SCEU: STEP-CONDITIONED EXOGENOUS UTILITY LEARNING FOR TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Exogenous variables provide valuable forecasting context, but their usefulness can vary across variables and future prediction steps. Most existing methods organize cross-variable information through channel-independent processing, channel-dependent interaction, or selective clustering, but these interaction strategies alone do not specify how exogenous information should be weighted at each future prediction step. To address this limitation, we propose **SCEU**, a **S**tep-**C**onditioned **E**xogenous **U**tility learning framework for time series forecasting. SCEU constructs an endogenous-only base forecast, assigns relative Channel Utility scores to exogenous variable-scale representations, and derives Temporal Utility representations from multi-scale endogenous historical patches. A gated bounded residual pathway selectively applies the correction obtained by fusing the two branches while controlling its magnitude at each forecast step. The learned utility weights serve as model-internal, step-conditioned scores for weighting endogenous and exogenous representations. Across diverse benchmarks, SCEU achieves strong forecasting performance. Ablation studies show that Channel Utility, Temporal Utility, and the adaptive response range generally improve forecasting accuracy in the evaluated settings. Overall, SCEU provides a step-conditioned mechanism for selectively utilizing exogenous information in time series forecasting. The code is available at https://anonymous.4open.science/r/sceu-EBCE.

1 INTRODUCTION

Time series forecasting plays a critical role in decision-making across financial, energy, transportation, and healthcare applications (Lu & Xu, 2024; Lago et al., 2021; Lin et al., 2024a; Wang et al., 2023; Williams, 2001; Vagropoulos et al., 2016; Karevan & Suykens, 2020; Wu et al., 2023b; Bao et al., 2017; Zeger et al., 2006). However, the endogenous target sequence typically captures only a partial view of the underlying system dynamics. External factors, such as weather conditions, calendar events, market indicators, and operational controls, can provide complementary information beyond the target's own history. Nevertheless, the predictive usefulness of exogenous information can vary across variables and forecast steps. A variable that provides valuable information for one future step may become redundant or even introduce noise for another step under the same observed system state and historical context. Therefore, effective exogenous forecasting requires not only modeling dependencies between variables, but also determining which external information to use and how to weight it at each forecast step.

Existing forecasting models differ mainly in how they represent and organize cross-variable interactions. The Channel-Independent (CI) strategy processes each variable separately, reducing potential interference among channels but limiting the explicit use of cross-variable dependencies (Zeng et al., 2023; Nie et al., 2023; Lin et al., 2024b; Xu et al., 2024; Stitsyuk & Choi, 2025; Dai et al., 2024a). In contrast, the Channel-Dependent (CD) strategy aggregates information across channels through attention or cross-variable transformations, enabling richer interactions but potentially propagating irrelevant signals and producing over-mixed representations (Ekambaram et al., 2023; Wu et al., 2020; Yi et al., 2023; Zhang & Yan, 2023; Liu et al., 2024b; Wu et al., 2021a; Zhou et al., 2022; Wu et al., 2023a; Cai et al., 2024; Wang et al., 2024a; Cirstea et al., 2022a; Liu et al., 2024a). Channel clustering, exemplified by the Channel Clustering Module (CCM), provides a selective alternative by organizing channels into data-dependent groups and structuring their interactions (Chen et al., 2024; Qiu et al., 2025c; Hu et al., 2025; Wu et al., 2021b; Qiu et al., 2025a). As detailed in Figure 1,

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

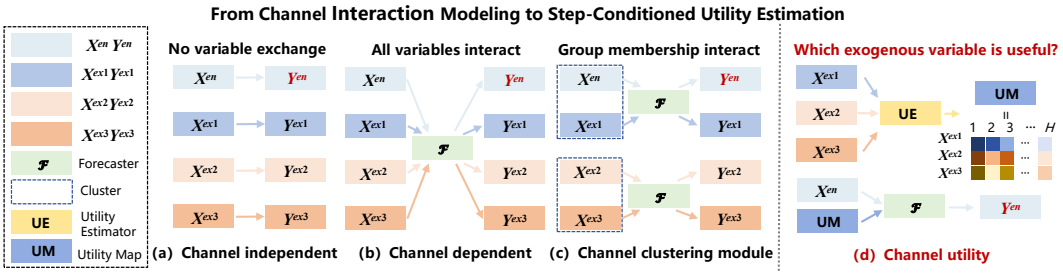

Figure 1: Comparison of channel interaction and utility learning mechanisms. Channel-Independent (CI) models process channels independently, Channel-Dependent (CD) models enable cross-channel interaction, and Channel Clustering Module (CCM) performs group-wise interaction. SCEU adds step-conditioned utility scores to adapt exogenous representation weighting to forecast positions.

CI, CD, and CCM address whether and how channels should interact, but these interaction strategies alone do not specify how exogenous information should be weighted at each future prediction step.

Recent exogenous forecasting methods improve interactions between endogenous and exogenous variables through variable selection, basis expansion, cross-attention, linear mappings, or dependency discovery (Lim et al., 2021; Olivares et al., 2023; Das et al., 2023; Tayal et al., 2024; Wang et al., 2024c; Zhou et al., 2025; Qiu et al., 2026; Li et al., 2026; Bennett et al., 2014). These approaches provide increasingly structured representations of cross-variable relationships and improve the ability to incorporate external information. Some support input-dependent variable selection. However, modeling dependency is not equivalent to identifying whether exogenous information provides additional predictive value beyond the endogenous history. A strongly correlated exogenous variable may provide little additional information beyond the endogenous history, while a globally important variable may only contribute to specific regions of the forecasting horizon. When relevance varies across forecast steps, applying horizon-invariant relevance weights may incorporate irrelevant information and reduce forecasting accuracy. In multi-step forecasting, a model predicts multiple future values from the information available at a common forecast origin (Taieb et al., 2012; Hyndman & Koehler, 2006). Thus, exogenous information should be weighted for each prediction step, rather than assumed to provide consistent predictive value throughout the forecasting horizon.

Consider electricity-load forecasting as an example. Temperature information available at the forecast origin may provide predictive information during extreme weather conditions but become less informative under mild conditions. Moreover, even under the same system state, its contribution to predicting load one hour ahead may differ from its contribution to predicting load 24 hours ahead. Thus, the usefulness of exogenous information can depend jointly on the current system state and the forecast position. Instead of assigning a fixed relevance score across the forecasting horizon, a forecasting model should adaptively weight external information for different forecast positions. This requirement introduces three challenges: *(1) learning exogenous representation weights conditioned on the current target state; (2) adapting information weighting to different forecast positions; and (3) controlling the magnitude of corrections to an endogenous-only base forecast.*

To address these challenges, we propose **SCEU**, a Step-Conditioned Exogenous Utility learning framework. In the standard setting, SCEU uses historical target and exogenous observations together with future exogenous trajectories assumed to be known at forecasting time. SCEU constructs an endogenous-only linear base forecast, which is refined through a selective correction derived jointly from endogenous and exogenous representations. The **Channel Utility** branch learns utility scores for exogenous variable-scale representations for different prediction steps under the current system state, combining relative importance with bounded signed modulation to aggregate these representations. Meanwhile, the **Temporal Utility** branch retrieves endogenous information from multiple temporal scales to provide step-dependent temporal evidence. The learned utility scores are model-internal relative weighting quantities over representations, optimized jointly with the forecasting objective without additional utility annotations. They should not be interpreted as identifiable estimates of individual variables' incremental predictive value. Finally, a gated bounded residual pathway controls the strength and magnitude of the correction obtained by fusing the two branches. SCEU thereby combines cross-variable interaction with step-conditioned representation weighting and bounded forecast correction. Our **contributions** are summarized as follows:

- We develop a step-conditioned utility learning approach to exogenous information utilization, where external representations are adaptively weighted for effective multi-step forecasting according to the current system state and forecast position.
- We propose SCEU with two complementary utility branches. Channel Utility learns relative scores over exogenous variable-scale representations, while Temporal Utility captures step-dependent temporal evidence from multi-scale endogenous history.
- We introduce a gated bounded residual correction mechanism that retains an endogenous-only forecast as an additive base and adaptively controls the magnitude of the fused correction. Experiments on diverse benchmarks show competitive accuracy across different forecasting scenarios, and component ablations show that the utility branches and adaptive response range generally improve forecasting accuracy in the evaluated settings.

2 Related Work

2.1 Channel Modeling Strategies

Channel modeling in multivariate forecasting has evolved from isolation to selective interaction. Channel-Independent (CI) methods, such as DLinear and PatchTST, process each variable independently to preserve channel-specific patterns and avoid interference (Zeng et al., 2023; Nie et al., 2023; Lin et al., 2024b; Xu et al., 2024; Challu et al., 2023). Channel-Dependent (CD) methods instead mix variables through MLPs, graph propagation, or attention to exploit cross-channel information (Ekambaram et al., 2023; Wu et al., 2020; Yi et al., 2023; Zhang & Yan, 2023; Liu et al., 2024b; Cirstea et al., 2019; Chen et al., 2023; Cai et al., 2024; Cirstea et al., 2022a; Zhao et al., 2023; Cirstea et al., 2022b; Wu et al., 2024). More recently, CCM groups related channels and restricts information exchange within learned clusters, providing a compromise between isolation and indiscriminate mixing (Chen et al., 2024; Qiu et al., 2025c; Hu et al., 2025; Qiu et al., 2025a). These strategies determine whether and how channels interact, whereas SCEU explores a different question: which exogenous information is useful for predicting the target at a forecast position.

2.2 Forecasting with Exogenous Variables

Forecasting with exogenous variables distinguishes the target series from auxiliary covariates. Earlier neural models incorporate historical, static, or known-future covariates through basis expansion, variable selection, or dense encoder-decoder structures (Lim et al., 2021; Olivares et al., 2023; Das et al., 2023; Tayal et al., 2024; Williams, 2001; Vagropoulos et al., 2016). Recent methods focus on endogenous-exogenous interaction. TimeXer combines endogenous patch-wise self-attention with variate-wise cross-attention (Wang et al., 2024c); CrossLinear introduces a lightweight cross-correlation embedding (Zhou et al., 2025); and DAG discovers and injects correlations along temporal and channel dimensions to incorporate historical and known future exogenous information (Qiu et al., 2026; Li et al., 2026). While these methods improve the representation and transfer of external information, strong cross-variable dependence does not necessarily imply that all external information should receive equal relevance across prediction steps. SCEU learns step-conditioned relative scores for exogenous representations and controls the correction magnitude from fused Channel Utility and Temporal Utility representations to refine an endogenous-only base forecast.

3 Preliminaries

Let T and H denote the lookback length and forecasting horizon, respectively. Given a multivariate time series with N variables, we denote the historical observations as $\mathbf{X}_{1:T,N} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T] \in \mathbb{R}^{N\times T}$, where $\mathbf{x}_t = [v_{t,1}, v_{t,2}, \ldots, v_{t,N}]^\top \in \mathbb{R}^N$ represents the observation vector at time step t.

