0.506 \pm 0.017}$
OGC-MoE	$\mathbf{0.627 \pm 0.021}$	$\mathbf{1.069 \pm 0.026}$	$\mathbf{0.522 \pm 0.022}$

4.2.3 Evaluation of Continual Expert Expansion

For each synthetic system, a separate $10{,}000$-step stream contains one routing-context segment and 49 randomly ordered scored segments: 34 novel and 15 historically explainable. Each decision uses the current and preceding segments, matching the main 400-step input and 200-step stride. We report detection metrics and compare the refined no-expansion mixture with the new-expert graph recovered after training on the corresponding segment; evaluation details appear in Appendix A.3.

Table 5: Continual expert expansion on the synthetic datasets. Higher is better.

	Novel-window detection			Graph AUROC on novel windows	
Dataset	Precision	Recall	F1	No expansion	After training
VAR	0.956 ± 0.025	1.000 ± 0.000	0.977 ± 0.013	0.617 ± 0.011	0.905 ± 0.026
Lorenz-96	0.951 ± 0.044	1.000 ± 0.000	0.975 ± 0.023	0.605 ± 0.015	0.922 ± 0.021

OGC-MoE detects all novel windows, with precision/F1 of $0.956/0.977$ on VAR and $0.951/0.975$ on Lorenz-96. The refined mixtures achieve AUROCs of 0.617 and 0.605, whereas the trained new-

expert graphs achieve 0.905 and 0.922, respectively, indicating substantially improved structural recovery after new-expert training.

4.2.4 Ablation Studies

We ablate graph mixture, local adaptation, routing refinement, and the residual score through *Hard Routing*, *w/o Local Adaptation*, *w/o Routing Refinement*, and *w/o Residual Score*, respectively.

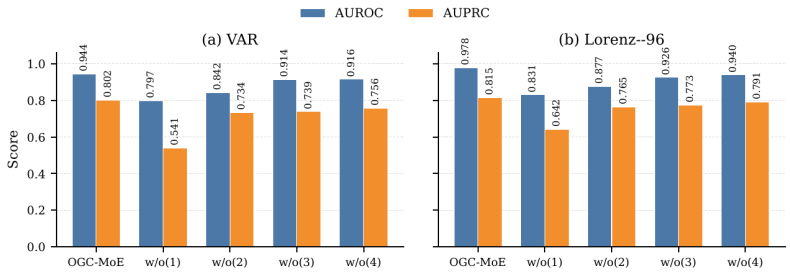

Figure 2: Ablation results on the synthetic datasets. From *w/o(1)* to *w/o(4)*, the variants correspond to hard routing, w/o local adaptation, w/o routing refinement, and w/o residual score, respectively. Higher values are better.

Hard routing and removing local adaptation cause the largest losses, while routing refinement and residual matching provide smaller but consistent gains.

For the THS dataset, Table 6 reports Lift>1 Rate.

Table 6: THS ablation results: mean Lift>1 Rate (higher is better).

Variant	Top-50	Top-100	Top-200
Hard Routing	0.600	0.617	0.621
w/o Local Adaptation	0.603	0.620	0.617
w/o Routing Refinement	0.614	0.620	0.625
w/o Residual Score	0.659	0.633	0.604
OGC-MoE	**0.665**	**0.635**	**0.627**

OGC-MoE performs best for every Top-K setting. At Top-50, hard routing and removing local adaptation reduce Lift>1 Rate from 0.665 to 0.600 and 0.603, respectively. Removing the residual score has the largest effect at Top-200, reducing the rate from 0.627 to 0.604. These results suggest that the components contribute differently across edge-selection budgets.

5 Conclusion

We presented OGC-MoE, an online Granger-causal mixture-of-experts framework for high-dimensional nonstationary time series. Guided by prediction and causal-residual compatibility, it combines historical causal graphs, adapts to local structural changes, and learns new experts for unseen states while preserving existing experts. Experiments on nonstationary VAR and Lorenz-96 systems, including the 256-variable setting, demonstrate strong graph recovery; financial results provide complementary evidence of predictive relevance. Ablation and expansion studies support graph mixture, local adaptation, and new-expert learning. Together, these findings support combining historical mechanism reuse with current-window structural learning, allowing the framework to accommodate both familiar causal patterns and newly encountered states. Future work will explore uncertainty-aware routing and expert-library management to balance adaptation, knowledge retention, and computational cost in longer streams.

