CAUSALEXOFORMER: END-TO-END LAG-AWARE CAUSAL STRUCTURE LEARNING FOR EXOGENOUS TIME-SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Exogenous time-series forecasting requires identifying relevant external variables, capturing their delayed effects, and modeling interactions among them. These dependencies are often learned implicitly, leaving the structure of external information use difficult to inspect and control. We propose CausalExoFormer, an end-to-end Transformer that jointly learns variable relevance, temporal allocation, and directed cross-channel structure for forecasting. A factorized source–lag gate selects historical observations, a directed graph learned with acyclicity regularization models interactions among external variables, and causal-gated cross-attention integrates the resulting representations with the target history. The learned structure directly controls information flow: closing a source gate removes both its direct and graph-mediated contributions. For multivariate forecasting, a shared graph supports target-specific predictions through exact self-exclusion, preserving the separation between each target's own history and its external inputs. Experiments on six standard benchmarks and six lake dissolved-oxygen forecasting tasks demonstrate the effectiveness of this approach. CausalExoFormer achieves the lowest horizon-averaged MAE on all six standard benchmarks and the lowest horizon-averaged MSE on four, among the evaluated methods, while consistently improving over TimeXer on the six lake tasks. Ablation studies establish the contribution of learned source–lag allocation, and environmental case studies reveal the variables and historical offsets used for prediction. Code is available at `https://anonymous.4open.science/r/CausalExoFormer-C42C`.

1 INTRODUCTION

Exogenous variables are central to environmental forecasting. Water temperature, biological activity, and water chemistry jointly influence dissolved oxygen; atmospheric variables interact across multiple timescales; and electricity series exhibit dependencies between consumers. These systems pose three connected questions: *which* external variables matter, *which historical observations* carry their information, and *how* the external variables interact. A forecasting architecture that exposes these choices can connect predictive performance to process-oriented analysis of environmental systems (Reichstein et al., 2019).

Recent Transformers improve long-horizon forecasting through patching and variate-level representations (Nie et al., 2023; Liu et al., 2024). TimeXer (Wang et al., 2024) introduces a particularly useful interface: a global target token queries external variate tokens through cross-attention. This interface learns content-dependent relevance, but does not explicitly separate dataset-level variable relevance from lag allocation. CrossLinear (Zhou et al., 2025) models direct cross-variable dependence and extends to multivariate forecasting through weight sharing, while GCGNet (Li et al., 2026) and DAG (Qiu et al., 2026) model structured temporal and channel interactions for TSF-X. In parallel, PCMCI+ (Runge, 2020), NOTEARS (Zheng et al., 2018), and DYNOTEARS (Pamfil et al., 2020) provide tools for discovering directed dependencies. Using a discovered graph as fixed preprocessing, however, separates its estimation from the representations optimized for the downstream forecast.

We propose **CausalExoFormer**, which learns lag-aware structure directly inside the forecasting pathway. A shared source-relevance score and historical-offset distribution determine which obser-

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.

vations can enter the cross-channel pathway, while a single directed graph organizes cross-channel propagation at each observed historical offset. Target-specific content attention then decides how the shared structured source bank contributes to each forecast output. Both structural components receive gradients from the joint prediction objective.

Making a structural gate inspectable also requires making it functional. Source gating precedes graph propagation, so closing a source removes both its direct and mediated external paths. For multi-target forecasting, we additionally remove each target's own source contribution by an exact rank-one correction before its query attends to the external bank; this preserves the separation between endogenous inputs and external source values without learning a separate graph for every output. A fixed null alternative anchors the absolute scale of gated attention, and a fixed-norm graph projection removes the reciprocal scalar rescaling of adjacency and message transformation. These choices give the structural parameters a concrete role in the computation and make information access directly testable by closing a source gate. This operational criterion complements the debate on attention-based explanations (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019).

Our contributions are threefold:

1. **Lag-aware causal structure on the prediction path.** A factorized source–lag gate and a globally shared contemporaneous graph jointly control how historical observations enter multi-target forecasts, with exact target self-exclusion preserving the target-relative external input bank.
2. **Exact structural control with a shared graph.** Source gating controls direct and graph-mediated paths, null-augmented attention anchors gate scale, and a fixed-scale projection makes adjacency strength operational. The shared graph requires one $S \times S$ acyclicity constraint rather than one graph per forecast target.
3. **Broad forecasting evidence and structural analysis.** Experiments cover six standard benchmarks and six lake tasks, with component ablations, sensitivity analyses, computational measurements, a lake case study, and a controlled synthetic target-source and lag-recovery protocol.

2 Problem and Structural Representation

Let $X_{1:T} \in \mathbb{R}^{T\times D}$ contain the observed histories of $D \geq 2$ channels. For a nonempty target set $\mathcal{M} \subseteq \{1,\dots,D\}$ with $M = |\mathcal{M}|$, we predict $\widehat{Y} \in \mathbb{R}^{H\times M}$ from observations available by the forecast origin T. Standard multivariate forecasting uses $\mathcal{M} = \{1,\dots,D\}$; a designated-target application uses $M = 1$. Write $Y_{h,m} = X_{T+h,m}$ for $h = 1,\dots,H$, with output columns ordered by $\mathcal{M}$. No future observations are supplied as covariates.

The source set is

$$\mathcal{S} = \bigcup_{m\in\mathcal{M}} \big(\{1,\dots,D\} \setminus \{m\}\big), \qquad S = |\mathcal{S}|. \tag{1}$$

Thus $S = D$ for all-channel forecasting and $S = D - 1$ for a single designated target. Let $\sigma(r)$ map source-bank row $r \in \{1,\dots,S\}$ to its original channel identity, and write $r(c) = \sigma^{-1}(c)$ for a channel $c \in \mathcal{S}$. Every target sees exactly $C = D - 1$ external channels. If $m \in \mathcal{S}$, let $r_m = \sigma^{-1}(m)$ and let $\Pi_m \in \{0,1\}^{C\times S}$ delete row r_m; if $m \notin \mathcal{S}$ (the single-target case), $S = C$ and $\Pi_m = I_C$.

CausalExoFormer uses one forecasting parameter set θ and one structural parameter set $\psi = (a, L, B_{\mathrm{raw}})$ for the dataset. The forecasting backbone produces one endogenous target token $g^{(m)}_{\mathrm{tar}}$ from each target's own history. The structural parameters are shared across targets and forecasting windows: $a \in \mathbb{R}^S$ controls source relevance, $L \in \mathbb{R}^{S\times K}$ allocates relevance over historical offsets, and $B_{\mathrm{raw}} \in \mathbb{R}^{S\times S}$ parameterizes a single directed source graph. Channel embeddings are indexed by original channel identity through σ. The joint output is

$$\widehat{Y}_{:,m} = f_\theta\Big(g^{(m)}_{\mathrm{tar}},\, \{Z^{(m)}_k, G^{(m)}_{:,k}\}_{k=0}^{K-1}; \psi\Big), \qquad m \in \mathcal{M}, \tag{2}$$

where $Z^{(m)}_k$ is an exactly self-excluded view of the shared structured source bank and $G^{(m)} = \Pi_m G$ is the corresponding target-visible gate. The superscript (m) denotes a deterministic view, not a

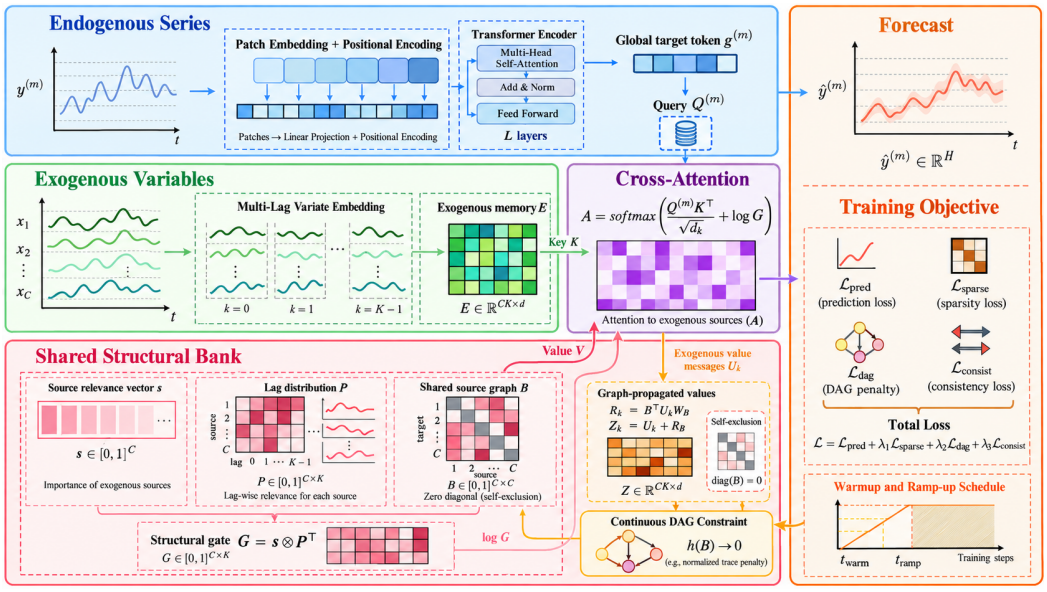

Figure 1: CausalExoFormer architecture. The endogenous branch produces a target query, while lag-aware exogenous embeddings pass through the shared source–lag gate and directed structural bank before cross-attention. The structural terms are activated with the warmup and ramp-up schedule described in Section 4.6.

separately parameterized structure. Targets exchange no predicted future values and are trained jointly with a loss averaged over target channels.

We use B_{ij} for the pathway from source row i to destination row j at the same historical offset. This direction and channel identity are shared across outputs. In the all-channel setting, a channel can act as a historical source for other targets while its own target query removes that source path. In a designated-target task, the target is absent from $\mathcal{S}$ from the outset. The causal structure is parameterized by directed, acyclicity-regularized information pathways. Source availability, historical allocation, and channel propagation have distinct roles in this representation; Proposition 1 establishes how they control the forecast computation.