For forecasting with exogenous variables, the observed sequence is divided into an *endogenous target sequence* $\mathbf{X}^{\text{en}} \in \mathbb{R}^{1\times T}$, which contains the historical observations of the endogenous target variable, and an exogenous sequence $\mathbf{X}^{\text{ex}}_{\text{his}} \in \mathbb{R}^{C\times T}$, where $C = N - 1$ denotes the number of exogenous variables. We assume that the corresponding future exogenous trajectories $\mathbf{X}^{\text{ex}}_{\text{fut}} \in \mathbb{R}^{C\times H}$ are known at forecasting time. **The objective is to predict the future endogenous sequence:**

$$\widehat{\mathbf{Y}}^{\text{en}} = \mathcal{M}(\mathbf{X}^{\text{en}}, \mathbf{X}^{\text{ex}}_{\text{his}}, \mathbf{X}^{\text{ex}}_{\text{fut}}) \in \mathbb{R}^{1\times H}. \tag{1}$$

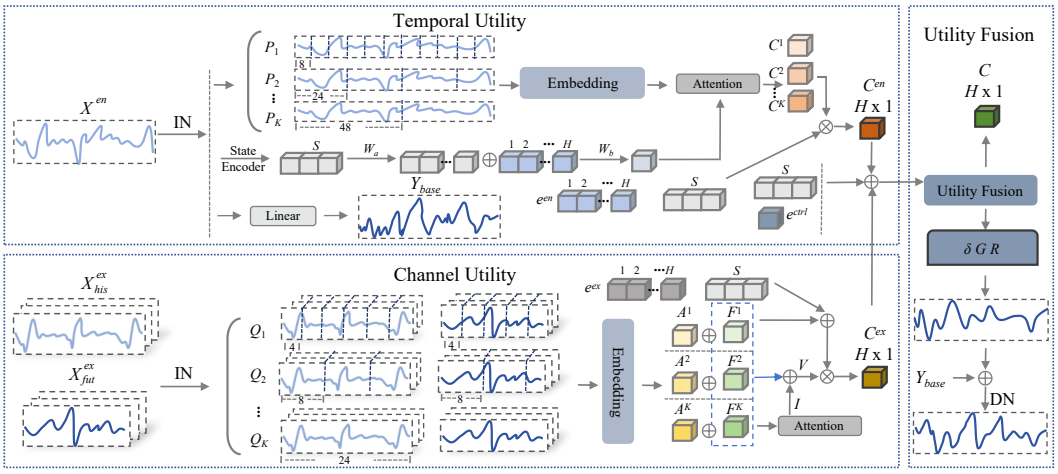

Figure 2: Overview of SCEU. Variables are normalized within the input window. A linear endogenous-only forecast is corrected using Temporal Utility over multi-scale endogenous historical patches and Channel Utility over forecast-step-dependent variable-scale representations. Utility fusion, a correction gate, and an adaptive response range regulate the correction process.

Here, $\mathcal{M}$ denotes the forecasting model, and forecast step $h \in \{1, \ldots, H\}$ corresponds to $T + h$.

Unlike multivariate forecasting, where multiple variables are considered forecasting targets, exogenous forecasting focuses on predicting an endogenous target variable while treating other variables as auxiliary information. These exogenous variables provide historical observations and, when available, known future covariates (Wang et al., 2024c; Zhou et al., 2025; Qiu et al., 2026).

4 Methodology

4.1 Model Overview

Given endogenous history $\mathbf{X}^{\text{en}} \in \mathbb{R}^{1\times T}$, historical exogenous variables $\mathbf{X}^{\text{ex}}_{\text{his}} \in \mathbb{R}^{C\times T}$, and known future trajectories $\mathbf{X}^{\text{ex}}_{\text{fut}} \in \mathbb{R}^{C\times H}$, SCEU predicts the future endogenous sequence over the forecasting horizon H. Figure 2 summarizes the model: an endogenous-only linear forecast is refined by fused Temporal and Channel Utility representations, while a correction gate and a step-adaptive response bound regulate the fused correction at each forecast step. The complete training and inference procedure is provided in Appendix A.

For a variable $\mathbf{x} \in \mathbb{R}^T$, instance-wise normalization is:

$$\widetilde{\mathbf{x}} = \frac{\mathbf{x} - \mu(\mathbf{x})}{\sigma(\mathbf{x}) + \epsilon}, \qquad \epsilon > 0. \tag{2}$$

Here, $\mu(\mathbf{x})$ and $\sigma(\mathbf{x})$ denote the mean and standard deviation computed over that variable's historical input window. Future exogenous trajectories are normalized using historical statistics, and the output is de-normalized using endogenous statistics (Passalis et al., 2020; Liu et al., 2022; Fan et al., 2023; Dai et al., 2024b). The current state is encoded from the latest $W \le T$ normalized target values:

$$\mathbf{s} = \mathrm{LN}\Big(\mathrm{GELU}\Big(\mathbf{W}_s\,\mathrm{Vec}(\widetilde{\mathbf{X}}^{\text{en}}_{T-W+1:T}) + \mathbf{b}_s\Big)\Big) \in \mathbb{R}^D. \tag{3}$$

Here, $\mathbf{W}_s \in \mathbb{R}^{D\times W}$ and $\mathbf{b}_s \in \mathbb{R}^D$. D denotes the hidden dimension, LN denotes layer normalization, and Vec converts the selected target values into a column vector. All utility estimation and correction control modules share the same state representation $\mathbf{s}$.

4.2 Endogenous-Only Base Forecasting

A learnable linear projection produces the endogenous-only base forecast in the normalized space:

$$\widehat{\mathbf{Y}}^{\text{base}} = \mathcal{P}_{\text{base}}(\widetilde{\mathbf{X}}^{\text{en}}) \in \mathbb{R}^{1\times H}. \tag{4}$$

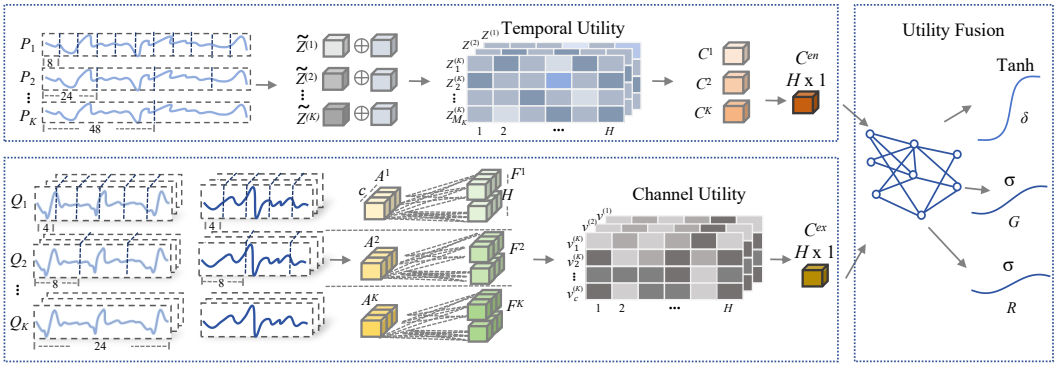

Figure 3: Detailed architecture of the utility modeling and correction pipeline. Temporal and Channel Utility estimate step-dependent representations from multi-scale endogenous historical patches and exogenous variable-scale inputs, respectively. The resulting utility representations are fused to generate gated and bounded forecast corrections.

It is modified only through the bounded correction pathway rather than being directly replaced by external information (Zeng et al., 2023; Lin et al., 2024b; Xu et al., 2024; Stitsyuk & Choi, 2025).

4.3 Temporal Utility

Figure 3 illustrates the utility modeling and correction pipeline of SCEU. The upper branch corresponds to Temporal Utility, which models historical endogenous patterns across multiple resolutions. For multi-scale endogenous patch lengths $\mathcal{P}_{\text{en}} = \{P_1, \ldots, P_K\}$, scale k retains $M_k = \lfloor T/P_k \rfloor$ non-overlapping patches. Patch $\mathbf{p}_i^{(k)} \in \mathbb{R}^{P_k}$ is embedded and contextualized as:

$$\mathbf{z}_i^{(k)} = \mathbf{W}_p^{(k)} \mathbf{p}_i^{(k)} + \mathbf{e}_k^{\text{en,scale}}, \tag{5}$$

$$\widetilde{\mathbf{Z}}^{(k)} = \text{Enc}^{(k)}([\mathbf{z}_1^{(k)}, \ldots, \mathbf{z}_{M_k}^{(k)}]). \tag{6}$$

Here, $\mathbf{e}_k^{\text{en,scale}}$ denotes the scale embedding, and $\text{Enc}^{(k)}$ denotes the encoder at scale k. The multi-scale construction captures both local variations and long-term temporal structures (Nie et al., 2023; Wang et al., 2024a; Hu et al., 2024; Yu et al., 2023; Wu et al., 2025b). Short patches preserve recent variations, whereas longer patches capture slower temporal dynamics. This design avoids imposing a single temporal resolution across different forecasting steps. For forecast step h, a state-conditioned query is:

$$\mathbf{q}_h = \mathbf{W}_b(\mathbf{e}_h^{\text{en}} + \mathbf{W}_a \mathbf{s}). \tag{7}$$

The step embedding encodes the relative forecasting horizon, while $\mathbf{s}$ conditions the query on the current operating regime. Their combination allows different historical patterns to contribute to different forecasting steps. Let $\mathbf{k}_i^{(k)}$ and $\mathbf{v}_i^{(k)}$ denote the key and value projections of the contextualized patch i at scale k. With the multi-head index omitted, the patch-level utility weighting is:

$$a_{h,i}^{(k)} = \text{Softmax}_i\left(\frac{\mathbf{q}_h^\top \mathbf{k}_i^{(k)}}{\sqrt{d}}\right), \qquad \mathbf{c}_h^{(k)} = \sum_{i=1}^{M_k} a_{h,i}^{(k)} \mathbf{v}_i^{(k)}. \tag{8}$$

Here, d denotes the query/key dimension of an attention head, and the Softmax is taken over the M_k historical patches at scale k. The attention operation follows standard scaled dot-product attention (Vaswani et al., 2017; Zhou et al., 2021; Wu et al., 2021a; Zhou et al., 2022; Liu et al., 2021). With $\bar{\mathbf{c}}_h = K^{-1} \sum_k \mathbf{c}_h^{(k)}$, the scale weights and final Temporal Utility representation are:

$$\boldsymbol{\beta}_h = \text{Softmax}(\text{MLP}_{\text{scale}}([\mathbf{s}; \mathbf{e}_h^{\text{en}}; \bar{\mathbf{c}}_h])), \tag{9}$$

$$\mathbf{c}_h^{\text{en}} = \mathbf{W}_O^{\text{en}} \sum_{k=1}^{K} \beta_{h,k} \mathbf{c}_h^{(k)}. \tag{10}$$

The notation $[\cdot\,;\,\cdot]$ denotes concatenation, and $\boldsymbol{\beta}_h \in \mathbb{R}^K$ contains weights normalized across scales. The learned scale weights depend on both the current state and the forecasting step, enabling the model to adaptively weight temporal resolutions for different forecasting horizons.

4.4 CHANNEL UTILITY

As shown in the lower branch of Figure 3, Channel Utility learns forecast-step-dependent relative utility scores over exogenous variables and temporal scales. For exogenous patch lengths $\mathcal{P}_{\text{ex}} = \{Q_1, \ldots, Q_K\}$, historical and future trajectories are partitioned at each scale. The right boundary is padded with the last observed value when the sequence length is not divisible by the patch length. For variable c, patch j, and scale k, the corresponding token is:

$$\mathbf{z}_{c,j}^{(k)} = \mathbf{W}_{p,\text{ex}}^{(k)} \mathbf{p}_{c,j}^{(k)} + \mathbf{e}_j^{\text{pos}} + \mathbf{e}_c^{\text{var}} + \mathbf{e}_k^{\text{ex,scale}}. \tag{11}$$

The three embeddings encode patch position, variable identity, and temporal scale, respectively. A scale-specific Transformer encodes patches of each exogenous variable independently with shared parameters (Wang et al., 2024c; Zhou et al., 2025; Zhang & Yan, 2023; Liu et al., 2024b).

