ECRI: EVIDENCE-BASED INTERACTION-AWARE TIME-SERIES VARIABLE SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Time-series variable selection for multivariate forecasting must distinguish redundant variables from complementary ones while accounting for recurring variable costs. Selection rules that require each variable to provide an individual gain can miss jointly predictive variables, while predictive relevance alone does not determine whether a variable is worth acquiring. We propose Evidence-based Contextual Relevance Inference (ECRI), a framework that evaluates variable acquisition as a conditional, cost-aware joint decision. ECRI uses interaction features in a lightweight Bayesian surrogate and scores individual and joint acquisition actions by their conditional marginal evidence under separate variable and interaction costs. This permits a variable pair to satisfy the acquisition criterion without either variable first satisfying a single-variable criterion, while avoiding repeated training of downstream forecasting models. We characterize this mechanism under orthogonal-design conditions, explicitly accounting for the evidence and structural penalties of joint acquisition. The resulting variable-selection mask can be reused across forecasting backbones. Across three real-world sensor-network datasets and four forecasting backbones, ECRI removes 47.0–97.4% of optional variables and achieves the lowest prediction time and FLOPs among the six evaluated variable selectors, with competitive forecasting accuracy. These results support interaction-aware variable selection as a practical approach to reducing recurring forecasting computation.

1 INTRODUCTION

Multivariate time-series forecasting plays a central role in applications such as traffic management, industrial monitoring, energy dispatch, and environmental warning, where predictions often rely not only on the target history but also on contextual information provided by other time-series variables, such as measurements from other sensors. As sensing infrastructures become increasingly data-rich, more candidate variables become available for forecasting. Yet more is not always better: additional variables may fail to deliver meaningful predictive gains, while each selected variable introduces recurring operational costs in data acquisition, transmission, storage, processing, and inference. This creates a fundamental time-series variable-selection problem: how to select a compact subset of variables that preserves forecasting quality.

Concretely, given a target variable and a set of candidate variables, we refer to a subset of these candidate variables as a context. Variable selection therefore aims to identify a compact context. Achieving this objective poses three main challenges. First, it is difficult to identify promising candidate contexts, because useful information may be distributed across interacting variables rather than captured by any variable alone. Second, determining which candidate contexts benefit forecasting is costly because modern forecasting models typically require training, so comparing alternative contexts involves repeated retraining, which becomes impractical as the candidate space grows. Third, predictive benefit alone does not determine whether a variable is worth selecting: many variables may offer small gains, yet using them all incurs recurring operational costs. The variable selection problem is further complicated by the need to identify a compact context that balances predictive benefit against recurring variable costs. Thus, an effective variable selector should account for cross-variable interactions, evaluate candidate contexts efficiently, and select only variables worth their recurring cost.

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.

Existing variable-selection methods, including Group Lasso (Yuan & Lin, 2006), ARD (Tipping, 2001), RMP (Ament & Gomes, 2021), and ChronoEpilogi (Vareille et al., 2024), offer a natural starting point. However, their selection decisions are primarily made at the level of individual variables, making it difficult to identify variables whose predictive value emerges only when considered jointly. A natural alternative is to evaluate candidate contexts directly through downstream forecasting performance. Cherepanova et al. (Cherepanova et al., 2023) adopt this approach by evaluating candidate subsets through downstream neural-network performance. However, extending such evaluation across many candidate contexts can require repeated training, becoming computationally prohibitive as the search space grows (Pad et al., 2025). Even when beneficial contexts are identified, predictive benefit alone is insufficient to determine which variables to select. Feature-acquisition work (Ghosh & Lan, 2023; Valancius et al., 2024) therefore weighs predictive utility against acquisition cost. Taken together, existing approaches lack a unified variable-selection framework that accounts for cross-variable interactions, enables efficient context evaluation, and balances predictive benefit against recurring variable costs.

Our key insight is that joint predictive value need not be measured by repeated downstream training: it can be captured through interaction features in a lightweight Bayesian surrogate, allowing acquisition decisions to account for separate variable and interaction costs. Existing approaches typically fall short on one of these fronts, by overlooking cross-variable interactions, relying on costly retraining, or neglecting recurring variable costs. We therefore propose Evidence-based Contextual Relevance Inference (ECRI), which introduces interaction features within a four-action context-expansion framework to capture cross-variable interactions. ECRI evaluates individual and joint acquisition actions using conditional marginal evidence under separate variable and interaction costs, yielding a compact context for downstream forecasting. The resulting variable-selection mask can be reused across forecasting backbones for faster inference with little or no loss in accuracy.

We evaluate ECRI on three real-world sensor-network forecasting datasets, four forecasting backbones, and five representative variable-selection baselines. Across the 12 dataset–backbone settings, ECRI consistently selects the smallest context among the six variable selectors, reducing optional variables by 47.0–97.4%, and achieves the lowest prediction time and FLOPs in every setting. ECRI also lies on the observed MAE–FLOPs Pareto frontier across all 12 settings, indicating a consistently favorable accuracy–efficiency trade-off. Ablation studies further show that conditional marginal evidence, structural costs, and interaction features play complementary roles, supporting the effectiveness of ECRI in selecting compact contexts reusable across forecasting backbones.

This paper makes the following contributions:

- We identify the recurring cost of acquiring and processing multiple variables as a key challenge in multivariate time-series forecasting, with the key insight that joint predictive value can be captured through interaction features and efficiently assessed with a lightweight Bayesian surrogate.
- We propose ECRI, an evidence-based variable-selection framework that uses interaction features to assess joint variable value under separate variable and interaction costs. ECRI produces compact variable-selection masks reusable across forecasting backbones, and we theoretically characterize its selection behavior.
- We conduct extensive experiments on three real-world sensor-network forecasting datasets with four forecasting backbones, showing that ECRI removes 47.0–97.4% of optional variables and achieves the lowest prediction time and FLOPs among six evaluated variable selectors across all 12 dataset–backbone settings, while maintaining competitive forecasting accuracy.

2 PRELIMINARIES

Definition 1 (Multivariate Time-Series Forecasting with a Selected Context). At forecast origin t, target variable i provides the mandatory history $X_{i,t} = [y_i^{t-L+1}, \ldots, y_i^t]$ with lookback length L. Let $\mathcal{U}_i$ denote the set of M_i candidate variables available as optional forecasting inputs. Each candidate variable $u \in \mathcal{U}_i$ provides a history $H_{i,u,t}$. We define a context as a subset of candidate variables. Given training windows, a variable selector returns a context $S_i \subseteq \mathcal{U}_i$. For a forecast horizon $h > 0$, the downstream forecasting model predicts the target value at time $t + h$ using the

A five-variable example. Let $r = aU + bVW + \epsilon$, with independent symmetric binary U, V, W, a duplicate variable $U' = U$, an independent noise variable Q, and independent zero-mean noise ϵ of variance σ^2.

Variable(s)	Population role
U	Direct signal aU
U' given U	No additional information
V alone; W alone	Zero conditional mean of r
V, W together	Joint signal bVW
Q	No predictive information

Selected context	Minimum expected loss
$\emptyset$	$a^2 + b^2 + \sigma^2$
$\{U\}$ or $\{U, U'\}$	$b^2 + \sigma^2$
$\{U, V\}$ or $\{U, W\}$	$b^2 + \sigma^2$
$\{U, V, W\}$	σ^2

Figure 1: A variable's predictive contribution depends on the selected context. In this example, adding a duplicate of an already selected variable leaves the loss unchanged, whereas jointly selecting V and W explains the interaction component that neither variable alone can predict.

mandatory target history and the histories of the variables in the selected context:

$$\widehat{y}_i^{t+h} = f_{i,S_i}\big(X_{i,t}, \{H_{i,u,t} : u \in S_i\}\big). \tag{1}$$

Definition 2 (Time-Series Variable Selection). For a specified forecasting model, let $\mathcal{L}_i(S)$ denote the expected loss on future observations of a predictor trained using the mandatory target history and a context $S \subseteq \mathcal{U}_i$. Let $C_i(S)$ denote the associated resource cost, such as prediction time or computational usage. The variable-selection problem is

$$S_i^\star \in \arg\min_{S \subseteq \mathcal{U}_i} \{\mathcal{L}_i(S) + \gamma C_i(S)\}, \qquad \gamma \geq 0, \tag{2}$$

where γ controls the trade-off between predictive loss and resource cost. Each variable's predictive contribution depends on the mandatory target history and the other variables in the selected context. Evaluating candidate contexts by repeatedly training the predictor is computationally expensive. This motivates a lightweight variable-selection method that approximates the objective without repeatedly fitting the downstream forecasting model.