3 CausalExoFormer

3.1 Explicit Historical-Offset Tokens

Let $\delta_k = k\Delta$, $k = 0, \dots, K-1$, with integer $\Delta \geq 1$ and $\delta_{K-1} < T$. For source-bank row r, define

$$e_r^{(k)} = \mathrm{LN}\big(W_x X_{T-\delta_k,\sigma(r)} + u_{\sigma(r)} + r_k\big), \qquad W_x \in \mathbb{R}^{d\times 1}, \quad u_{\sigma(r)}, r_k \in \mathbb{R}^d. \tag{3}$$

Each token refers to one specific observed time, while $u_{\sigma(r)}$ and r_k encode channel and offset identity. Let $E_k \in \mathbb{R}^{S\times d}$ collect the transposed source tokens as rows. Layer normalization acts over the d features of each token; preprocessing uses fixed training-set statistics, so an external token depends only on its indicated channel and time. Endogenous target branches retain their full look-back windows; the structured source bank uses the specified candidate offsets. For a forecast at $T+h$, the elapsed time from $X_{T-\delta_k,\sigma(r)}$ is $h+\delta_k$, so reported offsets remain tied to the forecast origin.

3.2 Shared Source–Lag Gate and Directed Propagation

For relevance logits $a \in \mathbb{R}^S$ and lag logits $L \in \mathbb{R}^{S\times K}$, set

$$s_r = \mathrm{sigmoid}(a_r), \qquad P_{r,:} = \mathrm{softmax}(L_{r,:}/\tau), \qquad G_{r,k} = s_r P_{r,k}. \tag{4}$$

Thus $\sum_k G_{r,k} = s_r$: source relevance and allocation across historical observations are represented separately. The dominant offset is $\delta_{\arg\max_k P_{r,k}}$. In a designated-target task this relevance is directly

target-specific because the source bank contains only that target's external channels; in all-channel forecasting the target query supplies output-specific routing. We initialize $a_r = 5$ and $L = 0$.

For $S > 1$, the shared source adjacency is

$$B = \mathrm{sigmoid}(B_{\mathrm{raw}} - \log(S-1)) \circ (\mathbf{1}\mathbf{1}^\top - I_S). \tag{5}$$

With $B_{\mathrm{raw}} = 0$, each off-diagonal weight is $1/S$, giving initial weighted out-degree $(S-1)/S$. For $S = 1$, set $B = 0$. Nonnegative weights represent message strengths, with B_{ij} routing source i to destination j; the shared feature transformation and prediction head produce signed responses.

The source gate is applied before any cross-channel operation:

$$U_k = \mathrm{Diag}(G_{:,k})E_k, \qquad R_k = B^\top U_k W_B, \qquad \|W_B\|_F = 1. \tag{6}$$

The graph message R_k is computed once and shared by all target queries. When target m is itself present in the source bank, its direct row and all messages emitted from that row must be excluded from its own external pathway. Because propagation is linear in source rows, this can be done exactly without recomputing a target-specific graph:

$$Z_k^{(m)} = \begin{cases} \Pi_m \left[U_k + R_k - B^\top_{r_m,:} \left(U_{k,r_m:} W_B \right) \right], & m \in \mathcal{S}, \\ U_k + R_k, & m \notin \mathcal{S}. \end{cases} \tag{7}$$

For $m \in \mathcal{S}$, the subtracted rank-one term is exactly the contribution of source m to every destination row, and Π_m then removes the target row itself. Therefore Eq. (7) equals graph propagation on the target-excluded source set, while B remains a single system-level graph. If B is acyclic, every target-visible vertex-deleted subgraph is acyclic as well.

The constraint on $W_B \in \mathbb{R}^{d \times d}$ fixes the scale of the shared message transformation. It is implemented by projection onto the Frobenius unit sphere, initialized at $I_d/\sqrt{d}$; an exactly zero proposed update retains the preceding feasible matrix. Source gating occurs before Eq. (6), so closing a source controls all of its outgoing paths. Token normalization occurs before gating; no row normalization of B or token-wise normalization of the propagated bank removes these amplitude controls.

3.3 Null-Augmented Cross-Attention

For target m, let $G^{(m)} = \Pi_m G$. Flatten $Z_k^{(m)}$ and $G^{(m)}$ in the same order, obtaining $Z^{(m)} \in \mathbb{R}^{CK \times d}$ and $g^{(m)} \in [0,1]^{CK}$. For one head of width d_k, use row-vector queries and projections $W^Q, W^K, W^V \in \mathbb{R}^{d \times d_k}$ with no key/value biases:

$$\begin{aligned} q_m &= g_{\mathrm{tar}}^{(m)} W^Q, \quad K_m = Z^{(m)} W^K, \qquad V_m = Z^{(m)} W^V, \\ z^{(m)} &= \frac{q_m K_m^\top}{\sqrt{d_k}}, \quad \widetilde{z}_i^{(m)} = z_i^{(m)} - \frac{1}{CK} \sum_{j=1}^{CK} z_j^{(m)}. \end{aligned} \tag{8}$$

The fixed null alternative has prior weight $C = D - 1$ and value zero. Attention is

$$A_0^{(m)} = \frac{C}{C + \sum_j g_j^{(m)} e^{\widetilde{z}_j^{(m)}}}, \qquad A_i^{(m)} = \frac{g_i^{(m)} e^{\widetilde{z}_i^{(m)}}}{C + \sum_j g_j^{(m)} e^{\widetilde{z}_j^{(m)}}}, \qquad o_m = \sum_{i=1}^{CK} A_i^{(m)} V_{m,i}. \tag{9}$$

Equivalently, apply softmax to $[\log C, \widetilde{z}^{(m)} + \log g^{(m)}]$. Stable log-sigmoid and log-softmax evaluate $\log g^{(m)}$; an explicitly zero gate uses a masked logit of $-\infty$. With neutral content logits and fully active target-visible relevance, $\sum_i g_i^{(m)} = C$ and the total external attention mass is $1/2$, independent of C and K.

The shared gate controls source availability before propagation and destination routing after exact self-exclusion. Target-specific content logits retain output-specific adaptation even though the structural source gate and graph are shared. Each attention head has its own content projections and shares G and B; head outputs are concatenated and passed through a bias-free output projection W^O. With $o_m^{\mathrm{MH}} = \mathrm{Concat}_{\mathrm{feat}}(o_m^{(1)}, \ldots, o_m^{(n_{\mathrm{heads}})}) W^O$, the fused target representation is $\bar{g}_m = g_{\mathrm{tar}}^{(m)} + o_m^{\mathrm{MH}}$. A shared forecasting head maps this representation, together with the target-local backbone states, to the H predictions. All cross-channel inputs to that head pass through the self-excluded bank.

Proposition 1 (Structural control and exact self-exclusion) *Fix a target m, finite network parameters, and the preprocessing statistics. A gate closure means replacing the indicated effective gate by zero in both its uses, with all other effective gates fixed. (i) For any source channel $c \neq m$, setting its gate $G_{r(c),k} = 0$ removes the computational dependence of target m's external pathway on $X_{T-\delta_k,c}$, including all graph-mediated paths. (ii) Equation* (7) *is exactly equal to first zeroing target m as a source, applying the shared graph, and then deleting its destination row; hence the source bank contributes no target-m value through B; the target-local query remains endogenous. (iii) If all target-visible source relevances are zero, then $o_m = 0$. (iv) With content logits fixed and $Q_m = \sum_i g_i^{(m)} e^{\tilde{z}_i^{(m)}} > 0$, scaling the target-visible gate by $\lambda > 0$ within its domain changes total external attention mass to $\lambda Q_m/(C + \lambda Q_m)$, which is strictly increasing in λ.*

Proofs and the scope of the fixed-content statement are given in Appendix B. They hold for the complete external attention computation and an arbitrary downstream target-local prediction head.

3.4 Structural Objective and Optimization

The source gate and graph are shared across forecast targets, so one acyclicity regularizer acts on the shared directed structure. Define the normalized NOTEARS residual (Zheng et al., 2018)

$$h_S(B) = \frac{\mathrm{Tr}(e^{B \circ B}) - S}{S}, \qquad \mathcal{R}_{\mathrm{AL}} = \alpha h_S(B) + \frac{\rho}{2} h_S(B)^2. \tag{10}$$

The residual has the NOTEARS zero set: $h_S(B) = 0$ exactly for acyclic weighted graphs. Optimization uses the soft adjacency in Eq. (5); exact DAGs lie in its closure. The continuous graph is used for forecasting, while any discrete DAG readout is obtained by retaining thresholded edges in decreasing-weight order only when they preserve acyclicity. A single global readout induces every target-visible subgraph by vertex deletion.

For a minibatch $\mathcal{B}$ of N_b windows, with standardized prediction errors $e_{n,h,m} = \widehat{Y}_{n,h,m} - Y_{n,h,m}$, the objective is

$$\mathcal{J} = \mathbb{E}_{\mathcal{B}}[\mathcal{L}_{\mathrm{pred}}(\mathcal{B})] + w(t)\left[\lambda_1 \mathcal{R}_{\mathrm{sparse}} + \mathcal{R}_{\mathrm{AL}}\right], \tag{11}$$

$$\mathcal{L}_{\mathrm{pred}}(\mathcal{B}) = \frac{1}{N_b H M} \sum_{n=1}^{N_b} \sum_{h=1}^{H} \sum_{m \in \mathcal{M}} e_{n,h,m}^2, \tag{12}$$

$$\mathcal{R}_{\mathrm{sparse}} = \frac{1}{S} \sum_{r=1}^{S} s_r + \frac{\mu}{S} \sum_{i \neq j} B_{ij}, \qquad \mu = 0.5. \tag{13}$$

The regularizer penalizes average source relevance and average weighted out-degree once per shared structure. The Frobenius constraint on W_B excludes the reciprocal rescaling $B \mapsto \lambda B$, $W_B \mapsto W_B/\lambda$ that would preserve graph messages while reducing adjacency penalties. For $M = 1$, $\mathcal{S}$ contains exactly the external channels of the designated target and the formulation reduces to the single-target model.

Structural parameters are frozen during warmup. We then activate their learning rate and ramp the structural objective with

$$w(t) = \min\left\{1, \max\left\{0, \frac{t - t_{\mathrm{warm}}}{t_{\mathrm{ramp}}}\right\}\right\}, \qquad \tau_t = \max\left\{\tau_{\min}, \tau_{\mathrm{init}} r_\tau^{\max(0, t - t_{\mathrm{warm}})}\right\}. \tag{14}$$

We use an operator-split schedule: minibatches update prediction and sparsity terms, followed by one adjacency update on $w(t)\mathcal{R}_{\mathrm{AL}}$ per epoch at learning rate η_D. This evaluates a single $S \times S$ matrix-exponential gradient per active epoch. After the ramp, update $\alpha \leftarrow \alpha + \rho h_S(B)$ and increase ρ only when the shared residual fails the progress test. Early stopping selects the complete forecasting, structural, and temperature checkpoint using validation MSE averaged over all forecast targets. Appendix C specifies the update order.