After scale-specific encoding, the historical tokens are summarized to provide variable-specific context. If $\widetilde{\mathbf{z}}_{c,j}^{(k)}$ denotes an encoded historical token and M_k^{his} is the number of historical patches, then:

$$\mathbf{a}_c^{(k)} = \text{MLP}_{\text{his}}\left(\left[\frac{1}{M_k^{\text{his}}} \sum_{j=1}^{M_k^{\text{his}}} \widetilde{\mathbf{z}}_{c,j}^{(k)}; \widetilde{\mathbf{z}}_{c,M_k^{\text{his}}}^{(k)} \right] \right). \tag{12}$$

And future step h uses $\mathbf{f}_{h,c}^{(k)} = \widetilde{\mathbf{z}}_{c,\lceil h/Q_k \rceil}^{(k)}$. Variable-wise self-attention models cross-variable dependencies before step-conditioned utility scoring:

$$\mathbf{I}_h^{(k)} = \text{MHA}_{\text{var}}(\mathbf{F}_h^{(k)} + \mathbf{A}^{(k)}, \mathbf{F}_h^{(k)} + \mathbf{A}^{(k)}, \mathbf{F}_h^{(k)} + \mathbf{A}^{(k)}). \tag{13}$$

Here, $\mathbf{F}_h^{(k)}, \mathbf{A}^{(k)} \in \mathbb{R}^{C \times D}$ collect the future tokens and historical summaries. For the row $\mathbf{i}_{h,c}^{(k)}$ of $\mathbf{I}_h^{(k)}$, the value path retains both cross-variable and variable-specific information:

$$\mathbf{v}_{h,c,k}^{\text{ex}} = \mathbf{W}_V^{\text{ex}} [\mathbf{i}_{h,c}^{(k)}; \mathbf{f}_{h,c}^{(k)}]. \tag{14}$$

The scoring feature $\mathbf{g}_{h,c,k} = [\mathbf{f}_{h,c}^{(k)}; \mathbf{a}_c^{(k)}; \mathbf{s}; \mathbf{e}_h^{\text{ex}}]$ combines future, historical, state, and forecast-step information. Importance and direction yield the signed utility score:

$$a_{h,c,k}^{\text{ex}} = \text{Softmax}_{c,k}(\text{MLP}_a(\mathbf{g}_{h,c,k})), \tag{15}$$

$$d_{h,c,k}^{\text{ex}} = \tanh(\text{MLP}_d(\mathbf{g}_{h,c,k})), \tag{16}$$

$$u_{h,c,k}^{\text{ex}} = a_{h,c,k}^{\text{ex}} d_{h,c,k}^{\text{ex}}. \tag{17}$$

The Softmax covers all $C \times K$ variable–scale combinations for each forecast step (Chen et al., 2024; Qiu et al., 2025c; Hu et al., 2025; Wu et al., 2025a). Unlike generic attention, the proposed utility score combines relative importance with bounded signed modulation to capture conditional predictive usefulness across variable–scale representations. The bounded direction is latent and does not imply causality. The resulting Channel Utility representation is:

$$\mathbf{c}_h^{\text{ex}} = \mathbf{W}_O^{\text{ex}} \sum_{c=1}^{C} \sum_{k=1}^{K} u_{h,c,k}^{\text{ex}} \mathbf{v}_{h,c,k}^{\text{ex}}. \tag{18}$$

4.5 UTILITY FUSION AND GATED BOUNDED CORRECTION

As illustrated on the right of Figure 3, the Temporal and Channel Utility representations are fused to generate a step-specific correction to the endogenous-only base forecast. The fusion weights are conditioned on the current state, the two utility representations, and their element-wise interaction:

$$[\gamma_h^{\text{en}}, \gamma_h^{\text{ex}}] = \text{Softmax}(\text{MLP}_{\text{fus}}[\mathbf{c}_h^{\text{en}}; \mathbf{c}_h^{\text{ex}}; \mathbf{s}; \mathbf{c}_h^{\text{en}} \odot \mathbf{c}_h^{\text{ex}}]), \tag{19}$$

$$\mathbf{c}_h^{\text{util}} = \gamma_h^{\text{en}} \mathbf{c}_h^{\text{en}} + \gamma_h^{\text{ex}} \mathbf{c}_h^{\text{ex}}, \tag{20}$$

$$\delta_h = \tanh(\text{MLP}_\delta(\mathbf{c}_h^{\text{util}})) \in [-1, 1]. \tag{21}$$

Here, Softmax normalizes over the two branches, and $\odot$ denotes element-wise multiplication. A control-specific step embedding determines how strongly this correction affects the base forecast:

$$g_h = \sigma\big(\text{MLP}_g[\mathbf{c}_h^{\text{en}}; \mathbf{c}_h^{\text{ex}}; \mathbf{s}; \mathbf{e}_h^{\text{ctrl}}]\big) \in [0, 1]. \tag{22}$$

Here and in Equation 23, σ denotes the logistic sigmoid function. The gate network is initialized to produce values close to zero, so that training starts close to the endogenous-only base forecast (Du et al., 2021; Cheng et al., 2023; Wang et al., 2024b). To further control the correction magnitude, SCEU learns a step-adaptive response bound. Given fixed $0 \le r_{\min} < r_{\max}$,

$$r_h = r_{\min} + (r_{\max} - r_{\min})\sigma\big(\mathrm{MLP}_r[\mathbf{s}; \mathbf{c}_h^{\mathrm{ex}}; \mathbf{e}_h^{\mathrm{ctrl}}]\big)\,, \tag{23}$$

$$\widehat{y}_h^{\mathrm{norm}} = \widehat{y}_h^{\mathrm{base}} + g_h r_h \delta_h, \qquad h = 1, \dots, H. \tag{24}$$

Thus, g_h controls the correction strength, while r_h bounds its step-specific magnitude. Finally, instance-wise de-normalization restores the forecast to the original scale. De-normalization uses the historical endogenous statistics from Equation 2. Formal properties and proofs of the utility and correction pathways are provided in Appendix B.

4.6 Forecasting Objective

SCEU is trained end to end using only the forecasting loss, without explicit utility supervision or auxiliary losses. For a batch of size B, we minimize the mean absolute error:

$$\mathcal{L} = \frac{1}{BH} \sum_{b=1}^{B} \sum_{h=1}^{H} \left| y_{b,h}^{\mathrm{en}} - \widehat{y}_{b,h}^{\mathrm{en}} \right|. \tag{25}$$

The learned utilities should therefore be interpreted as model-internal, step-conditioned scores (Qiu et al., 2025b; Zhang et al., 2024). Because the exogenous value representations incorporate cross-variable interactions, the utility assigned to a variable weights its context-dependent representation.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate SCEU on 13 real-world datasets: Weather, Electricity, Traffic, ETTh1, ETTh2, ETTm1, ETTm2, BE, DE, Energy, FR, NP, and PJM (Lago et al., 2021; Qiu et al., 2026; Wang et al., 2024c; Olivares et al., 2023). Following DAG (Qiu et al., 2026), the last channel is treated as the endogenous target and the others as exogenous variables. Dataset statistics and chronological splits are summarized in Appendix C.

Baselines and Implementation. We compare SCEU with DAG (Qiu et al., 2026), TimeXer (Wang et al., 2024c), TFT (Lim et al., 2021), TiDE (Das et al., 2023), DUET (Qiu et al., 2025c), CrossLinear (Zhou et al., 2025), and PatchTST (Nie et al., 2023). For Weather, Electricity, Traffic, and ETT, we use a lookback window of 96 and prediction horizons of $\{96, 192, 336, 720\}$. For the remaining datasets, the main comparisons use lookback windows of 168 and 720 for $H = 24$ and $H = 360$, respectively, following DAG (Qiu et al., 2026). Dataset-specific configurations follow standard forecasting protocols, with consistent evaluation within each dataset. We report normalized MSE and MAE. Detailed model configurations and training settings are provided in Appendix D. For Traffic, we use an 8-factor latent variant to efficiently process its 861 exogenous variables. Details, complexity analysis, and an accuracy comparison are provided in Appendix E. For ETT, we use the complete original files rather than the conventional 20-month truncation, resulting in a longer and more heterogeneous evaluation interval. Therefore, these results are not directly comparable with those under the truncated protocol.

5.2 Main Results

Table 1 reports the average results, horizon-wise results and qualitative examples are provided in Appendix F and Appendix I, respectively. Overall, SCEU achieves the best MSE and MAE on 9 of 13 datasets, demonstrating strong performance across diverse forecasting scenarios. Compared with DAG, SCEU reduces MSE and MAE by 30.2% and 18.2% on Traffic, and reduces MSE by 27.8% and 14.0% on ETTh1 and ETTm1, respectively. The strong ETT results are obtained using the complete original sequences rather than the commonly used 20-month truncation. SCEU also achieves the best overall performance on Weather, Electricity, BE, DE, Energy, and NP, and the best MSE on PJM. These results demonstrate the effectiveness of step-conditioned exogenous utility learning across datasets with diverse characteristics.

Table 1: Average forecasting results. Best results are shown in red and second-best results in blue with underlining. Lower is better. Apparent ties can result from rounding to three decimals.

	SCEU		DAG		TimeXer		TFT		TiDE		DUET		CrossLinear		PatchTST	
Dataset	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Weather	0.001	0.021	0.002	0.026	0.001	0.027	0.002	0.031	0.002	0.030	0.001	0.029	0.008	0.074	0.009	0.076
Electricity	0.193	0.321	0.194	0.329	0.208	0.337	0.375	0.471	0.313	0.417	0.281	0.411	0.264	0.399	0.269	0.404
Traffic	0.090	0.166	0.129	0.203	0.142	0.238	0.132	0.220	0.154	0.255	0.263	0.322	0.232	0.291	0.233	0.295
ETTh1	0.104	0.252	0.144	0.297	0.135	0.284	0.163	0.312	0.127	0.277	0.134	0.283	0.134	0.282	0.141	0.291
ETTh2	0.327	0.442	0.434	0.511	0.338	0.452	0.366	0.485	0.317	0.435	0.336	0.446	0.326	0.443	0.327	0.447
ETTm1	0.080	0.208	0.093	0.222	0.091	0.221	0.127	0.265	0.096	0.229	0.091	0.219	0.091	0.220	0.092	0.221
ETTm2	0.225	0.340	0.214	0.331	0.254	0.345	0.443	0.512	0.225	0.340	0.225	0.335	0.219	0.333	0.226	0.342
BE	0.420	0.274	0.424	0.280	0.452	0.290	0.454	0.291	0.498	0.325	0.515	0.354	0.479	0.337	0.577	0.432
DE	0.355	0.370	0.370	0.370	0.475	0.418	0.489	0.446	0.499	0.447	0.482	0.430	0.485	0.452	0.501	0.455
Energy	0.121	0.266	0.124	0.268	0.163	0.315	0.130	0.283	0.153	0.302	0.203	0.367	0.239	0.402	0.226	0.377
FR	0.422	0.232	0.415	0.224	0.428	0.241	0.509	0.264	0.485	0.282	0.496	0.327	0.483	0.298	0.588	0.410
NP	0.322	0.319	0.362	0.344	0.418	0.371	0.379	0.375	0.443	0.400	0.411	0.408	0.371	0.387	0.390	0.396
PJM	0.091	0.181	0.094	0.181	0.108	0.199	0.114	0.212	0.142	0.247	0.107	0.197	0.112	0.223	0.133	0.259

Table 2: Ablation results. BE, Energy, NP, and PJM are evaluated at $H = 24$; Electricity and Weather are evaluated at $H = 96$. Weather values are multiplied by 10 for readability.