Definition 3 (Bayesian-Evidence-Based Loss Approximation). We use Bayesian evidence from a lightweight Bayesian surrogate to approximate predictive loss. Let $\mathcal{D}_i$ denote the response observations in the N_i training windows available for variable selection, and let X_i collect the corresponding mandatory target histories $X_{i,t}$. The histories of the candidate variables are treated as conditioning variables and suppressed in the evidence notation below. For a context S, the surrogate evidence is

$$Z_i(S) = \int p_\vartheta(\mathcal{D}_i \mid X_i, S)\, \Pi_S(d\vartheta), \tag{3}$$

where Π_S is the parameter prior for the corresponding surrogate model. All candidate contexts are evaluated using the same observations and a consistent prior specification. We define the evidence gain relative to the target-history-only model as

$$\mathcal{G}_i(S) = \log Z_i(S) - \log Z_i(\varnothing). \tag{4}$$

A larger $\mathcal{G}_i(S)$ indicates stronger evidence for including context S relative to the target-history-only model. By the sequential factorization of the marginal likelihood, negative log evidence equals the cumulative posterior predictive log loss of the surrogate.

The following two assumptions connect this sequential predictive quantity to the expected loss of the downstream forecasting model.

Assumption 1 The historical observations are sufficient and the candidate variable set is controlled. Distribution shifts or limited effective sample sizes may limit the applicability of the approximation.

Assumption 2 The surrogate captures the predictive relationships that primarily determine the usefulness of candidate contexts, and the discrepancy between the surrogate and the downstream forecasting model varies only moderately across contexts after an affine calibration. This assumption does not require identical model architectures.

The formal assumptions are provided in Appendix A. Here, $\kappa_i > 0$ is the affine calibration scale, $\varepsilon_{\mathrm{align},i}$ is the alignment-error bound, and $\varepsilon_{\mathrm{ev},i}$ is the evidence-error bound.

Under these assumptions, with probability at least $1 - \delta$, simultaneously for all $S \subseteq \mathcal{U}_i$,

$$\left| \mathcal{L}_i(S) - \mathcal{L}_i(\varnothing) + \frac{\kappa_i}{N_i} \mathcal{G}_i(S) \right| \leq 2\varepsilon_{\mathrm{align},i} + 2\kappa_i \varepsilon_{\mathrm{ev},i}. \tag{5}$$

Consequently, up to a constant independent of the context, $-\kappa_i \mathcal{G}_i(S)/N_i$ approximates $\mathcal{L}_i(S)$.

Definition 4 (Cost Approximation). The resource cost of a context depends on the downstream forecasting architecture and its implementation. To obtain a selection criterion independent of the downstream forecasting model, we use a cardinality-based cost surrogate that assigns a positive cost to each selected variable. For a fixed history length $L > 0$, we define

$$\widehat{C}(S) = |S| \log(1 + L). \tag{6}$$

This surrogate increases monotonically with context size. It therefore favors smaller contexts when their predictive contributions are comparable.

Definition 5 (Individual and Interaction Evidence Scores). Each candidate variable has a feature representation. To distinguish individual evidence gains from departures from additive evidence, we define

$$b_{i,u} = \mathcal{G}_i(\{u\}), \qquad b_{i,uv} = \mathcal{G}_i(\{u, v\}) - \mathcal{G}_i(\{u\}) - \mathcal{G}_i(\{v\}), \qquad u \neq v. \tag{7}$$

The individual evidence score measures the evidence gained by adding variable u to the target-history-only model. The interaction evidence score measures the departure of the joint evidence gain from the sum of the individual evidence gains: positive values indicate complementary contributions on the evidence scale, whereas negative values indicate subadditivity. Figure 1 illustrates a setting in which neither variable alone improves prediction, whereas their interaction does. The three components in the example have zero pairwise cross moments, yielding the expected loss values shown in the figure.

3 Evidence-Based Interaction-Aware Variable Selection

ECRI uses Bayesian evidence to evaluate candidate selection states, greedily expands the selected context and its auxiliary interaction set under explicit structural costs, and accelerates selection through residual-based variable screening and action shortlisting. Throughout this section, we fix a target variable i and suppress its index.

3.1 Selection-State Scoring with Bayesian Evidence

Let $A = (S, E)$ denote the selection state, where S is the selected context and E is the set of variable pairs whose interaction features are included in the surrogate. Thus, S specifies the selected variables, whereas E specifies their auxiliary interaction features.

Let $\mathbf{h}_u \in \mathbb{R}^N$ denote the standardized feature column representing candidate variable u across the N training windows. It averages the standardized historical features. The interaction feature for variables u and v is

$$\mathbf{p}_{uv} = \mathcal{Z}(\mathbf{h}_u \odot \mathbf{h}_v), \tag{8}$$

where $\odot$ denotes element-wise multiplication and $\mathcal{Z}$ denotes z-score standardization. Let X be the standardized training-window matrix formed from the mandatory target histories $X_{i,t}$, and let $\mathbf{z} \in \mathbb{R}^N$ contain the target increments $z^t = y^{t+h} - y^t$ for the same N training windows. We construct the surrogate design matrix by concatenating the mandatory target-history columns, the columns representing the selected variables, and the auxiliary interaction columns:

$$\Phi_A = \big[X, \{\mathbf{h}_u : u \in S\}, \{\mathbf{p}_{uv} : (u, v) \in E\}\big]. \tag{9}$$

We use Bayesian linear regression as a lightweight Bayesian surrogate that supports evidence-based scoring. Following the Gaussian evidence model used in Bayesian model assessment (Tipping, 2001), we place an isotropic Gaussian prior on the regression coefficients and assume independent Gaussian observation noise:

$$\mathbf{z} = \Phi_A \mathbf{w} + \boldsymbol{\epsilon}, \qquad \mathbf{w} \sim \mathcal{N}(0, \alpha^{-1} I), \boldsymbol{\epsilon} \sim \mathcal{N}(0, \beta^{-1} I), \qquad \alpha, \beta > 0, \tag{10}$$

where D_A is the number of columns in Φ_A and $\mathbf{w} \in \mathbb{R}^{D_A}$ contains their regression coefficients. The prior precision α controls coefficient shrinkage; larger values concentrate the coefficients more strongly around zero. The noise precision β is the inverse observation-noise variance. Instead of evaluating only a fitted coefficient vector, we integrate over $\mathbf{w}$. This gives the marginal distribution $\mathbf{z} \sim \mathcal{N}(0, \Sigma_A)$, with

$$\Sigma_A = \beta^{-1} I + \alpha^{-1} \Phi_A \Phi_A^\top. \tag{11}$$

where Σ_A includes both observation noise and target variation induced by the coefficient prior through the surrogate design. In particular, $\Phi_A \Phi_A^\top$ describes how the training samples relate to one another through the current design. The resulting log evidence is

$$\ell(A; \alpha, \beta) = -\tfrac{1}{2}\big[N \log(2\pi) + \log\det \Sigma_A + \mathbf{z}^\top \Sigma_A^{-1} \mathbf{z}\big]. \tag{12}$$

The quadratic term measures the size of the observed target-increment vector relative to the covariance supported by the surrogate. It is small when the design assigns sufficient variation to directions aligned with the observations. The log-determinant term measures the overall covariance volume and penalizes spreading probability mass too broadly. Consequently, a selection state receives strong evidence when its surrogate explains the observations without introducing excessive covariance in unsupported directions. Because all designs use the same N observations, the first term is constant across comparisons. For each selection state, a finite number of Type-II iterations estimates the prior and noise precisions, yielding positive values $\widehat{\theta}_A = (\widehat{\alpha}_A, \widehat{\beta}_A)$. These iterations seek to increase the marginal likelihood with respect to the two precisions. We denote the Gaussian log evidence evaluated at the returned precisions by $\ell_{\mathrm{EB}}(A) = \ell(A; \widehat{\theta}_A)$. To evaluate a candidate acquisition action, let $A \oplus a$ denote the selection state obtained by applying an admissible action a to the current state A. The implementation assigns the action the conditional log-evidence gain

$$\Delta(a \mid A) = \ell_{\mathrm{EB}}(A \oplus a) - \ell_{\mathrm{EB}}(A). \tag{13}$$

This gain evaluates the additional explanatory value of the action given the current context and auxiliary interaction set. Appendix C develops its geometric interpretation and the relation between fixed and refitted precisions.