4 EXPERIMENTS

4.1 SETUP

We evaluate all-channel multivariate forecasting on ETTh1, ETTh2, ETTm1, ETTm2 (Zhou et al., 2021), Weather, and ECL, with $M = D = 7, 7, 7, 7, 21, 321$, respectively. Every channel is predicted and included in the reported errors. These all-channel models learn one shared $S = D$ source gate and graph; each target query uses the exact self-excluded view in Eq. (7), so it attends to $C = D - 1$ external channels. Six lake tasks (Erie1–2 and Huron1–4) evaluate designated-target forecasting with dissolved oxygen as the response and the remaining measured channels as historical covariates ($M = 1$, $S = C = D - 1$). The lake series are derived from publicly available real-time limnological observations released by the U.S. NOAA Great Lakes Environmental Research Laboratory (GLERL); Erie1–2 correspond to distinct Lake Erie station series and Huron1–4 to distinct Lake Huron station series after removal of a duplicated series. Dataset dimensions, sampling intervals, and lake split counts are reported in Appendix D.1.

Data are split chronologically before fitting preprocessing statistics. Channels are standardized using training means and standard deviations. Missing lake observations are forward-filled within each split; leading gaps use the corresponding training median. In validation and testing, forward filling uses preceding observations and the fallback constants are fitted on the training split. We compare TimeXer (Wang et al., 2024), CrossLinear (Zhou et al., 2025), iTransformer (Liu et al., 2024), PatchTST (Nie et al., 2023), DLinear (Zeng et al., 2023), and Autoformer (Wu et al., 2021). The prediction horizons are $H \in \{96, 192, 336, 720\}$. Standard-benchmark baseline entries are published multivariate results with look-back 96. CrossLinear entries follow the many-to-many comparison in Table 3 of the author manuscript of Zhou et al. (2025) (`arXiv:2505.23116v1`), under the same four horizons; its horizon-wise entries are transcribed and averages are recomputed. DAG (Qiu et al., 2026) and GCGNet (Li et al., 2026) are closely related structured TSF-X methods, but their published primary experiments use different target/data protocols: DAG is formulated around designated endogenous target variables and additionally studies future exogenous inputs, whereas GCGNet is evaluated on a different collection of exogenous-forecasting datasets. Their temporal/channel modeling choices are compared in Appendix G; the numerical benchmark table uses the common all-channel protocol. Lake comparisons are application results from this study; CrossLinear lake entries are single-run results from our implementation under the same look-back and preprocessing and are marked with †.

For N test windows, we report

$$\mathrm{MSE} = \frac{\sum_{n,h,m} e^2_{n,h,m}}{NHM}, \qquad \mathrm{MAE} = \frac{\sum_{n,h,m} |e_{n,h,m}|}{NHM}. \tag{15}$$

Errors are computed on the channel-wise standardized scale and include all selected targets. Each horizon is evaluated separately; "Avg" is the arithmetic mean over the four horizons. We do not pool sample-level errors across datasets; cross-dataset summaries use an unweighted arithmetic mean of the dataset-level horizon averages.

The default configuration uses look-back $T = 96$, patch length 16, hidden dimension 256, four heads, two encoder layers, and $K = 14$ offsets with $\Delta = 1$. We use $\lambda_1 = 0.01$, five warmup epochs, five ramp epochs, base learning rate 10^{-4}, structural rate multiplier $r = 10$, and $\eta_D = r\eta_{\mathrm{pred}}$. Training lasts up to 20 epochs with post-ramp patience 5. Temperature starts at 1, decays by 0.95 per post-warmup epoch, and has lower bound 0.1. The hyperparameter configuration is in Appendix C.2.

4.2 FORECASTING PERFORMANCE

Table 1 summarizes the four horizons; the complete results are in Appendix D. Under the all-channel multivariate protocol, CausalExoFormer has the lowest horizon-averaged MAE on all six standard benchmarks. CausalExoFormer also has the lowest horizon-averaged MSE on four of the six benchmarks; CrossLinear is lower on ETTh1 and ETTm1. The six-dataset macro-average MSE/MAE is 0.3054/0.3395 for CausalExoFormer and 0.3087/0.3470 for CrossLinear, computed from the displayed horizon-wise entries. The largest average-MSE gain over TimeXer remains on Weather (8.1% computed before rounding the averages). All six lake datasets also improve over TimeXer, with average-MSE reductions of approximately 1.2–2.6%. CrossLinear lake runs are included in

Table 1: Horizon-averaged forecasting results (MSE / MAE): all-channel prediction on the six standard benchmarks (upper block), and dissolved-oxygen prediction on the six lake tasks (lower block). CrossLinear standard-benchmark entries are the published many-to-many results of Zhou et al. (2025); lake entries marked † are single-run results from our implementation under the same look-back and preprocessing because Zhou et al. (2025) does not report these lake datasets. Results are arithmetic means of the four displayed horizon-wise values; best in bold and second best underlined.

Dataset	CausalExoFormer	CrossLinear	TimeXer	iTransformer	PatchTST	DLinear	Autoformer
ETTh1	0.436 / **0.427**	**0.431** / 0.432	0.437 / 0.437	0.454 / 0.448	0.469 / 0.455	0.456 / 0.452	0.496 / 0.487
ETTh2	**0.363 / 0.391**	0.368 / 0.396	0.368 / 0.396	0.383 / 0.407	0.387 / 0.407	0.569 / 0.515	0.450 / 0.459
ETTm1	0.378 / **0.385**	**0.371** / 0.393	0.382 / 0.397	0.410 / 0.410	0.387 / 0.400	0.403 / 0.407	0.588 / 0.517
ETTm2	**0.267 / 0.313**	0.272 / 0.321	0.274 / 0.322	0.288 / 0.332	0.281 / 0.326	0.350 / 0.401	0.327 / 0.371
Weather	**0.221 / 0.259**	0.238 / 0.269	0.241 / 0.271	0.258 / 0.278	0.259 / 0.281	0.265 / 0.317	0.338 / 0.382
ECL	**0.168 / 0.264**	0.173 / 0.270	0.171 / 0.270	0.178 / 0.270	0.216 / 0.304	0.212 / 0.300	0.227 / 0.338
Erie1	**1.165 / 0.634**	1.172 / 0.639†	1.179 / 0.645	1.198 / 0.652	1.220 / 0.661	1.258 / 0.675	1.313 / 0.700
Erie2	**0.684 / 0.622**	0.692 / 0.628†	0.701 / 0.634	0.712 / 0.641	0.726 / 0.650	0.748 / 0.663	0.781 / 0.687
Huron1	**0.501 / 0.532**	0.505 / 0.534†	0.511 / 0.537	0.519 / 0.543	0.529 / 0.551	0.545 / 0.563	0.569 / 0.582
Huron2	**0.405 / 0.450**	0.410 / 0.454†	0.416 / 0.459	0.422 / 0.464	0.431 / 0.470	0.444 / 0.481	0.463 / 0.498
Huron3	**1.007 / 0.752**	1.016 / 0.758†	1.028 / 0.767	1.043 / 0.776	1.063 / 0.786	1.096 / 0.803	1.144 / 0.832
Huron4	**2.266 / 0.918**	2.282 / 0.923†	2.300 / 0.929	2.336 / 0.939	2.380 / 0.953	2.454 / 0.973	2.561 / 1.008

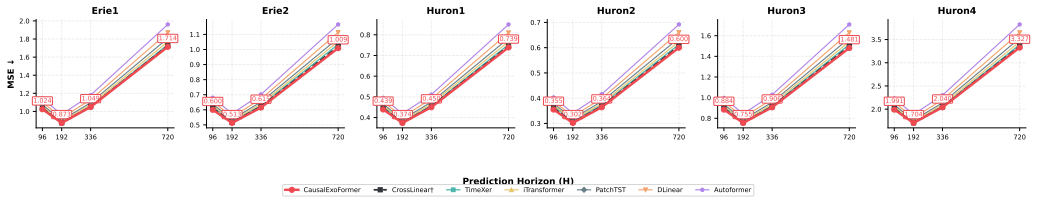

Figure 2: Horizon-wise forecasting MSE on the six lake dissolved-oxygen tasks. The figure visualizes Appendix Table 8, including the single-run CrossLinear† reruns. Lower is better.

Table 1 and Appendix Table 8, marked with †. Figure 2 shows the horizon-wise ordering of the lake methods. This pattern motivates examining which structural components produce the gains.

4.3 Component Ablations

Table 2 compares four variants at $H = 96$. **Uniform Gate** fixes $s_c = 1$ and $P_{c,k} = 1/K$ in both source gating and attention. **Current-Only** uses $K = 1$, restricting external observations to the forecast origin while retaining the full endogenous history. **No DAG Penalty** retains graph propagation and removes its acyclicity penalty. **No Exo Graph** sets $B = 0$ and retains gated external tokens. Each modification is applied to the shared structural bank; all target queries and all-target metric aggregation are retained. These tests compare the complete allocation interface, its historical coverage, graph propagation, and acyclicity regularization, respectively.

Table 2: All-channel ablation study (MSE / MAE) on standard benchmarks, $H = 96$. **Bold**: best; underline: second best.