AI USE STATEMENT

All ideas and contributions presented in this work originate from the authors, with each author making a substantial contribution to the research. Generative AI was not used to develop the research ideas, methodology, experiments, analysis, or conclusions. AI-assisted tools were used exclusively for language editing and polishing to improve the clarity and readability of the manuscript.

REFERENCES

Charles K Assaad, Emilie Devijver, and Eric Gaussier. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 73:767–819, 2022.

Sumanta Basu, Ali Shojaie, and George Michailidis. Network granger causality with inherent grouping structure. *The Journal of Machine Learning Research*, 16(1):417–453, 2015.

Alexis Bellot, Kim Branson, and Mihaela van der Schaar. Neural graphical modelling in continuous-time: consistency guarantees and algorithms. *arXiv preprint arXiv:2105.02522*, 2021.

Zachris Björkman, Jorge Loría, Sophie Wharrie, and Samuel Kaski. Incorporating expert knowledge into bayesian causal discovery of mixtures of directed acyclic graphs. *arXiv preprint arXiv:2510.06735*, 2025.

Bart Bussmann, Jannes Nys, and Steven Latré. Neural additive vector autoregression models for causal discovery in time series. In *International Conference on Discovery Science*, pp. 446–460. Springer, 2021.

Yuxiao Cheng, Lianglong Li, Tingxiong Xiao, Zongren Li, Jinli Suo, Kunlun He, and Qionghai Dai. Cuts+: High-dimensional causal discovery from irregular time-series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11525–11533, 2024.

Omar Faruque, Sahara Ali, Xue Zheng, and Jianwu Wang. Ts-causalnn: Learning temporal causal relations from non-linear non-stationary time series data. In *2025 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 323–332. IEEE, 2025.

Muhammad Hasan Ferdous and Md Osman Gani. Dcd: Decomposition-based causal discovery from autocorrelated and non-stationary temporal data. *arXiv preprint arXiv:2602.01433*, 2026.

Shanyun Gao, Raghavendra Addanki, Tong Yu, Ryan Rossi, and Murat Kocaoglu. Causal discovery in semi-stationary time series. *Advances in Neural Information Processing Systems*, 36:46624–46657, 2023.

Wei Gao and Haizhong Yang. Time-varying group lasso granger causality graph for high dimensional dynamic system. *Pattern Recognition*, 130:108789, 2022.

Wenbo Gong, Joel Jennings, Cheng Zhang, and Nick Pawlowski. Rhino: Deep causal temporal relationship learning with history-dependent noise. *arXiv preprint arXiv:2210.14706*, 2022.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969.

Wiebke Günther, Urmi Ninad, and Jakob Runge. Causal discovery for time series from multiple datasets with latent contexts. In *Uncertainty in Artificial Intelligence*, pp. 766–776. PMLR, 2023.

Stefan Haufe, Klaus-Robert Müller, Guido Nolte, and Nicole Krämer. Sparse causal discovery in multivariate time series. In *causality: objectives and assessment*, pp. 97–106. PMLR, 2010.

Biwei Huang, Kun Zhang, Mingming Gong, and Clark Glymour. Causal discovery and forecasting in nonstationary environments with state-space models. In *International conference on machine learning*, pp. 2901–2910. Pmlr, 2019.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020.

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

Alexis Kafantaris. Enes causal discovery. *arXiv preprint arXiv:2603.24436*, 2026.

Saurabh Khanna and Vincent YF Tan. Economy statistical recurrent units for inferring nonlinear granger causality. *arXiv preprint arXiv:1911.09879*, 2019.

Bo Liu, Di Dai, Jingwei Liu, Jiarui Jin, Xiaocheng Fang, Guangkun Nie, Hongyan Li, and Shenda Hong. Causalmoe: A billion-scale multimodal foundation model for granger causal discovery with pattern-routed heterogeneous experts. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 11422–11431, 2026.