	BE		Energy		NP		PJM		Electricity		Weather ($\times 10$)	
Variant	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Base	0.437	0.292	0.126	0.276	0.297	0.304	0.095	0.195	0.410	0.446	0.011	0.238
w/o Temporal Utility	0.338	0.232	0.079	0.219	0.204	0.231	0.062	0.150	0.176	0.304	0.007	0.185
w/o Channel Utility	0.402	0.273	0.121	0.269	0.274	0.284	0.087	0.183	0.187	0.312	0.011	0.227
w/o Adaptive Range	0.356	0.244	0.081	0.221	0.221	0.241	0.067	0.154	0.220	0.325	0.008	0.197
Full	0.351	0.228	0.077	0.213	0.187	0.216	0.058	0.146	0.161	0.294	0.007	0.176

5.3 Model Analyses

5.3.1 Ablation Studies

Table 2 reports the component ablations at $H = 24$ for BE, Energy, NP, and PJM, and $H = 96$ for Weather and Electricity. All variants use identical settings except for the ablated component. Full SCEU achieves the best result on 11 of 12 metrics, demonstrating the effectiveness of the components in step-conditioned utility learning and adaptive correction control.

Removing Channel Utility degrades accuracy, especially on Weather and Electricity. On Electricity, MSE and MAE increase from 0.161/0.294 to 0.187/0.312, with similar degradation on Weather. These results demonstrate the effectiveness of Channel Utility in organizing exogenous information. Temporal Utility improves metrics through multi-scale target history, while the Adaptive Response Range controls correction magnitude. Together, the three components integrate historical and exogenous information. Extended ablations across datasets and horizons are provided in Appendix G.

5.3.2 Parameter Sensitivity

We further examine the sensitivity of SCEU to the response upper bound, state window, number of temporal scales, and lookback window using a one-factor-at-a-time protocol (Wang et al., 2024c; Zhou et al., 2025; Qiu et al., 2026; Wang et al., 2024a). To display the four datasets on a common ordinate, the plotted MSE values of ETTh1 and Energy are shifted upward by 0.2. Filled markers identify configurations reused from the main experiments and ablations.

As shown in Figure 4, BE and Energy remain relatively stable across response upper bounds, whereas NP benefits from a larger bound. The preferred state window and number of scales vary across datasets. Increasing the lookback window generally improves ETTh1, NP, and BE, while Energy saturates at $T = 168$. These results indicate that the preferred configuration is dataset dependent, with longer historical contexts being generally beneficial within the tested range. Additional sensitivity results on DE, PJM, and FR can be found in Appendix J.

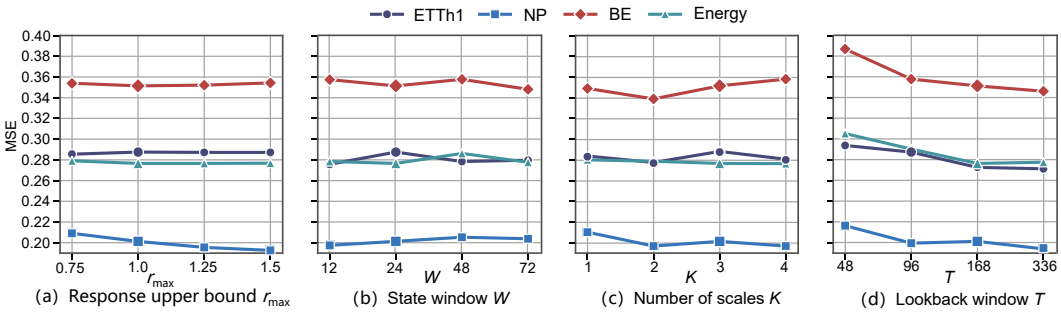

Figure 4: Parameter sensitivity of SCEU. Panels (a)–(d) vary $r_{\max}$, W, K, and T, respectively. ETTh1 uses $H = 96$; NP, BE and Energy use the shorter $H = 24$ horizon.

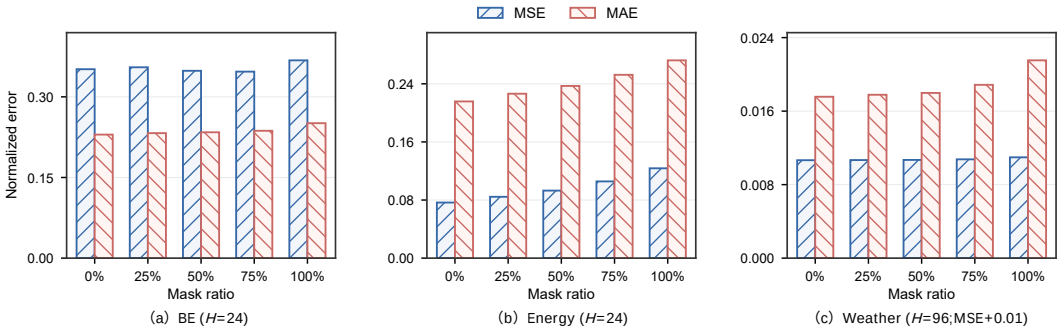

Figure 5: Forecasting performance under different levels of exogenous information missingness. BE and Energy are evaluated at $H = 24$, and Weather at $H = 96$. For visualization only, the Weather MSE values are shifted upward by 0.01; all analyses use the original values.

5.3.3 Forecasting with Missing Exogenous Information

A practical concern is whether SCEU remains applicable when exogenous observations over the forecasting horizon are incomplete or unavailable. We randomly mask 25%, 50%, 75%, and 100% of future exogenous inputs while keeping the target and exogenous histories observed, training and evaluating separately at each missingness level. Missing values are zeroed after normalization, with an availability encoding provided to the exogenous branch.

As detailed in Figure 5, SCEU remains effective as future exogenous information becomes incomplete. BE remains stable across missing ratios, while Weather shows limited degradation under partial missingness. Even when all forecast-horizon exogenous observations are unavailable, SCEU still produces meaningful forecasts using historical information. Overall, SCEU benefits from future exogenous information when available while remaining applicable when it is partially or completely unavailable. Additional results are provided in Appendix H.

6 Conclusion

We propose SCEU, a step-conditioned exogenous utility learning framework that jointly exploits multi-scale target history and exogenous information through Temporal Utility and Channel Utility. A gated bounded correction mechanism further enables controlled refinement of the endogenous-only base forecast. Experiments across 13 real-world datasets demonstrate strong forecasting performance, while ablation studies validate the effectiveness of both multi-scale utility components and the Adaptive Response Range. SCEU also remains effective when forecast-horizon exogenous information is partially or completely unavailable. Overall, SCEU provides an effective approach to selectively exploiting exogenous information for multi-step time series forecasting.

AI USE STATEMENT

In this work, we used generative AI tools for debugging, improving code readability, and manuscript language editing. We have not used generative AI tools for generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, proposing or refining research hypotheses, designing experimental methodologies, or interpreting research results.

Additionally, we used generative AI tools to improve manuscript readability, refine the organization of content, and proofread the manuscript. All AI-assisted outputs were reviewed and verified by the authors. AI-assisted code was tested within our experimental pipeline for correctness and reproducibility. Language and organizational suggestions were checked against the original scientific content to ensure that they preserved the meaning and did not introduce unsupported claims.

We take responsibility for the final content of this work, including all text, claims, experimental results, and artifacts produced with the aid of generative AI.

REFERENCES

Wei Bao, Jun Yue, and Yulei Rao. A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PLOS ONE*, 12(7):e0180944, 2017.

Christopher Bennett, Rodney A Stewart, and Junwei Lu. Autoregressive with exogenous variables and neural network short-term load forecast models for residential low voltage distribution networks. *Energies*, 7(5):2938–2960, 2014.

Wanlin Cai, Yuxuan Liang, Xianggen Liu, Jianshuai Feng, and Yuankai Wu. Msgnet: Learning multi-scale inter-series correlations for multivariate time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11141–11149, 2024.

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 *AAAI*, volume 37, pp. 6989–6997, 2023.

Jialin Chen, Jan Eric Lenssen, Aosong Feng, Weihua Hu, Matthias Fey, Leandros Tassiulas, Jure Leskovec, and Rex Ying. From similarity to superiority: Channel clustering for time series forecasting. *arXiv preprint arXiv:2404.01340*, 2024.

Ling Chen, Donghui Chen, Zongjiang Shang, Binqing Wu, Cen Zheng, Bo Wen, and Wei Zhang. Multi-scale adaptive graph neural network for multivariate time series forecasting. *TKDE*, 35(10): 10748–10761, 2023.

Yunyao Cheng, Peng Chen, Chenjuan Guo, Kai Zhao, Qingsong Wen, Bin Yang, and Christian S Jensen. Weakly guided adaptation for robust time series forecasting. *Proc. VLDB Endow.*, 17(4): 766–779, 2023.

Razvan-Gabriel Cirstea, Bin Yang, and Chenjuan Guo. Graph attention recurrent neural networks for correlated time series forecasting. In *MileTS19@KDD*, 2019.

Razvan-Gabriel Cirstea, Chenjuan Guo, Bin Yang, Tung Kieu, Xuanyi Dong, and Shirui Pan. Triformer: Triangular, variable-specific attentions for long sequence multivariate time series forecasting. In *IJCAI*, pp. 1994–2001, 2022a.

Razvan-Gabriel Cirstea, Bin Yang, Chenjuan Guo, Tung Kieu, and Shirui Pan. Towards spatio-temporal aware traffic time series forecasting. In *ICDE*, pp. 2900–2913, 2022b.

Tao Dai, Beiliang Wu, Peiyuan Liu, Naiqi Li, Jigang Bao, Yong Jiang, and Shu-Tao Xia. Periodicity decoupling framework for long-term series forecasting. In *ICLR*, 2024a.

Tao Dai, Beiliang Wu, Peiyuan Liu, Naiqi Li, Xue Yuerong, Shu-Tao Xia, and Zexuan Zhu. Ddn: Dual-domain dynamic normalization for non-stationary time series forecasting. In *NeurIPS*, 2024b.

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

Yuntao Du, Jindong Wang, Wenjie Feng, Sinno Jialin Pan, Tao Qin, Renjun Xu, and Chongjun Wang. Adarnn: Adaptive learning and forecasting of time series. In *CIKM*, pp. 402–411, 2021.

Vijay Ekambaram, Arindam Jati, Nam Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. Tsmixer: Lightweight mlp-mixer model for multivariate time series forecasting. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 459–469, 2023.

Wei Fan, Pengyang Wang, Dongkun Wang, Dongjie Wang, Yuanchun Zhou, and Yanjie Fu. Dish-ts: A general paradigm for alleviating distribution shift in time series forecasting. In *AAAI*, pp. 7522–7529, 2023.

Yifan Hu, Peiyuan Liu, Peng Zhu, Dawei Cheng, and Tao Dai. Adaptive multi-scale decomposition framework for time series forecasting. *arXiv preprint arXiv:2406.03751*, 2024.

Yifan Hu, Guibin Zhang, Peiyuan Liu, Disen Lan, Naiqi Li, Dawei Cheng, Tao Dai, Shu-Tao Xia, and Shirui Pan. Timefilter: Patch-specific spatial-temporal graph filtration for time series forecasting. *ICML*, 2025.

Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.

Zahra Karevan and Johan AK Suykens. Transductive lstm for time-series prediction: An application to weather forecasting. *Neural Networks*, 125:1–9, 2020.

Jesus Lago, Grzegorz Marcjasz, Bart De Schutter, and Rafał Weron. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, 293:116983, 2021. doi: 10.1016/j.apenergy.2021.116983.

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 *ICLR*, 2026.