Model Variant	ETTh1		ETTh2		ETTm1		ETTm2		Weather		ECL	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
CausalExoFormer (Full)	**0.373**	**0.389**	**0.281**	**0.333**	**0.314**	**0.351**	**0.160**	**0.246**	**0.146**	**0.193**	0.137	0.235
Uniform Gate	0.388	0.402	0.294	0.341	0.329	0.360	0.171	0.254	0.157	0.201	0.147	0.243
Current-Only	0.390	0.404	0.296	0.343	0.331	0.362	0.173	0.257	0.159	0.204	0.149	0.245
No DAG Penalty	0.377	0.392	0.284	0.336	0.318	0.354	0.162	0.248	0.148	0.196	0.140	0.237
No Exo Graph	0.379	0.394	0.286	0.338	0.320	0.356	0.163	0.249	0.149	0.197	**0.136**	**0.234**

Uniform gating raises MSE by 4.0–7.5% across the six benchmarks, averaging 5.9%. Current-Only raises it by 4.6–8.9%, averaging 6.8%, supporting the full historical-offset interface over its current-observation restriction. Removing the shared source graph increases MSE on five of six benchmarks, with a 1.4% mean relative increase across all six. On ECL, the no-graph variant is slightly lower at $0.136/0.234$ versus $0.137/0.235$ for the full model. Removing the acyclicity penalty

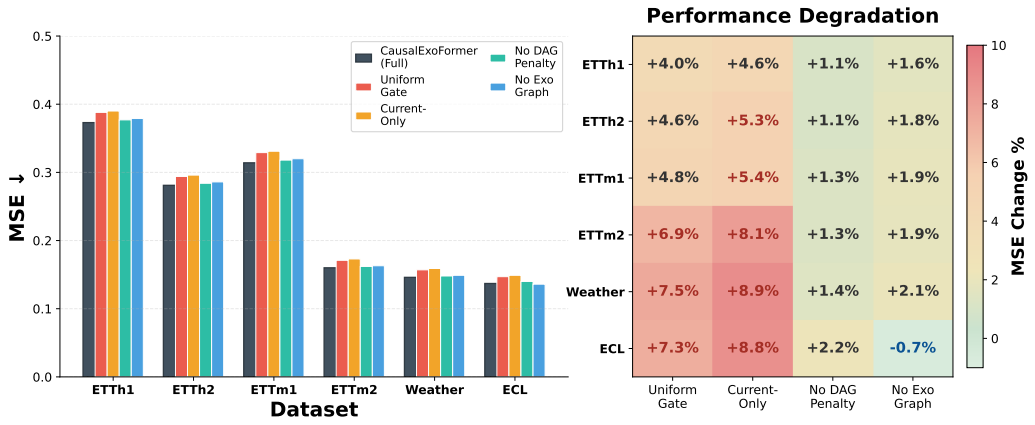

Figure 3: Ablation visualization at $H = 96$. The left panel reports the benchmark MSE values from Table 2; the right heatmap reports the relative MSE change with respect to the full model. Positive values indicate degradation, while the negative ECL cell for No Exo Graph corresponds to the small improvement already reported in Table 2.

Table 3: Synthetic target-source and lag recovery over 20 random evaluation SCMs (mean $\pm$ std). Source-F1 evaluates recovery of the four injected target-driving sources; Lag-Acc and Lag-MAE evaluate the recovered delay among correctly identified sources.

Method	Source-F1↑	Lag-Acc.↑	Lag-MAE↓
DYNOTEARS	0.78 ± 0.06	0.70 ± 0.07	0.52 ± 0.12
PCMCI+	0.82 ± 0.05	0.74 ± 0.06	0.43 ± 0.10
CausalExoFormer	$\mathbf{0.89 \pm 0.04}$	$\mathbf{0.83 \pm 0.05}$	$\mathbf{0.27 \pm 0.07}$

raises MSE by 1.4% on average. Figure 3 visualizes these relative changes. The lake ablations in Appendix D.4 provide a complementary application-level comparison. The largest contributions come from external temporal information and learned allocation, with graph propagation supplying an additional structural pathway.

4.4 Synthetic Target-Source and Lag Recovery

To evaluate whether the learned source–lag gate recovers target-relevant historical inputs, we additionally use controlled dynamic structural-causal models (SCMs) with known lagged source-to-target relations. Each system contains one designated target and $C = S = 11$ observed source variables. The sources interact through a random contemporaneous DAG, while four source variables are selected to affect the target at ground-truth delays in $\{1, \ldots, 5\}$. The designated-target task evaluates recovery of these injected sources and their delays, while the contemporaneous graph supplies interacting source dynamics. Each evaluation SCM contributes 5,000 observations after a 500-step burn-in and uses a chronological 70/10/20 train/validation/test split. Results are aggregated over 20 independently generated evaluation SCMs; five additional development SCMs are used only to select method-specific source-recovery thresholds. Appendix A gives the generator equations, recovery protocol, and reporting rules.

4.5 Interpreting Variable and Temporal Selection

The factorization $G_{r,k} = s_r P_{r,k}$ provides separate readouts of source relevance and historical allocation. On the lake tasks ($M = 1$), the source bank contains only the dissolved-oxygen target's external channels, so these scores are directly interpretable as target-specific external relevance. Table 4 shows the highest-relevance external variables for the Lake-Erie2 dissolved-oxygen case study. The selected panel includes water temperature, phycocyanin, turbidity, and pH. These measurements

Table 4: Highest-relevance external variables in the Lake-Erie2 dissolved-oxygen case study. Offset times are relative to the forecast origin on the 15-minute sampling grid.

Variable	Relevance s_r	Offset index k	Historical offset
water_temp	0.93	4	60 minutes
phycocyanin	0.89	4	60 minutes
turbidity	0.84	5	75 minutes
pH	0.71	2	30 minutes

describe thermal conditions, biological activity, suspended material, and water chemistry, making the learned selection directly inspectable in environmental terms.

The dominant offsets differ across variables rather than collapsing to a common temporal position. Their meaning is tied to the forecast origin by Eq. (3), and their pathways remain controlled after propagation by Proposition 1. This combination links the displayed source–lag parameters to actual information access in the model.

4.6 Sensitivity and Computational Cost

The lake sweeps characterize the common configuration $\lambda_1 = 0.01$, $K = 14$, and five warmup epochs (Appendix E). Fourteen offsets give the lowest MSE on five datasets and are within 0.002 of the best on Erie2. Warmup improves forecast error compared with immediate structural regularization. The Lake-Erie2 measurements report 1.02M parameters, 10.3 seconds per training epoch, and 6.8 ms inference per batch, compared with 0.88M, 9.1 seconds, and 6.1 ms for TimeXer. Appendix E.1 gives the measurements and the separate costs of propagation and acyclicity.

5 Conclusion

We introduced **CausalExoFormer**, which learns source–lag selection and a directed cross-channel graph as components of a Transformer forecasting computation. A single system-level graph is shared across forecast targets, while exact self-exclusion preserves the target-relative external input bank without duplicating acyclicity constraints. Source gating controls direct and graph-mediated information, a null attention alternative anchors the external branch, and a fixed-scale message transformation removes reciprocal graph/message rescaling. Across twelve datasets, the results combine competitive forecasting with explicit structural readouts. The central methodological result is an end-to-end interface in which historical source selection, temporal allocation, and directed interactions govern the pathways that produce the forecast.

Reproducibility statement. The formulation, optimization algorithm, hyperparameter configuration, full horizon-wise result tables, sensitivity sweeps, and synthetic target-source and lag-recovery protocol are provided in the paper and appendices. Appendix B proves the stated computational properties.

REFERENCES

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. DAGMA: Learning DAGs via M-matrices and a log-determinant acyclicity characterization. In *Advances in Neural Information Processing Systems*, volume 35, pages 8226–8239, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/36e2967f87c3362e37cf988781a887ad-Abstract-Conference.html.

Cristian Challu, Kin G. Olivares, Boris N. Oreshkin, Federico Garza Ramirez, Max Mergenthaler Canseco, and Artur Dubrawski. NHITS: Neural hierarchical interpolation for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6989–6997, 2023. URL https://ojs.aaai.org/index.php/AAAI/article/view/25854.

Abhimanyu Das, Weihao Kong, Andrew Leach, Shaan Mathur, Rajat Sen, and Rose Yu. Long-term forecasting with TiDE: Time-series dense encoder. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=pCbC3aQB5W.

Clive W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969. URL https://doi.org/10.2307/1912791.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019. URL https://aclanthology.org/N19-1357/.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*, 2017. URL https://research.google/pubs/categorical-reparameterization-with-gumbel-softmax/.

Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *Proceedings of the 35th International Conference on Machine Learning*, 2018. URL https://proceedings.mlr.press/v80/kipf18a.html.

Zhengyu Li, Xiangfei Qiu, Yuhan Zhu, Xingjian Wu, Jilin Hu, Chenjuan Guo, and Bin Yang. GCGNet: Graph-consistent generative network for time series forecasting with exogenous variables. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2f27964513a28d034530bfdd117ea31d-Abstract-Conference.html.

Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4): 1748–1764, 2021. URL https://doi.org/10.1016/j.ijforecast.2021.03.012.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=JePfAI8fah.

Christos Louizos, Max Welling, and Diederik P. Kingma. Learning sparse neural networks through L_0 regularization. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1712.01312.

Sindy Löwe, David Madras, Richard Zemel, and Max Welling. Amortized causal discovery: Learning to infer causal graphs from time-series data. In *Proceedings of the First Conference on Causal Learning and Reasoning (CLeaR)*, volume 177, pages 509–525. PMLR, 2022. URL https://proceedings.mlr.press/v177/lowe22a.html.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1611.00712.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2211.14730.

Kin G. Olivares, Cristian Challu, Grzegorz Marcjasz, Rafał Weron, and Artur Dubrawski. Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with NBEATSx. *International Journal of Forecasting*, 39(2):884–900, 2023. URL https://doi.org/10.1016/j.ijforecast.2022.03.001.

Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020. URL https://www.servicenow.com/research/publication/boris-n.-oreshkin-n-be-iclr2020.html.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. DYNOTEARS: Structure learning from time-series data. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108, pages 1595–1605, 2020. URL https://proceedings.mlr.press/v108/pamfil20a.html.

Xiangfei Qiu, Yuhan Zhu, Zhengyu Li, Xingjian Wu, Bin Yang, and Jilin Hu. DAG: A dual correlation network for time series forecasting with exogenous variables. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2509.14933.

Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019. URL https://www.nature.com/articles/s41586-019-0912-1.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 1388–1397, 2020. URL https://proceedings.mlr.press/v124/runge20a.html.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 2019. URL https://doi.org/10.1126/sciadv.aau4996.

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, 2022. URL https://doi.org/10.1109/TPAMI.2021.3065601.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. TimeXer: Empowering transformers for time series forecasting with exogenous variables. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0113ef4642264adc2e6924a3cbbdf532-Abstract-Conference.html.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019. URL https://aclanthology.org/D19-1002/.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pages 22419–22430, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/bcc0d400288793e8bdcd7c19a8ac0c2b-Abstract.html.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023. URL https://github.com/thuml/TimesNet.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph WaveNet for deep spatial-temporal graph modeling. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019. URL https://www.ijcai.org/proceedings/2019/264.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2020. URL https://arxiv.org/abs/2005.11650.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11121–11128, 2023. URL https://ojs.aaai.org/index.php/AAAI/article/view/26317.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations*, 2023. URL https://github.com/Thinklab-SJTU/Crossformer.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/e347c51419ffb23ca3fd5050202f9c3d-Abstract.html.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric DAGs. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 2020. URL https://proceedings.mlr.press/v108/zheng20a.html.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11106–11115, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17325.