3.2 Cost-aware Search over Variables and Interactions

Starting from the mandatory target history, ECRI expands the selected context and auxiliary interaction set through four acquisition actions: **Single**: add one unselected variable. **Pair**: add two unselected variables and their interaction feature. **Expand**: add one unselected variable and its interaction feature with an already selected variable. **Edge**: add an interaction feature between two already selected variables.

These actions preserve the inclusion constraint that an interaction feature can be activated only when both participating variables belong to the selected context. The Pair action allows ECRI to select jointly useful variables without first requiring either variable to satisfy a single-variable acquisition criterion. It remains available throughout the search whenever both variables are unselected and the resulting action is feasible. Actions are restricted to the candidate space and respect the context-size cap $C_{\max}$. An edge action remains possible after this cap is reached because it does not add another variable.

To account for auxiliary interaction features, we extend the variable-cost surrogate $\widehat{C}(S)$ to the selection state $A = (S, E)$:

$$\widehat{C}(A) = |S| \log(1 + L) + \eta |E| \log 2, \quad \eta \geq 0. \tag{14}$$

The first term penalizes selected variables, while the second penalizes auxiliary interaction features in the surrogate. Together, they define the structural cost of the selection state. For an action adding a set of variables $U(a)$ and at most one interaction $e(a)$, the incremental structural cost is

$$c(a) = \widehat{C}(A \oplus a) - \widehat{C}(A) = |U(a)| \log(1 + L) + \eta \mathbb{I}[e(a) \neq \emptyset] \log 2. \tag{15}$$

We define the net value V and ranking score ρ of an action as follows. The surrogate trade-off weight λ absorbs the loss calibration and resource-cost scaling from Section 2; it need not equal the downstream objective weight γ.

$$V(a \mid A) = \Delta(a \mid A)/N - \lambda c(a), \qquad \lambda \geq 0, \tag{16}$$

Algorithm 1 ECRI Action Selection

Require: Initial selection state A, acceptance threshold τ
Ensure: Final selection state A
Fit the Bayesian surrogate model for A
loop
  $a^\star \leftarrow \text{null}$, $\rho^\star \leftarrow -\infty$
  for all feasible actions a under A, in a fixed order **do**
    Fit the Bayesian surrogate model using the selection obtained by applying a to A
    Compute the evidence gain and scores $V(a \mid A)$ and $\rho(a \mid A)$
    if $V(a \mid A) > \tau$ **and** $\rho(a \mid A) > \rho^\star$ **then**
      $a^\star \leftarrow a$
      $\rho^\star \leftarrow \rho(a \mid A)$
    end if
  end for
  if $a^\star = \text{null}$ **then**
    break
  end if
  Update A by applying $a^\star$
  Adopt the surrogate model fitted for the accepted action
end loop
return A

$$\rho(a \mid A) = V(a \mid A)/n_a, \qquad n_a = \max(1, |U(a)|). \tag{17}$$

The net value V determines whether an action is worth accepting. The ranking score ρ compares eligible actions by their net value per newly selected variable, since a Pair action adds two variables at once. For an Edge action, the denominator is one.

Algorithm 1 performs iterative greedy selection by exhaustively evaluating all feasible actions at each step. Starting from the initial state A, ECRI fits a Bayesian surrogate and evaluates each action by refitting the surrogate for the resulting selection state. Among actions satisfying $V(a \mid A) > \tau$, it accepts the action with the highest ranking score $\rho(a \mid A)$ and updates both the selection state and the surrogate. A fixed evaluation order resolves ties deterministically. This process repeats until no feasible action passes the acceptance threshold. The final context S defines the variable-selection mask used by downstream forecasting models.

3.3 Computational Speedup

The exhaustive procedure in Algorithm 1 requires a surrogate refit for every feasible action at each iteration. To reduce this cost, we apply residual-based variable screening to form a smaller candidate pool, followed by residual-based action shortlisting. Bayesian evidence is evaluated only for shortlisted actions.

We first fit an L_2-regularized linear regression of $\mathbf{z}$ on the mandatory target history with ridge parameter $\lambda_{\mathrm{r}} > 0$ and obtain the residual $\mathbf{r}_{\mathrm{init}}$. The subscript r denotes ridge regularization. The regularizer is applied to all coefficients. For each candidate variable u, we use the same standardized feature column $\mathbf{h}_u$ as in the evidence model. We compute the residual-based score

$$m(u) = \frac{(\mathbf{h}_u^\top \mathbf{r}_{\mathrm{init}})^2}{N(\mathbf{h}_u^\top \mathbf{h}_u + \lambda_{\mathrm{r}})}. \tag{18}$$

This score measures how well the feature representation of variable u explains the residual left by the mandatory target history under a regularized linear fit. We retain the candidate variable pool $\mathcal{P} = \mathrm{Top}_k\{m(u) : u \in \mathcal{U}\}$, where $k = \min(K_s, M)$ and Top_k returns the indices of the highest-scoring variables. The candidate variable pool remains fixed during selection.

At each iteration, we further reduce the number of evidence evaluations using the current regression residual. For the current selection state A, we compute

$$\mathbf{r}_A = \mathbf{z} - \Phi_A \left(\Phi_A^\top \Phi_A + \lambda_{\mathrm{r}} I\right)^{-1} \Phi_A^\top \mathbf{z}. \tag{19}$$

For each feasible action a within the screened candidate space, let H_a contain its newly introduced columns, with $U(a)$ and n_a defined as above. Let P denote the regularized residual projection energy and s the inexpensive screening score:

$$P(a \mid A) = \mathbf{r}_A^\top H_a \left(H_a^\top H_a + \lambda_{\mathrm{r}} I\right)^{-1} H_a^\top \mathbf{r}_A, \qquad s(a \mid A) = \frac{P(a \mid A)}{N n_a^2} - \frac{\lambda c(a)}{n_a}, \tag{20}$$

where $c(a)$ is the incremental complexity cost of the action. The projection is normalized by n_a^2.

We use the B highest-scoring feasible actions to form a shortlist $\mathcal{B}(A)$, or all feasible actions if fewer than B are available. The evidence evaluation and acceptance steps of Algorithm 1 are applied only to this shortlist: among actions satisfying $V(a \mid A) > \tau$, we accept the one maximizing $\rho(a \mid A)$. After acceptance, we update the selection state, recompute the residual, and construct a new shortlist. Appendix D specifies the four action blocks and exact tie-breaking rules. Selection stops when no shortlisted action passes the acceptance threshold.

Computational complexity Variable screening and action shortlisting reduce evidence evaluations per iteration from $O(M^2)$ to at most B, while precomputed sufficient statistics reduce the per-evaluation cost from $O(ND_{\text{full}}^2 + RD_{\text{full}}^3)$ to $O(RD^3)$, where R is the number of evidence-fitting updates and $D \le d_0 + k + k(k-1)/2$ for $k = \min(K_s, M)$.

4 Experiments

We evaluate ECRI's accuracy–efficiency trade-off, variable reduction, recurring prediction cost, and selection overhead, with ablations study on component contributions.

Datasets. We use Beijing Multi-Site Air Quality (Zhang et al., 2017), METR-LA, and PEMS-BAY (Li et al., 2018), treating each station's PM2.5 series or sensor's speed series in turn as the target. Target history is mandatory; all remaining variables are selectable. Forecast horizons are 1 for Beijing Air and 12 (one hour) for both traffic datasets, with chronological 60/20/20 train/validation/test splits.

Forecasting Backbones. We evaluate TCN (Bai et al., 2018) and three Transformer-based models: PatchTST (Nie et al., 2023), TimeXer (Wang et al., 2024), and iTransformer (Liu et al., 2024). A separate model is trained for each target, selection method, and backbone. ECRI fixes each target's selection mask before training and reuses it across backbones. We use at most 5,000 rows per split and target, 100 training epochs, early-stopping patience 10, and base seeds 291/1291/2291 with deterministic target offsets.

Baselines. **Target only** uses target history and **Full context** adds all candidate histories. Five selection baselines use the same candidate inputs: **Group Lasso** for group sparsity (Yuan & Lin, 2006), **ChronoEpilogi-FB** for conditional forward selection and backward elimination (Vareille et al., 2024), **GOGA-HDAIC** for greedy group selection with an information criterion (Huang & Tsay, 2024), and Bayesian relevance-pruning methods **ARD** and **RMP** using shared precision and variance parameters, respectively (Tipping, 2001; Ament & Gomes, 2021).