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. doi: 10.1016/j.ijforecast.2021.03.012. URL https://doi.org/10.1016/j.ijforecast.2021.03.012.

ChungYi Lin, Shen-Lung Tung, Hung-Ting Su, and Winston H Hsu. Teltrans: Applying multi-type telecom data to transportation evaluation and prediction via multifaceted graph modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 22927–22933, 2024a.

Shengsheng Lin, Weiwei Lin, Wentai Wu, Haojun Chen, and Junjie Yang. SparseTSF: Modeling long-term time series forecasting with 1k parameters. In *ICML*, pp. 30211–30226, 2024b.

Qinshuo Liu, Yanwen Fang, Pengtao Jiang, and Guodong Li. Dgcformer: Deep graph clustering transformer for multivariate time series forecasting. *arXiv preprint arXiv:2405.08440*, 2024a.

Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International conference on learning representations*, 2021.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In *NeurIPS*, 2022.

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*, 2024b.

Minrong Lu and Xuerong Xu. Trnn: An efficient time-series recurrent neural network for stock price prediction. *Information Sciences*, 657:119951, 2024.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

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.

Nikolaos Passalis, Anastasios Tefas, Juho Kanniainen, Moncef Gabbouj, and Alexandros Iosifidis. Deep adaptive input normalization for time series forecasting. *IEEE Trans. Neural Networks Learn. Syst.*, 31(9):3760–3765, 2020.

Xiangfei Qiu, Hanyin Cheng, Xingjian Wu, Jilin Hu, and Chenjuan Guo. A comprehensive survey of deep learning for multivariate time series forecasting: A channel strategy perspective. *arXiv preprint arXiv:2502.10721*, 2025a.

Xiangfei Qiu, Xingjian Wu, Hanyin Cheng, Xvyuan Liu, Chenjuan Guo, Jilin Hu, and Bin Yang. DBLoss: Decomposition-based loss function for time series forecasting. In *NeurIPS*, 2025b.

Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. DUET: Dual clustering enhanced multivariate time series forecasting. In *SIGKDD*, pp. 1185–1196, 2025c.

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 *Proceedings of the 43rd International Conference on Machine Learning*, 2026. arXiv:2509.14933.

Artyom Stitsyuk and Jaesik Choi. xpatch: Dual-stream time series forecasting with exponential seasonal-trend decomposition. In *AAAI*, volume 39, pp. 20601–20609, 2025.

Souhaib Ben Taieb, Gianluca Bontempi, Amir F Atiya, and Antti Sorjamaa. A review and comparison of strategies for multi-step ahead time series forecasting based on the nn5 forecasting competition. *Expert systems with applications*, 39(8):7067–7083, 2012.

Kshitij Tayal, Arvind Renganathan, Xiaowei Jia, Vipin Kumar, and Dan Lu. Exotst: Exogenous-aware temporal sequence transformer for time series prediction. In *ICDM*, pp. 857–862, 2024.

Stylianos I Vagropoulos, GI Chouliaras, Evaggelos G Kardakos, Christos K Simoglou, and Anastasios G Bakirtzis. Comparison of sarimax, sarima, modified sarima and ann-based models for short-term pv generation forecasting. In *ENERGYCON*, pp. 1–6, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pp. 5998–6008, 2017.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. In *ICLR*, 2024a.

Yihang Wang, Yuying Qiu, Peng Chen, Kai Zhao, Yang Shu, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. Rose: Register assisted general time series forecasting with decomposed frequency learning. *arXiv preprint arXiv:2405.17478*, 2024b.

Yihe Wang, Yu Han, Haishuai Wang, and Xiang Zhang. Contrast everything: A hierarchical contrastive framework for medical time-series. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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 *NeurIPS*, volume 37, pp. 469–498, 2024c.

Billy M Williams. Multivariate vehicular traffic flow prediction: Evaluation of arimax modeling. 1776(1):194–200, 2001.

Binqing Wu, Zongjiang Shang, Jianlong Huang, and Ling Chen. Millgnn: Learning multi-scale lead-lag dependencies for multi-variate time series forecasting. In *CIKM*, pp. 3344–3354, 2025a.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *NeurIPS*, pp. 22419–22430, 2021a.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *ICLR*, 2023a.

Haixu Wu, Hang Zhou, Mingsheng Long, and Jianmin Wang. Interpretable weather forecasting for worldwide stations with a unified deep model. *Nature Machine Intelligence*, 2023b.

Xingjian Wu, Xiangfei Qiu, Hanyin Cheng, Zhengyu Li, Jilin Hu, Chenjuan Guo, and Bin Yang. Enhancing time series forecasting through selective representation spaces: A patch perspective. In *NeurIPS*, 2025b.

Xinle Wu, Dalin Zhang, Chenjuan Guo, Chaoyang He, Bin Yang, and Christian S Jensen. AutoCTS: Automated correlated time series forecasting. *Proc. VLDB Endow.*, 15(4):971–983, 2021b.

Xinle Wu, Xingjian Wu, Bin Yang, Lekui Zhou, Chenjuan Guo, Xiangfei Qiu, Jilin Hu, Zhenli Sheng, and Christian S Jensen. AutoCTS++: zero-shot joint neural architecture and hyperparameter search for correlated time series forecasting. *The VLDB Journal*, 33(5):1743–1770, 2024.

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*, pp. 753–763, 2020.

Zhijian Xu, Ailing Zeng, and Qiang Xu. FITS: modeling time series with 10k parameters. In *ICLR*, 2024.

Kun Yi, Qi Zhang, Wei Fan, Hui He, Liang Hu, Pengyang Wang, Ning An, Longbing Cao, and Zhendong Niu. Fouriergnn: Rethinking multivariate time series forecasting from a pure graph perspective. *Advances in Neural Information Processing Systems*, 36, 2023.

Chengqing Yu, Fei Wang, Zezhi Shao, Tao Sun, Lin Wu, and Yongjun Xu. Dsformer: A double sampling transformer for multivariate time series long-term prediction. In *CIKM*, pp. 3062–3072, 2023.

Scott L Zeger, Rafael Irizarry, and Roger D Peng. On time series analysis of public health and biomedical data. *Annual Review of Public Health*, 27:57–79, 2006.

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*, 2023.

Jiawen Zhang, Xumeng Wen, Zhenwei Zhang, Shun Zheng, Jia Li, and Jiang Bian. ProbTS: Benchmarking point and distributional forecasting across diverse prediction horizons. In *NeurIPS*, 2024.

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

Kai Zhao, Chenjuan Guo, Yunyao Cheng, Peng Han, Miao Zhang, and Bin Yang. Multiple time series forecasting with dynamic graph modeling. *Proc. VLDB Endow.*, 17(4):753–765, 2023.

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 *AAAI*, volume 35, pp. 11106–11115, 2021.

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 *SIGKDD*, 2025.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *ICML*, pp. 27268–27286, 2022.

A End-to-End SCEU Algorithm

Algorithm 1 summarizes the complete optimization and inference procedure. It follows the computational graph in Section 4: the endogenous-only prediction is formed first, the two utility branches construct step-conditioned contexts, and the gated bounded pathway supplies only a residual correction. All quantities in the utility and control branches are learned from the forecasting loss; no utility labels or auxiliary supervision are required.

Algorithm 1: End-to-end training and inference for SCEU

Input: Training set $\mathcal{D}$; endogenous history $\mathbf{X}^{\mathrm{en}}$; historical and known-future exogenous variables $\mathbf{X}^{\mathrm{ex}}_{\mathrm{his}}$ and $\mathbf{X}^{\mathrm{ex}}_{\mathrm{fut}}$; horizon H; learning rate η; epochs N; model parameters Θ

Output: Forecast $\widehat{\mathbf{Y}}^{\mathrm{en}}$

Training phase:
for $i \leftarrow 1$ **to** N **do**
  foreach *mini-batch* $(\mathbf{X}^{\mathrm{en}}, \mathbf{X}^{\mathrm{ex}}_{\mathrm{his}}, \mathbf{X}^{\mathrm{ex}}_{\mathrm{fut}}, \mathbf{Y}^{\mathrm{en}}) \in \mathcal{D}$ **do**
    Normalize each input variable using its historical-window statistics
    Encode the current state $\mathbf{s}$ using Equation 3
    Form the endogenous-only base forecast $\widehat{\mathbf{Y}}^{\mathrm{base}}$ using Equation 4
    Construct $\mathbf{c}^{\mathrm{en}}_h$ for every h with the multi-scale Temporal Utility branch (Equations 5–10)
    Construct $\mathbf{c}^{\mathrm{ex}}_h$ for every h with the variable–scale Channel Utility branch (Equations 11–18)
    Fuse the two contexts and compute (δ_h, g_h, r_h) using Equations 19–23
    Compute $\widehat{y}^{\mathrm{norm}}_h = \widehat{y}^{\mathrm{base}}_h + g_h r_h \delta_h$ for $h = 1, \ldots, H$, then de-normalize the forecast
    Evaluate $\mathcal{L}$ with Equation 25 and apply an Adam update to Θ with learning rate η
Inference phase:
Repeat the normalized forward pass without gradient updates and **return** $\widehat{\mathbf{Y}}^{\mathrm{en}}$

B Theoretical Properties of the Utility and Correction Pathways

The following results are deterministic consequences of the SCEU parameterization. They do not require distributional assumptions and should not be interpreted as a generalization or optimization guarantee. Instead, they formalize two design properties: the utility aggregation cannot acquire arbitrarily large coefficient mass, and the final prediction remains inside a known neighborhood of the endogenous-only base forecast.

B.1 Bounded signed Channel Utility

Proposition 1 (Signed utility mass and context bound)**.** *For a fixed prediction step h, let $u^{\mathrm{ex}}_{h,c,k} = a^{\mathrm{ex}}_{h,c,k} d^{\mathrm{ex}}_{h,c,k}$ be defined by Equations 15–17. Then:*

$$\sum_{c=1}^{C} \sum_{k=1}^{K} \left| u^{\mathrm{ex}}_{h,c,k} \right| \le 1. \tag{26}$$

Consequently, for any vector norm and its induced operator norm:

$$\left\| \mathbf{c}^{\mathrm{ex}}_h \right\| \le \left\| \mathbf{W}^{\mathrm{ex}}_O \right\|_{\mathrm{op}} \max_{c,k} \left\| \mathbf{v}^{\mathrm{ex}}_{h,c,k} \right\|. \tag{27}$$

Proof. The Softmax in Equation 15 gives $a^{\mathrm{ex}}_{h,c,k} \ge 0$ and $\sum_{c,k} a^{\mathrm{ex}}_{h,c,k} = 1$. The hyperbolic tangent in Equation 16 gives $|d^{\mathrm{ex}}_{h,c,k}| \le 1$. Hence $|u^{\mathrm{ex}}_{h,c,k}| = a^{\mathrm{ex}}_{h,c,k} |d^{\mathrm{ex}}_{h,c,k}| \le a^{\mathrm{ex}}_{h,c,k}$, and summing proves

Equation 26. Applying the triangle inequality and submultiplicativity to Equation 18 then yields:

$$\begin{aligned}\|\mathbf{c}_h^{\mathrm{ex}}\| &\le \|\mathbf{W}_O^{\mathrm{ex}}\|_{\mathrm{op}} \sum_{c,k} |u_{h,c,k}^{\mathrm{ex}}| \left\|\mathbf{v}_{h,c,k}^{\mathrm{ex}}\right\| \\ &\le \|\mathbf{W}_O^{\mathrm{ex}}\|_{\mathrm{op}} \max_{c,k} \left\|\mathbf{v}_{h,c,k}^{\mathrm{ex}}\right\|.\end{aligned}$$

The proposition bounds the aggregate coefficient mass even though the signed scores may cancel. It does not make an individual utility identifiable: each score still weights a representation that already contains cross-variable interaction through Equation 13.