Pengfei Zhou, Yunlong Liu, Junli Liang, Qi Song, and Xiangyang Li. CrossLinear: Plug-and-play cross-correlation embedding for time series forecasting with exogenous variables. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, 2025. doi: 10.1145/3711896.3736899. URL https://arxiv.org/abs/2505.23116v1.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*, pages 27268–27286, 2022. URL https://proceedings.mlr.press/v162/zhou22g.html.

A Synthetic Target-Source and Lag-Recovery Protocol

The synthetic benchmark evaluates the source–lag allocation G in a controlled setting with interacting source variables. Let $x_t \in \mathbb{R}^C$ denote the $C = 11$ observed source variables and y_t the designated target. For each SCM, we sample a topological order and independently include each admissible forward edge with probability $p = 2/(C-1)$, yielding C expected directed edges, equivalently an expected mean total degree of two. The resulting signed contemporaneous coefficient matrix $B_\star$ follows this DAG, with nonzero absolute coefficients sampled uniformly from $[0.4, 0.8]$. Source dynamics follow

$$x_t = B_\star^\top x_t + A_\star x_{t-1} + \varepsilon_t, \qquad x_t = (I - B_\star^\top)^{-1}(A_\star x_{t-1} + \varepsilon_t), \tag{16}$$

where $A_\star = \mathrm{Diag}(a_1, \ldots, a_C)$ with $a_c \sim \mathrm{Unif}(0.1, 0.4)$ and $\varepsilon_t \sim \mathcal{N}(0, I_C)$. In topological order, $(I - B_\star^\top)^{-1} A_\star$ is triangular with diagonal entries a_c; its spectral radius is therefore at most 0.4, ensuring stable source dynamics.

Four distinct source variables form the target-driver set $\mathcal{D}_\star$. Each driver receives a ground-truth lag $\ell_c \sim \mathrm{Unif}\{1, 2, 3, 4, 5\}$ and a signed coefficient with absolute value sampled from $[0.5, 1.0]$. The target obeys

$$y_t = \phi y_{t-1} + \sum_{c \in \mathcal{D}_\star} \beta_c x_{c,t-\ell_c} + \varepsilon_t^y, \qquad \phi \sim \mathrm{Unif}(0.1, 0.4), \quad \varepsilon_t^y \sim \mathcal{N}(0, 1). \tag{17}$$

There is no target-to-source edge. All source variables are observed, innovations are mutually independent, and the contemporaneous source graph is acyclic by construction. We discard 500 burn-in steps and retain 5,000 observations, split chronologically into 70%/10%/20% train/validation/test segments. Standardization statistics are fitted on the training segment only.

The synthetic task uses one-step forecasting ($H = 1$) to isolate target-source and lag recovery. CausalExoFormer uses $K = 6$ and $\Delta = 1$, giving candidate offsets $\{0, 1, 2, 3, 4, 5\}$ relative to the forecast origin. For a one-step forecast of y_{T+1}, a true source lag ℓ_c uses $x_{c,T+1-\ell_c}$, hence its matching offset is $k_c^\star = \ell_c - 1$. Offsets $0, \ldots, 4$ cover the injected lags $1, \ldots, 5$; offset 5 is the additional candidate. DYNOTEARS and PCMCI+ are evaluated on their lagged source-to-target scores under the same observed system. Method-specific source-score thresholds (and the PCMCI+ significance level) are selected once on five independently generated development SCMs and then frozen for the 20 evaluation SCMs. No evaluation-SCM ground truth is used for threshold selection.

For target-source recovery, CausalExoFormer predicts source r as relevant when s_r exceeds the frozen development threshold; DYNOTEARS and PCMCI+ use their corresponding lagged source-to-target scores. Source-F1 compares the selected sources with the four injected drivers in $\mathcal{D}_\star$, matching the role of s_r as a source-availability score in the forecasting pathway. For a correctly recovered source, CausalExoFormer's predicted delay is $\widehat{\ell}_c = 1 + \arg\max_{k \in \{0,\ldots,5\}} P_{r(c),k}$. Lag-Acc compares $\widehat{\ell}_c$ with ℓ_c and Lag-MAE averages $|\widehat{\ell}_c - \ell_c|$, in sampling steps. Equivalently, both can be evaluated in offset coordinates against $\ell_c - 1$; the same coordinates are used for predictions and ground truth. Reported values are means and standard deviations over the 20 evaluation SCMs, with one metric value per evaluation SCM.

B Computational Properties

Exact self-exclusion. Assume $m \in \mathcal{S}$ and let r_m be its source-bank row. The jth row of $R_k = B^\top U_k W_B$ is $\sum_i B_{ij} U_{k,i:} W_B$. The contribution of source m is exactly $B_{r_m,j} U_{k,r_m:} W_B$, which is the jth row of $B_{r_m,:}^\top (U_{k,r_m:} W_B)$. Subtracting this rank-one matrix therefore gives the same messages as first setting row r_m of U_k to zero. Applying Π_m then removes the target's direct row and its destination row. Hence Eq. (7) is algebraically identical to propagation on the target-excluded source set, while the expensive shared product $B^\top U_k W_B$ is computed only once.

Source-path exclusion. Fix target m and a different source channel $c \neq m$. The observation $X_{T-\delta_k,c}$ enters the structured source bank only through its channel-local token. If $G_{r(c),k} = 0$, the corresponding row of U_k is independent of that observation. The shared message R_k and the self-exclusion correction are both linear functions of rows of U_k; therefore every direct and graph-mediated contribution of that source to $Z_k^{(m)}$ vanishes. Subsequent keys, values, and external attention

output for target m are independent of the closed observation. The source channel's own endogenous target branch, when it is also forecast, is intentionally unaffected.

Endogenous fallback. If all target-visible source relevance scores are zero, then all rows visible to target m vanish after source gating and self-exclusion, so $Z^{(m)} = 0$ and $V_m = 0$. Equation (9) gives $A_0^{(m)} = 1$ and $o_m = 0$. The bias-free multi-head output projection also maps zero to zero, leaving the endogenous residual pathway intact.

Absolute gate scale in the attention rule. With target-m content logits held fixed, let $Q_m = \sum_i g_i^{(m)} e^{\widetilde{z}_i^{(m)}}$. Under a positive common scaling $g^{(m)} \mapsto \lambda g^{(m)}$ that remains in the gate domain, the external attention mass is

$$M_m(\lambda) = \frac{\lambda Q_m}{C + \lambda Q_m}, \qquad M'_m(\lambda) = \frac{C Q_m}{(C + \lambda Q_m)^2} > 0 \quad (Q_m > 0). \tag{18}$$

In the complete network, source gating also changes keys and content logits, so this derivative is a property of the attention rule at fixed content rather than a global monotonicity statement for the predictor. At neutral logits and fully active target-visible relevance, $\sum_i g_i^{(m)} = C$, giving external mass $1/2$.

Scale of graph messages. Both U_k and $R_k = B^\top U_k W_B$ have S source rows before target-specific selection. With $\|W_B\|_F = 1$, the reparameterization $B' = \lambda B$, $W'_B = W_B/\lambda$ is feasible only for $\lambda = 1$ when $\lambda > 0$. The constraint removes precisely this reciprocal scalar symmetry. For fixed U_k, W_B, setting $B = 0$ removes graph messages exactly.

Acyclicity and target views. NOTEARS establishes $\mathrm{Tr}(e^{B \circ B}) - S = 0$ if and only if the weighted adjacency is acyclic (Zheng et al., 2018). Division by positive S preserves the zero set. A finite residual is an optimization quantity, not a dimension-independent distance to a DAG. Every vertex-deleted target view of a DAG is also a DAG, so one global constraint is sufficient for all self-excluded views. For discrete visualization, eligible edges are thresholded, sorted by decreasing weight, and retained only if they preserve acyclicity. This reporting operation remains distinct from the soft adjacency used for forecasting.

Equivalent vertex-deleted operator. Let $D_m = \Pi_m^\top \Pi_m$, $\widetilde{U}_k = \Pi_m U_k$, and $\widetilde{B}_m = \Pi_m B \Pi_m^\top$. For $m \in \mathcal{S}$, Eq. (7) can be written as

$$Z_k^{(m)} = \Pi_m D_m U_k + \Pi_m B^\top D_m U_k W_B = \widetilde{U}_k + \widetilde{B}_m^\top \widetilde{U}_k W_B, \tag{19}$$

This identity establishes self-exclusion for the single propagation operator in Eq. (6).

Source and destination roles of the gate. Before propagation, $G_{r,k}$ controls whether an observation can emit a message. In Eq. (9), $G_{r,k}^{(m)}$ controls routing through the corresponding destination token, which can contain messages from other open sources. These are separate operations using shared structural parameters. In particular, the neutral-logit mass of $1/2$ normalizes the attention prior; value amplitudes still depend on the source gate. The Uniform Gate and Current-Only experiments compare these complete interfaces.

C Optimization Details

Let $\psi = (a, L, B_{\mathrm{raw}})$ denote the shared structural parameters and θ the forecasting parameters. A training batch contains N_b time windows and all M target losses. The source tokens, gate, and graph messages are constructed once per minibatch; each target then applies Eq. (7) and its own content query. Each forecast horizon uses its own fitted model.

Algorithm 1: Joint training with a shared structural graph.

1. Initialize $a = 5$, $L = 0$, $B_{\mathrm{raw}} = 0$, $\alpha = 0$, $\rho = 1$, $h_{\mathrm{prev}} = +\infty$, and $W_B = I_d/\sqrt{d}$.

2. For each epoch t, compute $w(t)$ and τ_t by Eq. (14). Freeze ψ for $t \le t_{\text{warm}}$; otherwise use structural learning rate $r\eta_{\text{pred}}$.
3. For each minibatch, construct E_k, G, U_k, R_k once. Form each target's self-excluded bank $Z_k^{(m)}$, compute the averaged loss in Eq. (12), and update trainable parameters on $\mathcal{L}_{\text{pred}} + w(t)\lambda_1\mathcal{R}_{\text{sparse}}$. Project W_B onto the Frobenius unit sphere after the update.
4. If $w(t) > 0$, update only B_{raw} on $w(t)\mathcal{R}_{\text{AL}}$ at learning rate $\eta_D = r\eta_{\text{pred}}$.
5. After the ramp, recompute $h_t = h_S(B)$ from the updated graph. Set $\alpha \leftarrow \alpha + \rho h_t$. If $h_t > \kappa h_{\text{prev}}$, set $\rho \leftarrow \min(\gamma\rho, \rho_{\max})$. Then store $h_{\text{prev}} \leftarrow h_t$.
6. Evaluate all-target validation MSE after each post-ramp epoch and save the best complete (θ, ψ, τ_t) checkpoint. Stop at the epoch limit or when patience is exhausted; restore the saved temperature and parameters for inference.