Metrics. We report context size $1 + |S_i|$ (including the target), per-sample forward FLOPs and prediction time for recurring cost, selection time for one-time overhead, and MAE for forecasting error. MAE–FLOPs Pareto frontiers summarize the accuracy–efficiency trade-off.

Implementation Details. ECRI uses residual-based marginal screening with $K_s = 80$, a shared shortlist of at most $B = 4$ actions, an optional-variable cap $C_{\max} = \min(30, M)$, and lookback $L = 12$. We set residual regularization $\lambda_{\mathrm{r}} = 10^{-8}$, Type-II initialization $(\alpha^{(0)}, \beta^{(0)}) = (1, 10^6)$, and at most $R = 30$ updates per selection-state fit. The interaction-cost weight, structural-cost weight, and acceptance threshold are $\eta = 1$, $\lambda = 10^{-3}$, and $\tau = 0$, respectively. The hardware and software environment is in Appendix D.

Main Results Table 1 reports results for two representative forecasting backbones, TCN and iTransformer, while the complete results for PatchTST and TimeXer are provided in Appendix E. **Claim 1:** ECRI consistently produces the smallest context among the six selection methods, retaining only 6.83 of 12 inputs on Beijing Air, 6.28 of 207 on METR-LA, and 14.37 of 325 on PEMS-BAY. These correspond to reductions of 47.0%, 97.4%, and 95.9% of the optional context relative to Full context. **Claim 2:** ECRI achieves the lowest prediction time and FLOPs among the six selection methods.

Table 1: Forecasting results with selected variables on three real-world datasets. Selection and prediction times are in seconds. Parenthesized values are relative to Full context. The results on the other two backbones are in Appendix E.

Method	Context Size	Select. Time	TCN			iTransformer		
			Pred. Time	MFLOPs	MAE	Pred. Time	MFLOPs	MAE
Dataset					**Beijing Air**			
Target only	1.00	0.000	0.015	0.298	10.022	0.028	0.198	10.027
Full context	12.00	0.000	0.021	0.332	9.409	0.131	2.370	9.950
ChronoEpilogi-FB	10.25 (↓15.9%)	171.238	0.019 (↓9.5%)	0.327 (↓1.5%)	9.585 (+1.87%)	0.119 (↓9.2%)	2.024 (↓14.6%)	9.948 (-0.02%)
GOGA-HDAIC	11.75 (↓2.3%)	0.906	0.021 (↓0.0%)	0.331 (↓0.3%)	9.415 (+0.06%)	0.130 (↓0.8%)	2.321 (↓2.1%)	9.961 (+0.11%)
Group Lasso	12.00 (↓0.0%)	1.141	0.021 (↓0.0%)	0.332 (↓0.0%)	9.409 (0.00%)	0.131 (↓0.0%)	2.370 (↓0.0%)	9.950 (0.00%)
ARD	12.00 (↓0.0%)	0.275	0.022 (↑4.8%)	0.332 (↓0.0%)	9.409 (0.00%)	0.131 (↓0.0%)	2.370 (↓0.0%)	9.950 (0.00%)
RMP	11.17 (↓7.5%)	1.255	0.021 (↓0.0%)	0.329 (↓0.9%)	9.502 (+0.99%)	0.128 (↓2.3%)	2.205 (↓7.0%)	9.941 (-0.09%)
ECRI	**6.83 (↓47.0%)**	2.408	**0.013 (↓38.1%)**	**0.316 (↓4.8%)**	9.701 (+3.10%)	**0.090 (↓31.3%)**	**1.350 (↓43.0%)**	9.963 (+0.13%)
Dataset					**METR-LA**			
Target only	1.00	0.000	0.283	0.298	7.925	0.551	0.198	7.759
Full context	207.00	0.000	1.272	0.931	7.764	55.822	40.883	7.817
ChronoEpilogi-FB	18.79 (↓91.4%)	4,286.457	0.373 (↓70.7%)	0.353 (↓62.1%)	7.862 (+1.26%)	3.643 (↓93.5%)	3.711 (↓90.9%)	7.677 (-1.79%)
GOGA-HDAIC	68.67 (↓67.2%)	2,332.486	0.599 (↓52.9%)	0.506 (↓45.6%)	7.821 (+0.73%)	13.968 (↓75.0%)	13.562 (↓66.8%)	7.663 (-1.97%)
Group Lasso	25.08 (↓88.3%)	91.547	0.403 (↓68.3%)	0.372 (↓60.0%)	7.990 (+2.91%)	5.418 (↓90.3%)	4.954 (↓87.9%)	7.847 (+0.38%)
ARD	61.84 (↓70.5%)	100.873	0.565 (↓55.6%)	0.485 (↓47.9%)	7.898 (+1.73%)	12.177 (↓78.2%)	12.213 (↓70.1%)	7.681 (-1.74%)
RMP	41.51 (↓80.3%)	503.256	0.478 (↓62.4%)	0.423 (↓54.6%)	7.901 (+1.76%)	7.407 (↓86.7%)	8.199 (↓79.9%)	7.723 (-1.20%)
ECRI	**6.28 (↓97.4%)**	165.405	**0.214 (↓83.2%)**	**0.314 (↓66.3%)**	7.842 (+1.00%)	**1.491 (↓97.3%)**	**1.240 (↓97.0%)**	**7.631 (-2.38%)**
Dataset					**PEMS-BAY**			
Target only	1.00	0.000	0.316	0.298	2.951	0.776	0.198	3.266
Full context	325.00	0.000	2.983	1.293	2.700	150.963	64.189	2.810
ChronoEpilogi-FB	46.47 (↓86.0%)	40,919.642	0.674 (↓77.4%)	0.438 (↓66.1%)	2.737 (+1.37%)	11.686 (↓92.3%)	9.179 (↓85.7%)	2.975 (+5.87%)
GOGA-HDAIC	80.77 (↓75.4%)	5,196.710	0.901 (↓69.8%)	0.543 (↓58.0%)	2.729 (+1.07%)	23.312 (↓84.6%)	15.953 (↓75.1%)	2.927 (+4.16%)
Group Lasso	73.12 (↓77.7%)	159.748	0.854 (↓71.4%)	0.520 (↓59.8%)	2.614 (-3.19%)	21.003 (↓86.1%)	14.442 (↓77.5%)	2.922 (+3.99%)
ARD	71.59 (↓78.2%)	145.219	0.854 (↓71.4%)	0.515 (↓60.2%)	2.610 (-3.33%)	20.339 (↓86.5%)	14.139 (↓78.0%)	2.958 (+5.27%)
RMP	50.36 (↓84.8%)	1,255.272	0.691 (↓76.8%)	0.450 (↓65.2%)	2.594 (-3.93%)	12.299 (↓91.9%)	9.946 (↓84.5%)	2.996 (+6.62%)
ECRI	**14.37 (↓95.9%)**	466.803	**0.412 (↓86.2%)**	**0.339 (↓73.8%)**	**2.556 (-5.33%)**	**4.183 (↓97.2%)**	**2.837 (↓95.6%)**	2.914 (+3.70%)

On the two representative backbones reported in Table 1, ECRI reduces forecasting time by 38.1%, 83.2%, and 86.2% for TCN, and by 31.3%, 97.3%, and 97.2% for iTransformer across the three datasets. **Claim 3:** Despite this aggressive compression, the corresponding MAE changes relative to Full context remain within -5.33% to $+3.70\%$. These results show that ECRI can substantially reduce the forecasting interface and recurring computation without a proportional loss in predictive performance. **Claim 4:** ECRI is faster than recent complicated selectors but slower than the traditional selectors. It requires 2.408 s, 165.405 s, and 466.803 s to construct target-specific contexts for Beijing Air, METR-LA, and PEMS-BAY, respectively. For the largest PEMS-BAY, its selection time is approximately $87.5\times$ lower than ChronoEpilogi-FB and $11.1\times$ lower than GOGA-HDAIC. This selection cost is paid only once during context construction, whereas the reductions in context size and forecasting computation recur throughout deployment.