B.2 Convex utility fusion

Proposition 2 (Fusion does not amplify context norms)**.** *For every prediction step h and every norm:*

$$\left\|\mathbf{c}_h^{\mathrm{util}}\right\| \le \gamma_h^{\mathrm{en}} \left\|\mathbf{c}_h^{\mathrm{en}}\right\| + \gamma_h^{\mathrm{ex}} \left\|\mathbf{c}_h^{\mathrm{ex}}\right\| \le \max\{\left\|\mathbf{c}_h^{\mathrm{en}}\right\|, \left\|\mathbf{c}_h^{\mathrm{ex}}\right\|\}. \tag{28}$$

Proof. Equation 19 uses a two-way Softmax, so $\gamma_h^{\mathrm{en}}, \gamma_h^{\mathrm{ex}} \ge 0$ and $\gamma_h^{\mathrm{en}} + \gamma_h^{\mathrm{ex}} = 1$. The first inequality is the triangle inequality applied to Equation 20; the second follows because the right-hand side is a convex combination of the two context norms.

B.3 Bounded deviation from the endogenous base

Proposition 3 (Prediction and MAE stability)**.** *Let $0 \le r_{\min} < r_{\max}$. For every input and prediction step h, the normalized SCEU forecast satisfies:*

$$\left|\widehat{y}_h^{\mathrm{norm}} - \widehat{y}_h^{\mathrm{base}}\right| \le r_h \le r_{\max}. \tag{29}$$

If inverse normalization uses endogenous scale $s_{\mathrm{en}} = \sigma(\mathbf{X}^{\mathrm{en}}) + \epsilon$, then:

$$\left|\widehat{y}_h^{\mathrm{en}} - \widehat{y}_h^{\mathrm{base,en}}\right| \le s_{\mathrm{en}} r_{\max}. \tag{30}$$

Moreover, for the normalized per-step absolute losses $\ell_h = |y_h - \widehat{y}_h^{\mathrm{norm}}|$ and $\ell_h^{\mathrm{base}} = |y_h - \widehat{y}_h^{\mathrm{base}}|$:

$$|\ell_h - \ell_h^{\mathrm{base}}| \le r_{\max}. \tag{31}$$

The same bound holds after averaging over forecast steps and samples.

Proof. By construction, $g_h \in [0,1]$, $r_h \in [r_{\min}, r_{\max}]$, and $\delta_h \in [-1,1]$. Equation 24 gives:

$$\left|\widehat{y}_h^{\mathrm{norm}} - \widehat{y}_h^{\mathrm{base}}\right| = |g_h r_h \delta_h| \le r_h \le r_{\max}.$$

Inverse normalization multiplies a normalized difference by s_{en}, which proves Equation 30. Finally, the reverse triangle inequality gives:

$$|\ell_h - \ell_h^{\mathrm{base}}| \le |\widehat{y}_h^{\mathrm{norm}} - \widehat{y}_h^{\mathrm{base}}| \le r_{\max}.$$

Averaging preserves the inequality.

Proposition 3 is a safety property rather than an accuracy guarantee: the correction may increase or decrease the error, but it cannot move the normalized forecast more than $r_{\max}$ away from the base. The gate gives a sharper sample- and step-specific certificate, $|\widehat{y}_h^{\mathrm{norm}} - \widehat{y}_h^{\mathrm{base}}| \le g_h r_h$.

Table 3: Dataset statistics. "Exog." denotes the number of exogenous variables.

Dataset	Exog.	Endogenous target	Frequency	Length	Train/Val/Test	Horizons
Weather	20	CO_2 concentration	10 minutes	52,696	7:1:2	96, 192, 336, 720
Electricity	320	Electricity consumption	1 hour	26,304	7:1:2	96, 192, 336, 720
Traffic	861	Road occupancy rate	1 hour	17,544	7:1:2	96, 192, 336, 720
ETTh1/ETTh2	6	Oil temperature	1 hour	17,420	6:2:2	96, 192, 336, 720
ETTm1/ETTm2	6	Oil temperature	15 minutes	69,680	6:2:2	96, 192, 336, 720
BE	2	Belgium electricity price	1 hour	52,416	7:1:2	24, 168, 360
DE	2	Germany electricity price	1 hour	52,416	7:1:2	24, 168, 360
Energy	5	Thermoelectric generation	1 hour	13,064	7:1:2	24, 168, 360
FR	2	France electricity price	1 hour	52,416	7:1:2	24, 168, 360
NP	2	Nord Pool electricity price	1 hour	52,416	7:1:2	24, 168, 360
PJM	2	PJM electricity price	1 hour	52,416	7:1:2	24, 168, 360

C Dataset Statistics

Table 3 summarizes the benchmark composition, sampling frequency, chronological split, and prediction horizons. The benchmark suite covers widely used long-horizon forecasting datasets and regional electricity-market and energy datasets, spanning diverse temporal resolutions, channel dimensions, and forecast horizons. The number of exogenous variables ranges from a few market signals to hundreds of channels, enabling evaluation under both compact and multivariate settings.

Common long-horizon benchmarks. Weather, Electricity, Traffic, and the four ETT subsets cover different temporal resolutions and channel counts. Following a unified formulation, the last channel is treated as the endogenous target and all remaining channels as exogenous variables, enabling consistent evaluation across different multivariate structures.

Weather provides dense observations at a 10-minute resolution. Electricity and Traffic contain 320 and 861 exogenous variables, respectively, and therefore test model scalability and selective information utilization in high-dimensional systems, where indiscriminate variable mixing may introduce redundant or weakly relevant information. In contrast, the ETT subsets contain six exogenous variables and provide compact systems at hourly and 15-minute resolutions. Together, these datasets span low- to high-dimensional forecasting scenarios.

For ETT, we use the complete original files rather than the conventional 20-month truncation adopted in many implementations. This changes the chronological training, validation, and test intervals relative to the commonly reported setting. Therefore, our ETT results should be compared only with methods evaluated using the same data range and splitting protocol, which is applied consistently to all competing methods.

Electricity-market and energy datasets. BE, DE, FR, NP, and PJM represent regional electricity-market forecasting tasks, with electricity prices as endogenous targets and market-related forecasts as exogenous variables. Although these datasets contain fewer exogenous channels, they provide practical settings for evaluating the use of future external information. The horizons of 24, 168, and 360 steps cover different forecasting ranges under hourly sampling.

Energy uses thermoelectric generation as the endogenous target and the remaining generation measurements as exogenous variables, extending the evaluation beyond electricity-price forecasting. Together, these datasets examine whether SCEU can selectively exploit external information across different targets and horizons.

For all datasets, train, validation, and test subsets are constructed chronologically without random shuffling across temporal boundaries. The same partitions and forecasting horizons are used for all compared methods, preventing future test observations from entering model fitting and ensuring a consistent evaluation protocol.

D Parameter Configuration

All experiments use a random seed of 2021. The model uses a hidden dimension of 64, a feed-forward dimension of 128, four attention heads, one encoder layer, no dropout, and MAE loss. We optimize it with Adam at a learning rate of 10^{-4} for at most 50 epochs, with early stopping patience set to five. For Weather, Electricity, Traffic, and ETT, the lookback length is 96 and the prediction horizons are $\{96, 192, 336, 720\}$. For BE, DE, Energy, FR, NP, and PJM, the main experiments use $(T, H) = (168, 24)$ and $(720, 360)$. The selected $(K, W, r_{\max})$ settings for NP, DE, PJM, and FR are $(2, 24, 1.5)$, $(2, 12, 2.0)$, $(3, 24, 1.75/2.0)$, and $(2/4, 24, 1.0)$, respectively, where slash-separated values correspond to $H = 24/360$. Traffic uses the latent variant with eight exogenous tokens. ETTh2 and ETTm2 use a fixed response range of 0.5.

E Latent Exogenous Variant for High-Dimensional Inputs

Traffic contains 861 exogenous variables, making direct variable-wise interaction substantially more expensive than on the other benchmarks. To reduce the training cost, we use a low-rank latent variant of the exogenous branch with $R = 8$ latent factors. Importantly, this variant does not discard or pre-compress the original variables. All exogenous variables are first independently encoded using the same multi-scale patch representation as in the standard model and subsequently participate in the latent interaction.

At each scale, let $\mathbf{I}_p \in \mathbb{R}^{C\times D}$ denote the variable representations for a future patch. Instead of constructing pairwise interactions among all C variables, we introduce R learnable latent queries conditioned on the current state and forecast position. Their attention over the original variables is

$$\mathbf{A}_p = \mathrm{Softmax}\left(\frac{\mathbf{Q}_p\mathbf{K}_p^\top}{\sqrt{D}}\right) \in \mathbb{R}^{R\times C},$$

where $\mathbf{Q}_p \in \mathbb{R}^{R\times D}$ contains the latent queries and $\mathbf{K}_p \in \mathbb{R}^{C\times D}$ contains the variable representations. The latent representations are then obtained by

$$\mathbf{L}_p = \mathbf{A}_p\mathbf{V}_p \in \mathbb{R}^{R\times D}.$$

The resulting latent factors interact through self-attention and are aggregated using state- and horizon-conditioned gates. This replaces the $C \times C$ all-variable interaction with an $R \times C$ latent-variable interaction, reducing the corresponding variable-interaction complexity from $O(C^2)$ to $O(CR)$, where $R \ll C$. For Traffic, we set $C = 861$ and $R = 8$.

The latent variant is used for computational efficiency rather than to improve forecasting accuracy. In fact, retaining the standard exogenous branch achieves better performance: at $H = 96$, the uncompressed SCEU obtains an MSE of 0.076 and an MAE of 0.149. We nevertheless use the latent variant in the main Traffic benchmark to reduce training time and maintain a practical computational budget. Thus, the reported Traffic result reflects a deliberate trade-off between computational efficiency and forecasting accuracy rather than an accuracy gain from latent compression.

F Full Forecasting Results

The complete horizon-wise forecasting results are reported in Tables 4 and 5. Presenting each horizon separately exposes variations that are hidden by the averages in the main text and permits comparisons under exactly matched forecast lengths.

Common forecasting benchmarks. The horizon-wise results show that SCEU is consistently competitive on Weather, Electricity, Traffic, ETTh1, and ETTm1. The margins vary with the dataset and horizon, and the results on ETTh2 and ETTm2 are more mixed. This heterogeneity indicates that step-conditioned exogenous correction is most useful when the external channels contain information that is not already captured by the endogenous history. It also cautions against interpreting an average improvement as uniform superiority at every horizon.

Table 4: Results on ETT using the complete original files and on other common forecasting datasets. ETT results are obtained from local reruns under the unified protocol. Best results are shown in red and second-best results in blue with underlining. Lower values are better.