Sindy Löwe, David Madras, Richard Zemel, and Max Welling. Amortized causal discovery: Learning to infer causal graphs from time-series data. In *Conference on causal learning and reasoning*, pp. 509–525. PMLR, 2022.

Ričards Marcinkevičs and Julia E Vogt. Interpretable models for granger causality using self-explaining neural networks. *arXiv preprint arXiv:2101.07600*, 2021.

Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):19, 2019.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International conference on artificial intelligence and statistics*, pp. 1595–1605. Pmlr, 2020.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. *Advances in neural information processing systems*, 26, 2013.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on uncertainty in artificial intelligence*, pp. 1388–1397. Pmlr, 2020.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5 (11):eaau4996, 2019.

Ali Shojaie and George Michailidis. Discovering graphical granger causality using the truncating lasso penalty. *Bioinformatics*, 26(18):i517–i523, 2010.

Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B Fox. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4267–4279, 2021.

Burak Varıcı, Dmitriy A Katz, Dennis Wei, Prasanna Sattigeri, and Ali Tajer. Interventional causal discovery in a mixture of dags. *Advances in Neural Information Processing Systems*, 37:86574–86601, 2024.

APPENDIX A SYNTHETIC DATASET GENERATION

We construct two nonstationary synthetic datasets based on linear VAR and nonlinear Lorenz-96 systems. For the main comparison, both datasets contain $N = 128$ variables, 3,000 training observations, and a 2,000-step online test stream. The training data are organized into $K = 4$ predefined domains of 750 observations. Within each training domain, the causal support and interaction strengths remain fixed, providing a relatively stable historical state from which one expert is learned. The domain labels are supplied only for offline training and are not inferred by OGC-MoE.

The test stream is generated separately as a sequence of window-level causal states without predefined test-domain labels. Within each evaluation window, the causal support and interaction strengths remain fixed; nonstationarity is introduced only between successive windows through edge deletion, edge addition, and coefficient perturbation. This process permits both the recurrence of previously observed relations and the emergence of new causal patterns during online evaluation. The complete data generator is fixed before model training and does not use routing weights or model outputs. Each reported run uses distinct data-generation and model-training seeds, while all methods within a run receive the same data streams.

APPENDIX A.1 NONSTATIONARY VAR DATASET

The VAR process is defined as

$$\mathbf{x}_t = \sum_{\ell=1}^{p} \mathbf{B}_{r(t)}^{(\ell)} \mathbf{x}_{t-\ell} + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}), \tag{14}$$

where $r(t)$ indexes the evaluation window containing time t, $\mathbf{B}_{r(t)}^{(\ell)}$ is its fixed lag-ℓ coefficient matrix, $p = 5$, and $\sigma = 0.1$. Following the convention that $i \to j$ denotes a causal influence from variable i to variable j, the ground-truth graph is

$$A_{r,ij}^{\star} = \mathbb{I}\left[\max_{1 \le \ell \le p} \left| B_{r,ji}^{(\ell)} \right| > 0\right]. \tag{15}$$

For each training domain, every target variable is independently assigned five distinct non-self causal parents. Each parent is assigned uniformly to one of the p lags, with coefficient magnitude sampled from $\mathcal{U}(0.2, 0.5)$ and sign sampled uniformly. The total self-regressive strength is set to 0.8, and the resulting coefficient matrices remain fixed within that domain. The test stream is initialized from a separately sampled coefficient matrix following the same base rule. Its causal structure is then updated chronologically: for each target, the number of replaced parents is sampled uniformly from $\{0, 1, 2, 3\}$, and removed parents are replaced by variables absent from the current parent set so that the expected in-degree remains fixed. For two successive test states, the update is

$$\mathbf{B}_r^{(\ell)} = \left(\mathbf{1} - \mathbf{D}_r^{(\ell)}\right) \odot \mathbf{S}_r^{(\ell)} \odot \mathbf{B}_{r-1}^{(\ell)} + \mathbf{U}_r^{(\ell)}, \tag{16}$$