The matrix exponential and residual subtraction are evaluated in float64; gradients return in the parameter dtype. Each active epoch uses one differentiable acyclicity update; post-ramp epochs additionally evaluate the updated residual without gradients for multiplier control. With 20 epochs and five warmup epochs, the temperature reaches $0.95^{15} \simeq 0.463$; 0.1 is a lower bound. Dominant-offset ties use the smallest offset index. For $S = 1$, the graph is identically zero and its acyclicity update is omitted.

C.1 Target-local backbone and output interface

For each target m, the endogenous backbone encodes only $X_{1:T,m}$. Nonoverlapping length-16 patches represent the look-back window; a target global token supplies $g_{\text{tar}}^{(m)}$ for the external query, following the endogenous/global-token interface of TimeXer (Wang et al., 2024). With $T = 96$, there are six patches. Denote the target-local patch states by H_m and the common prediction head by $\mathcal{H}_\theta$. The output interface is

$$\widehat{Y}_{:,m} = \mathcal{H}_\theta(H_m, \bar{g}_m), \qquad \bar{g}_m = g_{\text{tar}}^{(m)} + o_m^{\text{MH}}. \tag{20}$$

The forecasting parameters are shared over m; channel identity enters through the channel embeddings. Layer normalization in the source bank is token-local, key/value projections are linear, and the external output projection has no bias. Consequently, zero external values give $o_m^{\text{MH}} = 0$ and preserve the endogenous prediction interface exactly. Each horizon uses an independently fitted head and model checkpoint. The structural exclusion proof depends on this information flow, not on the internal parameterization of the target-local prediction head.

C.2 Hyperparameters

D Full Experimental Tables

D.1 Dataset Statistics

Dataset statistics are listed in Table 6. Variable counts give D, the total observed channels, and every target attends to $C = D - 1$ external channels. Standard all-channel benchmarks use $M = D$ and a shared source bank of size $S = D$; lake tasks use $M = 1$ and $S = C = D - 1$. The sampling interval determines the physical time associated with Δ and K. The ETT series follow the standard benchmark construction of Zhou et al. (2021); Weather and ECL use the standard processed long-horizon forecasting files. The lake series originate from the NOAA GLERL public data portal (`https://www.glerl.noaa.gov/data/`). Lake rows give chronological sequence counts. For the standard benchmarks, we report dimensions and sampling intervals and use the published benchmark protocol identified in Section 4.1; dataset-length and window-count conventions are not mixed in this table.

D.2 Standard Benchmark Results

Table 5: Model and structural-optimization hyperparameters.

Hyperparameter	Symbol	Value
Look-back window	T	96
Patch length	P	16
Model dimension	d	256
Attention heads	n_{heads}	4
Encoder layers	N_{enc}	2
Number of historical offsets	K	14
Lag step size	Δ	1
Sparsity weight	λ_1	0.01
Adjacency logit offset	$b_0(S)$	$\log(S-1)$
Initial edge weight	$B_{ij}^{(0)}$	$1/S$
Initial weighted degree	–	$(S-1)/S$
Initial DAG penalty	ρ_{init}	1.0
DAG growth factor	γ	10
AL progress ratio	κ	0.25
Maximum ρ	$\rho_{\max}$	10^4
Softmax temperature init	τ_{init}	1.0
Softmax temperature min	$\tau_{\min}$	0.1
Softmax temperature anneal rate	r_τ	0.95
Warmup epochs	t_{warm}	5
Ramp-up epochs	t_{ramp}	5
Structural LR multiplier	r	10
Causal learning rate	η_{causal}	10^{-3}
Base learning rate	η_{pred}	10^{-4}
Total epochs	$T_{\max}$	20
Batch size	–	32
Graph projection norm	$\|W_B\|_F$	1
Acyclicity-step learning rate	η_D	$r\eta_{\text{pred}}$

Table 6: Dataset dimensions and sampling intervals; chronological split sizes are shown for the lake tasks.

Dataset	Variables	Frequency	Train / Val / Test	Total	Forecast targets
Erie1	10	Hourly	10,950 / 1,566 / 3,128	15,644	Dissolved oxygen
Erie2	14	15 min	131,969 / 18,854 / 37,705	188,528	Dissolved oxygen
Huron1	15	20 min	22,191 / 3,171 / 6,340	31,702	Dissolved oxygen
Huron2	15	20 min	18,243 / 2,607 / 5,212	26,062	Dissolved oxygen
Huron3	15	20 min	7,148 / 1,022 / 2,042	10,212	Dissolved oxygen
Huron4	15	20 min	7,219 / 1,032 / 2,062	10,313	Dissolved oxygen
ETTh1	7	Hourly	Benchmark protocol	–	All channels
ETTh2	7	Hourly	Benchmark protocol	–	All channels
ETTm1	7	15 min	Benchmark protocol	–	All channels
ETTm2	7	15 min	Benchmark protocol	–	All channels
Weather	21	10 min	Benchmark protocol	–	All channels
ECL	321	Hourly	Benchmark protocol	–	All channels

Table 7: All-channel multivariate forecasting results (MSE / MAE) on standard benchmarks. CrossLinear values are transcribed from the published many-to-many multivariate comparison of Zhou et al. (2025); the other baseline values are the published multivariate results used in our comparison. All channels enter each metric, and Avg rows are arithmetic means of the four displayed horizons. **Bold**: best; underline: second best.

Methods		**CausalExoFormer**		**CrossLinear**		**TimeXer**		**iTransformer**		**PatchTST**		**DLinear**		**Autoformer**	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	**0.373**	**0.389**	0.374	0.393	0.382	0.403	0.386	0.405	0.414	0.419	0.386	0.400	0.449	0.459
	192	0.434	**0.421**	**0.422**	0.424	0.429	0.435	0.441	0.436	0.460	0.445	0.437	0.432	0.500	0.482
	336	0.470	**0.438**	**0.459**	0.447	0.468	0.448	0.487	0.458	0.501	0.466	0.481	0.459	0.521	0.496
	720	**0.467**	**0.458**	**0.467**	0.465	0.469	0.461	0.503	0.491	0.500	0.488	0.519	0.516	0.514	0.512
	Avg	0.436	**0.427**	**0.431**	0.432	0.437	0.437	0.454	0.448	0.469	0.455	0.456	0.452	0.496	0.487
ETTh2	96	**0.281**	**0.333**	0.282	0.337	0.286	0.338	0.297	0.349	0.302	0.348	0.374	0.387	0.346	0.388
	192	**0.358**	**0.384**	0.360	0.387	0.363	0.389	0.380	0.400	0.388	0.400	0.477	0.476	0.456	0.452
	336	0.409	**0.418**	**0.405**	0.422	0.414	0.423	0.428	0.432	0.426	0.433	0.594	0.541	0.482	0.486
	720	**0.403**	**0.427**	0.424	0.439	0.408	0.432	0.427	0.445	0.431	0.446	0.831	0.657	0.515	0.511
	Avg	**0.363**	**0.391**	0.368	0.396	0.368	0.396	0.383	0.407	0.387	0.407	0.569	0.515	0.450	0.459
ETTm1	96	0.314	**0.351**	**0.311**	0.354	0.318	0.356	0.334	0.368	0.329	0.367	0.345	0.372	0.505	0.475
	192	0.359	**0.374**	**0.352**	0.379	0.362	0.383	0.387	0.391	0.367	0.385	0.380	0.389	0.553	0.496
	336	0.392	**0.391**	**0.381**	0.401	0.395	0.407	0.426	0.420	0.399	0.410	0.413	0.413	0.621	0.537
	720	0.447	**0.423**	**0.439**	0.439	0.452	0.441	0.491	0.459	0.454	0.439	0.474	0.453	0.671	0.561
	Avg	0.378	**0.385**	**0.371**	0.393	0.382	0.397	0.410	0.410	0.387	0.400	0.403	0.407	0.588	0.517
ETTm2	96	**0.160**	**0.246**	0.170	0.254	0.171	0.256	0.180	0.264	0.175	0.259	0.193	0.292	0.255	0.339
	192	**0.230**	**0.291**	0.236	0.298	0.237	0.299	0.250	0.309	0.241	0.302	0.284	0.362	0.281	0.340
	336	**0.289**	**0.329**	0.294	0.336	0.296	0.338	0.311	0.348	0.305	0.343	0.369	0.427	0.339	0.372
	720	**0.387**	**0.386**	0.388	0.394	0.392	0.394	0.412	0.407	0.402	0.400	0.554	0.522	0.433	0.432
	Avg	**0.267**	**0.313**	0.272	0.321	0.274	0.322	0.288	0.332	0.281	0.326	0.350	0.401	0.327	0.371
Weather	96	**0.146**	**0.193**	0.154	0.202	0.157	0.205	0.174	0.214	0.177	0.218	0.196	0.255	0.266	0.336
	192	0.205	**0.244**	**0.200**	0.246	0.204	0.247	0.221	0.254	0.225	0.259	0.237	0.296	0.307	0.367
	336	0.260	**0.283**	**0.257**	0.286	0.261	0.290	0.278	0.296	0.278	0.297	0.283	0.335	0.359	0.395
	720	**0.273**	**0.316**	0.340	0.343	0.340	0.341	0.358	0.349	0.354	0.348	0.345	0.381	0.419	0.428
	Avg	**0.221**	**0.259**	0.238	0.269	0.241	0.271	0.258	0.278	0.259	0.281	0.265	0.317	0.338	0.382
ECL	96	**0.137**	**0.235**	0.139	0.237	0.140	0.242	0.148	0.240	0.195	0.285	0.197	0.282	0.201	0.317
	192	**0.154**	**0.251**	0.157	0.254	0.157	0.256	0.162	0.253	0.199	0.289	0.196	0.285	0.222	0.334
	336	**0.173**	**0.266**	0.176	0.275	0.176	0.275	0.178	0.269	0.215	0.305	0.209	0.301	0.231	0.338
	720	**0.208**	**0.302**	0.221	0.315	0.211	0.306	0.225	0.317	0.256	0.337	0.245	0.333	0.254	0.361
	Avg	**0.168**	**0.264**	0.173	0.270	0.171	0.270	0.178	0.270	0.216	0.304	0.212	0.300	0.227	0.338

D.3 Lake Forecasting Results

Table 8: Dissolved-oxygen forecasting results (MSE / MAE) on lake datasets. CrossLinear† values are single-run results from our implementation under the same look-back and preprocessing; Zhou et al. (2025) does not report these lake datasets. Avg rows are arithmetic means of the four displayed horizons. **Bold**: best; underline: second best.