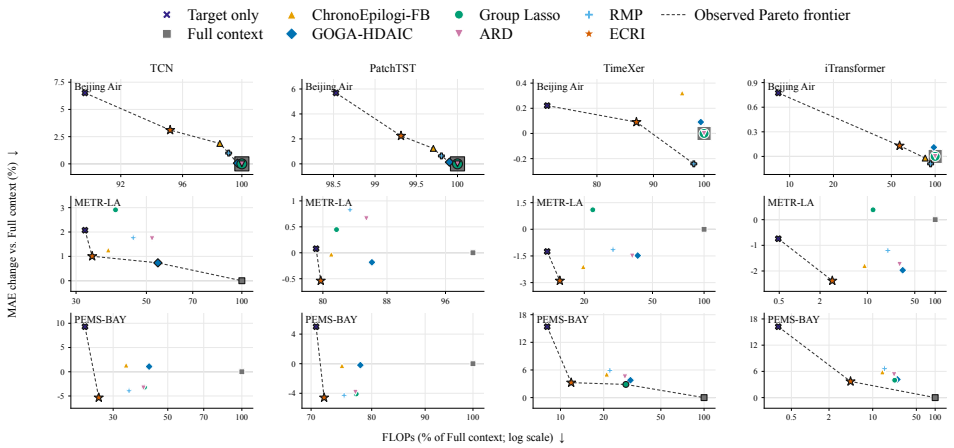


Figure 2: The MAE–FLOPs space and observed Pareto frontiers across three datasets and four forecasting backbones.

Figure 2 further compares the reported configurations of all methods in the MAE–FLOPs space across three datasets and four forecasting backbones. Both axes are normalized against Full context, and lower-left indicates a better accuracy–efficiency trade-off. **Claim 5:** ECRI lies on the observed Pareto frontier in all settings, indicating that none of the displayed alternatives simultaneously achieves both lower MAE and lower computational cost.

Ablation Study Table 2 evaluates the contribution of the main components of ECRI. Full ECRI achieves the lowest MAE, the smallest context size, and the lowest FLOPs among all variants, indicating that the three components play complementary roles. Overall, the three components play complementary roles: conditional evidence and structural cost are essential for compact context selection, while interaction modeling provides additional gains in predictive performance.

Table 2: Ablation study of the main ECRI components. The variants remove the evidence criterion, interaction modeling, and structural cost, respectively.

Variant	MAE	Context Size	Pred.Time	MFLOPs
Full ECRI	**2.5557**	**13.37**	**0.412**	**0.339**
w/o Evidence	2.6254	29.76	0.542	0.390
w/o Interaction	2.5651	14.18	0.571	0.342
w/o Cost	2.6144	26.92	0.511	0.381

5 Related Work

Time-series variable selection. ChronoEpilogi discovers informationally equivalent time-series Markov boundaries under composition, interchangeability, and perfect-test assumptions for FBE (Vareille et al., 2024). TSelect selects relevant, non-redundant channels independently of the classifier (Nuyts et al., 2025). Group Lasso and GOGA–HDAIC use group regularization and greedy path selection, respectively (Yuan & Lin, 2006; Huang & Tsay, 2024), while LA-NAM combines empirical Bayes feature selection with interaction ranking (Bouchiat et al., 2024). ECRI scores candidate contexts using conditional marginal evidence from a lightweight Bayesian surrogate, without training forecasting models during selection.

Interaction-aware variable selection. Prior work distinguishes variable inclusion costs from coefficient sparsity (Bien et al., 2013) and discovers interactions missed by marginal screening (Shah, 2016; Fan et al., 2016). OptFeature selects feature-field and feature-value interactions in deep sparse networks (Lyu et al., 2023); Meta-ANOVA screens interactions in a pre-trained black-box model for functional ANOVA approximation (Choi et al., 2025). ECRI jointly acquires variable pairs and activates their surrogate interaction, without requiring nonzero main-effect coefficients.

Variable acquisition. DiFA and dynamic time-series acquisition select measurements as observations arrive (Ghosh & Lan, 2023; Sergeev et al., 2024), with retrospective evaluation requiring explicit acquisition assumptions (von Kleist et al., 2025). Learning-to-Measure meta-learns acquisition policies with lookahead for pairwise synergies (Kobayashi et al., 2025). ECRI instead produces a fixed offline selection mask using conditional marginal evidence and separate variable and interaction costs. Separately trained forecasting backbones then use the selected context alongside the target history.

6 Conclusion

Time-series variable selection for multivariate forecasting must account for jointly predictive variables while balancing predictive benefit against recurring forecasting costs. ECRI addresses this problem through interaction features in a lightweight Bayesian surrogate, using conditional log-evidence gains under separate variable and interaction costs to evaluate individual and joint acquisition actions. Across three datasets and four backbones, ECRI selects the smallest contexts and achieves the lowest prediction time and FLOPs among the evaluated variable selectors, while maintaining competitive forecasting accuracy. These results support interaction-aware variable selection as a practical way to construct compact forecasting contexts whose variable-selection masks can be reused across backbones. Future work is to improve candidate screening reliability and apply to large time-series foundation models.

AI USE STATEMENT

ChatGPT was used to assist with language polishing and improve the clarity and readability of the manuscript. The technical ideas, experimental design and execution, and analysis of the results were carried out by the authors. The authors reviewed and edited all AI-assisted text and take full responsibility for the content of the manuscript.

REFERENCES

Sebastian E. Ament and Carla P. Gomes. Sparse Bayesian learning via stepwise regression. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 264–274. PMLR, 2021.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Jacob Bien, Jonathan Taylor, and Robert Tibshirani. A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141, 2013.

Kouroche Bouchiat, Alexander Immer, Hugo Yèche, Gunnar Ratsch, and Vincent Fortuin. Improving neural additive models with bayesian principles. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4416–4443, 2024.

Valeriia Cherepanova, Roman Levin, Gowthami Somepalli, Jonas Geiping, C. Bayan Bruss, Andrew G. Wilson, Tom Goldstein, and Micah Goldblum. A performance-driven benchmark for feature selection in tabular deep learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yongchan Choi, Seokhun Park, Chanmoo Park, Dongha Kim, and Yongdai Kim. Meta-ANOVA: Screening interactions for interpretable machine learning. *Journal of the Korean Statistical Society*, 54:478–495, 2025.

Yingying Fan, Yinfei Kong, Daoji Li, and Jinchi Lv. Interaction pursuit with feature screening and selection. *arXiv preprint arXiv:1605.08933*, 2016.

Aritra Ghosh and Andrew Lan. DiFA: Differentiable feature acquisition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7705–7713, 2023.

Shuo-Chieh Huang and Ruey S. Tsay. Time series forecasting with many predictors. *Mathematics*, 12(15):2336, 2024. doi: 10.3390/math12152336.

Yuta Kobayashi, Zilin Jing, Jiayu Yao, Hongseok Namkoong, and Shalmali Joshi. Learning-to-measure: In-context active feature acquisition, 2025.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018.

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.

Fuyuan Lyu, Xing Tang, Dugang Liu, Chen Ma, Weihong Luo, Liang Chen, Xiuqiang He, and Xue Liu. Towards hybrid-grained feature interaction selection for deep sparse network. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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.

Loren Nuyts, Lorenzo Perini, and Jesse Davis. Tselect: selecting relevant and non-redundant channels for multivariate time series classification. *Data Mining and Knowledge Discovery*, 39, 2025. doi: 10.1007/s10618-025-01132-4.

Pedram Pad, Hadi Hammoud, Mohamad Dia, Nadim Maamari, and Liza Andrea Dunbar. SAND: One-shot feature selection with additive noise distortion. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 47443–47455. PMLR, 2025.

Fedor Sergeev, Paola Malsot, Gunnar Ratsch, and Vincent Fortuin. Towards dynamic feature acquisition on medical time series by maximizing conditional mutual information. In *ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2024.

Rajen D. Shah. Modelling interactions in high-dimensional data with backtracking. *Journal of Machine Learning Research*, 17(207):1–31, 2016.

Michael E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

Michael E. Tipping and Anita C. Faul. Fast marginal likelihood maximisation for sparse bayesian models. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, volume R4 of *Proceedings of Machine Learning Research*, pp. 276–283. PMLR, 2003.

Michael Valancius, Maxwell Lennon, and Junier Oliva. Acquisition conditioned oracle for nongreedy active feature acquisition. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48957–48975. PMLR, 2024.

Etienne Vareille, Michele Linardi, Ioannis Tsamardinos, and Vassilis Christophides. Chronoepilogi: Scalable time series selection with multiple solutions. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4267.