where r indexes successive evaluation windows, $\mathbf{D}_r^{(\ell)}$ masks deleted coefficients, $\mathbf{U}_r^{(\ell)}$ contains newly added coefficients, and the nonzero entries of $\mathbf{S}_r^{(\ell)}$ independently perturb retained coefficients with magnitudes sampled from $\mathcal{U}(0.7, 1.3)$. Retained coefficients change sign with probability 0.1, while new coefficients follow the initialization rule above. Each resulting system is rescaled when necessary to satisfy the VAR stability condition. The updated coefficient matrices are fixed throughout window r and are changed only before generating the next window. Consequently, every evaluated window has a single ground-truth graph, while the sequence remains causally nonstationary across windows. Window-level mechanism identities are not provided to the evaluated methods.

APPENDIX A.2 NONSTATIONARY LORENZ-96 DATASET

The nonlinear dataset is generated from a generalized Lorenz-96 system:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = s_{r(t),i} \left(x_{p_{i,1}^{(r(t))}} - x_{p_{i,2}^{(r(t))}}\right) x_{p_{i,3}^{(r(t))}} - x_i + F, \tag{17}$$

where $\mathcal{P}_i^{(r)} = \{p_{i,1}^{(r)}, p_{i,2}^{(r)}, p_{i,3}^{(r)}\}$ contains the three causal parents of variable x_i. We set $F = 10$, and the ground-truth graph is defined by $A^{\star}_{r,ji} = \mathbb{I}[j \in \mathcal{P}_i^{(r)}]$.

Each training domain is generated independently from the standard Lorenz-96 neighborhood, with 55% of variable positions randomly permuted to obtain a sparse nonlinear interaction graph. Its parent sets and interaction strengths remain fixed within the domain. The online test stream begins from another independently sampled parent configuration and then evolves chronologically under stochastic rewiring. For each target variable, the number of replaced parents is sampled uniformly from $\{0, 1, 2\}$. The selected parents are removed and replaced by distinct non-self variables that are not already parents:

$$\mathcal{P}_i^{(r)} = \left(\mathcal{P}_i^{(r-1)} \setminus \mathcal{D}_{r,i}\right) \cup \mathcal{U}_{r,i}, \qquad |\mathcal{P}_i^{(r)}| = 3, \tag{18}$$

where r indexes successive evaluation windows and $\mathcal{D}_{r,i}$ and $\mathcal{U}_{r,i}$ are equally sized deletion and addition sets. Duplicate parents and self-relations are excluded. Interaction strengths evolve independently according to $s_{r,i} = \mathrm{clip}(s_{r-1,i} + \xi_{r,i}, 0.6, 1.4)$, where $\xi_{r,i} \sim \mathcal{N}(0, 0.1^2)$. The parent sets and strengths are held fixed within each evaluation window and updated only when moving to the next window. Thus, the trajectory evolves recursively across windows without within-window causal changes.

The system is integrated using the fourth-order Runge–Kutta method with a step size of 0.05. Independent Gaussian noise with standard deviation 0.01 is added after each integration step, and the first $1{,}000$ steps are discarded as burn-in. Each window is integrated under one fixed parent configuration and strength vector. The resulting active parent set defines the window-level ground-truth graph; its mechanism identity is not provided during online inference.

APPENDIX A.3 DIAGNOSTIC STREAM FOR CONTINUAL EXPERT EXPANSION

The continual-expansion study uses an additional $10{,}000$-step stream for each synthetic system; it is not used in the main method comparison. Each stream is divided into fifty 200-step mechanism segments. The first segment provides the historical context needed for the first 400-step model input and is not scored. Each of the remaining 49 segments is evaluated using that segment and its preceding 200 steps, which preserves the window length of 400 and stride of 200 used in the main experiment. Among the scored segments, 15 are non-novel and 34 are novel, and their temporal positions are randomized independently for each run.

A non-novel segment is generated from a historical training mechanism, a recombination of relations from multiple historical mechanisms, or such a structure with sparse local edge additions and deletions. These cases test reuse and local adaptation without requiring library expansion. For a recombined segment, the generator additionally records the normalized source-mechanism proportions $\boldsymbol{\alpha}_t^{\star} \in \Delta^{K-1}$; mechanisms absent from the combination receive zero weight. These proportions are retained only for the routing diagnostic below and are never provided to the model. Each novel segment uses a separately sampled causal support that is unavailable in the historical expert library, so different novel segments need not share the same causal mechanism. Both these supports and their temporal positions are resampled for every run. Consequently, novelty is determined by the source mechanism rather than by the model's matching score or expansion threshold.