Methods		CausalExoFormer		CrossLinear†		TimeXer		iTransformer		PatchTST		DLinear		Autoformer	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Erie1	96	**1.024**	**0.468**	1.035	0.476	1.044	0.483	1.057	0.487	1.073	0.493	1.101	0.501	1.143	0.517
	192	**0.873**	**0.546**	0.874	0.550	0.875	0.554	0.888	0.559	0.903	0.566	0.930	0.577	0.967	0.597
	336	**1.049**	**0.597**	1.053	0.602	1.058	0.607	1.075	0.615	1.096	0.623	1.130	0.637	1.180	0.660
	720	**1.714**	**0.924**	1.725	0.929	1.738	0.935	1.770	0.948	1.808	0.963	1.871	0.986	1.961	1.025
	Avg	**1.165**	**0.634**	1.172	0.639	1.179	0.645	1.198	0.652	1.220	0.661	1.258	0.675	1.313	0.700
Erie2	96	**0.600**	**0.459**	0.610	0.466	0.621	0.475	0.628	0.478	0.638	0.484	0.655	0.493	0.680	0.508
	192	**0.513**	**0.536**	0.517	0.540	0.521	0.544	0.528	0.549	0.537	0.556	0.553	0.567	0.575	0.586
	336	**0.615**	**0.586**	0.622	0.591	0.629	0.597	0.639	0.604	0.652	0.612	0.672	0.625	0.702	0.648
	720	**1.009**	**0.907**	1.020	0.913	1.034	0.919	1.053	0.932	1.075	0.946	1.113	0.968	1.166	1.007
	Avg	**0.684**	**0.622**	0.692	0.628	0.701	0.634	0.712	0.641	0.726	0.650	0.748	0.663	0.781	0.687
Huron1	96	**0.439**	**0.393**	0.445	0.397	0.452	0.402	0.458	0.405	0.465	0.410	0.477	0.418	0.495	0.430
	192	**0.374**	**0.458**	0.376	0.459	0.379	0.461	0.385	0.466	0.391	0.472	0.403	0.481	0.419	0.497
	336	**0.450**	**0.501**	0.454	0.503	0.458	0.506	0.466	0.512	0.475	0.519	0.490	0.530	0.511	0.549
	720	**0.739**	**0.777**	0.745	0.778	0.753	0.779	0.767	0.789	0.783	0.802	0.810	0.821	0.850	0.853
	Avg	**0.501**	**0.532**	0.505	0.534	0.511	0.537	0.519	0.543	0.529	0.551	0.545	0.563	0.569	0.582
Huron2	96	**0.355**	**0.331**	0.361	0.336	0.368	0.343	0.373	0.346	0.379	0.350	0.388	0.357	0.403	0.368
	192	**0.302**	**0.387**	0.305	0.390	0.309	0.394	0.313	0.398	0.319	0.403	0.328	0.411	0.341	0.424
	336	**0.364**	**0.423**	0.368	0.427	0.373	0.432	0.379	0.437	0.387	0.443	0.399	0.453	0.416	0.469
	720	**0.600**	**0.657**	0.606	0.661	0.613	0.665	0.624	0.674	0.638	0.685	0.660	0.701	0.692	0.729
	Avg	**0.405**	**0.450**	0.410	0.454	0.416	0.459	0.422	0.464	0.431	0.470	0.444	0.481	0.463	0.498
Huron3	96	**0.884**	**0.556**	0.896	0.564	0.910	0.574	0.921	0.579	0.935	0.586	0.960	0.596	0.996	0.615
	192	**0.755**	**0.648**	0.759	0.652	0.763	0.658	0.773	0.665	0.787	0.673	0.810	0.686	0.843	0.709
	336	**0.906**	**0.708**	0.913	0.714	0.922	0.722	0.937	0.731	0.955	0.741	0.985	0.757	1.028	0.784
	720	**1.481**	**1.095**	1.497	1.103	1.515	1.112	1.542	1.127	1.575	1.145	1.630	1.172	1.709	1.218
	Avg	**1.007**	**0.752**	1.016	0.758	1.028	0.767	1.043	0.776	1.063	0.786	1.096	0.803	1.144	0.832
Huron4	96	**1.991**	**0.679**	2.012	0.687	2.037	0.696	2.061	0.701	2.094	0.710	2.149	0.722	2.230	0.745
	192	**1.704**	**0.791**	1.706	0.794	1.708	0.797	1.732	0.805	1.762	0.816	1.813	0.832	1.887	0.860
	336	**2.040**	**0.865**	2.051	0.870	2.064	0.875	2.097	0.885	2.138	0.898	2.204	0.917	2.301	0.950
	720	**3.327**	**1.335**	3.357	1.341	3.391	1.347	3.452	1.366	3.527	1.387	3.649	1.420	3.825	1.476
	Avg	**2.266**	**0.918**	2.282	0.923	2.300	0.929	2.336	0.939	2.380	0.953	2.454	0.973	2.561	1.008

D.4 Lake Ablations

The lake variants use the definitions in Section 4.3.

Table 9: Ablation study (MSE / MAE) on lake datasets, $H = 96$. **Bold**: best; underline: second best.

Model Variant	Erie1		Erie2		Huron1		Huron2		Huron3		Huron4	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
CausalExoFormer (Full)	1.024	0.468	**0.600**	**0.459**	**0.439**	**0.393**	**0.355**	**0.331**	0.884	0.556	**1.991**	**0.679**
Uniform Gate	1.058	0.489	0.624	0.476	0.456	0.406	0.369	0.343	0.915	0.579	2.047	0.708
Current-Only	1.070	0.497	0.632	0.482	0.462	0.411	0.374	0.348	0.926	0.587	2.066	0.719
No DAG Penalty	**1.022**	**0.467**	0.603	0.461	0.441	0.395	0.357	0.333	0.886	0.558	1.996	0.681
No Exo Graph	1.027	0.471	0.602	0.460	0.442	0.396	0.356	0.332	**0.882**	**0.554**	1.994	0.680

E Sensitivity and Efficiency

Each sweep varies the indicated hyperparameter. The default configuration matches the corresponding full-model lake result at $H = 96$. Figures 4 and 5 provide graphical views of the sparsity-weight and lag-count sweeps.

Table 10: Forecasting sensitivity to the sparsity weight ($H = 96$). Only the indicated hyperparameter varies.

Dataset	λ_1	MSE	MAE
Erie1	0.001	1.067	0.486
	0.005	1.036	0.473
	0.01 (default)	**1.024**	**0.468**
	0.05	1.031	0.471
	0.1	1.058	0.482
Erie2	0.001	0.642	0.480
	0.005	**0.600**	**0.459**
	0.01 (default)	**0.600**	**0.459**
	0.05	0.618	0.471
	0.1	0.646	0.487
Huron1	0.001	0.468	0.411
	0.005	0.446	0.398
	0.01 (default)	0.439	0.393
	0.05	**0.435**	**0.390**
	0.1	0.456	0.404
Huron2	0.001	0.382	0.349
	0.005	**0.355**	**0.331**
	0.01 (default)	**0.355**	**0.331**
	0.05	0.366	0.340
	0.1	0.385	0.353
Huron3	0.001	0.933	0.584
	0.005	0.897	0.563
	0.01 (default)	0.884	0.556
	0.05	**0.879**	**0.553**
	0.1	0.912	0.572
Huron4	0.001	2.082	0.706
	0.005	2.011	0.686
	0.01 (default)	**1.991**	**0.679**
	0.05	2.003	0.684
	0.1	2.061	0.702

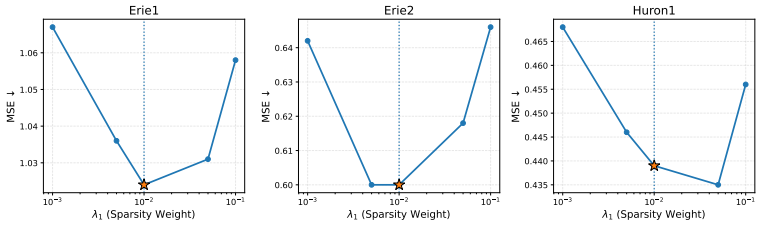

Figure 4: Sensitivity to the sparsity weight λ_1 at $H = 96$ on Erie1, Erie2, and Huron1. The dotted line and star indicate the default $\lambda_1 = 0.01$; the complete six-dataset values are reported in Table 10.

Table 11: Sensitivity to the number of candidate historical offsets K on lake datasets ($H = 96$).

Dataset	K	MSE	MAE
Erie1	1	1.083	0.495
	4	1.061	0.484
	7	1.038	0.474
	14 (default)	**1.024**	**0.468**
	21	1.046	0.479
Erie2	1	0.641	0.480
	4	0.625	0.471
	7	**0.598**	**0.458**
	14 (default)	0.600	0.459
	21	0.612	0.465
Huron1	1	0.464	0.409
	4	0.453	0.402
	7	0.444	0.397
	14 (default)	**0.439**	**0.393**
	21	0.448	0.399
Huron2	1	0.375	0.346
	4	0.367	0.339
	7	0.359	0.333
	14 (default)	**0.355**	**0.331**
	21	0.362	0.336
Huron3	1	0.912	0.578
	4	0.901	0.569
	7	0.890	0.561
	14 (default)	**0.884**	**0.556**
	21	0.898	0.564
Huron4	1	2.046	0.701
	4	2.024	0.692
	7	2.004	0.684
	14 (default)	**1.991**	**0.679**
	21	2.012	0.686

Across the lag-count sweep, $K = 14$ is the best setting on five of the six lake datasets and remains within 0.002 MSE of the best value on Erie2. We next vary the warmup duration while holding the remaining configuration fixed.