Henrik von Kleist, Alireza Zamanian, Ilya Shpitser, and Narges Ahmidi. Evaluation of active feature acquisition methods for time-varying feature settings. *Journal of Machine Learning Research*, 26 (60):1–84, 2025.

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.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67, 2006.

Shuyi Zhang, Bin Guo, Anlan Dong, Jing He, Ziping Xu, and Song Xi Chen. Cautionary tales on air-quality improvement in beijing. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2205):20170457, 2017.

A ASSUMPTIONS

Assumption 1 (Uniform Evidence–Loss Approximation). Let N_i denote the number of predictive observations in the training windows used to compute the evidence. For each context $S \subseteq \mathcal{U}_i$, define the evidence-based average log loss as

$$\mathcal{L}_i^{\mathrm{ev}}(S) := -\frac{1}{N_i} \log Z_i(S). \tag{21}$$

Let $\mathcal{L}_i^{\mathrm{sur}}(S)$ denote the surrogate's expected predictive log loss. Assume that, with probability at least $1 - \delta$,

$$\sup_{S \subseteq \mathcal{U}_i} |\mathcal{L}_i^{\mathrm{ev}}(S) - \mathcal{L}_i^{\mathrm{sur}}(S)| \leq \varepsilon_{\mathrm{ev},i}. \tag{22}$$

Assumption 1 requires the evidence-based average log loss to approximate the surrogate's expected predictive log loss uniformly over candidate contexts. A stable predictive mechanism, controlled temporal dependence, and stabilization of the surrogate's posterior predictive behavior can support this approximation.

Assumption 2 (Uniform Surrogate–Downstream Loss Alignment). There exist a calibration scale $\kappa_i > 0$ and an offset b_i, both independent of the context S, such that

$$\sup_{S \subseteq \mathcal{U}_i} |\mathcal{L}_i(S) - [b_i + \kappa_i \mathcal{L}_i^{\mathrm{sur}}(S)]| \leq \varepsilon_{\mathrm{align},i}. \tag{23}$$

Assumption 2 requires context-dependent changes in the surrogate's expected predictive log loss to track changes in the downstream forecasting loss after a common affine calibration. The surrogate and the downstream forecasting model use information from the same target history and selected context, although their representations and architectures may differ. The assumption requires the surrogate to capture the predictive relationships that primarily determine the usefulness of candidate contexts, with a uniformly bounded discrepancy after calibration.

B RUNNING EXAMPLE

Let U, V, W be independent symmetric random signs taking values in $\{-1, +1\}$ with equal probability, let $U' = U$, and let Q be a noise variable independent of U, V, W. Let ϵ be independent of these variables and the mandatory target history, with mean zero and variance σ^2. Assume that (U, V, W, Q, ϵ) is jointly independent of the mandatory target history. Consider the residual signal $r = aU + bVW + \epsilon$. For a context $S \subseteq \{U, U', V, W, Q\}$, write $\mathbb{E}[r \mid S]$ for the conditional expectation given the variables in S. The squared-loss oracle predictor is $\mathbb{E}[r \mid S]$, because any square-integrable predictor $f(S)$ satisfies

$$\mathbb{E}[(r - f(S))^2] = \mathbb{E}[(r - \mathbb{E}[r \mid S])^2] + \mathbb{E}[(f(S) - \mathbb{E}[r \mid S])^2]. \tag{24}$$

Zero means and independence give $\mathbb{E}[r] = 0$, $\mathbb{E}[r \mid U] = aU$, and $\mathbb{E}[r \mid V] = \mathbb{E}[r \mid W] = 0$. Moreover, $\mathbb{E}[r \mid U, V] = \mathbb{E}[r \mid U, W] = aU$, $\mathbb{E}[r \mid V, W] = bVW$, and $\mathbb{E}[r \mid U, V, W] = aU + bVW$. The components aU, bVW, and ϵ have zero pairwise cross moments, yielding the expected loss values in Figure 1. The duplicate variable U' carries the same information as U, while adding the independent noise variable Q to any context leaves the conditional mean unchanged. The population association with the interaction feature is $\mathbb{E}[rVW] = b$. In this special case, the interaction signal is representable by the surrogate's product interaction feature, and the candidate variable pool is assumed to retain both participating variables, V and W. The example illustrates the conditional predictive value of individual variables and their interactions; it does not establish recovery when variable screening excludes a relevant variable or the surrogate representation omits the interaction signal.

C WHEN JOINT ACQUISITION MATTERS

As in Section 3, we fix a target variable and suppress its index. We analyze when acquiring two variables together with their interaction can succeed even though acquiring either variable alone fails. All comparisons preserve the scaling of existing feature columns. We first use common precisions $\alpha, \beta > 0$, then discuss precision refitting.

Conditional log-evidence gain. Standard Gaussian marginal-likelihood identities express the gain from appending a block of input columns H as a fit improvement minus a covariance-volume penalty (Tipping, 2001; Tipping & Faul, 2003). Define

$$K_H = H^\top \Sigma_A^{-1} H, \qquad \mathbf{b}_H = H^\top \Sigma_A^{-1} \mathbf{z}. \tag{25}$$

When the appended coefficients have prior precision α, the fixed-precision gain is

$$\Delta^{\alpha,\beta}(H \mid A) = \frac{1}{2}\left[\mathbf{b}_H^\top (\alpha I + K_H)^{-1} \mathbf{b}_H - \log\det(I + \alpha^{-1} K_H)\right]. \tag{26}$$

This follows from the matrix determinant lemma and the Woodbury identity. For an action a that appends H_a, we write $\Delta^{\alpha,\beta}(a \mid A) = \Delta^{\alpha,\beta}(H_a \mid A)$. Unlike the refitted gain $\Delta(a \mid A)$ used in Section 3, this quantity holds both precisions fixed across the comparison.

The identity allows us to quantify the cost of adding the feature columns of the two participating variables alongside an informative interaction column.

Joint acquisition across a single-variable barrier. Fix a state A and two unselected variables with standardized columns $\mathbf{h}_u, \mathbf{h}_v \in \mathbb{R}^N$. Let $\mathbf{p}_{uv}$ be their standardized interaction column, and define

$$H_{uv} = [\mathbf{h}_u, \mathbf{h}_v, \mathbf{p}_{uv}]. \tag{27}$$

Suppose

$$H_{uv}^\top H_{uv} = N I_3, \qquad \Phi_A^\top H_{uv} = 0, \qquad \mathbf{z} = b_{\text{int}} \mathbf{p}_{uv} + \mathbf{r}, \qquad H_{uv}^\top \mathbf{r} = 0, \tag{28}$$

where b_{int} is the interaction coefficient and $\mathbf{r}$ is the response component orthogonal to the appended columns.

Consistent with the structural cost in Section 3, define $c_{\text{s}} = \log(1 + L)$, $c_{\text{e}} = \eta \log 2$. Thus, a Single action costs c_{s}, and the Pair action costs $2c_{\text{s}} + c_{\text{e}}$. Set

$$r = \frac{\beta N}{\alpha}, \qquad \omega = \frac{r}{1 + r}. \tag{29}$$

Let a_u and a_v denote the two Single actions, and let a_{uv} denote the Pair action. At common precisions, the Single gains are

$$\Delta^{\alpha,\beta}(a_u \mid A) = \Delta^{\alpha,\beta}(a_v \mid A) = -\frac{1}{2}\log(1 + r). \tag{30}$$

Both therefore fail the acceptance test when $\lambda, \tau \geq 0$.

The Pair gain is

$$\Delta^{\alpha,\beta}(a_{uv} \mid A) = \frac{1}{2}\left[\omega \beta N b_{\text{int}}^2 - 3\log(1 + r)\right]. \tag{31}$$

Define the fixed-precision net value by

$$V^{\alpha,\beta}(a \mid A) = \frac{\Delta^{\alpha,\beta}(a \mid A)}{N} - \lambda c(a). \tag{32}$$

The Pair passes the acceptance test $V^{\alpha,\beta}(a_{uv} \mid A) > \tau$ if and only if

$$b_{\text{int}}^2 > \frac{3\log(1 + r) + 2N\left[\lambda(2c_{\text{s}} + c_{\text{e}}) + \tau\right]}{\omega \beta N}. \tag{33}$$

If every other Single action also fails, Single-only search stops at A, whereas joint search has an eligible continuation whenever this Pair is feasible, survives screening, and is evaluated.

The interaction supplies the response alignment, but the Pair pays the evidence penalty for all three appended directions, as well as their structural cost.