The generator records a binary label for each evaluation window. These labels, together with the time-indexed ground-truth graphs, are used only to compute novel-window precision, recall, and F1 and new-expert AUROC. For each correctly detected novel segment, the expert trained from that segment is evaluated against its corresponding ground-truth graph, and the resulting AUROCs are averaged within each run. This measures post-training structural recovery for the observed segment. During online processing, the mixed graph is the provisional output for that window, and the new expert becomes available from the next window onward. The reported new-expert AUROC therefore does not measure the graph output before training or performance on subsequent unseen observations. The labels are not provided to OGC-MoE for routing, adaptation, or expert construction. Thus, the diagnostic stream separately evaluates whether the method distinguishes historically explainable variation from dispersed unseen causal structure and whether the resulting expert recovers that new structure.

Table 7: Routing-weight recovery on recombined non-novel segments. Results are reported as mean $\pm$ standard deviation over five runs.

Dataset	Alpha MAE ↓	Alpha Cosine ↑
VAR	0.0123 ± 0.0083	0.9988 ± 0.0013
Lorenz-96	0.0242 ± 0.0134	0.9957 ± 0.0034

Routing recovery on recombined segments. For the subset of non-novel segments generated by recombining historical mechanisms, we additionally examine whether the inferred routing weights recover the source proportions used by the generator. Let $\mathcal{M}$ denote the set of recombined segments and $\widehat{\boldsymbol{\alpha}}_t \in \Delta^{K-1}$ the routing weights estimated by OGC-MoE after expert selection and renormalization, with unselected experts assigned zero weight. We report the mean absolute weight error

$$\mathrm{AlphaMAE} = \frac{1}{|\mathcal{M}|} \sum_{t \in \mathcal{M}} \frac{1}{K} \left\| \widehat{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}_t^\star \right\|_1, \tag{19}$$

and the mean cosine similarity

$$\mathrm{AlphaCos} = \frac{1}{|\mathcal{M}|} \sum_{t \in \mathcal{M}} \frac{\widehat{\boldsymbol{\alpha}}_t^\top \boldsymbol{\alpha}_t^\star}{\left\| \widehat{\boldsymbol{\alpha}}_t \right\|_2 \left\| \boldsymbol{\alpha}_t^\star \right\|_2}. \tag{20}$$

Lower AlphaMAE and higher AlphaCos indicate better recovery of the generating expert combination.

The low AlphaMAE and near-unit cosine similarity show that the router closely recovers the generating expert proportions on the controlled recombined segments. The slightly larger error on Lorenz-96 suggests that routing is more difficult under nonlinear dynamics, although the estimated mixture direction remains highly consistent with the source combination. This diagnostic is supplementary and is not used in the primary graph-recovery or novelty-detection evaluation.

APPENDIX B ADDITIONAL EXPERIMENTAL DETAILS

APPENDIX B.1 REAL-WORLD FINANCIAL DATASET

The real-world dataset consists of 102 THS concept indices and $1{,}940$ chronologically ordered trading-day observations. The first $1{,}440$ observations are used for training, and the remaining 500 form the online test stream. This dataset evaluates expert routing and the temporal consistency of the discovered relations in a real financial system. Because complete and reliable ground-truth causal graphs are unavailable, graph-recovery metrics are not used for this dataset.

Historical-domain construction. The expert-training domains are constructed using only the $1{,}440$-day training period. We first fit a VAR model and form time-indexed features from its prediction residuals and graph-weighted residual statistics. After reducing these features with PCA, we apply PELT change-point detection and partition the training sequence into four contiguous intervals containing 494, 341, 240, and 365 trading days, respectively. One causal expert is trained on each interval. This partition is fixed before online evaluation and does not use any test observations.