Dataset	H	SCEU		DAG		TimeXer		TFT		TiDE		DUET		CrossLinear		PatchTST	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Weather	96	0.001	0.018	0.001	0.023	0.001	0.023	0.001	0.028	0.001	0.026	0.002	0.028	0.012	0.093	0.006	0.062
	192	0.001	0.020	0.001	0.022	0.001	0.026	0.002	0.030	0.001	0.028	0.002	0.030	0.005	0.056	0.008	0.073
	336	0.001	0.022	0.002	0.028	0.001	0.027	0.002	0.030	0.002	0.031	0.001	0.025	0.009	0.080	0.011	0.087
	720	0.001	0.025	0.002	0.030	0.002	0.032	0.002	0.035	0.002	0.036	0.002	0.033	0.007	0.067	0.012	0.084
	Avg	0.001	0.021	0.002	0.026	0.001	0.027	0.002	0.031	0.002	0.030	0.001	0.029	0.008	0.074	0.009	0.076
Electricity	96	0.161	0.294	0.158	0.298	0.160	0.296	0.365	0.456	0.312	0.414	0.248	0.385	0.281	0.416	0.274	0.406
	192	0.181	0.315	0.156	0.295	0.186	0.320	0.339	0.448	0.274	0.391	0.290	0.421	0.248	0.384	0.264	0.399
	336	0.197	0.322	0.217	0.353	0.218	0.346	0.367	0.471	0.313	0.417	0.295	0.426	0.260	0.396	0.271	0.408
	720	0.231	0.353	0.246	0.370	0.267	0.386	0.429	0.510	0.353	0.445	0.290	0.414	0.268	0.401	0.268	0.402
	Avg	0.193	0.321	0.194	0.329	0.208	0.337	0.375	0.471	0.313	0.417	0.281	0.411	0.264	0.399	0.269	0.404
Traffic	96	0.093	0.173	0.119	0.187	0.135	0.233	0.139	0.221	0.149	0.251	0.247	0.316	0.261	0.322	0.241	0.307
	192	0.104	0.189	0.121	0.191	0.135	0.230	0.116	0.212	0.149	0.246	0.240	0.297	0.232	0.289	0.246	0.305
	336	0.081	0.154	0.125	0.200	0.143	0.237	0.134	0.228	0.150	0.251	0.310	0.356	0.223	0.280	0.215	0.280
	720	0.083	0.148	0.152	0.232	0.156	0.250	0.138	0.220	0.169	0.271	0.253	0.318	0.213	0.272	0.229	0.286
	Avg	0.090	0.166	0.129	0.203	0.142	0.238	0.132	0.220	0.154	0.255	0.263	0.322	0.232	0.291	0.233	0.295
ETTh1	96	0.087	0.229	0.128	0.283	0.104	0.247	0.137	0.283	0.104	0.248	0.107	0.248	0.101	0.240	0.102	0.245
	192	0.096	0.241	0.124	0.273	0.128	0.274	0.165	0.309	0.124	0.271	0.133	0.276	0.123	0.268	0.138	0.285
	336	0.102	0.250	0.134	0.290	0.134	0.285	0.163	0.312	0.130	0.281	0.133	0.283	0.132	0.284	0.135	0.287
	720	0.131	0.289	0.188	0.344	0.175	0.331	0.186	0.344	0.149	0.307	0.165	0.325	0.180	0.337	0.190	0.345
	Avg	0.104	0.252	0.144	0.297	0.135	0.284	0.163	0.312	0.127	0.277	0.134	0.283	0.134	0.282	0.141	0.291
ETTh2	96	0.239	0.371	0.290	0.411	0.245	0.376	0.313	0.439	0.244	0.373	0.246	0.373	0.243	0.376	0.244	0.378
	192	0.300	0.428	0.349	0.467	0.298	0.427	0.333	0.456	0.290	0.418	0.286	0.415	0.279	0.414	0.281	0.416
	336	0.315	0.435	0.346	0.460	0.329	0.453	0.339	0.458	0.314	0.437	0.319	0.437	0.316	0.440	0.318	0.448
	720	0.455	0.533	0.750	0.707	0.480	0.551	0.478	0.586	0.418	0.513	0.493	0.560	0.466	0.542	0.463	0.545
	Avg	0.327	0.442	0.434	0.511	0.338	0.452	0.366	0.485	0.317	0.435	0.336	0.446	0.326	0.443	0.327	0.447
ETTm1	96	0.039	0.143	0.044	0.151	0.043	0.149	0.053	0.173	0.049	0.163	0.042	0.146	0.043	0.149	0.043	0.148
	192	0.065	0.189	0.072	0.200	0.073	0.201	0.099	0.242	0.074	0.204	0.073	0.199	0.073	0.200	0.072	0.199
	336	0.093	0.230	0.106	0.245	0.105	0.245	0.140	0.289	0.112	0.251	0.102	0.239	0.105	0.245	0.106	0.245
	720	0.124	0.269	0.149	0.294	0.145	0.289	0.216	0.357	0.151	0.296	0.145	0.290	0.143	0.287	0.146	0.291
	Avg	0.080	0.208	0.093	0.222	0.091	0.221	0.127	0.265	0.096	0.229	0.091	0.219	0.091	0.220	0.092	0.221
ETTm2	96	0.116	0.235	0.103	0.219	0.213	0.236	0.361	0.448	0.119	0.238	0.117	0.229	0.103	0.219	0.125	0.243
	192	0.184	0.310	0.181	0.308	0.180	0.308	0.404	0.491	0.175	0.301	0.172	0.294	0.167	0.294	0.169	0.297
	336	0.255	0.371	0.251	0.369	0.261	0.377	0.485	0.541	0.256	0.372	0.258	0.368	0.255	0.368	0.262	0.374
	720	0.345	0.445	0.321	0.428	0.360	0.459	0.521	0.569	0.351	0.448	0.351	0.450	0.352	0.450	0.347	0.453
	Avg	0.225	0.340	0.214	0.331	0.254	0.345	0.443	0.512	0.225	0.340	0.225	0.335	0.219	0.333	0.226	0.342

Electricity-market and energy benchmarks. SCEU is strongest on BE, DE, Energy, and NP. It obtains the best average MSE on PJM and ranks second overall on FR, where DAG remains stronger. The complete results therefore support a selective interpretation of exogenous utility: external variables can provide substantial gains in some systems, but their incremental value is not constant across markets, prediction horizons, or error metrics. The endogenous base forecast remains important in settings where the external inputs are weak, redundant, or less reliable.

G Supplementary Ablation Results

Protocol. We compare the full SCEU model with four variants: Base, w/o Temporal Utility, w/o Channel Utility, and w/o Adaptive Range. Base disables both utility branches, while the remaining variants remove one component at a time. We report BE, DE, Energy, FR, NP, and PJM at $H \in \{24, 168\}$ and Weather at $H \in \{96, 192\}$. The complete results are summarized in Table 6.

Observed Patterns. As shown in Table 6, the extended ablations provide additional diagnostic evidence for the proposed design. Channel Utility provides consistent gains across most settings, highlighting the effectiveness of the proposed utility formulation for organizing exogenous information. Temporal Utility contributes complementary multi-scale historical evidence in most settings,

Table 5: Results on the electricity-market and energy datasets. Competing results are taken from Table 5 of the DAG paper; in particular, the DAG results on NP and PJM are the published values. Best results are shown in red and second-best results in blue with underlining.

Dataset	H	SCEU		DAG		TimeXer		TFT		TiDE		DUET		CrossLinear		PatchTST	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
BE	24	0.351	0.228	0.362	0.229	0.392	0.253	0.426	0.272	0.426	0.285	0.432	0.272	0.391	0.259	0.452	0.326
	360	0.489	0.320	0.486	0.331	0.512	0.327	0.482	0.310	0.571	0.364	0.597	0.436	0.568	0.416	0.702	0.538
	Avg	0.420	0.274	0.424	0.280	0.452	0.290	0.454	0.291	0.498	0.325	0.515	0.354	0.479	0.337	0.577	0.432
DE	24	0.271	0.322	0.277	0.322	0.339	0.362	0.380	0.383	0.367	0.383	0.376	0.378	0.387	0.396	0.413	0.412
	360	0.438	0.417	0.462	0.418	0.610	0.474	0.599	0.509	0.630	0.511	0.589	0.482	0.583	0.507	0.589	0.498
	Avg	0.355	0.370	0.370	0.370	0.475	0.418	0.489	0.446	0.499	0.447	0.482	0.430	0.485	0.452	0.501	0.455
Energy	24	0.077	0.213	0.079	0.215	0.122	0.273	0.093	0.235	0.103	0.248	0.117	0.283	0.241	0.418	0.239	0.390
	360	0.166	0.319	0.169	0.320	0.204	0.357	0.167	0.331	0.202	0.355	0.288	0.452	0.237	0.385	0.214	0.363
	Avg	0.121	0.266	0.124	0.268	0.163	0.315	0.130	0.283	0.153	0.302	0.203	0.367	0.239	0.402	0.226	0.377
FR	24	0.342	0.181	0.357	0.179	0.366	0.208	0.543	0.253	0.418	0.255	0.384	0.251	0.390	0.226	0.518	0.368
	360	0.501	0.283	0.473	0.269	0.489	0.273	0.475	0.274	0.551	0.308	0.607	0.403	0.575	0.370	0.658	0.452
	Avg	0.422	0.232	0.415	0.224	0.428	0.241	0.509	0.264	0.485	0.282	0.496	0.327	0.483	0.298	0.588	0.410
NP	24	0.187	0.216	0.202	0.237	0.236	0.266	0.219	0.249	0.284	0.301	0.246	0.287	0.210	0.266	0.249	0.294
	360	0.457	0.422	0.521	0.451	0.600	0.475	0.539	0.501	0.601	0.498	0.576	0.528	0.531	0.508	0.530	0.498
	Avg	0.322	0.319	0.362	0.344	0.418	0.371	0.379	0.375	0.443	0.400	0.411	0.408	0.371	0.387	0.390	0.396
PJM	24	0.058	0.146	0.057	0.143	0.075	0.166	0.095	0.195	0.106	0.214	0.072	0.166	0.088	0.191	0.116	0.239
	360	0.124	0.217	0.130	0.218	0.140	0.231	0.133	0.229	0.177	0.279	0.141	0.228	0.135	0.254	0.149	0.279
	Avg	0.091	0.181	0.094	0.181	0.108	0.199	0.114	0.212	0.142	0.247	0.107	0.197	0.112	0.223	0.133	0.259

Table 6: Consolidated component ablation results. Weather values are multiplied by 10 for readability. Rankings are determined before rounding. Best results are shown in red and second-best results in blue with underlining. Lower values are better.

Dataset	H	Base		w/o Temporal Utility		w/o Channel Utility		w/o Adaptive Range		Full	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
BE	24	0.437	0.292	0.338	0.232	0.402	0.273	0.356	0.244	0.351	0.228
	168	0.588	0.356	0.455	0.289	0.579	0.359	0.475	0.300	0.451	0.286
DE	24	0.537	0.472	0.296	0.333	0.507	0.451	0.344	0.360	0.271	0.322
	168	0.977	0.635	0.491	0.420	0.918	0.613	0.621	0.475	0.492	0.421
Energy	24	0.126	0.276	0.079	0.219	0.121	0.269	0.081	0.221	0.077	0.213
	168	0.178	0.329	0.112	0.262	0.201	0.352	0.117	0.266	0.111	0.261
FR	24	0.407	0.239	0.348	0.186	0.379	0.219	0.361	0.199	0.342	0.181
	168	0.543	0.306	0.454	0.240	0.535	0.304	0.454	0.250	0.465	0.240
NP	24	0.297	0.304	0.204	0.231	0.274	0.284	0.221	0.241	0.187	0.216
	168	0.548	0.456	0.358	0.336	0.527	0.432	0.411	0.366	0.355	0.336
PJM	24	0.095	0.195	0.062	0.150	0.087	0.183	0.067	0.154	0.058	0.146
	168	0.188	0.272	0.108	0.194	0.181	0.265	0.127	0.209	0.106	0.191
Weather ($\times 10$)	96	0.011	0.238	0.007	0.185	0.011	0.227	0.008	0.197	0.007	0.176
	192	0.013	0.265	0.009	0.209	0.013	0.258	0.010	0.227	0.008	0.202

although limited reversals occur on BE, DE, and FR. The Adaptive Response Range further strengthens the model by regulating the magnitude of forecast corrections. Overall, Temporal Utility and Channel Utility capture complementary predictive information, while the adaptive controller determines how strongly this information should refine the endogenous base forecast.