The barrier can also persist when both precisions are optimized. Within this configuration, adding an orthogonal feature column whose projected response energy does not exceed the current fitted noise variance gives nonpositive profiled gain, while a sufficiently informative interaction can make the joint action pass the cost-adjusted acceptance test.

D Implementation Details

Four action scores. For Eq. (20), H_a contains only the newly appended columns. These columns are not first residualized against the current design matrix Φ_A. For compactness, write

$$P_a := P(a \mid A), \qquad c_{\mathrm{s}} = \log(1+L), \qquad c_{\mathrm{e}} = \eta \log 2.$$

Here, P_a is the regularized residual projection energy, and c_{s} and c_{e} are the incremental costs of one variable and one interaction, respectively.

The four action-specific screening scores are:

Action	Appended block H_a	n_a	Screening score $s(a \mid A)$
Single(u)	$\mathbf{h}_u$	1	$P_a/N - \lambda c_{\mathrm{s}}$
Pair(u,v)	$[\mathbf{h}_u, \mathbf{h}_v, \mathbf{p}_{uv}]$	2	$P_a/(4N) - \lambda(2c_{\mathrm{s}} + c_{\mathrm{e}})/2$
Expand(u,v)	$[\mathbf{h}_u, \mathbf{p}_{uv}]$	1	$P_a/N - \lambda(c_{\mathrm{s}} + c_{\mathrm{e}})$
Edge(u,v)	$\mathbf{p}_{uv}$	1	$P_a/N - \lambda c_{\mathrm{e}}$

For Expand, u is unselected and v is already selected. Edge adds no variable, so $n_a = \max(1, |U(a)|) = 1$.

The implementation first divides the projection energy by n_a, then normalizes by N and subtracts the action cost, and finally divides by n_a again:

$$s(a \mid A) = \frac{P_a/(Nn_a) - \lambda c(a)}{n_a} = \frac{P_a}{Nn_a^2} - \frac{\lambda c(a)}{n_a}.$$

Thus, the double normalization of the Pair projection term is part of the implemented heuristic. By comparison, the evidence-based ranking score

$$\rho(a \mid A) = \frac{V(a \mid A)}{n_a}$$

normalizes the net evidence value once.

Let $\mathcal{A}(A)$ denote the feasible actions within the screened candidate space at state A. The shortlist contains the globally highest-scoring $\min(B, |\mathcal{A}(A)|)$ actions. It neither reserves a slot for each action type nor filters actions by the sign of $s(a \mid A)$. The projection-based screening scores are not upper bounds on evidence gains.

Deterministic ordering. Variable-score ties are resolved by ascending variable identifiers; pair-score ties are resolved by ascending canonical variable pairs. Shortlisting uses a stable ascending sort on $\big(-s(a \mid A), \mathrm{kind}(a), U(a)\big)$, with action types ordered as Edge, Expand, Pair, Single. For this sorting key, $U(a)$ is represented as an ordered tuple of newly added variable identifiers. If the screening score and new-variable tuple both tie for Expand actions, their original enumeration order is retained: unselected variables follow pool order, and selected variables follow acceptance order. The implementation uses acceptance threshold $\tau = 0$. Among shortlisted actions satisfying $V(a \mid A) > 0$, the accepted action maximizes $\big(\rho(a \mid A), V(a \mid A), \Delta(a \mid A), s(a \mid A)\big)$ lexicographically. A complete tie retains the earlier shortlisted action.

Hardware and software environment for experiments. All experiments were conducted on a Linux server equipped with two Intel Xeon Platinum 8457C processors, 256 GB of system memory, and a single NVIDIA A100 PCIe GPU with 80 GB of GPU memory. The software environment used Python 3.12.7 and PyTorch 2.5.0 with CUDA 12.4.

E Experiment Results on the Omitted Backbones

Table 3 reports results for the other two forecasting backbones, PatchTST and TimeXer. The results lead to the same claim in the main text.

Table 3: Forecasting results with PatchTST and TimeXer on three real-world datasets. Selection and prediction times are in seconds. Parenthesized values are relative to Full context.

Method	Context Size	Select. Time	PatchTST			TimeXer		
			Pred. Time	MFLOPs	MAE	Pred. Time	MFLOPs	MAE
Dataset					**Beijing Air**			
Target only	1.00	0.000	0.080	1.005	9.985	0.074	0.955	9.960
Full context	12.00	0.000	0.089	1.020	9.448	0.107	1.324	9.938
ChronoEpilogi-FB	10.25 (↓15.9%)	171.238	0.086 (↓3.4%)	1.017 (↓0.3%)	9.565 (+1.24%)	0.102 (↓4.7%)	1.265 (↓4.5%)	9.970 (+0.32%)
GOGA-HDAIC	11.75 (↓2.3%)	0.906	0.088 (↓1.1%)	1.019 (↓0.1%)	9.461 (+0.14%)	0.106 (↓0.9%)	1.315 (↓0.7%)	9.947 (+0.09%)
Group Lasso	12.00 (↓0.0%)	1.141	0.090 (↑1.1%)	1.020 (↓0.0%)	9.448 (0.00%)	0.107 (↓0.0%)	1.324 (↓0.0%)	9.938 (0.00%)
ARD	12.00 (↓0.0%)	0.275	0.089 (↓0.0%)	1.020 (↓0.0%)	9.448 (0.00%)	0.106 (↓0.9%)	1.324 (↓0.0%)	9.938 (0.00%)
RMP	11.17 (↓7.5%)	1.255	0.087 (↓2.2%)	1.018 (↓0.2%)	9.508 (+0.64%)	0.105 (↓1.9%)	1.296 (↓2.1%)	9.914 (-0.24%)
ECRI	**6.83** (↓**47.0%**)	2.408	**0.085** (↓**4.5%**)	**1.013** (↓**0.7%**)	9.660 (+2.24%)	**0.094** (↓**12.1%**)	**1.150** (↓**13.1%**)	9.947 (+0.09%)
Dataset					**METR-LA**			
Target only	1.00	0.000	1.352	1.005	7.616	1.303	0.955	7.744
Full context	207.00	0.000	2.366	1.269	7.610	11.333	7.863	7.842
ChronoEpilogi-FB	18.79 (↓91.4%)	4,286.457	1.500 (↓36.6%)	1.028 (↓19.0%)	7.608 (-0.03%)	2.204 (↓80.6%)	1.551 (↓80.3%)	7.677 (-2.10%)
GOGA-HDAIC	68.67 (↓67.2%)	2,332.486	1.726 (↓27.0%)	1.092 (↓13.9%)	7.596 (-0.18%)	4.534 (↓60.0%)	3.224 (↓59.0%)	7.726 (-1.48%)
Group Lasso	25.08 (↓88.3%)	91.547	1.515 (↓36.0%)	1.036 (↓18.4%)	7.644 (+0.45%)	2.482 (↓78.1%)	1.762 (↓77.6%)	7.928 (+1.10%)
ARD	61.84 (↓70.5%)	100.873	1.694 (↓28.4%)	1.083 (↓14.7%)	7.660 (+0.66%)	4.214 (↓62.8%)	2.995 (↓61.9%)	7.724 (-1.50%)
RMP	41.51 (↓80.3%)	503.256	1.607 (↓32.1%)	1.057 (↓16.7%)	7.673 (+0.83%)	3.257 (↓71.3%)	2.313 (↓70.6%)	7.752 (-1.15%)
ECRI	**6.28** (↓**97.4%**)	165.405	**1.442** (↓**39.1%**)	**1.012** (↓**20.3%**)	**7.569** (**-0.54%**)	**1.583** (↓**86.0%**)	**1.132** (↓**85.6%**)	**7.615** (**-2.89%**)
Dataset					**PEMS-BAY**			
Target only	1.00	0.000	2.130	1.005	2.750	2.039	0.955	3.268
Full context	325.00	0.000	4.828	1.420	2.619	26.853	11.820	2.832
ChronoEpilogi-FB	46.47 (↓86.0%)	40,919.642	2.546 (↓47.3%)	1.064 (↓25.1%)	2.612 (-0.27%)	5.490 (↓79.6%)	2.480 (↓79.0%)	2.976 (+5.08%)
GOGA-HDAIC	80.77 (↓75.4%)	5,196.710	2.735 (↓43.4%)	1.108 (↓22.0%)	2.614 (-0.19%)	8.004 (↓70.2%)	3.630 (↓69.3%)	2.939 (+3.78%)
Group Lasso	73.12 (↓77.7%)	159.748	2.695 (↓44.2%)	1.098 (↓22.7%)	2.512 (-4.09%)	7.439 (↓72.3%)	3.373 (↓71.5%)	2.913 (+2.86%)
ARD	71.59 (↓78.2%)	145.219	2.740 (↓43.2%)	1.096 (↓22.8%)	2.518 (-3.86%)	7.321 (↓72.7%)	3.322 (↓71.9%)	2.960 (+4.52%)
RMP	50.36 (↓84.8%)	1,255.272	2.582 (↓46.5%)	1.069 (↓24.7%)	2.507 (-4.28%)	5.760 (↓78.5%)	2.610 (↓77.9%)	2.999 (+5.90%)
ECRI	**14.37** (↓**95.9%**)	466.803	**2.358** (↓**51.2%**)	**1.023** (↓**28.0%**)	**2.499** (**-4.58%**)	**3.113** (↓**88.4%**)	**1.403** (↓**88.1%**)	2.924 (+3.25%)