APPENDIX B.2 EVALUATION METRICS

For the VAR and Lorenz-96 datasets, AUROC measures discrimination between true and non-causal edges, whereas AUPRC emphasizes precision–recall performance under sparse graphs and class imbalance. Self-loops are excluded; scores are computed per test window and then averaged.

Financial-event definition. For concept index i on trading day u, let $C_{u,i}$ denote its closing price and define the log return $r_{u,i} = \log C_{u,i} - \log C_{u-1,i}$. We remove the common market movement using the cross-index median,

$$r_u^{\mathrm{mkt}} = \mathrm{median}_j(r_{u,j}), \qquad \widetilde{r}_{u,i} = r_{u,i} - r_u^{\mathrm{mkt}}. \tag{21}$$

Let $\mathcal{R}$ be a reference period preceding online evaluation. The concept-specific threshold and abnormal-movement indicator are

$$\xi_i = Q_{0.8}(\{|\widetilde{r}_{s,i}| : s \in \mathcal{R}\}), \qquad E_{u,i} = \mathbb{I}[|\widetilde{r}_{u,i}| > \xi_i]. \tag{22}$$

Here, $Q_{0.8}$ is the empirical quantile. Thus, threshold estimation does not use test or future observations.

To evaluate temporal directional support, a source event at day u is paired with whether the target experiences at least one event during the following five trading days:

$$F_{u,j}^{(1:5)} = \max_{h\in\{1,\dots,5\}} E_{u+h,j}, \qquad \mathrm{Lift}_{1:5}(i \to j) = \frac{P(F_{u,j}^{(1:5)} = 1 \mid E_{u,i} = 1)}{P(F_{u,j}^{(1:5)} = 1)}. \tag{23}$$

The reverse score swaps i and j while preserving this temporal order. Hence, Direction Support compares temporally ordered relations rather than same-day association. Future events are used only for post-hoc evaluation and never for online inference.

For financial data, Eq. 13 is evaluated on the Top-K directed edges ranked by estimated causal strength in each window. Direction Support uses only edges with valid forward and reverse cumulative lagged lifts, and source days without a complete five-day horizon are excluded. All metrics are computed at $K \in \{50, 100, 200\}$ and averaged across the online stream.

APPENDIX B.3 Optimization and Initialization Details

For routing refinement, we use $\boldsymbol{\alpha} = \mathrm{softmax}(\mathbf{u})$ with $\mathbf{u}^{(0)} = \log(\boldsymbol{\alpha}_t^{(0)} + \epsilon)$ and optimize only $\mathbf{u}$ for a fixed number of steps. The learning rate and optimization budget are selected on held-out historical windows. The smoothness term is omitted for the first online window; after a new expert is appended, a zero entry is added to the preceding routing vector before computing that term.

Two-stage local adaptation. The implementation first learns a sparse correction $\Delta\mathbf{A}_t$ and forms $\mathbf{A}_t^{\mathrm{cand}} = \mathrm{clip}(\mathbf{A}_t^{\mathrm{mix}} + \Delta\mathbf{A}_t, 0, 1)$, regularized by $\lambda_{\mathrm{anc}}^{\Delta} = 0.03$. Starting from this candidate, Eq. 11 refines the final graph using $\lambda_{\mathrm{anc}}^{\mathrm{adapt}} = 0.005$. These fixed stage-specific coefficients instantiate the abstract anchor strength λ_{anc} used in the main text.

Because expert graphs are parameterized by logits, a new graph is initialized as $\mathbf{Z}_{K+1}^{(0)} = \mathrm{logit}(\mathrm{clip}(\mathbf{A}_{K+1}^{(0)}, \epsilon, 1-\epsilon))$, with its diagonal masked to zero. This is only a warm start, and all off-diagonal logits remain trainable.