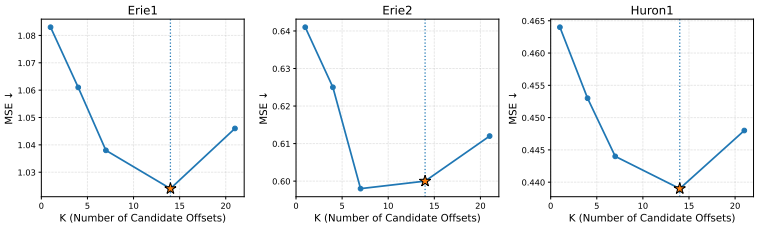

Figure 5: Sensitivity to the number of candidate historical offsets K at $H = 96$ on Erie1, Erie2, and Huron1. The dotted line and star indicate the default $K = 14$; the complete six-dataset values are reported in Table 11.

Table 12: Forecasting sensitivity to warmup length ($H = 96$).

Dataset	t_{warm}	MSE	MAE
Erie1	0	1.071	0.487
	2	1.039	0.474
	5 (default)	**1.024**	**0.468**
	10	1.029	0.470
Erie2	0	0.645	0.482
	2	0.608	0.464
	5 (default)	0.600	0.459
	10	**0.598**	**0.458**
Huron1	0	0.470	0.412
	2	0.444	0.397
	5 (default)	**0.439**	**0.393**
	10	0.441	0.394
Huron2	0	0.384	0.348
	2	0.359	0.334
	5 (default)	**0.355**	**0.331**
	10	0.357	0.332
Huron3	0	0.931	0.583
	2	0.892	0.560
	5 (default)	0.884	0.556
	10	**0.882**	**0.555**
Huron4	0	2.076	0.706
	2	2.008	0.684
	5 (default)	**1.991**	**0.679**
	10	1.997	0.681

E.1 Computational Cost

The shared-structure formulation removes the per-target graph factor from both storage and acyclicity. Constructing the gated source bank costs $O(KSd)$; shared graph propagation costs $O(KS^2d + KSd^2)$. Equation (7) adds only a rank-one source correction per target. Because key/value projections are linear, the shared component can be projected once and the correction applied in projected space; target-specific routing therefore scales with the target–source pairs, $O(MKCd)$ in model-width notation, rather than repeating an $S \times S$ graph multiplication. Structural storage is $O(SK + S^2)$, and the NOTEARS step uses one differentiated $S \times S$ matrix exponential, plus a post-update residual evaluation for multiplier control, with total cost $O(S^3)$ per active structural epoch. Inference uses no matrix exponential.

For ECL, $M = S = 321$ and $C = 320$: the model uses one 321×321 graph and one $321 \times K$ source–lag allocation, not separate graphs for the 321 forecast targets. With $K = 14$, the structural state is approximately 0.11M scalar parameters before the shared forecasting backbone. Table 13 reports the designated-target Lake-Erie2 timing comparison; the same shared-structure equations are used in the multivariate benchmarks.

Table 13: Single-target computational cost on Lake-Erie2 ($M = 1$, $T = 96$, $H = 96$). Training time is per epoch; inference time is per batch (batch size 32).

Method	#Params (M)	Train (s/epoch)	Infer (ms/batch)
DLinear	0.09	2.8	1.9
PatchTST	0.82	8.7	5.8
iTransformer	0.71	10.4	6.9
TimeXer	0.88	9.1	6.1
CausalExoFormer	1.02	10.3	6.8

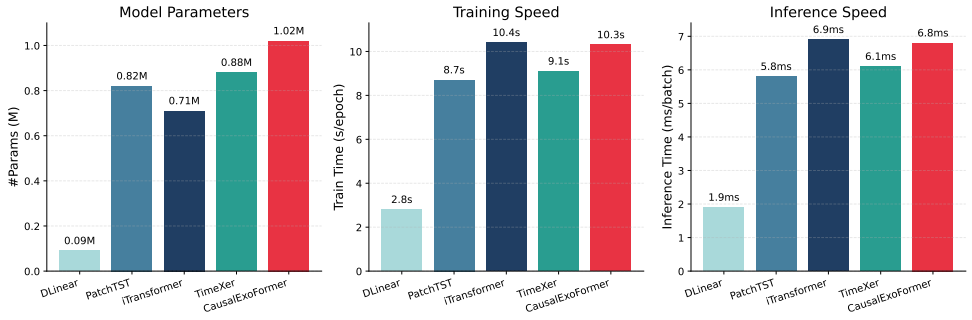

Figure 6: Computational-cost summary on Lake-Erie2 ($M = 1$, $T = 96$, $H = 96$), visualizing Table 13. The three panels report parameter count, training time per epoch, and inference latency per batch.

F Additional Structural Visualization

Figure 7 visualizes the Lake-Erie2 structural readouts reported in Table 4. The relevance values are the learned source-level scores s_r, while the offsets identify the dominant historical positions selected by $P_{r,k}$ on the 15-minute grid.

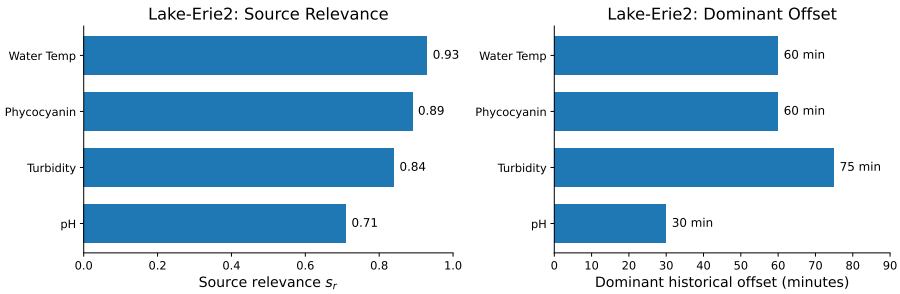

Figure 7: Lake-Erie2 dissolved-oxygen case study. Left: learned source relevance s_r for the four highest-relevance external variables. Right: the corresponding dominant historical offsets measured from the forecast origin.

G Related Work

Forecasting architectures and temporal representations. The Transformer (Vaswani et al., 2017) provides a general attention-based sequence architecture. Informer (Zhou et al., 2021) addresses long-sequence computation through sparse attention and distilling, while Autoformer (Wu et al., 2021) and FEDformer (Zhou et al., 2022) introduce decomposition, autocorrelation, and frequency-domain processing. PatchTST (Nie et al., 2023) represents local temporal segments as patches, and iTransformer (Liu et al., 2024) attends over variate tokens. Crossformer (Zhang and Yan, 2023) explicitly separates temporal and cross-dimension interactions; TimesNet (Wu et al., 2023) organizes temporal variation through two-dimensional representations. Linear and multilayer-perceptron alternatives, including DLinear (Zeng et al., 2023), N-BEATS (Oreshkin et al., 2020), and NHITS (Challu et al., 2023), emphasize decomposition, basis expansion, and hierarchical interpolation. These works motivate separating temporal representation from the external information interface. CausalExoFormer retains a target-local temporal backbone and uses explicit historical observations as its externally gated tokens.

Forecasting with exogenous variables. TFT (Lim et al., 2021) combines variable selection with multi-horizon forecasting. NBEATSx (Olivares et al., 2023) extends basis-expansion forecasting with covariates, and TiDE (Das et al., 2023) uses a dense encoder–decoder that handles covariates

and nonlinear dependencies. TimeXer (Wang et al., 2024) provides the closest attention interface: endogenous patch representations interact with external variate representations through a global token. CrossLinear (Zhou et al., 2025) introduces cross-correlation embeddings and supports many-to-many forecasting through shared parameters. CausalExoFormer separates dataset-level source relevance, historical-offset allocation, and target-specific content routing, while learning one directed propagation graph across the source bank.

Structured exogenous forecasting. GCGNet (Li et al., 2026) combines a variational generator, graph-based consistency alignment, and refinement to model temporal and channel correlations. DAG (Qiu et al., 2026) uses temporal and channel causal modules and considers available future exogenous inputs. We cite the accepted ICML 2026 version; the corresponding preprint is arXiv:2509.14933. Our input contract supplies historical observations only. Our shared graph acts on gated historical tokens before target-specific attention, and exact vertex deletion removes the target's source values from its own external bank. These differences concern where structure acts in the forecasting computation and which observations are available at prediction time.

Learned interaction graphs. Graph WaveNet (Wu et al., 2019) learns adaptive dependencies for spatial–temporal prediction, while MTGNN (Wu et al., 2020) learns graph structure for multivariate forecasting without requiring a predefined graph. Neural relational inference (Kipf et al., 2018) jointly infers latent interactions and dynamics. Amortized causal discovery (Löwe et al., 2022) transfers shared dynamical information across systems to infer their graphs. In CausalExoFormer, one graph is fitted per dataset and shared across forecast targets; the target-dependent view is an algebraic self-exclusion operation, rather than a separately fitted graph. This sharing gives consistent edge indexing and avoids a per-target matrix-exponential constraint.

Causal discovery from time series. Granger's predictive formulation (Granger, 1969) connects temporal dependence to improvements in prediction. Neural Granger causality (Tank et al., 2022) uses structured sparsity in nonlinear forecasting models. PCMCI (Runge et al., 2019) and PCMCI+ (Runge, 2020) use conditional-independence tests for lagged and contemporaneous relationships. NOTEARS (Zheng et al., 2018) provides a differentiable acyclicity characterization; nonparametric extensions (Zheng et al., 2020), DYNOTEARS (Pamfil et al., 2020), and DAGMA (Bello et al., 2022) broaden differentiable graph learning through nonlinear models, temporal structure, and alternative constraints. We use the NOTEARS characterization as a structural regularizer inside a forecasting interface. The synthetic experiment evaluates the separately parameterized target-source selection and historical-offset readout against known injected drivers.

Gating and operational interpretability. Concrete and Gumbel–Softmax relaxations (Maddison et al., 2017; Jang et al., 2017) enable differentiable optimization of discrete-variable models, and hard-concrete gates support sparse neural networks (Louizos et al., 2018). Our source–lag gates are deterministic sigmoid and temperature-softmax parameters with separate relevance and allocation roles. Work on attention-based explanations (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019) motivates specifying what a structural readout does to the computation. We establish exact source-path removal, target self-exclusion, and an endogenous fallback when external sources are closed. This operational connection is useful for environmental forecasting, where process-oriented analysis complements prediction accuracy (Reichstein et al., 2019).