H Forecasting with Incomplete Future Exogenous Information

Protocol. We further evaluate SCEU under incomplete future exogenous information. Electricity is evaluated at $H = 96$, ETTh1 and ETTh2 at $H \in \{96, 336\}$, and DE, FR, NP, and PJM at $H = 24$. For each sample, we randomly mask 0%, 25%, 50%, 75%, or 100% of the forecast-step–exogenous-channel grid, while keeping the target history and historical exogenous observations fully available.

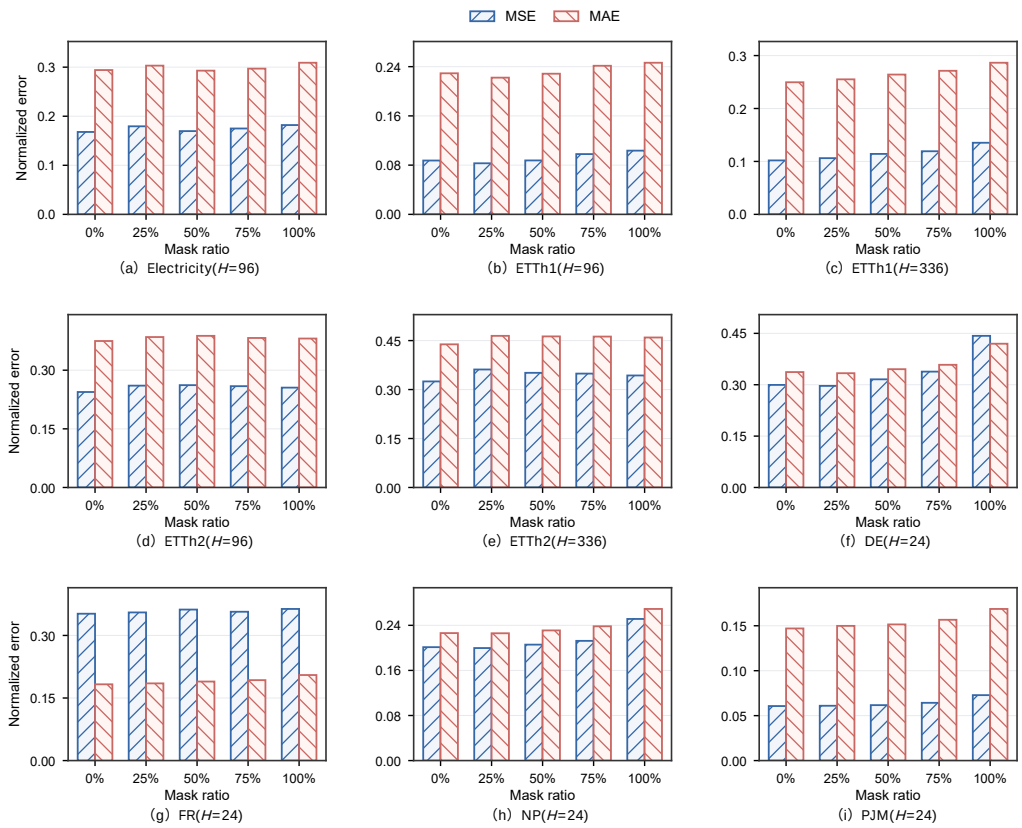

Figure 6: Forecasting performance under different levels of future exogenous information availability. Each panel reports normalized MSE and MAE for one dataset–horizon pair across future exogenous mask ratios. Only exogenous values within the forecasting horizon are masked; historical target and exogenous observations remain available. Lower values are better.

Masked values are set to zero after normalization, and an availability encoding indicates which future entries are observed.

For each nonzero mask ratio, the model is trained and evaluated under the corresponding missingness level. During training, masks are resampled for each batch, whereas validation and test masks are fixed using seed 2021. Thus, this experiment evaluates SCEU under different levels of future exogenous availability rather than unexpected test-time missingness.

Observed Patterns. As shown in Figure 6, SCEU remains effective as future exogenous information becomes increasingly incomplete. The sensitivity to missingness varies across datasets and forecast horizons, reflecting differences in the predictive information provided by future exogenous variables. Performance changes moderately under partial masking in several settings, while larger degradation generally occurs when future exogenous information is completely unavailable.

Even at the 100% mask ratio, where no forecast-horizon exogenous observations are available, SCEU can still produce meaningful forecasts using the available historical information. Overall, the results show that SCEU can benefit from future exogenous information when available while remaining applicable when such information is partially or completely unavailable.

I Qualitative Forecasting Examples

Figures 7 and 8 provide qualitative comparisons on NP under short- and long-horizon forecasting settings, respectively. Each panel shows the ground truth together with the prediction produced by SCEU, DAG, TimeXer, DUET, CrossLinear, or PatchTST over the same test window. The $H = 24$

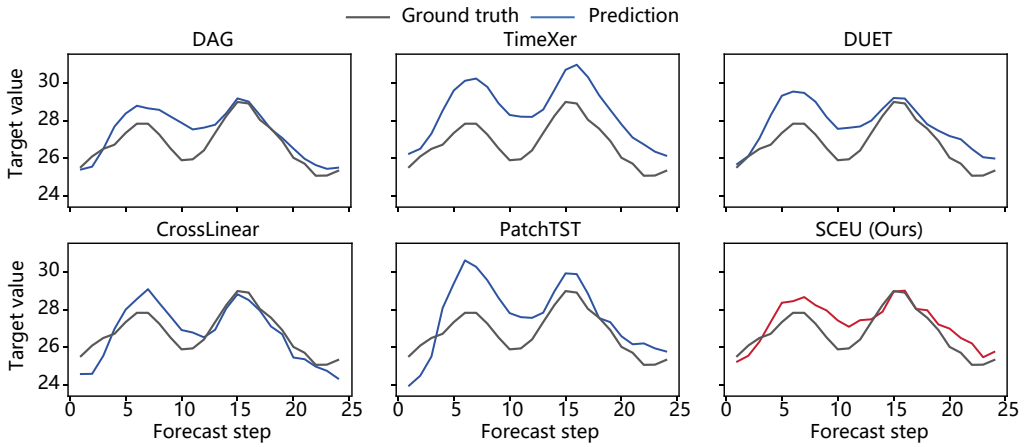

Figure 7: Qualitative forecasting comparison on NP at $H = 24$. All models are evaluated on the same test window with a lookback length of 168.

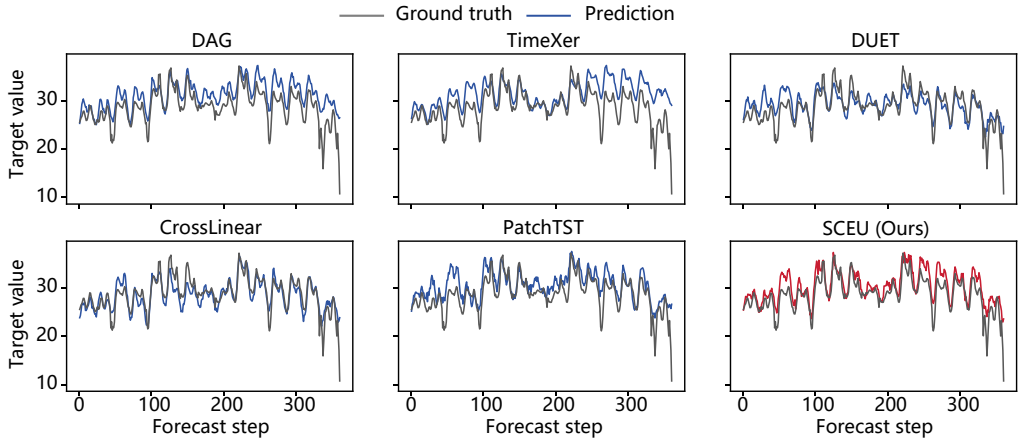

Figure 8: Qualitative forecasting comparison on NP at $H = 360$. All models are evaluated on the same test window with a lookback length of 720.

experiment uses a lookback length of 168, whereas the $H = 360$ experiment uses a lookback length of 720. By presenting different forecasting ranges, these examples complement the aggregate error metrics and provide a direct view of how each model tracks local variations, recovers turning points, and maintains trajectory consistency as the prediction horizon increases.

At $H = 24$, the comparison mainly reflects the ability to capture short-term variations and rapidly adapt to local changes in the target trajectory. In contrast, $H = 360$ places greater emphasis on maintaining the overall temporal pattern over an extended forecasting interval, where small local errors can accumulate and lead to noticeable long-horizon drift. Considering both settings therefore provides a more complete qualitative assessment than either horizon alone. In particular, the visualizations help reveal prediction characteristics that may be obscured by averaged MSE and MAE values, such as delayed responses around turning points, excessive smoothing, and deviations in trajectory shape. These observations provide additional insight into model behavior across different temporal scales and forecasting conditions.

J Supplementary Parameter Sensitivity

We complement the parameter analysis in the main text with additional results on DE, PJM, and FR. Following the same one-factor-at-a-time protocol, we vary the response upper bound $r_{\max}$, state window W, number of temporal scales K, and lookback window T, while keeping the re-

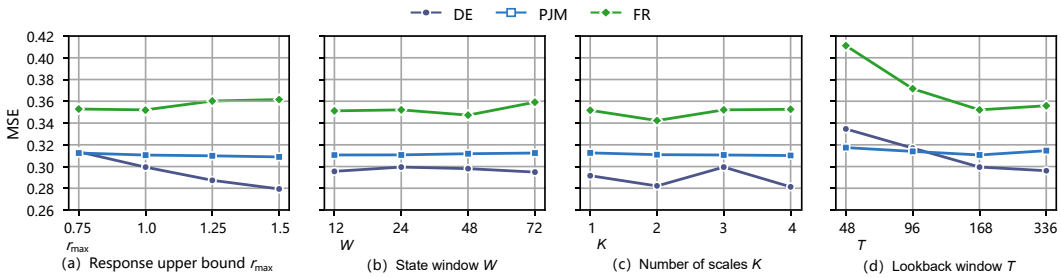

Figure 9: Parameter sensitivity of SCEU on DE, PJM, and FR at $H = 24$. The four panels vary $r_{\max}$, W, K, and T, respectively. For visualization on a common ordinate, 0.25 is added to the MSE values of PJM. Filled markers denote configurations reused from the main experiments.

maining settings fixed. These parameters respectively control the allowable correction strength, the amount of recent context used to characterize the current state, the granularity of multi-scale temporal modeling, and the total historical context available to the model. The supplementary experiments therefore examine different aspects of SCEU rather than variations of a single model capacity factor.

Figure 9 shows that SCEU maintains stable behavior across the evaluated parameter ranges, although the preferred configuration varies with dataset characteristics. Increasing $r_{\max}$ improves DE and PJM over the evaluated range, whereas FR is comparatively stable and obtains its lowest MSE at $r_{\max} = 1.0$. This suggests that the appropriate correction range depends on how strongly useful external information contributes to the target dynamics.

The influence of historical context exhibits a similar dataset-specific pattern. A longer lookback window benefits DE, while PJM and FR perform best at $T = 168$, indicating that additional history is not uniformly beneficial once sufficient temporal context has been captured. The non-monotonic trends for W and K further suggest that the appropriate state context and temporal resolution depend on the underlying dynamics of each dataset. Overall, these results support the robustness of SCEU across a practical range of parameter choices while also motivating dataset-specific tuning for optimal performance.