APPENDIX B.4 Post-Mixture and New-Expert Reference Statistics

Each historical expert k stores prediction-loss statistics (μ_k^T, σ_k^T) and residual-profile statistics $(\boldsymbol{\mu}_k^R, \boldsymbol{\sigma}_k^R)$ computed from its historical calibration windows. For the selected expert set $\mathcal{S}_t$, let

$$\ell_t^{\mathrm{mix}} = \mathcal{L}_{\mathrm{pred}}\Big(\mathbf{Y}_t, \widehat{\mathbf{Y}}_t^{\mathrm{graph}}(\mathbf{A}_t^{\mathrm{mix}})\Big) \tag{24}$$

denote the prediction loss under the mixed causal-strength matrix. Using the refined weights after selection and renormalization, we construct a routing-weighted reference mean and a root-mean-square reference scale:

$$\mu_t^{T,\mathrm{mix}} = \sum_{k\in\mathcal{S}_t} \alpha_{t,k}\mu_k^T, \qquad \sigma_t^{T,\mathrm{mix}} = \left(\sum_{k\in\mathcal{S}_t} \alpha_{t,k}(\sigma_k^T)^2\right)^{1/2}. \tag{25}$$

These quantities provide a normalization reference; the scale is not an estimate of the standard deviation of the mixed prediction loss. The normalized mixed-prediction score is

$$z_t^{\mathrm{mix}} = \max\left\{0, \frac{\ell_t^{\mathrm{mix}} - \mu_t^{T,\mathrm{mix}}}{\sigma_t^{T,\mathrm{mix}} + \epsilon}\right\}. \tag{26}$$

For the residual term, each selected expert is evaluated using $f_{\theta_k}(\mathcal{W}_t, \mathbf{A}_k)$ and compared with its own reference profile via Eq. 6. The post-mixture residual discrepancy is

$$d_t^{\text{mix}} = \sum_{k \in \mathcal{S}_t} \alpha_{t,k} d_{t,k}^R. \tag{27}$$

Thus, z_t^{mix} assesses prediction under the shared mixed graph, while d_t^{mix} summarizes the current window's residual compatibility with the selected historical experts under their own graphs. The two scores enter Eq. 10 with the historical calibration scales ρ_z and ρ_d.

After a new expert $\mathcal{E}_{K+1}$ is trained on an unseen-state window, we construct its reference statistics using the same profiling procedure. The observed training window is partitioned into reference subwindows, from which the trained expert's prediction losses and residual profiles are collected to estimate μ_{K+1}^T, σ_{K+1}^T, $\boldsymbol{\mu}_{K+1}^R$, and $\boldsymbol{\sigma}_{K+1}^R$. These subwindows come from the data used to train the new expert. Minimum scale floors prevent degenerate normalization when few reference subwindows are available. The new expert and its reference statistics are appended to the library together and are used only from window $t+1$ onward.

APPENDIX B.5 NOVELTY-THRESHOLD CALIBRATION

Rather than using a universal numerical threshold across datasets, we calibrate γ_{novel} on an independent validation stream containing only historical mechanisms and their valid combinations. This stream represents windows that should be explainable by the existing expert library. Let $\mathcal{V}_{\text{known}}$ denote these validation windows. We set

$$\gamma_{\text{novel}} = 1.10 \max\left\{1, Q_{0.95}\left(\{q_t : \mathcal{W}_t \in \mathcal{V}_{\text{known}}\}\right)\right\}, \tag{28}$$

where $Q_{0.95}$ denotes the empirical 95th percentile. The factor 1.10 provides a fixed safety margin and guarantees $\gamma_{\text{novel}} > 1$. A test window is treated as an unseen state when $q_t \geq \gamma_{\text{novel}}$. The same calibration rule is used for all datasets without access to test windows or test labels.

APPENDIX B.6 FIXED EXPERT-SELECTION AND TRANSFER SETTINGS

To avoid dataset-specific tuning, we use the same expert-selection and transfer settings in all experiments. Each online window retains the two experts with the largest refined routing weights, and new-expert graph initialization uses $\lambda_c = 0.3$. Table 8 summarizes these settings.

Table 8: Fixed expert-selection and transfer settings.