F Parameter Sensitivity Analysis

To evaluate the sensitivity of ECRI to the cost coefficient λ, we vary λ using TCN across three datasets, as shown in Figure 3. Increasing λ consistently reduces the number of selected sources and forward FLOPs, while the mean validation MAE changes mainly in an increasing trend. These results demonstrate that λ provides predictable control over pruning strength and computational cost, and that the default value $\lambda = 10^{-3}$ offers a balanced trade-off between accuracy and efficiency across the evaluated settings.

G Which Actions Are Used?

We audit the frozen per-target search traces underlying Table 1. Counts include only accepted additions and are checked against the saved variable-selection masks. The same frozen mask is reused by all four backbones; action counts are therefore reported once per target.

Table 4: Action usage in the variable selectors underlying Table 1. Each action column gives accepted actions / targets using that action. Joint-var. is the percentage of selected optional variables introduced by Pair or Expand. This counts actions, not their causal predictive benefit.

Dataset	Targets	Single	Pair	Expand	Edge	Joint-var.
Beijing	12	41 / 12	0 / 0	29 / 11	0 / 0	41.4%
METR-LA	207	718 / 196	39 / 38	297 / 159	0 / 0	34.3%
PEMS-BAY	325	2633 / 325	25 / 22	1661 / 309	1 / 1	39.4%

Expand is the most frequently accepted interaction action. Pair appears in 38 of 207 METR-LA targets and 22 of 325 PEMS-BAY targets; Edge is accepted only once across all 571 targets. Neither wind dataset accepts any interaction action. Their compression therefore does not demonstrate successful interaction-driven acquisition. Unevaluated or rejected interaction candidates can still affect the shortlist, so absence of accepted interactions alone does not establish equivalence to Single-only search.

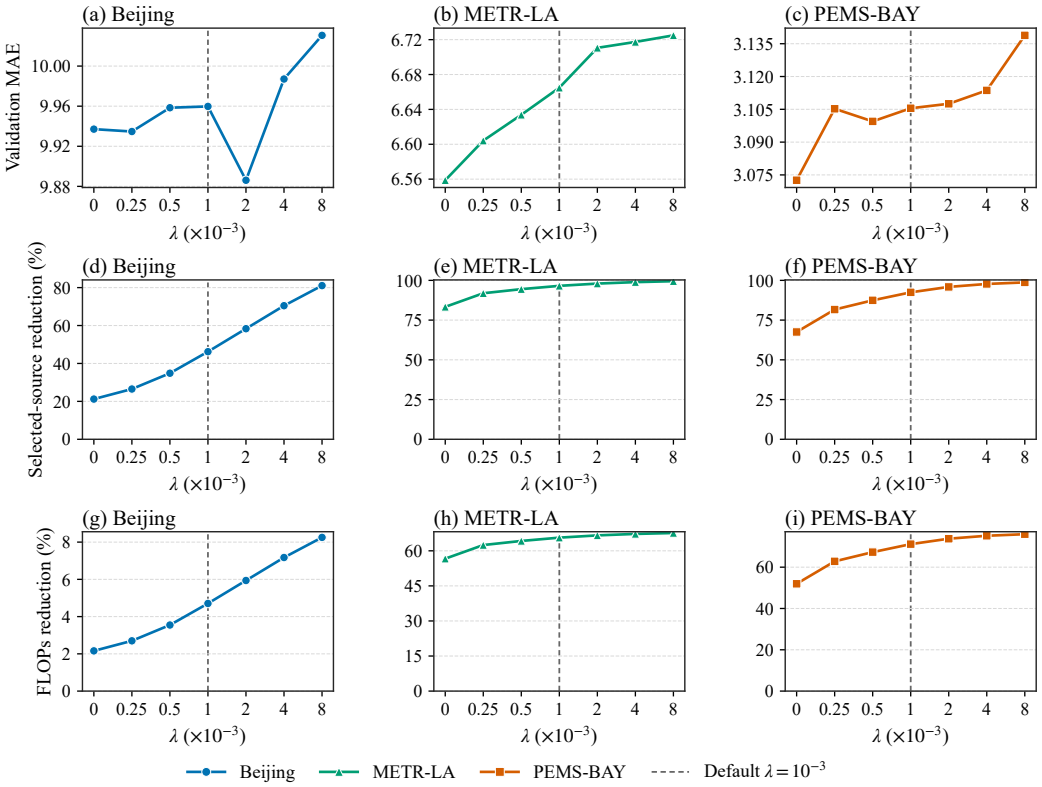


Figure 3: The sensitivity analysis on the cost coefficient λ using TCN as the backbone.

For the PEMS-BAY Single-only comparison, variable screening is held at marginal-only and every target has the same candidate pool as the reference. Restoring all actions lowers mean MAE by 0.0091, but the per-target median difference is zero and the interquartile range is $[-0.0706, 0.0592]$. ECRI has lower MAE on 158 targets, identical MAE on 16, and higher MAE on 151; the 16 identical results use identical variable-selection masks. These are observed paired differences, not confidence intervals treating network targets as independent. The selected variable counts differ between the two paths, so this comparison does not isolate interaction representation at fixed acquisition cost.

H Evidence Geometry on Real Selection Paths

Table 5: Accepted interaction actions on the frozen real-data paths. Endpoint rejection means that all newly acquired endpoints have nonpositive refitted Single scores at the same state (two for Pair, one for Expand). All-Single rejection is checked only for Pair and includes every remaining source in the pool. Fixed rejection holds the returned current-state precisions constant. Counts condition on accepted actions; they are not independent replications or downstream accuracy effects.

Dataset	Action	Count	Endpoints rejected	All Singles rejected	Fixed rejection
Beijing	Expand	29	3	–	1
METR-LA	Expand	297	119	–	1
METR-LA	Pair	39	15	3	1
PEMS-BAY	Edge	1	–	–	0
PEMS-BAY	Expand	1661	337	–	2
PEMS-BAY	Pair	25	3	0	0

Coverage and counterfactuals. We inspect the frozen main-configuration paths for all Beijing, METR-LA, and PEMS-BAY targets. Of these, 496 accept an interaction action, giving 2,052 accepted actions: 64 Pair, 1,987 Expand, and one Edge. For every included target, the production selector reproduces the recorded candidate variable pool, selected context, full action sequence, and scores. We retain the actual design and sufficient statistics constructed by that replay.

At each accepted action, we add each new variable separately to the same current selection state, retaining its existing auxiliary interactions, and refit the original evidence score with the same settings. A variable is rejected when its net Single score is nonpositive. For Pair actions with both variables rejected, we additionally evaluate every remaining Single in the candidate variable pool to test whether all would be rejected. For other Pair actions, a positive Single score for either variable already disproves that full stopping condition.

What the observed barriers establish. Both variables have rejected Single scores in 18/64 Pair actions: 15 on METR-LA and three on PEMS-BAY. A further 28 Pair actions have exactly one rejected variable; the remaining 18 have neither variable rejected. Among Expand actions, the new variable has a rejected Single score in 459/1,987 cases (3 Beijing, 119 METR-LA, 337 PEMS-BAY). Thus interaction proposals can change variable acceptance on these correlated real designs.

Only three Pair states reject *every* available Single, all on METR-LA; two of those states have no prior auxiliary interactions. Rejection of the participating variables alone is therefore not evidence that Single-only search must stop. Even the full stopping check conditions on ECRI's current state, which a separate Single-only trajectory need not visit.