Setting	VAR	Lorenz-96	THS
Selected experts	Top-2	Top-2	Top-2
λ_c	0.3	0.3	0.3

APPENDIX B.7 HYPERPARAMETER SENSITIVITY

We vary λ_R and the overall local-adaptation anchor strength by $x \in \{0, 0.25, 0.5, 1, 2, 4\}$, one component at a time, with $x = 1$ as the default. Each point is the mean over five runs; error bars are omitted. The default λ_R is 0.1 on VAR and 0.5 on Lorenz-96. For the anchor analysis, the same x scales both implementation coefficients: $\lambda_{\text{anc}}^{\text{adapt}} = 0.005x$ and $\lambda_{\text{anc}}^{\Delta} = 0.03x$.

Figure 3 shows that the default residual weight achieves the best joint AUROC–AUPRC result on both datasets, while moderate changes retain similar AUROC. Removing the residual term reduces performance, particularly on VAR. The joint anchor multiplier has a larger effect: eliminating or strongly weakening it substantially lowers AUROC, whereas the default gives the best AUPRC and either the best or tied-best AUROC. Excessively large anchor multipliers reduce AUPRC because they restrict adaptation to changing causal relations. These results support using both signals and show that the selected defaults provide a balanced setting across linear and nonlinear dynamics.

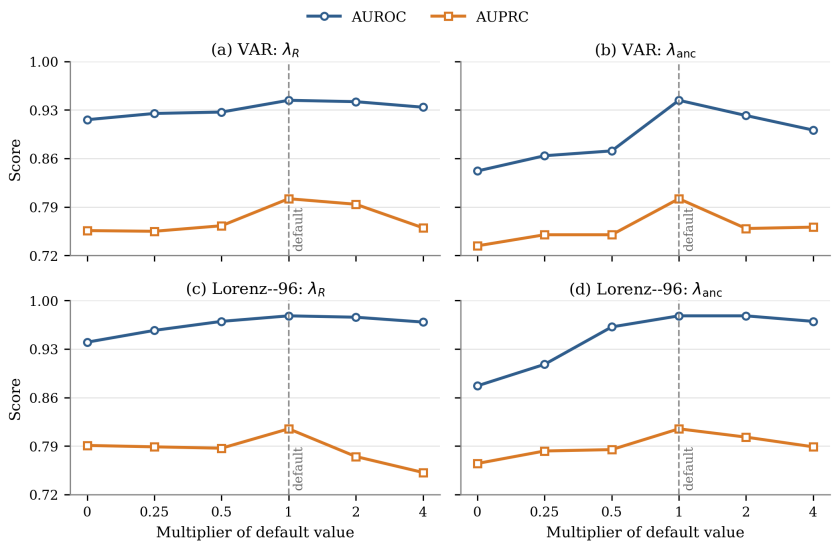

Figure 3: Sensitivity to λ_R and the local-adaptation anchor strength on the synthetic datasets. For the anchor analysis, the horizontal axis gives the common multiplier applied to both implementation-level coefficients; the dashed line marks the default configuration.

APPENDIX B.8 BASELINE METHODS AND ONLINE EVALUATION PROTOCOL

We include three general time-series causal discovery baselines. eSRU (Khanna & Tan, 2019) uses structured group sparsity in component-wise recurrent predictors to identify nonlinear Granger-causal relations. NGM (Bellot et al., 2021) learns a sparse continuous-time vector field with neural ordinary differential equations and extracts local dependencies from the learned dynamics. CUTS+ (Cheng et al., 2024) combines coarse-to-fine causal discovery with graph-neural message passing for high-dimensional and irregularly sampled time series.

We also compare with methods designed for heterogeneous or nonstationary causal mechanisms. PCMCI$_\Omega$ (Gao et al., 2023) models alternating and recurring mechanisms in semi-stationary time series. TS-CausalNN (Faruque et al., 2025) uses parallel causal convolutional layers to discover nonlinear contemporaneous and lagged relations. DCD (Ferdous & Gani, 2026) decomposes a non-stationary series into trend, seasonal, and residual components and integrates their causal relations. CausalMoE (Liu et al., 2026) routes temporal patches to heterogeneous representation experts and recovers a Granger-causal graph after representation-level aggregation, whereas OGC-MoE routes and combines predictor–causal-strength-matrix experts.

At window t, static-graph baselines are refitted on the training data and test observations available through t, while CausalMoE and OGC-MoE follow their native inference procedures. Thus, no method accesses future observations or test ground-truth graphs.